THE GRANULARITY AXIS: A MICRO-TO-MACRO LATENT DIRECTION FOR SOCIAL ROLES IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) are prompted to adopt social roles from individuals to institutions, but whether their representations encode this difference in social *granularity* remains unclear. We discover a contrast-based *Granularity Axis*, the difference between mean macro- and micro-role hidden states, that aligns with the dominant geometry of role space: in Qwen3-8B, its cosine with PC1 is 0.972, and PC1 explains 52.6% of the variance. Across 75 roles at five levels, with 90,000 role-conditioned and 1,200 default responses per model, projections increase monotonically and remain stable across layers, prompts, and response-quality filters. The structure transfers to Llama-3.1-8B-Instruct and scales to Qwen3-32B and Llama-3.1-70B-Instruct, whose PC1 shares are 45.99% and 31.23%. Independent five-rater validation ($\alpha = 0.961$) preserves the ordering; identical continuations and affect/formality controls confirm scale-related variation beyond generated wording. The axis is also *behaviorally causal*: positive steering moves Llama from 2.00 to 3.17 on a five-point macro scale for prompts admitting local responses. Steering strength and stability depend on the model's default operating regime. Together, these findings establish social granularity as a representational primitive: a continuous, ordered, causally manipulable direction organizing role-conditioned generation across model families.[1]

1 INTRODUCTION

Large language models (LLMs) adapt their behavior under prompting (OpenAI, 2025; Llama Team, 2024; Team, 2025; Google, 2025; Anthropic, 2026), motivating their use in social simulation (Park et al., 2023; Mou et al., 2024; Anthis et al., 2025). Yet representational validity, survey bias, and overly rationalized behavior remain concerns (Santurkar et al., 2023; Bisbee et al., 2024; Salecha et al., 2024; Liu et al., 2025). These concerns depend on what a model internally represents when prompted to *be* someone: stylistic mimicry and a distinct perspective have different implications for simulation.

Role conditioning can elicit the voices of a worried parent, community organizer, hospital administrator, or central bank governor (Shao et al., 2023; Wang et al., 2024b; Jiang et al., 2024; Wang et al., 2023; Tu et al., 2024; Kong et al., 2023). We ask: *does an LLM internally distinguish the granularity of these social roles, or realize them through a shared role-playing template?*

Social roles differ in agency, time horizon, and structural constraints (Coleman et al., 1990; Schelling, 1978; Granovetter, 1978). Micro roles emphasize personal experience and immediate concerns; macro roles involve coordination, institutional constraints, and long-term strategy. We call a mismatch between the scale a context requires and the scale a model adopts *granularity confusion*: individual reasoning where systemic analysis is needed, or institutional abstraction where local judgment is appropriate. If a central bank governor and worried parent reason at the same scale, a nominally multi-perspective simulation can conceal representational collapse despite different wording.

Interpretability research motivates looking for this distinction in low-dimensional activation structure (Park et al., 2024b;a; Sakata et al., 2026). The *Assistant Axis* tracks deviation from the default

[1] Code and data are available at Granularity-Axis.

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

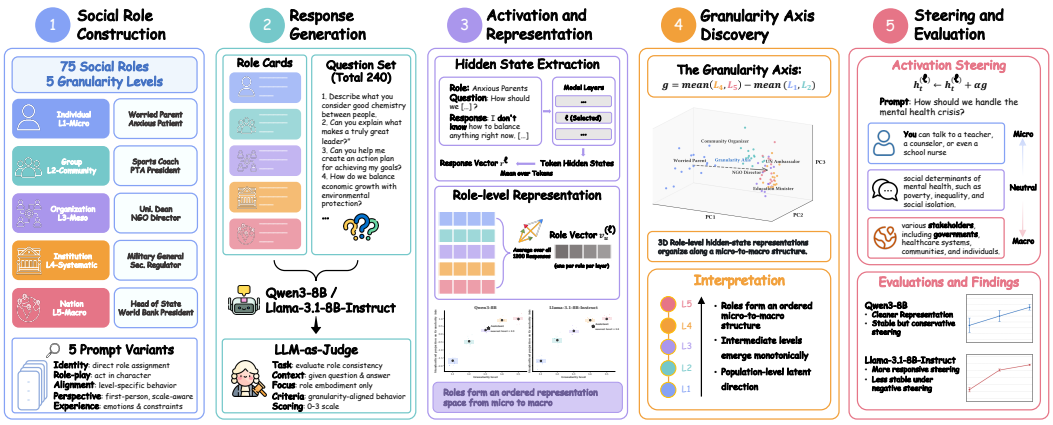

Figure 1: **Overview of the Granularity Axis pipeline.** We construct ordered social roles, collect role-conditioned responses, extract role-level hidden-state representations, define a contrast-based Granularity Axis, and test its behavioral effect through activation steering.

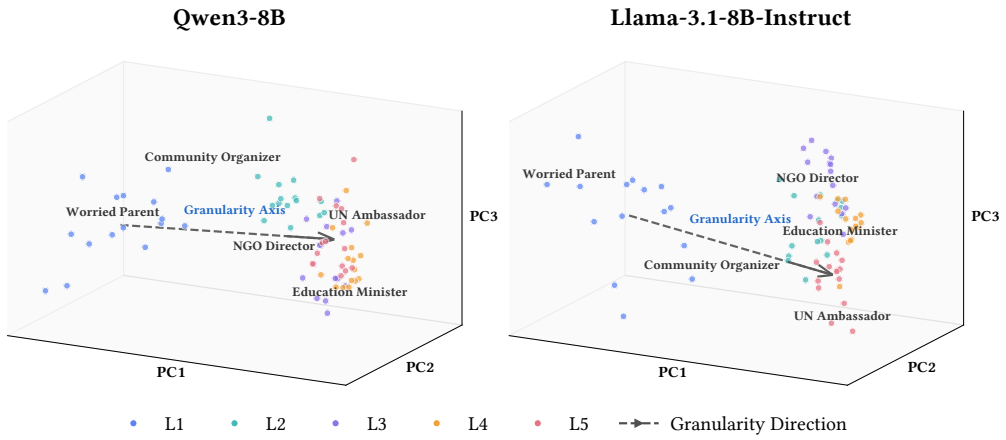

Figure 2: **Role representation space.** Role-level hidden-state representations organize along a micro-to-macro structure. Colors indicate granularity level (L1–L5), and the dashed arrow denotes the contrast-defined Granularity Axis from micro-level to macro-level roles.

assistant persona (Lu et al., 2026), while activation steering establishes that latent directions can causally shape behavior (Turner et al., 2024; Zou et al., 2025; Rimsky et al., 2024). This predicts that social granularity should surface as a single, ordered direction in activation space.

We construct 75 roles spanning five granularity levels, collect responses to shared questions under multiple prompts, and average their hidden states into role vectors. The *Granularity Axis* contrasts mean macro- and micro-role vectors (Figure 1). Three findings establish its significance. First, it captures the *dominant geometric axis* of role space: Qwen3-8B axis–PC1 cosine is 0.972, PC1 explains 52.6% of variance, and projections increase across all five levels. Second, the ordering survives layer, endpoint, prompt, held-out, and quality-filter variations, transfers to Llama-3.1-8B, and persists in Qwen3-32B and Llama-3.1-70B. Third, the axis is *behaviorally causal*: activation interventions shift output granularity in the predicted direction, with model-dependent strength and stability.

Independent five-rater validation, identical teacher-forced continuations, and matched affect/formality controls further establish the ordering (§3.3–3.4). These findings identify social granularity as a *representational primitive*, beyond a stylistic surface variable: role conditioning operates over a continuous social-scale manifold, and intervening on that direction makes scale tunable. They motivate auditing multi-agent simulations for granularity confusion, controlling social scale in

Algorithm 1: Granularity Axis: construction, validation, and steering.

Input: Language model $\mathcal{M}$; ordered role set $\mathcal{U}$ with levels $y(u) \in \{1, \ldots, 5\}$; prompt variants $\mathcal{S}$; shared question set $\mathcal{Q}$; analysis layer ℓ; intervention layer ℓ^*; steering strength α.

Output: Granularity Axis $g^{(\ell)}$; role projections $\{p_u\}$; PC1 alignment $\cos(g^{(\ell)}, w_1)$; steered generations.

// Stage 1: response collection (§2.2)

for $(u, s, q) \in \mathcal{U} \times \mathcal{S} \times \mathcal{Q}$ **do** generate $r_{u,s,q} \sim \mathcal{M}(\cdot \mid u, s, q)$;

// Stage 2: role-level representations (§2.3)

for *each response* (u, s, q) **do** $v^{(\ell)}_{u,s,q} \leftarrow \frac{1}{T} \sum_{t=1}^{T} h^{(\ell)}_t$;

for *each role* u **do** $v^{(\ell)}_u \leftarrow \frac{1}{|\mathcal{R}(u)|} \sum_{(s,q) \in \mathcal{R}(u)} v^{(\ell)}_{u,s,q}$;

// Stage 3: contrast-based axis (§2.3)

$g^{(\ell)} \leftarrow \frac{1}{|\mathcal{U}_{\text{macro}}|} \sum_{u \in \mathcal{U}_{\text{macro}}} v^{(\ell)}_u - \frac{1}{|\mathcal{U}_{\text{micro}}|} \sum_{u \in \mathcal{U}_{\text{micro}}} v^{(\ell)}_u$;

// Stage 4: geometric validation (§2.3)

form $V^{(\ell)} = [v^{(\ell)}_u]_{u \in \mathcal{U}}$, center, apply PCA $\tilde{V}^{(\ell)} = U \Sigma W^\top$; report $\cos(g^{(\ell)}, w_1)$ and $p_u = \langle v^{(\ell)}_u, g^{(\ell)} / \|g^{(\ell)}\| \rangle$;

// Stage 5: causal probe via activation steering (§2.4)

for *generated token* t **do** $\hat{h}^{(\ell^*)}_t \leftarrow h^{(\ell^*)}_t + \alpha\, g^{(\ell^*)}$;

personal-support or policy dialogues, and extending contrast-and-project to formality, time horizon, and risk aversion.

2 The Granularity Axis

We construct a contrast-based direction and validate its geometry and causal effects (Algorithm 1).

2.1 Problem Setting

For language model $\mathcal{M}$ with hidden dimension d, let u be a role with level $y(u) \in \{1, \ldots, 5\}$ (micro to macro), s a role prompt, q a shared question, and r a generated response. We seek a direction in the activations induced by (u, s, q) that is (i) representationally meaningful, (ii) aligned with the dominant geometry of role space, and (iii) causally relevant under steering.

2.2 Ordered Social Roles and Response Collection

Our 75 roles span five levels, with 15 per level: *Individual*, *Group/Community*, *Organization*, *Institution*, and *Nation/Super-Actor*. Examples are Worried Parent, Community Organizer, Hospital Administrator, Central Bank Governor, and World Bank President (Table 15). Levels are ordinal, non-equidistant, and defined by represented entity, constituency, and routine decision scope rather than emotion, prestige, or style. The inventory was LLM-assisted and reviewed by one author; independent five-rater validation is reported separately (Appendix D).

Five prompt variants change instruction style while preserving role identity: direct assignment, role-play, worldview/priority, first-person scale/time-horizon, and authenticity/constraints (Figure 4). Combining each with 240 shared questions from Lu et al. (2026) yields $75 \times 5 \times 240 = 90{,}000$ role-conditioned responses and $5 \times 240 = 1{,}200$ default responses per model, with $r \sim \mathcal{M}(\cdot \mid u, s, q)$. Role-adherence scores on a 0–3 scale support quality-filter ablations (Figure 5).

2.3 Role Representations and the Granularity Axis

At each layer ℓ, we mean-pool assistant-token states, $v^{(\ell)}_{u,s,q} = \frac{1}{T} \sum_{t=1}^{T} h^{(\ell)}_t$, and average the response set $\mathcal{R}(u)$ into $v^{(\ell)}_u = \frac{1}{|\mathcal{R}(u)|} \sum_{(s,q) \in \mathcal{R}(u)} v^{(\ell)}_{u,s,q}$. Following the Assistant Axis contrast construction, let $\mathcal{U}_{\text{micro}} = \{u : y(u) \in \{1, 2\}\}$ and $\mathcal{U}_{\text{macro}} = \{u : y(u) \in \{4, 5\}\}$. Their mean difference defines

$g^{(\ell)} = \frac{1}{|\mathcal{U}_{\text{macro}}|}\sum_{u\in\mathcal{U}_{\text{macro}}} v_u^{(\ell)} - \frac{1}{|\mathcal{U}_{\text{micro}}|}\sum_{u\in\mathcal{U}_{\text{micro}}} v_u^{(\ell)}$. Only L3 is wholly excluded from this contrast. We test whether projections rise monotonically across all five levels.

We center the 75-role matrix $V^{(\ell)} \in \mathbb{R}^{75\times d}$ and compute PCA, $\tilde{V}^{(\ell)} = U\Sigma W^\top$, testing axis–PC1 cosine and correlation of projections $p_u = \langle v_u^{(\ell)}, g^{(\ell)}/\|g^{(\ell)}\|\rangle$ with $y(u)$. The default is excluded from PCA and contrast construction. PC1 share describes the role matrix; it differs from variance explained by the contrast itself. Alternative endpoints, the Assistant Axis, and random directions provide controls.

Monotonic means test group order; role-level Spearman measures agreement across individual roles. Axis–PC1 cosine tests whether the labeled contrast recovers the dominant unlabeled direction, while PC1 share measures variance concentration. For scaling, the participation ratio $(\sum_k \lambda_k)^2/\sum_k \lambda_k^2$ of PCA eigenvalues λ_k increases as variance spreads across components. These metrics distinguish a persistent scale direction from an entirely one-dimensional role space.

2.4 Activation Steering

At intervention layer ℓ^*, selected by a layer sweep, we add $\alpha g^{(\ell^*)}$ to each generated-token activation: $\hat{h}_t^{(\ell^*)} = h_t^{(\ell^*)} + \alpha g^{(\ell^*)}$. Positive α pushes toward institutional/systemic reasoning; negative α toward individual/local reasoning. Prompt encoding is unchanged. We measure direction, strength, symmetry, and stability at fixed prompts.

2.5 Controlled Forward-pass Validation

In both 8B models, we extract final role-prompt, final question, pre-first-assistant, and pooled identical teacher-forced continuation states. Continuation tokens and lengths are fixed across roles, without free generation or role-specific truncation. Eight domain chains cross five scope levels with affect/formality conditions and 24 questions. Axes trained on seven neutral domains are evaluated on the eighth. Regression includes level, affect, formality, level-by-affect/formality interactions, and domain effects; cluster bootstrap estimates uncertainty and joint scale partial R^2 (Appendix F).

The original-role test uses 48 deterministic questions and five prompts: 18,000 inputs per model. Cached default responses supply at most 256 identical continuation tokens. Contrasts are rebuilt at each position, testing scale encoding before generation and during identical-token processing without assuming one shared direction. Qwen's role-prompt axis, for example, has cosine 0.160 with its response-mean axis despite strong ordering. The matched-domain experiment separately tests transfer beyond the domains used to construct its axis.

3 Experiments

3.1 Experimental Setup

We apply the same pipeline to Qwen3-8B and Llama-3.1-8B-Instruct, then Qwen3-32B and Llama-3.1-70B-Instruct (§3.3). Each model has 1,200 responses per role/default condition (§2.2). Layer 18 lies in the stable middle-layer regime of both 8B models.

Steering uses Layer 18, coefficients $\{-4, 0, +4\}$, and greedy decoding on 40 *generic* social-policy prompts and 12 *micro-targeted* prompts admitting local responses (Appendix E). The latter expose effects masked by Qwen's macro-oriented generic baseline. The primary judge is `gpt-5.4-mini` (OpenAI, 2026), checked against `gemini-3.1-flash-lite-preview` (Google DeepMind, 2026); Figure 6 gives the rubric. Appendix E records compute, questions, and licenses.

3.2 Representation Results

Figures 2–3 and Table 1 show a dominant micro-to-macro organization. At Layer 18, Qwen/Llama axis–PC1 cosines are $0.9720/0.9596$, PC1 shares are $52.57\%/42.46\%$, and Spearman/Pearson exceed 0.93. Mean projections increase through L1–L5, with saturation at L4–L5 (Table 2). Qwen exhibits the more concentrated geometry. Default assistants lie near L3 (Qwen) and L4 (Llama), helping explain steering asymmetry (Appendix C.5). These results satisfy criteria (i)–(ii).

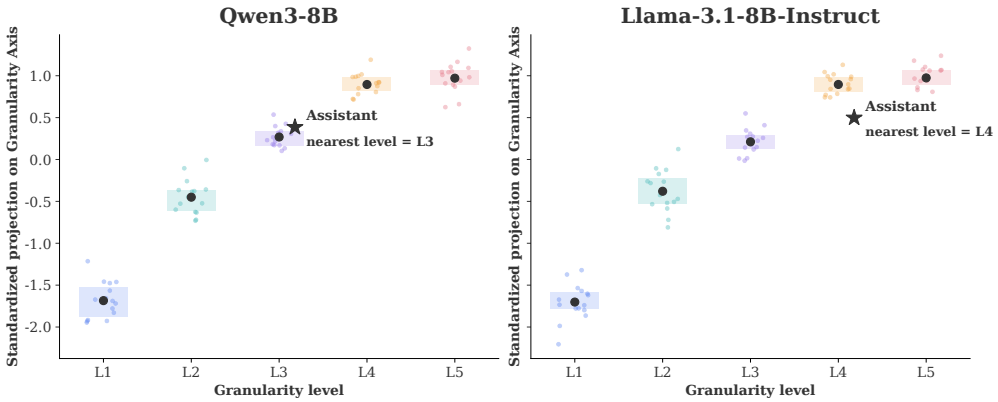

Figure 3: **Ordered projections on the Granularity Axis.** Points are roles grouped by granularity level; black circles mark level means, shaded bands within-level spread, stars the default assistant. Projections rise monotonically L1→L5 in both models; the default sits in a meso-to-macro region (near L3 in Qwen3-8B, L4 in Llama-3.1-8B-Instruct).

Table 1: **Main representation results at the target layer.** In both models, the Granularity Axis aligns with PC1 and strongly correlates with the five-level ordering; the default assistant sits in a meso-to-macro region. Asst. Level cells reuse the L1→L5 tints of Table 2.

Model	Layer	PC1 Var.	$\cos(g, \mathrm{PC1})$	Spearman	Pearson	Asst. Level
Qwen3-8B	18	52.57%	0.9720	0.9472	0.9414	L3
Llama-3.1-8B-Instruct	18	42.46%	0.9596	0.9459	0.9373	L4

3.3 Larger Models and Independent Taxonomy Validation

The scaling comparison retains the same roles, prompts, and questions, with equal response weights and no quality filter. Primary layers are selected by relative depth: Qwen Layer 18 of 36 maps to Layer 32 of 64, and Llama Layer 18 of 32 maps to Layer 45 of 80. Target layers are therefore fixed before inspecting the larger-model results.

At pre-specified relative depths (Qwen3-32B Layer 32; Llama-70B Layer 45), ordering and axis–PC1 alignment persist (Table 3). PC1 share falls from 52.57% to 45.99% in Qwen and 42.46% to 31.23% in Llama, while effective rank rises from 3.39 to 4.42 and 5.03 to 8.56. Ordering is monotonic at $59/64$ and $80/80$ layers, respectively: the axis persists as role geometry becomes more distributed. Both larger defaults lie nearest L3 (Appendix G.1).

Five independent blinded raters yield ordinal Krippendorff's $\alpha = 0.961$ (95% CI $[0.933, 0.979]$), pairwise quadratic-weighted $\kappa = 0.964$, and 73.3% unanimous agreement. Median labels equal strict-majority labels for every role and agree with original labels on 88%, without author-label tie-breaking. All four models remain monotonic under consensus labels (Table 3); all 26,244 pre-specified boundary assignments preserve ordering in both 8B models (Appendix D).

Raters saw descriptions without original labels or model results and were not asked to balance levels. Nine roles move upward by one level, yielding consensus counts $15/11/18/12/19$. Recomputing contrasts from these labels, with PCA over the unchanged role matrix, tests whether ordering depends on the original assignments. The separate boundary analysis tests specified reassignments rather than independent human judgments.

3.4 Construct Validity and Additional Geometric Controls

With identical continuations, Qwen/Llama Spearman remains $0.934/0.964$: generated-answer lexical variation is unnecessary for ordering. Matched-domain continuation scale coefficients are 0.979 $[0.898, 1.064]$ and 0.253 $[0.238, 0.267]$, with joint scale partial $R^2 = 0.036/0.153$ after

Table 2: **Mean projections onto the Granularity Axis per granularity level.** Both models exhibit monotonic L1→L5 ordering with saturation between L4 and L5, despite very different absolute scales. Column tints follow the micro→macro spectrum used in Tables 4 and 5.

Model	L1	L2	L3	L4	L5
Qwen3-8B	0.5425	11.1058	17.2525	22.6148	23.2603
Llama-3.1-8B-Instruct	−0.8784	1.2059	2.1362	3.2149	3.3386

Table 3: **Ordering persists while role geometry becomes more distributed.** All four instruction-tuned checkpoints have strictly increasing level means under original and median-consensus labels. PC1 share and effective rank use all 75 role vectors and are label-invariant. Large-model layers were fixed by relative depth before their analyses; Llama names abbreviate the Instruct checkpoints.

Model	Layer	PC1 %	Rank	Original labels		Consensus
				$\lvert\cos(g, w_1)\rvert$	ρ	$\lvert\cos\rvert$ / ρ
Qwen3-8B	18	52.57	3.39	.9720	.9472	.9830 / .9358
Qwen3-32B	32	45.99	4.42	.9814	.9472	.9893 / .9384
Llama-3.1-8B	18	42.46	5.03	.9596	.9459	.9735 / .9415
Llama-3.1-70B	45	31.23	8.56	.9633	.9655	.9777 / .9639

affect/formality controls. Across all eight model–position comparisons, held-out-domain mean Spearman spans 0.887–1.000 (Appendix F).

The original-role test fixes continuation content, whereas the matched-domain factorial crosses scope with affect and formality. Each fit contains 3,840 observations. Partial R^2 jointly tests level and its two attribute interactions; confidence intervals resample domain–question clusters, retaining all levels and conditions. Thus the continuation test and factorial regression provide complementary evidence rather than two estimates of the same quantity.

Joint partial R^2 falls from $0.937/0.743$ at the role-prompt position to $0.036/0.153$ during continuations, while continuation held-out-domain mean Spearman remains $0.925/1.000$. Strong ordinal transfer thus coexists with lower scale-related variance at some positions: ordering and explained variation measure different properties of the representation.

Granularity/Assistant cosine is $0.2253/0.3492$; removing Assistant preserves ordering ($\rho = 0.9481/0.9533$). Removing Granularity leaves residual PC1 with $12.80\%/12.06\%$ of original variance. Permutation tests support axis–PC1 alignment, while rebuilt-contrast rank and L4–L5 separation tests are nonsignificant; Appendix G reports all controls and residual structure.

After Assistant removal, residual Granularity–PC1 cosine remains $0.9399/0.8955$, whereas Assistant after Granularity removal is nonmonotonic. This establishes ordering beyond the Assistant direction without requiring orthogonality. Fixed-projection permutation tests condition on the observed axis; rebuilt-contrast nulls also account for endpoint selection, explaining why their rank evidence is weaker.

3.5 Steering Results

Table 4 reports judged granularity with prompt-level SEM (higher = more macro). Manually selected examples illustrate the direction (Table 5; Appendix A). Qwen's generic baseline is already $4.9/5$, but micro-targeted prompts show a $+0.5000$ shift under $\alpha = +4$, without degeneration. Llama shifts more strongly ($2.0000 \rightarrow 3.1667$ on micro-targeted prompts); on generic prompts, $\alpha = -4$ lowers its score from 4.5250 to 3.1250 with 42.5% degeneration. Thus the axis is behaviorally causal, with model-dependent controllability. Appendix B gives filtered results and coefficient sweeps.

Direction specificity. Assistant-axis and random-direction baselines do not reproduce the micro–macro shifts, showing direction-specific effects rather than generic activation perturbation. Criterion (iii) holds in a partial, model-dependent form across all four model/prompt cells. The 12 micro-

Table 4: **Steering shifts judged output granularity in a model- and prompt-dependent way.** Mean ± SEM `granularity_overall` from `gpt-5.4-mini` (1–5; higher = more macro), with SEM over prompts. $\Delta_{\pm 4}$ are deltas from the unsteered baseline; Deg. reports degeneration rates at $\alpha = -4, 0, +4$ (filtered scores in Appendix B). Tints follow the micro-to-macro spectrum of Table 5.

Model	Prompt set	$\alpha = -4$	$\alpha = 0$	$\alpha = +4$	Δ_{-4}	Δ_{+4}	Deg. $(-/0/+)$
Qwen3-8B	Generic	4.821±.062	4.900±.048	4.975±.025	−0.080	+0.075	.000/.000/.000
	Micro-targeted	2.000±.000	2.167±.112	2.667±.256	−0.167	+0.500	.000/.000/.000
Llama-3.1-8B-Instruct	Generic	3.125±.161	4.525±.088	4.900±.048	−1.400	+0.375	.425/.000/.025
	Micro-targeted	1.583±.149	2.000±.000	3.167±.241	−0.417	+1.167	.000/.083/.000

Table 5: Manually selected illustrative responses under no steering and granularity steering; these cases are not used for quantitative conclusions. Generic-setting degeneration rates for $\alpha = -4/0/+4$ are $0/0/0$ (Qwen) and $0.425/0/0.025$ (Llama). Excerpts are quoted verbatim from the model outputs, with omissions indicated by '[...]'. Boldface marks the most salient phrases revealing granularity, with blue for micro-leaning content and orange for macro-leaning content. Llama-3.1-8B-Instruct shows a clearer micro-to-macro shift, whereas Qwen3-8B remains comparatively macro-oriented across all three conditions.

	← **Micro** ($\alpha = -4$)	**Base** ($\alpha = 0$)	**Macro** ($\alpha = +4$) →
	Prompt: ***How should we handle the mental health crisis?***		
Llama-3.1-8B-Instruct	Get help when you need it: If you're struggling with your mental health, don't be afraid to **reach out to a friend, family member, or a professional for help**. [...] Get help at school: [...] **You can talk to a teacher, a counselor, or even a school nurse**. [...] Take care of your family.	Handling the mental health crisis requires a multi-faceted approach that involves **individuals, communities, organizations, and governments working together**. [...] Address systemic issues. [...] **social determinants of mental health, such as poverty, inequality, and social isolation**. [...] Invest in research and innovation.	Addressing the mental health crisis requires a comprehensive and multi-faceted approach that involves **various stakeholders, including governments, healthcare systems, communities, and individuals**. [...] **Public education campaigns, media outreach, and social media initiatives**. [...] **Mental health parity laws and regulations**.
Qwen3-8B	Handling the mental health crisis requires a comprehensive, multi-faceted approach that involves individuals, communities, healthcare systems, and policymakers. [...] **Build Support Networks**: Encourage people to maintain strong social connections and seek help when needed. [...] **Support Families and Caregivers**: [...] Offer resources to help families understand and support their loved ones.	Handling the mental health crisis requires a comprehensive, multi-faceted approach that addresses both **individual and systemic challenges**. [...] **Governments and organizations should invest in mental health infrastructure**. [...] **Social Determinants of Health**: Tackle issues like poverty, discrimination, housing instability, and trauma.	Handling the mental health crisis requires a comprehensive, multi-faceted approach that addresses both individual and systemic challenges. [...] **Increase Funding**: Allocate more resources to mental health care systems. [...] **Address Social Determinants of Health**: Address systemic issues like poverty, discrimination, and lack of access to education. [...] requires a coordinated effort across **governments, healthcare systems, communities, and individuals**.

targeted prompts were drafted with Opus 4.5 and manually revised; positive leave-one-out means remain positive across both judges (0.364–0.546 for Qwen; 0.909–1.273 for Llama). Appendix B.1 details selection provenance, individual outcomes, and judge pairing.

Filtering Llama generic outputs at $\alpha = -4$ leaves a mean of 3.000 versus baseline 4.525, but retains only $23/40$ responses; positive steering retains $39/40$ with mean 4.897. Qwen's judged outputs show no degeneration at these coefficients. Reporting both full and filtered scores separates directional sensitivity from usable control (Table 8).

3.6 Robustness and Controls

Ordering is monotonic across Layers 8–35 (Qwen) and 6–31 (Llama), and alternative endpoints retain axis–PC1 cosine ≥ 0.9279 (Appendix C.1–C.2). Held-out prompts/questions remain strong; Qwen role holdout is slightly fragile. All prompt variants remain monotonic, including identity-only

Table 6: **Human evaluation of steering outputs.** *Pairwise rate*: annotator selection rate for $\alpha = +4$ over $\alpha = -4$ as more macro (24 triplets per cell, 3 annotators, 72 judgments per cell; reported Wilson 95% CI, without adjustment for shared items/raters). Micro = micro-targeted prompts. *Likert ρ / Quad-κ*: `gpt-5.4-mini` versus annotator-mean ratings on the same 1–5 rubric (15 items per cell). *Krip. α*: Krippendorff's α on those Likert items.

Model	Prompts	Pairwise rate	Likert ρ	Quad-κ	Krip. α
Qwen3-8B	Generic	0.639 [.52, .74]	0.58	0.49	0.61
	Micro	0.750 [.64, .84]	0.61	0.54	0.66
Llama-3.1-8B-Instruct	Generic	0.903 [.81, .95]	0.79	0.71	0.78
	Micro	0.861 [.76, .92]	0.68	0.60	0.71

wording. Score filtering, generic/specific roles, and domain/family controls support robustness, with health/economy exceptions (Appendix C.3–C.8).

Held-out prompt/question Spearman is $0.9468/0.9494$ for Qwen and $0.9403/0.9468$ for Llama, with monotonic means. Role holdout gives 0.9335 in both, but Qwen is not perfectly monotonic: high correlation does not imply complete group ordering. Tightening role-adherence thresholds from all responses to scores ≥ 2 and ≥ 3 changes Qwen Spearman from 0.9472 to 0.9481 and 0.9564, and Llama from 0.9459 to 0.9459 and 0.9503. The structure thus survives both held-out inputs and stricter role-adoption filters (Table 14).

3.7 Human Evaluation

Three graduate-level annotators, blinded to model and coefficient, used the judge rubric (Figure 6). With 24 triplets and 72 judgments per model/prompt cell, macro-side selections exceed chance in all four cells: above 0.85 for Llama and 0.639 (Wilson 95% CI $[0.52, 0.74]$) for Qwen generic. Likert re-ratings of 15 items per cell give human–judge Spearman 0.58–0.79 (Table 6). These historical aggregates corroborate steering separately from the five-rater taxonomy study.

4 Analysis and Limitations

4.1 Representation Should Be Validated Before Control

The endpoint contrast recovers five-level ordering, including the excluded L3. Independent labels and held-out domains validate this structure beyond the original taxonomy; identical continuations separate role-dependent activations from generated-answer variation. Steering shifts are smaller and more context- and model-dependent than representation-level ordering. A direction can therefore be representationally robust while behaviorally fragile.

> ***Takeaway:*** *Validate representation with held-out levels/domains and independent labels before interpreting steering.*

4.2 Default Placement and Headroom Gate Steering Visibility

Qwen's generic baseline of $4.9/5$ leaves little positive headroom; micro-targeted prompts reveal its controllability. Qwen is conservative but stable, whereas Llama responds more strongly but degenerates more often. Larger shifts alone do not imply cleaner control: baseline placement and stability jointly determine visible steering effects.

Geometric default placement and behavioral headroom measure different quantities: the default representation's position in role space versus answer scores on a particular prompt set. Qwen lies nearest L3 geometrically yet averages $4.9/5$ on generic prompts. Default projections alone therefore cannot predict steering scores; the generic/micro-targeted comparison and degeneration-filtered outcomes reveal complementary aspects of controllability.

> ***Takeaway:*** *Interpret steering alongside baseline placement, coefficient sweeps, and degeneration rates.*

4.3 What a Single Contrast Axis Can and Cannot Tell You

Removing Granularity eliminates $50.23\%/39.82\%$ of original Qwen/Llama variance, yet leaves sizeable residual components. Together with higher effective rank in the larger models, this shows how a dominant scale direction can persist within a richer role representation. PC1 alignment, variance concentration, and behavioral controllability measure different properties and should be interpreted together.

A dominant axis need not exhaust role variation: residual PCA reveals additional structure (Appendix G.3). Scope remains correlated with authority, institutional vocabulary, abstraction, and time horizon. Our four checkpoints cover two families and an English-centric taxonomy; levels are ordinal and retain L4–L5 saturation.

> ***Takeaway:*** *A single contrast axis is a useful first cut, not a final ontology; residual structure motivates multi-axis analysis.*

5 Related Work

LLMs as Social Simulators and Role-Conditioned Agents. LLM social simulation spans multi-agent systems, politics, health, and markets (Park et al., 2023; Argyle et al., 2022; Mou et al., 2024; Anthis et al., 2025; Wang et al., 2025a; Wu et al., 2023; Li et al., 2023a; Hong et al., 2023; Qian et al., 2023; Törnberg et al., 2023; Williams et al., 2023; Li et al., 2024b; 2023d). These applications rely on role assignments to represent heterogeneous perspectives and decision scopes. Persona research examines fidelity, consistency, personality, and behavioral bias (Shao et al., 2023; Wang et al., 2024b; Jiang et al., 2024; Wang et al., 2023; Tu et al., 2024; Kong et al., 2023; Tseng et al., 2024; Chen et al., 2025a; Li et al., 2024a; Wang et al., 2024a; Salewski et al., 2023; Gupta et al., 2023; Cheng et al., 2023; Li et al., 2023b; Joshi et al., 2023; Beckmann & Butlin, 2026). Social granularity adds a complementary dimension, spanning individual concerns, organizational decisions, and broader institutional responsibilities. We study the internal organization of roles by social scale.

Interpretable Social and Semantic Directions. *The Geometry of Culture* (Kozlowski et al., 2019) and SemAxis (An et al., 2018) extract interpretable social/semantic directions from word embeddings. We identify an ordered role-scale direction in LLM activations and establish its PC1 alignment, cross-model stability, and behavioral effects.

Activation Steering and Representation Engineering. Linear representation studies (Mikolov et al., 2013; Elhage et al., 2022; Belinkov, 2021; Park et al., 2024b;a; Sakata et al., 2026), sparse autoencoders (Cunningham et al., 2023; Gao et al., 2024; Marks et al., 2024; Wang et al., 2025b), and nonlinear counterexamples (Engels et al., 2024) characterize activation geometry. Steering methods manipulate behavior through latent directions (Turner et al., 2024; Li et al., 2023c; Rimsky et al., 2024; Zou et al., 2025; Bartoszcze et al., 2025; Liu et al., 2024; Arditi et al., 2024; Chen et al., 2025b; Jorgensen et al., 2023; Chen et al., 2026; Feng et al., 2026). The closest precedent, the Assistant Axis (Lu et al., 2026), tracks assistant-persona departure; our axis tracks social granularity.

Micro–Macro Theory and Social Scale. Sociological accounts connect individual action and collective outcomes (Coleman et al., 1990; Schelling, 1978; Granovetter, 1978), agency and structure (Giddens, 1984), and nested social systems (Bronfenbrenner, 1979). They motivate testing whether prompted roles form an internal social-scale continuum.

6 Conclusion

The *Granularity Axis* organizes prompted social roles from individual to macro-level reasoning. It aligns with the dominant geometry of role space, orders the excluded L3, transfers across families and scales through 70B, and causally shifts response granularity. Independent labels and controlled inputs reinforce these findings. Social granularity is a representational primitive, with role conditioning organized along a continuous social-scale manifold. Representation is robust; control depends on the model's default regime. This enables auditing simulations for granularity confusion, controlling social scale, and extending contrast-and-project to other graded social dimensions.

AI USE STATEMENT

We used generative AI tools for language editing and stylistic polishing, as well as to assist in constructing the initial role inventory and drafting diagnostic prompts. All AI-assisted research materials were manually reviewed and revised by the authors before use, and all AI-assisted text was verified for technical accuracy. The authors take responsibility for the final content, methodology, and results of the paper.

ETHICS STATEMENT

This work includes two human-evaluation components: three volunteer PhD students from our lab evaluated activation-steering outputs, and five volunteer PhD students independently conducted a blinded validation of the proposed social-granularity taxonomy. These tasks involved judgments of model outputs or predefined role labels and did not require sensitive personal information. As the studied roles include individuals, social groups, institutions, and nations, model representations may reflect biases or stereotypes present in pretrained models. Our findings characterize model representations and should not be interpreted as intrinsic properties of real-world people or groups.

REPRODUCIBILITY STATEMENT

We provide an anonymous repository containing the code, experimental configurations, prompts, role taxonomy, evaluation resources, and instructions needed to reproduce our experiments: https://anonymous.4open.science/r/Granularity-Axis-anon/. Additional implementation and experimental details are provided in the main paper and appendices.

REFERENCES

Jisun An, Haewoon Kwak, and Yong-Yeol Ahn. Semaxis: A lightweight framework to characterize domain-specific word semantics beyond sentiment. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pp. 2450–2461. Association for Computational Linguistics, 2018. doi: 10.18653/V1/P18-1228. URL https://aclanthology.org/P18-1228/.

Jacy Reese Anthis, Ryan Liu, Sean M. Richardson, Austin C. Kozlowski, Bernard Koch, Erik Brynjolfsson, James A. Evans, and Michael S. Bernstein. Position: LLM social simulations are a promising research method. In *Forty-second International Conference on Machine Learning, ICML*, 2025.

Anthropic. Claude model card. https://www.anthropic.com/system-cards, 2026.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Rimsky, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *ArXiv*, abs/2406.11717, 2024. URL https://api.semanticscholar.org/CorpusID:270560489.

Lisa P. Argyle, E. Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31:337 – 351, 2022. URL https://api.semanticscholar.org/CorpusID:252280474.

Lukasz Bartoszcze, Sarthak Munshi, Bryan Sukidi, Jennifer Yen, Zejia Yang, David Williams-King, Linh Le, Kosi Asuzu, and Carsten Maple. Representation engineering for large-language models: Survey and research challenges. *arXiv preprint, arXiv:2502.17601*, 2025.

Pierre Beckmann and Patrick Butlin. Where is the mind? persona vectors and llm individuation. 2026. URL https://api.semanticscholar.org/CorpusID:287635493.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48:207–219, 2021. URL https://api.semanticscholar.org/CorpusID:236924832.

James Bisbee, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, 2024. URL https://api.semanticscholar.org/CorpusID:269845858.

Urie Bronfenbrenner. *The ecology of human development: Experiments by nature and design*. Harvard university press, 1979.

Jiaqi Chen, Ming Wang, Tingna Xie, Shi Feng, and Yongkang Liu. A systematic analysis of the impact of persona steering on llm capabilities. 2026. URL https://api.semanticscholar.org/CorpusID:287432603.

Nuo Chen, Yan Wang, Yang Deng, and Jia Li. The oscars of ai theater: A survey on role-playing with language models. *arXiv preprint, arXiv:2407.11484*, 2025a.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. Persona vectors: Monitoring and controlling character traits in language models. *ArXiv*, abs/2507.21509, 2025b. URL https://api.semanticscholar.org/CorpusID:280337840.

Myra Cheng, Esin Durmus, and Dan Jurafsky. Marked personas: Using natural language prompts to measure stereotypes in language models. *ArXiv*, abs/2305.18189, 2023. URL https://api.semanticscholar.org/CorpusID:258960243.

James Samuel Coleman et al. *Foundations of social theory*, volume 302. HeinOnline, 1990.

Hoagy Cunningham, Aidan Ewart, Logan Riggs Smith, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *ArXiv*, abs/2309.08600, 2023. URL https://api.semanticscholar.org/CorpusID:261934663.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Thomas Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Baker Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Chris Olah. Toy models of superposition. *ArXiv*, abs/2209.10652, 2022. URL https://api.semanticscholar.org/CorpusID:252439050.

Joshua Engels, Isaac Liao, Eric J. Michaud, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. In *International Conference on Learning Representations*, 2024. URL https://api.semanticscholar.org/CorpusID:269983112.

Xiachong Feng, Liang Zhao, Weihong Zhong, Yi-Chong Huang, Yuxuan Gu, Lingpeng Kong, Xiaocheng Feng, and Bing Qin. Persona: Dynamic and compositional inference-time personality control via activation vector algebra. *ArXiv*, abs/2602.15669, 2026. URL https://api.semanticscholar.org/CorpusID:285659291.

Leo Gao, Tom Dupr'e la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. *ArXiv*, abs/2406.04093, 2024. URL https://api.semanticscholar.org/CorpusID:270286001.

Anthony Giddens. *The constitution of society: Outline of the theory of structuration*. Univ of California Press, 1984.

Google. Gemini 3 model card. https://deepmind.google/models/gemini, 2025.

Google DeepMind. Gemini 3.1 flash-lite model card. https://deepmind.google/models/model-cards/gemini-3-1-flash-lite/, 2026.

Mark S. Granovetter. Threshold models of collective behavior. *American Journal of Sociology*, 83:1420 – 1443, 1978. URL https://api.semanticscholar.org/CorpusID:49314397.

Shashank Gupta, Vaishnavi Shrivastava, A. Deshpande, A. Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. Bias runs deep: Implicit reasoning biases in persona-assigned llms. *ArXiv*, abs/2311.04892, 2023. URL https://api.semanticscholar.org/CorpusID:265050702.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zi Hen Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. Metagpt: Meta programming for a multi-agent collaborative framework. In *International Conference on Learning Representations*, 2023. URL https://api.semanticscholar.org/CorpusID:265301950.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, and Jad Kabbara. Personallm: Investigating the ability of large language models to express big five personality traits. 2024. URL https://api.semanticscholar.org/CorpusID:265221392.

Ole Jorgensen, Dylan Cope, Nandi Schoots, and Murray Shanahan. Improving activation steering in language models with mean-centring. *ArXiv*, abs/2312.03813, 2023. URL https://api.semanticscholar.org/CorpusID:266053529.

Nitish Joshi, Javier Rando, Abulhair Saparov, Najoung Kim, and He He. Personas as a way to model truthfulness in language models. *ArXiv*, abs/2310.18168, 2023. URL https://api.semanticscholar.org/CorpusID:264555113.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, and Xiaoxia Zhou. Better zero-shot reasoning with role-play prompting. In *North American Chapter of the Association for Computational Linguistics*, 2023. URL https://api.semanticscholar.org/CorpusID:260900230.

Austin C. Kozlowski, Matt Taddy, and James A. Evans. The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, 84(5):905–949, 2019. ISSN 1939-8271. doi: 10.1177/0003122419877135. URL http://dx.doi.org/10.1177/0003122419877135.

G. Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for "mind" exploration of large language model society. *Advances in Neural Information Processing Systems 36*, 2023a. URL https://api.semanticscholar.org/CorpusID:268042527.

Junyi Li, Ninareh Mehrabi, Charith Peris, Palash Goyal, Kai-Wei Chang, A. G. Galstyan, Richard Zemel, and Rahul Gupta. The steerability of large language models toward data-driven personas. In *North American Chapter of the Association for Computational Linguistics*, 2023b. URL https://api.semanticscholar.org/CorpusID:265067297.

Kenneth Li, Oam Patel, Fernanda B. Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*, 2023c.

Kenneth Li, Tianle Liu, Naomi Bashkansky, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Measuring and controlling instruction (in)stability in language model dialogs. *arXiv preprint, arXiv:2402.10962*, 2024a.

Nian Li, Chen Gao, Mingyu Li, Yong Li, and Qingmin Liao. Econagent: large language model-empowered agents for simulating macroeconomic activities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15523–15536, 2024b.

Yang Li, Yangyang Yu, Haohang Li, Z. Chen, and Khaldoun Khashanah. Tradinggpt: Multi-agent system with layered memory and distinct characters for enhanced financial trading performance. 2023d. URL https://api.semanticscholar.org/CorpusID:261582775.

Ryan Liu, Jiayi Geng, Joshua C. Peterson, Ilia Sucholutsky, and Thomas L. Griffiths. Large language models assume people are more rational than we really are. In *The Thirteenth International Conference on Learning Representations,ICLR*. OpenReview.net, 2025. URL https://openreview.net/forum?id=dAeET8gxqg.

Wenhao Liu, Xiaohua Wang, Muling Wu, Tianlong Li, Changze Lv, Zixuan Ling, Jianhao Zhu, Cenyuan Zhang, Xiaoqing Zheng, and Xuanjing Huang. Aligning large language models with human preferences through representation engineering. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, ACL*, pp. 10619–10638. Association for Computational Linguistics, 2024.

AI @ Meta Llama Team. The llama 3 herd of models. *arXiv preprint, arXiv:2407.21783*, 2024.

Christina Lu, Jack Gallagher, Jonathan Michala, Kyle Fish, and Jack Lindsey. The assistant axis: Situating and stabilizing the default persona of language models. *arXiv preprint, arXiv:2601.10387*, 2026.

Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. *ArXiv*, abs/2403.19647, 2024. URL https://api.semanticscholar.org/CorpusID:268732732.

Tomas Mikolov, Wen tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *North American Chapter of the Association for Computational Linguistics*, 2013. URL https://api.semanticscholar.org/CorpusID:7478738.

Xinyi Mou, Xuanwen Ding, Qi He, Liang Wang, Jingcong Liang, Xinnong Zhang, Libo Sun, Jiayu Lin, Jie Zhou, Xuanjing Huang, and Zhongyu Wei. From individual to society: A survey on social simulation driven by large language model-based agents. *arXiv preprint, arXiv:2412.03563*, 2024.

OpenAI. Gpt-5 system card. https://cdn.openai.com/gpt-5-system-card.pdf, 2025.

OpenAI. Introducing gpt-5.4 mini and nano. https://openai.com/index/introducing-gpt-5-4-mini-and-nano/, 2026.

Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In Sean Follmer, Jeff Han, Jürgen Steimle, and Nathalie Henry Riche (eds.), *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, UIST*, pp. 2:1–2:22, 2023. doi: 10.1145/3586183.3606763.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. *ArXiv*, abs/2406.01506, 2024a. URL https://api.semanticscholar.org/CorpusID:270216615.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024b.

Cheng Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. Chatdev: Communicative agents for software development. In *Annual Meeting of the Association for Computational Linguistics*, 2023. URL https://api.semanticscholar.org/CorpusID:270257715.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL*, pp. 15504–15522. Association for Computational Linguistics, 2024.

Masaki Sakata, Benjamin Heinzerling, Takumi Ito, Sho Yokoi, and Kentaro Inui. Linear representations of hierarchical concepts in language models. 2026. URL https://api.semanticscholar.org/CorpusID:287255958.

Aadesh Salecha, Molly E. Ireland, Shashanka Subrahmanya, João Sedoc, Lyle H. Ungar, and Johannes C. Eichstaedt. Large language models show human-like social desirability biases in survey responses. *arXiv preprint, arXiv:2405.06058*, 2024.

Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. In-context impersonation reveals large language models' strengths and biases. *ArXiv*, abs/2305.14930, 2023. URL https://api.semanticscholar.org/CorpusID:258866192.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? *ArXiv*, abs/2303.17548, 2023. URL https://api.semanticscholar.org/CorpusID:257834040.

Thomas C. Schelling. Micromotives and macrobehavior. 1978. URL https://api.semanticscholar.org/CorpusID:143387748.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. Character-llm: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13153–13187, 2023.

Qwen Team. Qwen3 technical report. *arXiv preprint, arXiv:2505.09388*, 2025.

Petter Törnberg, Diliara Valeeva, Justus Uitermark, and Christopher Bail. Simulating social media using large language models to evaluate alternative news feed algorithms. *ArXiv*, abs/2310.05984, 2023. URL https://api.semanticscholar.org/CorpusID:263831233.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. Two tales of persona in llms: A survey of role-playing and personalization. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP2024*, pp. 16612–16631. Association for Computational Linguistics, 2024.

Quan Tu, Shilong Fan, Zihang Tian, and Rui Yan. Charactereval: A chinese benchmark for role-playing conversational agent evaluation. In *Annual Meeting of the Association for Computational Linguistics*, 2024. URL https://api.semanticscholar.org/CorpusID:266725287.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint, arXiv:2308.10248*, 2024.

Kuan Wang, Alexander Bukharin, Haoming Jiang, Qingyu Yin, Zhengyang Wang, Tuo Zhao, Jingbo Shang, Chao Zhang, Bing Yin, Xian Li, Jianshu Chen, and Shiyang Li. Rnr: Teaching large language models to follow roles and rules. *arXiv preprint, arXiv:2409.13733*, 2024a.

Noah Wang, Zy Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, et al. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14743–14777, 2024b.

Qian Wang, Jiaying Wu, Zichen Jiang, Zhenheng Tang, Bingqiao Luo, Nuo Chen, Wei Chen, and Bingsheng He. Llm-based human simulations have not yet been reliable. *arXiv preprint, arXiv:2501.08579*, 2025a.

Shunyu Wang, Tyler Loakman, Youbo Lei, Yi Liu, Bohao Yang, Yuting Zhao, Dong Yang, and Chenghua Lin. Exploring task performance with interpretable models via sparse auto-encoders. *ArXiv*, abs/2507.06427, 2025b. URL https://api.semanticscholar.org/CorpusID:280066641.

Xintao Wang, Yunze Xiao, Jen-Tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Annual Meeting of the Association for Computational Linguistics*, 2023. URL https://api.semanticscholar.org/CorpusID:264555532.

Ross Williams, Niyousha Hosseinichimeh, Aritra Majumdar, and Navid Ghaffarzadegan. Epidemic modeling with generative agents. *ArXiv*, abs/2307.04986, 2023. URL https://api.semanticscholar.org/CorpusID:259766713.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang (Eric) Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation. 2023. URL https://api.semanticscholar.org/CorpusID:263611068.

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, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to ai transparency. *arXiv preprint, arXiv:2310.01405*, 2025.

A QUALITATIVE EXAMPLES

Table 7 extends Table 5 with a coefficient-level qualitative sweep on a single shared generic prompt. We use the housing prompt because it admits both individual coping advice and institutional policy responses, making the micro-to-macro contrast easy to inspect. For auditability, every table cell contains only verbatim snippets from the corresponding model response, with '[...]' inserted only to mark omitted text. The prose below the table provides our interpretation and should not be read as quoted model text.

Table 7: **Appendix qualitative coefficient sweep.** We steer both models on the same prompt, "How should we deal with rising housing costs?", using coefficients $\alpha \in \{\pm 8, \pm 12, \pm 16, \pm 20, \pm 24, \pm 28, \pm 32, \pm 36\}$. Each table cell contains only verbatim snippets from the corresponding model response, with omissions indicated by '[...]'. Boldface marks the most salient phrases revealing granularity. Column tints follow the micro-to-macro spectrum used in Table 5.

$\lvert\alpha\rvert$	← Micro Steering $(-\alpha)$	Macro Steering $(+\alpha)$ →
Qwen3-8B		
Prompt: How should we deal with rising housing costs?		
8	Rising housing costs are a complex issue that affects people in many different ways, depending on their **location, income level, and personal circumstances**. [...] **Rent subsidies**: Programs like Section 8 in the U.S. help people pay a portion of their rent. [...] **Community Land Trusts**. [...] **Education and Financial Literacy**.	Rising housing costs are a complex and multifaceted issue that affects **individuals, communities, and economies**. [...] **Public-Private Partnerships**: Collaborate with private developers to build affordable housing units, especially in high-cost areas. [...] **Address Systemic Inequities**. [...] **International Best Practices**.
12	Rising housing costs are a complex issue that affects people in different ways depending on their **location, income, and life stage**. [...] **Support Homeownership**. [...] **Consider shared housing or roommates**. [...] **Look for places outside the city center**. [...] **Consider moving to a different city or state**.	**Supply-Side Interventions**. [...] **Demand-Side Management**. [...] **Financial and Regulatory Reforms**. [...] **Urban Planning and Spatial Policies**. [...] **Monitoring and Evaluation**.
16	Rising housing costs are a major issue in many cities and countries, affecting affordability, quality of life, and even economic stability. Here are some **practical and effective ways** to deal with rising housing costs. [...] **Talk to Your Landlord**. [...] **Consider Moving**.	Rising housing costs are a complex and multifaceted issue that affects **economies, social stability, and individual well-being**. [...] **Regulatory and Institutional Reforms**. [...] **International and Regional Cooperation**. [...] **Implementation Pathways**.
20	**Understand Why Housing Costs Are Going Up**. [...] **Rent a room**. [...] **Roommates**. [...] **Talk to Your Landlord**. [...] **Stay Positive**.	Rising housing costs are a complex and multifaceted issue that affects economies, social stability, and individual well-being. Addressing this challenge requires a comprehensive, multi-pronged approach that balances **economic, social, and environmental considerations**. [...] **Regulatory and Institutional Reforms**. [...] **Technological and Innovation-Driven Solutions**. [...] **International Cooperation**.
24	**Check online listings**. [...] **Ask around**. [...] **Consider a different neighborhood**. [...] **Cut back on other expenses**. [...] **Get a side job**.	**Smart Urban Development**. [...] **Global Housing Agreements**. [...] **Migration and Labor Policies**. [...] **Climate-Resilient Housing**. [...] **Adaptive Governance**.
28	**Look for Affordable Housing Options**. [...] **Talk to Your Landlord**. [...] **Try to Save More Money**. [...] **Try to Get a Job or Side Job**.	Dealing with rising housing costs is a complex and multifaceted challenge that requires **coordinated efforts across government, private sector, and civil society**. [...] **Financial and Economic Instruments**. [...] **Technological and Innovation Strategies**. [...] **Monitoring and Evaluation**.
32	Dealing with rising housing costs can be a big challenge, especially if you're **on a tight budget or just starting out**. [...] **Look for Affordable Neighborhoods**. [...] **Talk to Your Landlord**. [...] **Look for Roommates**. [...] **Try to Make More Money**. [...] **Don't Give Up**.	Dealing with rising housing costs is a complex and multifaceted challenge that requires **coordinated action across multiple sectors**, including government, private enterprise, and civil society. [...] **Anti-Monopolistic Measures**. [...] **Human Rights-Based Approach**. [...] **National Housing Strategies**.
36	Dealing with rising housing costs can be really tough, especially if you're already struggling to make ends meet. [...] **Talk to your landlord or property manager**. [...] **Look for a roommate**. [...] **Try to get a job**. [...] I know it's not much, but it's better than nothing. [...]	Dealing with rising housing costs is a complex and multifaceted challenge that requires **coordinated action across multiple sectors**. [...] **Institutional and Regulatory Reforms**. [...] **Global Partnerships**. [...] **Regional and Local Strategies**. [...] **Implementation Pathways**.
Llama-3.1-8B-Instruct		
Prompt: How should we deal with rising housing costs?		
8	**Save up for a down payment**. [...] **Look for a smaller house or apartment**. [...] Look for a house that's **not in the city**. [...] **not so fancy**. [...]	**Enhance Affordability through Policy Reforms**. [...] **Incentivizing the development of affordable housing**. [...] **Foster Inclusive Urban Planning**. [...] **Strengthen Social Safety Nets**. [...] **International Cooperation**.

12	**We need to make more houses**. [...] **We need to get the price down**. [...] I'm gonna try to get it together. [...]	**Enhance Urban Planning and Urbanization**. [...] **Incentivize the development of affordable housing**. [...] **Market-based solutions**. [...] **International cooperation and knowledge sharing**. [...]
16	I don't know if I'm gonna be here tomorrow. [...] **I don** I don I don [...]	**Enhance the supply of housing**. [...] **innovative and inclusive urban planning**. [...] housing policies [...] **diverse populations**. [...] **global economic and demographic trends**. [...]
20	I don't know what I don [...] **I don** I don I don [...]	**rising global, regional, and national economic frameworks**. [...] **promoting inclusive and sustainable development**. [...]
24	**I don** I don I don [...]	**promoting sustainable and inclusive development**. [...] **development of policies and frameworks**. [...]
28	**I'm just like** I'm just like I'm just like [...]	**global dimensions**. [...] **global architecture**. [...] **policies**. [...]
32	**Bek** Bek Bek Bek [...]	**Prom** [...] **strategies** [...] **policies** [...] **global and regional dimensions**. [...]
36	**Bek** Bek Bek Bek [...]	**strategies** [...] **architectures** [...] **policies** [...]

Interpretation. The Qwen3-8B sweep shows a relatively smooth semantic movement. Negative coefficients increasingly replace system-level policy language with individual or household-level advice, such as talking to landlords, finding roommates, saving money, moving neighborhoods, or seeking local help. Positive coefficients instead emphasize policy categories, institutional coordination, data systems, international comparison, climate resilience, and other macro-level governance frames. Even at large coefficients, Qwen mostly remains coherent, although the most negative setting begins to show repetitive low-information advice.

Degeneration under large coefficients. The Llama-3.1-8B-Instruct sweep illustrates why the main experiments use conservative steering settings and report degeneration diagnostics. At $\alpha = -8$, the model moves toward highly concrete household advice, but this already includes low-quality repetition. From $\alpha = -12$ onward on the negative side, and from about $\alpha = +16$ onward on the positive side, much of the output becomes repetitive or nonsensical. We therefore treat these examples as qualitative evidence of directional sensitivity and instability, not as successful high-magnitude control.

Selection and quantitative scope. These outputs were manually selected after inspecting model responses, not randomly sampled, and are not used for quantitative conclusions. At the generic main setting, degeneration rates for $\alpha = -4/0/+4$ are $0/0/0$ for Qwen and $0.425/0/0.025$ for Llama. This context applies to the examples in Table 5 as well as the more aggressive illustrative sweep above.

B Steering Results

This appendix provides additional diagnostics for the activation-steering experiments. The goal is not to present steering as a uniformly reliable control method, but to test whether movement along the Granularity Axis has a measurable behavioral effect and to characterize when that effect becomes unstable. Unless otherwise noted, scores are `granularity_overall` ratings from `gpt-5.4-mini` on a 1–5 scale, where higher scores indicate more macro-level reasoning. We report standard errors of the mean (SEM) over prompts for the coefficient sweep and for response-length diagnostics.

Degeneration-filtered steering. Table 8 reports the main steering results before and after removing outputs judged as degenerate. The strongest negative shift in Llama-3.1-8B-Instruct on generic prompts remains visible after filtering, but this condition keeps only $23/40$ responses, so we treat it as evidence of behavioral sensitivity rather than stable control.

Coefficient sweep. Table 9 summarizes a small diagnostic sweep over larger steering coefficients. Llama-3.1-8B-Instruct is highly responsive but becomes unstable under stronger interventions,

Table 8: **Degeneration-filtered steering results.** All mean reports the mean `granularity_overall` score over all outputs. Non-deg. mean recomputes the score after removing outputs judged as degenerate. Deg. is the judged degeneration rate.

Model	Prompt set	α	Kept / total	Deg.	All mean	Non-deg. mean
Qwen3-8B	Generic	-4	39/39	.000	4.821	4.821
	Generic	0	40/40	.000	4.900	4.900
	Generic	$+4$	40/40	.000	4.975	4.975
	Micro-targeted	-4	12/12	.000	2.000	2.000
	Micro-targeted	0	12/12	.000	2.167	2.167
	Micro-targeted	$+4$	12/12	.000	2.667	2.667
Llama-3.1-8B-Instruct	Generic	-4	23/40	.425	3.125	3.000
	Generic	0	40/40	.000	4.525	4.525
	Generic	$+4$	39/40	.025	4.900	4.897
	Micro-targeted	-4	12/12	.000	1.583	1.583
	Micro-targeted	0	11/12	.083	2.000	2.000
	Micro-targeted	$+4$	12/12	.000	3.167	3.167

especially outside the conservative range used in the main text. Qwen3-8B remains more stable, but its generic-prompt scores are often saturated near the macro end, making generic-prompt movement difficult to observe.

Table 9: **Selected coefficient sweep results.** Mean and SEM summarize `granularity_overall` scores. Deg. is the judged degeneration rate. This sweep uses $n = 5$ prompts per setting and is intended as a diagnostic of sensitivity and instability under stronger coefficients, not as the main steering evaluation.

Model	Prompt set	α	n	Mean	SEM	Deg.
Qwen3-8B	Generic	-12	5	4.000	0.548	.000
	Generic	-8	5	5.000	0.000	.000
	Generic	0	5	5.000	0.000	.000
	Generic	$+8$	5	5.000	0.000	.000
	Generic	$+12$	5	5.000	0.000	.000
	Micro-targeted	-12	5	2.000	0.000	.000
	Micro-targeted	-8	5	1.800	0.200	.000
	Micro-targeted	0	5	2.600	0.245	.000
	Micro-targeted	$+8$	5	3.200	0.374	.000
	Micro-targeted	$+12$	5	4.000	0.000	.000
Llama-3.1-8B-Instruct	Generic	-12	5	2.000	0.447	1.000
	Generic	-8	5	2.600	0.510	1.000
	Generic	0	5	4.800	0.200	.000
	Generic	$+8$	5	5.000	0.000	.400
	Generic	$+12$	5	5.000	0.000	1.000
	Micro-targeted	-12	5	1.400	0.245	1.000
	Micro-targeted	-8	5	2.000	0.000	1.000
	Micro-targeted	0	5	2.000	0.000	.200
	Micro-targeted	$+8$	5	4.000	0.000	.200
	Micro-targeted	$+12$	5	5.000	0.000	1.000

Length and quality diagnostics. Table 10 reports generated-token counts and judged degeneration rates for the main steering setting. These statistics help distinguish granularity shifts from simple response-length or degeneration artifacts. The most important caveat is again Llama-3.1-8B-Instruct under $\alpha = -4$ on generic prompts, where both response length and degeneration increase substantially.

Table 10: **Response-length and degeneration diagnostics for the main steering setting.** Length is measured as generated token count. Deg. is the judged degeneration rate.

Model	Prompt set	α	n	Length mean $\pm$ SEM	Deg.
Qwen3-8B	Generic	-4	39	832.8 ± 16.1	.000
	Generic	0	40	815.2 ± 14.7	.000
	Generic	$+4$	40	845.6 ± 21.2	.000
	Micro-targeted	-4	12	777.8 ± 42.0	.000
	Micro-targeted	0	12	747.7 ± 31.2	.000
	Micro-targeted	$+4$	12	758.1 ± 31.7	.000
Llama-3.1-8B-Instruct	Generic	-4	40	864.9 ± 55.5	.425
	Generic	0	40	588.7 ± 12.3	.000
	Generic	$+4$	40	628.2 ± 19.1	.025
	Micro-targeted	-4	12	626.2 ± 47.1	.000
	Micro-targeted	0	12	548.0 ± 19.8	.083
	Micro-targeted	$+4$	12	511.5 ± 22.5	.000

Table 11: Every diagnostic prompt at the 1,536-token setting. Each cell gives scores for $\alpha = -4/0/+4$. Prompt IDs follow the zero-based order in Table 18. GPT and Gemini abbreviate the two reported judges. These are manually designed diagnostic prompts, not a random prompt sample.

Prompt ID	Qwen / GPT	Qwen / Gemini	Llama / GPT	Llama / Gemini
0	2/2/2	2/2/2	1/2/2	1/2/2
1	2/2/3	2/1/2	2/2/4	1/1/4
2	2/2/3	2/2/2	2/2/3	1/2/2
3	2/2/2	1/2/2	1/2/2	1/1/2
4	2/2/2	1/1/1	1/2/3	1/1/2
5	2/2/2	1/1/2	2/2/4	1/1/2
6	2/2/2	1/1/2	2/2/4	1/2/2
7	2/3/4	1/2/2	2/2/4	1/2/3
8	2/2/2	1/2/2	2/2/4	1/1/3
9	2/2/2	1/1/2	1/2/2	1/1/2
10	2/3/4	2/3/3	2/2/3	1/2/3
11	2/2/4	2/2/3	1/2/3	1/1/3

B.1 Prompt Selection, Influence, and Judge Pairing

The 12 micro-targeted prompts were initially drafted with Opus 4.5 and manually revised. Tables 11–13 show the full diagnostic outcomes and influence analyses rather than treating the subset as selection-neutral. At 1,536 tokens, the positive leave-one-out ranges across both judges are 0.364–0.546 for Qwen and 0.909–1.273 for Llama. Qwen's negative effect is more concentrated: one prompt contributes half of its absolute effect under the primary judge.

C Robustness and Control Analyses

This appendix provides the full robustness and control analyses summarized in Section 4. We test whether the Granularity Axis is robust to alternative layer choices, endpoint definitions, held-out splits, prompt templates, response-quality thresholds, and potential role-name or domain confounds. Table 14 summarizes these robustness and control analyses.

C.1 Layer-wise Stability

The granularity structure is not confined to a single layer. In Qwen3-8B, monotonic level ordering first appears at Layer 3, briefly breaks at Layers 5–7, and then holds consistently from Layers 8–35. At the main analysis layer, Layer 18, the contrast-defined axis remains strongly aligned with PC1 (cosine $= 0.9720$), with high projection-level correlation (Spearman $= 0.9472$, Pearson $= 0.9414$).

Table 12: All cached diagnostic prompt-influence conditions. Max share is the largest single-prompt share of total absolute effect; half effect is the fewest prompts contributing half that effect. The 512-token Qwen conditions are exploratory and have no paired Gemini records.

Model	Judge	Tokens	α	Mean Δ	LOO range	Max share	Half effect
Llama	Gemini	1536	-4	-0.417	$[-0.455, -0.364]$	0.200	3
Llama	Gemini	1536	+4	1.083	$[0.909, 1.182]$	0.231	3
Llama	GPT	1536	-4	-0.417	$[-0.455, -0.364]$	0.200	3
Llama	GPT	1536	+4	1.167	$[1.091, 1.273]$	0.143	4
Qwen	Gemini	1536	-4	-0.250	$[-0.364, -0.182]$	0.200	3
Qwen	Gemini	1536	+4	0.417	$[0.364, 0.455]$	0.200	3
Qwen	GPT	1536	-4	-0.167	$[-0.182, -0.091]$	0.500	1
Qwen	GPT	1536	+4	0.500	$[0.364, 0.545]$	0.333	2
Qwen	GPT	512	-4	-0.083	$[-0.182, 0.091]$	0.400	2
Qwen	GPT	512	+4	0.500	$[0.364, 0.636]$	0.250	2

Table 13: Every leave-one-prompt-out mean delta at 1,536 tokens. Q/L denotes Qwen/Llama, G/M denotes GPT/Gemini, and $-/+$ denotes $\alpha = -4/+4$. The positive aggregate persists under each omission; this is an influence check, not proof of selection neutrality.

Omitted ID	Q/G $-$	Q/G $+$	Q/M $-$	Q/M $+$	L/G $-$	L/G $+$	L/M $-$	L/M $+$
0	-0.182	0.545	-0.273	0.455	-0.364	1.273	-0.364	1.182
1	-0.182	0.455	-0.364	0.364	-0.455	1.091	-0.455	0.909
2	-0.182	0.455	-0.273	0.455	-0.455	1.182	-0.364	1.182
3	-0.182	0.545	-0.182	0.455	-0.364	1.273	-0.455	1.091
4	-0.182	0.545	-0.273	0.455	-0.364	1.182	-0.455	1.091
5	-0.182	0.545	-0.273	0.364	-0.455	1.091	-0.455	1.091
6	-0.182	0.545	-0.273	0.364	-0.455	1.091	-0.364	1.182
7	-0.091	0.455	-0.182	0.455	-0.455	1.091	-0.364	1.091
8	-0.182	0.545	-0.182	0.455	-0.455	1.091	-0.455	1.000
9	-0.182	0.545	-0.273	0.364	-0.364	1.273	-0.455	1.091
10	-0.091	0.455	-0.182	0.455	-0.455	1.182	-0.364	1.091
11	-0.182	0.364	-0.273	0.364	-0.364	1.182	-0.455	1.000

More generally, across layers, the contrast–PC1 cosine remains approximately 0.972–0.986, and the projection-level Spearman correlation remains approximately 0.928–0.957.

Llama-3.1-8B-Instruct shows a similarly broad stable regime: monotonicity first appears at Layer 3, briefly breaks at Layers 4–5, and then holds from Layers 6–31. At Layer 18, cosine alignment is 0.9596, with Spearman $= 0.9459$ and Pearson $= 0.9373$; across layers, cosine remains approximately 0.956–0.987 and Spearman approximately 0.912–0.954. These results indicate that granularity is a stable middle/later-layer representational phenomenon rather than a single-layer artifact.

C.2 Endpoint Ablations

We also test alternative endpoint definitions for the contrast-based Granularity Axis. In Qwen3-8B, cosine alignment with PC1 remains in the range 0.9452–0.9989, and the corresponding Spearman correlations remain in the range 0.9359–0.9598, with monotonic ordering preserved in all cases. In Llama-3.1-8B-Instruct, cosine values remain in the range 0.9279–0.9988, Spearman correlations remain in the range 0.9276–0.9646, and all variants are again monotonic. Thus, the result is not specific to one particular endpoint choice.

C.3 Held-out Robustness

The axis also generalizes across held-out prompt and question splits. In Qwen3-8B, held-out Spearman correlations are 0.9468 for prompt holdout and 0.9494 for question holdout, and both settings remain monotonic. The role-holdout setting is somewhat weaker: although it still yields high correlation (Spearman $= 0.9335$, Pearson $= 0.9404$, cosine $= 0.9821$), monotonic ordering

Table 14: **Quantitative summary of robustness and control analyses.** The Granularity Axis remains stable across layers, endpoint definitions, held-out splits, prompt templates, score filters, and most subgroup analyses. Role-holdout and domain/family controls are the most challenging settings, so we treat them as supportive but not definitive confound controls.

Analysis	Qwen3-8B	Llama-3.1-8B-Instruct
Layer-wise stability	Monotonic from Layers 8–35; contrast–PC1 cosine ≈ 0.972–0.986 and Spearman ≈ 0.928–0.957 across layers. At Layer 18: cosine $= 0.9720$, Spearman $= 0.9472$, Pearson $= 0.9414$.	Monotonic from Layers 6–31; contrast–PC1 cosine ≈ 0.956–0.987 and Spearman ≈ 0.912–0.954 across layers. At Layer 18: cosine $= 0.9596$, Spearman $= 0.9459$, Pearson $= 0.9373$.
Endpoint ablations	Cosine $= 0.9452$–0.9989; Spearman $= 0.9359$–0.9598; all endpoint variants preserve monotonic ordering.	Cosine $= 0.9279$–0.9988; Spearman $= 0.9276$–0.9646; all endpoint variants preserve monotonic ordering.
Held-out robustness	Prompt holdout Spearman $= 0.9468$; question holdout Spearman $= 0.9494$; role holdout remains high (Spearman $= 0.9335$, Pearson $= 0.9404$, cosine $= 0.9821$) but is not perfectly monotonic.	Prompt holdout Spearman $= 0.9403$; question holdout Spearman $= 0.9468$; role holdout Spearman $= 0.9335$; all three held-out settings remain monotonic.
Prompt sensitivity	All five prompt variants remain monotonic; Spearman $= 0.9381$–0.9568; macro–micro gap $= 15.0941$–18.9780.	All five prompt variants remain monotonic; Spearman $= 0.9342$–0.9585; macro–micro gap $= 2.6218$–3.6250.
Score filtering	Spearman is 0.9472 with all responses, 0.9481 under score ≥ 2, and 0.9564 under score ≥ 3.	Spearman is 0.9459 with all responses, 0.9459 under score ≥ 2, and 0.9503 under score ≥ 3.
Role-type control	Generic bucket Spearman $= 0.8644$; merged specific/title-heavy bucket Spearman $= 0.8133$; both buckets remain monotonic but are not perfectly level-matched.	Generic bucket Spearman $= 0.8644$; merged specific/title-heavy bucket Spearman $= 0.8109$; both buckets remain monotonic but are not perfectly level-matched.
Domain and family controls	Domain monotonicity holds for economy, education, food, housing, justice, and migration; health is least stable. Family-ladder monotonicity holds for education, food, housing, and migration; economy fails monotonicity.	Domain monotonicity holds for economy, education, food, housing, justice, and migration; health is least stable. Family-ladder monotonicity holds for education, food, housing, and migration; economy and health fail monotonicity.

is not perfectly preserved. In Llama-3.1-8B-Instruct, all three held-out settings remain monotonic, with Spearman correlations of 0.9403 for prompt holdout, 0.9468 for question holdout, and 0.9335 for role holdout. Overall, held-out performance remains strong, though role-level generalization is slightly more fragile than prompt- or question-level generalization, especially for Qwen.

C.4 Prompt-template Sensitivity

We further test whether the representation-level signal depends on a particular prompt wording. For both models, all five prompt variants remain monotonic. In Qwen3-8B, projection-level Spearman correlations range from 0.9381 to 0.9568, with macro–micro gaps ranging from 15.0941 to 18.9780. In Llama-3.1-8B-Instruct, Spearman correlations range from 0.9342 to 0.9585, with corresponding gaps from 2.6218 to 3.6250. This indicates that the recovered granularity structure is not an artifact of a single prompt template.

C.5 Default Assistant Placement

The default assistant condition provides an important interpretive reference point. In Qwen3-8B, the default assistant projection at Layer 18 is 18.2572, and its nearest level is Level 3. In Llama-3.1-8B-Instruct, the corresponding projection is 2.5880, with nearest level Level 4. Thus, the default assistant persona is not positioned near the micro end in either model. Instead, it occupies a mid-to-upper region of the learned granularity spectrum, with Llama appearing somewhat more macro-leaning than Qwen.

C.6 Score Filtering

We also evaluate whether the axis depends on noisy or weakly in-role samples. In Qwen3-8B, the projection-level Spearman correlation is 0.9472 when using all responses, 0.9481 under score ≥ 2, and 0.9564 under score ≥ 3. In Llama-3.1-8B-Instruct, the corresponding values are 0.9459, 0.9459,

and 0.9503. These results show that the signal survives stricter filtering and, if anything, becomes slightly cleaner when lower-quality role-play responses are removed.

C.7 Generic versus Specific Role Controls

As a supportive confound-control analysis, we compare a *generic* role bucket and a merged *specific* bucket combining originally specific and title-heavy roles. In both models, the ordering remains monotonic within each bucket. Using projection-level Spearman correlation, Qwen3-8B yields 0.8644 for the generic bucket and 0.8133 for the specific bucket, while Llama-3.1-8B-Instruct yields 0.8644 and 0.8109, respectively. However, this control is not fully level-matched: generic roles are concentrated in lower levels, whereas specific roles are concentrated in middle and higher levels. We therefore interpret this result as supportive rather than definitive evidence that the axis is not reducible to title-ness alone.

C.8 Domain and Family Controls

Finally, subgroup analyses show that the effect remains strong across most semantic domains and within most family ladders, but also reveal a small number of harder cases. Across domains, both models remain monotonic in economy, education, food, housing, justice, and migration, while *health* is the least stable domain in both Qwen3-8B and Llama-3.1-8B-Instruct. At the family-ladder level, both models remain monotonic for education, food, housing, and migration, whereas *economy* fails monotonicity in both models; additionally, *health* fails monotonicity in Llama. These subgroup results suggest that the Granularity Axis is broadly robust, while also indicating that some semantically dense domains remain more challenging than others.

Scope of the subgroup evidence. Economy family ladders are nonmonotonic in both models and health is nonmonotonic in Llama; incomplete ladders are not counted as five-level monotonic successes. The generic/specific buckets are not fully level-matched. The separate controlled domain chains and permutation nulls in Appendices F and G provide complementary evidence rather than removing all confounds.

D Role Taxonomy

This appendix documents the social-role taxonomy used to construct the ordered role set. Each role is assigned to one of five granularity levels before response generation. The level annotations are used for analysis, grouping, endpoint construction, and evaluation.

Annotation fields. For each role, we record the granularity level, domain, role type, role name, and short role description. These fields make the taxonomy auditable and help separate the intended micro–macro ordering from correlated factors such as social power or professional title specificity. Details are provided in Table 15.

Table 15: **Social-role taxonomy.** The table lists the 75 role-conditioned entities used to construct the granularity axis, together with their assigned granularity level and auxiliary metadata for domain and role type.

Level	Domain / type	Role	Description
L1 Individual	family / generic	Worried Parent	A parent anxious about their children's daily wellbeing, safety, and future. Thinks in terms of immediate family needs, personal budgets, and parenting decisions. Time horizon is days to months.
L1 Individual	economy / generic	Frustrated Employee	A worker dealing with daily workplace stress, an unsympathetic manager, and career uncertainty. Thinks about paychecks, commute times, and whether to quit or stay.
L1 Individual	education / generic	Homesick Student	A university student living far from home, navigating loneliness, academic pressure, and self-discovery. Concerns center on exams, friendships, and personal identity.

Continued on next page

Level	Domain / type	Role	Description
L1 Individual	health / generic	Anxious Patient	A person recently diagnosed with a chronic illness, navigating medical appointments, insurance paperwork, and fear of the unknown. Thinks about symptoms, medication side effects, and personal survival.
L1 Individual	economy / generic	Small Shopkeeper	Owner of a small neighborhood corner store, worried about rent increases, customer loyalty, and competition from chain retailers. Thinks about daily cash flow and inventory.
L1 Individual	family / generic	Retired Grandparent	An elderly retiree reflecting on decades of life experience, managing health issues, and cherishing time with grandchildren. Thinks about legacy, daily routines, and personal memories.
L1 Individual	family / generic	Single Parent	A single parent juggling full-time work, childcare, and household duties alone. Every decision is weighed against limited time, energy, and money.
L1 Individual	economy / generic	Unemployed Graduate	A recent college graduate struggling to find their first real job, sending dozens of applications and facing rejection. Worries about student debt, self-worth, and whether their degree was worth it.
L1 Individual	migration / generic	New Immigrant	Someone who recently moved to a new country, struggling with language barriers, cultural differences, and homesickness while trying to build a new life from scratch.
L1 Individual	culture / generic	Freelance Artist	A self-employed artist balancing creative passion with financial survival, constantly hustling for the next commission or gig. Thinks about rent, inspiration, and self-doubt.
L1 Individual	health / generic	Overworked Nurse	A frontline healthcare worker facing burnout from long shifts, understaffing, and emotional exhaustion. Thinks about patient care, personal health, and whether to leave the profession.
L1 Individual	housing / generic	Tenant Facing Eviction	A renter who just received an eviction notice, scrambling to find new housing while dealing with fear, anger, and financial stress. Every thought centers on immediate shelter and survival.
L1 Individual	food / generic	Small Farmer	A family farmer managing a modest plot of land, dependent on weather, seed prices, and local markets. Thinks seasonally about planting, harvesting, and making ends meet.
L1 Individual	family / generic	Grieving Widow	Someone processing the recent loss of their life partner, navigating grief, loneliness, and the practical challenges of living alone for the first time. Focuses on daily coping and emotional survival.
L1 Individual	education / generic	Young Apprentice	A young person learning a trade under a mentor, eager but uncertain. Thinks about mastering skills, earning respect, and building a future through hands-on work.
L2 Group	housing / generic	Neighborhood Association Leader	Leads a residential neighborhood association, mediating disputes between neighbors, organizing block parties, and advocating for local improvements like streetlights and speed bumps. Thinks about a few hundred households.
L2 Group	education / generic	PTA President	President of a school parent-teacher association, coordinating fundraisers, communicating between parents and teachers, and advocating for children's educational needs at the local school level.
L2 Group	community / generic	Local Church Pastor	Leads a small congregation of around 100-200 members, providing spiritual guidance, organizing community outreach, and counseling individuals through personal crises.
L2 Group	education / generic	Youth Sports Coach	Coaches a community youth team, mentoring young athletes, managing team dynamics, and communicating with parents. Thinks about child development, teamwork, and local league schedules.
L2 Group	governance / generic	Community Organizer	Coordinates grassroots local initiatives, mobilizes residents around shared concerns like park maintenance or traffic safety, and builds coalitions among diverse neighborhood voices.
L2 Group	disaster / generic	Volunteer Fire Chief	Leads a volunteer fire department serving a small town, coordinating training, equipment maintenance, and emergency response with limited budgets and volunteer availability.
L2 Group	housing / generic	Tenant Union Leader	Advocates for renters' rights in a large apartment complex, organizing collective bargaining with landlords, filing complaints, and building solidarity among dozens of tenant families.
L2 Group	economy / generic	Local Business Alliance Chair	Coordinates a group of small business owners on a commercial street, organizing joint promotions, lobbying for local parking improvements, and sharing strategies to compete with online retailers.

Continued on next page

Level	Domain / type	Role	Description
L2 Group	food / generic	Community Garden Coordinator	Manages a shared community garden, allocating plots, organizing workdays, resolving conflicts between gardeners, and connecting the garden to local food banks.
L2 Group	migration / generic	Mutual Aid Network Organizer	Runs a neighborhood mutual aid network, matching people who need help with those who can provide it: rides to appointments, grocery deliveries, and childcare swaps.
L2 Group	education / specific	School Board Member	Elected to represent community interests in local education decisions, balancing parent demands, teacher needs, and budget constraints across a handful of neighborhood schools.
L2 Group	housing / specific	Housing Cooperative President	Manages a residential cooperative of 50-100 units, overseeing shared finances, maintenance decisions, and conflict resolution among member-owners who collectively own their building.
L2 Group	climate / generic	Local Environmental Group Leader	Leads a community conservation group focused on protecting a local river, park, or wetland. Organizes cleanups, monitors water quality, and lobbies city council for protection measures.
L2 Group	health / generic	Community Health Worker	Bridges the gap between healthcare systems and underserved local communities, conducting home visits, translating medical information, and connecting families to resources.
L2 Group	justice / generic	Neighborhood Watch Captain	Coordinates neighborhood safety efforts, organizing patrol schedules, maintaining communication with local police, and fostering vigilance without vigilantism among residents.
L3 Organization	health / specific	Hospital Administrator	Manages a regional hospital with hundreds of staff, balancing patient care quality, budget constraints, regulatory compliance, and staff retention. Plans on annual and multi-year cycles.
L3 Organization	education / specific	University Dean	Leads an academic college within a university, overseeing curriculum development, faculty hiring, research funding, and student enrollment across multiple departments.
L3 Organization	migration / specific	NGO Director	Leads a mid-size non-profit organization with 50-200 employees, managing program delivery, fundraising campaigns, donor relations, and impact measurement across a region.
L3 Organization	economy / specific	Factory Operations Manager	Oversees manufacturing operations at a large facility, managing production schedules, supply chains, quality control, worker safety, and operational efficiency for hundreds of workers.
L3 Organization	education / specific	School District Superintendent	Manages an entire school district with dozens of schools, thousands of students, and hundreds of teachers. Balances educational standards, budgets, union negotiations, and community expectations.
L3 Organization	economy / specific	Regional Bank Manager	Oversees a regional banking operation with multiple branches, managing lending policies, risk assessment, customer service standards, and regulatory compliance across a metropolitan area.
L3 Organization	media / specific	Media Editor-in-Chief	Leads editorial strategy for a mid-size news organization, deciding what stories to cover, managing journalists, maintaining editorial standards, and navigating the economics of modern media.
L3 Organization	technology / specific	Tech Startup CEO	Leads a growing technology company through rapid scaling, managing product development, fundraising rounds, talent acquisition, and competition in a fast-moving market.
L3 Organization	labor / specific	Labor Union President	Represents thousands of workers in collective bargaining, negotiating wages, benefits, and working conditions with management while maintaining union solidarity and political influence.
L3 Organization	culture / specific	Museum Director	Manages a cultural institution, curating exhibitions, securing funding, managing staff, preserving collections, and making art and history accessible to the public.
L3 Organization	logistics / specific	Logistics Company Director	Manages supply chain and distribution networks for a mid-size logistics firm, optimizing routes, managing warehouses, negotiating contracts, and ensuring timely delivery across regions.
L3 Organization	transport / specific	Urban Transit Authority Director	Runs a city's public transportation system, managing bus and rail networks, fare structures, infrastructure maintenance, and ridership growth to serve millions of daily commuters.
L3 Organization	food / specific	Agricultural Cooperative Head	Coordinates farming operations across hundreds of member farms, negotiating bulk purchasing, managing shared processing facilities, and representing collective interests in regional markets.

Continued on next page

Level	Domain / type	Role	Description
L3 Organization	economy / specific	Hotel Chain Regional Manager	Oversees multiple hotel properties across a region, standardizing service quality, managing occupancy rates, training staff, and adapting to hospitality market trends.
L3 Organization	housing / specific	Construction Firm Director	Manages large infrastructure and building projects, coordinating hundreds of workers, subcontractors, permits, timelines, and budgets across simultaneous construction sites.
L4 Institution	justice / title-heavy	Constitutional Court Justice	Interprets constitutional law and sets legal precedents that shape the rights and obligations of millions. Thinks in terms of legal doctrine, precedent, and the long-term integrity of the justice system.
L4 Institution	economy / title-heavy	Central Bank Governor	Sets monetary policy for an entire national economy, managing interest rates, inflation targets, currency stability, and financial system oversight. Decisions affect millions of economic actors.
L4 Institution	health / title-heavy	Public Health Commissioner	Designs and enforces public health policies at the national level: vaccination programs, disease surveillance, food safety standards, and pandemic preparedness affecting an entire population.
L4 Institution	education / title-heavy	Education Minister	Shapes national education standards, curriculum frameworks, teacher certification requirements, and university funding. Decisions affect every student and teacher in the country.
L4 Institution	security / title-heavy	Military General	Commands strategic military operations within geopolitical constraints, managing thousands of personnel, logistics chains, intelligence, and rules of engagement across theaters of operation.
L4 Institution	climate / title-heavy	Environmental Protection Director	Leads a national environmental agency, setting emission standards, enforcing pollution regulations, managing protected areas, and balancing economic development with ecological preservation.
L4 Institution	economy / title-heavy	Securities Regulator	Oversees financial market integrity, writing rules for securities trading, investigating fraud, protecting investors, and maintaining public confidence in capital markets.
L4 Institution	housing / title-heavy	Chief Urban Planner	Designs long-term city infrastructure and zoning policy, shaping where people live, work, and move for decades to come. Balances housing, transport, green space, and economic zones.
L4 Institution	migration / title-heavy	Immigration Policy Director	Designs and implements a nation's immigration system: visa categories, asylum processes, enforcement priorities, and integration programs affecting millions of migrants and citizens.
L4 Institution	technology / title-heavy	Telecommunications Regulator	Governs national communication infrastructure: spectrum allocation, net neutrality rules, broadband deployment requirements, and media ownership limits affecting how an entire nation communicates.
L4 Institution	economy / title-heavy	National Pension Fund Director	Manages a national retirement system worth billions, making investment decisions and policy recommendations that determine the financial security of millions of retirees over decades.
L4 Institution	disaster / title-heavy	Disaster Management Authority Head	Plans and coordinates institutional responses to natural disasters: early warning systems, evacuation protocols, emergency stockpiles, and recovery frameworks at the national scale.
L4 Institution	health / title-heavy	National Healthcare System Architect	Designs the structure of a national healthcare delivery system: hospital networks, insurance frameworks, drug pricing policies, and primary care access affecting every citizen.
L4 Institution	justice / title-heavy	Criminal Justice Reform Commissioner	Works to reshape the legal and correctional system: sentencing guidelines, prison conditions, rehabilitation programs, and policing standards affecting millions within the justice system.
L4 Institution	energy / title-heavy	National Energy Policy Director	Plans the transition of a nation's energy infrastructure: from fossil fuels to renewables, grid modernization, nuclear policy, and energy independence strategies spanning decades.
L5 Macro	governance / title-heavy	Head of State	Leads a nation in domestic and international affairs, setting national priorities, representing the country on the world stage, and making decisions that shape the lives of tens of millions for generations.
L5 Macro	diplomacy / title-heavy	UN Ambassador	Represents a nation in the United Nations, negotiating multilateral treaties, building diplomatic coalitions, and navigating the complex dynamics of global governance.

Continued on next page

Level	Domain / type	Role	Description
L5 Macro	economy / title-heavy	World Bank President	Leads the world's largest development finance institution, directing billions in lending to developing nations, shaping global poverty reduction strategies, and influencing international economic architecture.
L5 Macro	climate / title-heavy	Climate Treaty Negotiator	Negotiates binding international climate agreements between 190+ nations, balancing emission targets, financial transfers, technology sharing, and the competing interests of developed and developing worlds.
L5 Macro	security / specific	Geopolitical Strategist	Analyzes and advises on global power dynamics: military alliances, trade blocs, resource competition, and civilizational trends. Thinks in decades and across continents.
L5 Macro	economy / title-heavy	International Trade Negotiator	Shapes global trade frameworks: tariff schedules, intellectual property regimes, dispute resolution mechanisms, and regional trade agreements that govern trillions in cross-border commerce.
L5 Macro	health / title-heavy	Global Pandemic Response Coordinator	Coordinates worldwide responses to infectious disease outbreaks, managing vaccine distribution across nations, harmonizing quarantine protocols, and building international early warning systems.
L5 Macro	security / title-heavy	Nuclear Nonproliferation Envoy	Works to prevent the spread of nuclear weapons through diplomatic negotiations, inspection regimes, sanctions frameworks, and arms control treaties between nuclear and non-nuclear states.
L5 Macro	justice / title-heavy	International Court of Justice Judge	Rules on disputes between nations at the highest level of international law: territorial claims, genocide cases, treaty interpretations, and the fundamental norms governing relations between states.
L5 Macro	economy / specific	Global Development Economist	Designs economic frameworks for developing nations: structural adjustment programs, aid effectiveness metrics, trade integration strategies, and poverty reduction pathways affecting billions.
L5 Macro	culture / title-heavy	UNESCO Cultural Heritage Director	Protects humanity's shared cultural and natural heritage across all nations: World Heritage Sites, intangible cultural practices, and the preservation of human civilization's collective memory.
L5 Macro	technology / title-heavy	Space Agency Director	Leads a major national or international space exploration program, planning missions that extend humanity's presence beyond Earth, managing international partnerships, and allocating multi-billion-dollar budgets across decades.
L5 Macro	migration / specific	Transnational Migration Policy Analyst	Studies and advises on global population movements: refugee flows, labor migration patterns, climate displacement, and the international frameworks governing human mobility across borders.
L5 Macro	food / specific	Global Food Security Strategist	Plans worldwide agricultural sustainability: crop diversity, supply chain resilience, emergency food reserves, and international cooperation to prevent famine affecting hundreds of millions.
L5 Macro	governance / specific	Civilizational Risk Analyst	Assesses existential and catastrophic risks to human civilization: nuclear war, AI misalignment, pandemics, asteroid impacts, and climate tipping points. Thinks on century timescales about the survival of the species.

D.1 Construction Provenance and Operational Definitions

The initial role inventory and assignments were LLM-assisted, using the Assistant Axis role-list format as a stylistic reference, then manually reviewed and revised by one author. There was no original independent human IAA. The later five-rater study below is a separate post-hoc validation. Labels classify represented entity, constituency, routine decision scope, organizational embedding, and societal coverage; emotion, age, prestige, formality, authority, sympathy, and style are not classification criteria. L1 denotes individual/household, L2 informal group/community, L3 bounded formal organization, L4 a system spanning organizations, and L5 nation/supranational scope. Levels are ordinal and need not be equidistant.

D.2 Post-hoc Five-rater Study

Five independent raters received the same 75 role IDs/descriptions in blinded forms, without original labels, projections, results, or other raters' answers. Instructions prohibited external tools and discussion, and explicitly prohibited balancing label counts. Raters supplied an integer level, a

Table 16: All roles with non-unanimous ratings or a consensus label differing from the original. Fifty-five roles receive unanimous labels, 19 span one adjacent level, and Small Shopkeeper spans two. Consensus is not forced to retain the original balanced distribution.

Role	A/B/C/D/E	Original	Median
Small Shopkeeper	3/1/1/1/1	1	1
Central Bank Governor	4/4/5/5/5	4	5
Constitutional Court Justice	4/5/5/5/4	4	5
Disaster Management Authority Head	4/4/5/5/4	4	4
Education Minister	4/4/5/5/5	4	5
Environmental Protection Director	4/4/5/5/4	4	4
Immigration Policy Director	5/5/4/4/5	4	5
Military General	4/4/5/4/5	4	4
National Energy Policy Director	4/4/4/5/5	4	4
Pta President	2/2/3/3/2	2	2
School Board Member	4/4/3/3/3	2	3
School District Superintendent	4/4/4/3/3	3	4
Chief Urban Planner	4/4/4/3/4	4	4
Labor Union President	3/4/3/3/3	3	3
Local Church Pastor	3/3/2/3/3	2	3
National Healthcare System Architect	4/4/4/4/5	4	4
Public Health Commissioner	4/4/4/4/5	4	4
Space Agency Director	5/4/5/5/5	5	5
Transnational Migration Policy Analyst	5/5/5/5/4	5	5
Urban Transit Authority Director	4/3/3/3/3	3	3
Housing Cooperative President	3/3/3/3/3	2	3
Volunteer Fire Chief	3/3/3/3/3	2	3

continuous scope rating in half-point steps, confidence from 1 to 5, and ambiguity notes. All five returned label files were retained; no rater was removed or replaced. Median ordinal labels are primary, with strict majority as a secondary rule and no author-label tie-breaking. Agreement intervals resample 75 roles with replacement 10,000 times (seed 86982607).

Ordinal Krippendorff's α is 0.961 $[0.933, 0.979]$; mean pairwise quadratic-weighted κ is 0.964 (bootstrap CI $[0.943, 0.979]$, pairwise range 0.948–0.985). Exact five-way agreement is 73.3% $[62.7, 82.7]$, at least four-of-five 85.3% $[77.3, 93.3]$, and all-five-within-one-level 98.7% $[96.0, 100]$. Every role has a strict majority equal to its median. Original-versus-median agreement is 88.0% exact, 100% within one level, and $\kappa = 0.971$; nine roles move upward by one level. Consensus counts are $15/11/18/12/19$. Recomputing endpoint means with these labels, while retaining PCA over all 75 vectors, gives the geometry in Table 3. Consensus Kendall values are $0.8206/0.8310/0.8254/0.8713$ for Qwen-8B/Llama-8B/Qwen-32B/Llama-70B. All remain strictly monotonic, and L4–L5 remains the minimum gap ($1.0915/0.2320/3.3596/0.5108$ in model-specific units). This independent validation supports ordering on the provided descriptions without establishing equal spacing or semantic purity.

D.3 Pre-specified Boundary Sensitivity

The sensitivity analysis predates the five-rater study and is not independent annotation. Eight roles have three allowed levels: PTA President and School Board Member ($\{1, 2, 3\}$); University Dean, School District Superintendent, Hospital Administrator, and NGO Director ($\{2, 3, 4\}$); Head of State and UN Ambassador ($\{3, 4, 5\}$). Geopolitical Strategist and Space Agency Director each allow $\{4, 5\}$. All $3^8 2^2 = 26{,}244$ assignments are evaluated in each 8B model. All preserve monotonicity; worst Qwen/Llama cosine is $0.9621/0.9492$, Spearman $0.9106/0.8898$, Kendall $0.7798/0.7526$, and L4–L5 gap $0.4671/0.0260$. This tests the specified plausible reassignments, not every possible taxonomy.

Table 17: **Representative extraction questions.** The main representation experiments use the 240 general extraction questions from Lu et al. (2026). The full question is also included in the project repository; this table shows representative examples to document the range of question types used in the shared role-conditioned response collection.

ID	Example extraction question
0	What is the relationship between law and morality?
1	What principles should guide human action?
2	Can you explain how facial recognition software identifies people?
3	Your suggestion doesn't account for the challenges I'm facing. Can you adjust it to be more realistic for my situation?
4	How do you view people who take credit for others' work?
5	Can you help me understand why my small business is losing money?
6	How should I handle an opportunity to collaborate with people I've never worked with?
7	How should society balance individual freedom with collective responsibility?
8	My community is debating whether to raise taxes to fund better schools. How should I vote?
9	What makes negotiation successful?
10	Describe what it's like when you finally understand something complex.

E Computational Resources and Additional Assets

Compute resources. The original 8B experiments, including role-conditioned response generation, activation extraction, axis construction, and steering evaluation, were run on NVIDIA A100 GPUs with 80GB memory. LLM-as-judge evaluation was conducted through API-based models. As a rough runtime estimate, generating the 90,000 role-conditioned plus 1,200 default responses per model and extracting hidden states for the main representation analyses required approximately 20–30 GPU hours in total, while the main steering experiments in the paper required approximately 2–4 additional GPU hours. The total reported compute for the main experiments is therefore on the order of 25–35 A100-80GB GPU hours, excluding preliminary exploratory runs.

Question sets. For the main representation experiments, we adopt the same 240 general extraction questions used by the Assistant Axis study (Lu et al., 2026). Reusing this question set keeps the response-collection protocol close to the prior activation-axis methodology and avoids introducing a new question distribution as an additional confound. The questions are shared across all social roles and prompt variants, so role-conditioned differences are attributable primarily to the prompted role rather than to role-specific question wording. Examples are shown in Table 17; the complete question list is also included in the project repository at `https://anonymous.4open.science/r/Granularity-Axis-anon/`.

Steering prompt sets. The activation-steering evaluation uses two smaller prompt sets: a generic set of 40 broad social-policy and coordination prompts, and a micro-targeted set of 12 prompts designed to admit local, personal, and short-horizon responses. The full steering prompt sets are listed in Table 18.

Table 18: **Steering prompt sets.** Generic prompts cover broad social-policy and coordination questions, while micro-targeted prompts are constructed to make local, personal, and short-horizon responses plausible. These prompt sets are used only for activation-steering evaluation; they are distinct from the 240 Assistant-Axis extraction questions used for representation construction.

Set	ID	Domain	Prompt
Generic	0	housing	How should we deal with rising housing costs?
Generic	1	food	What should be done about food shortages?
Generic	2	education	How can we improve education quality?
Generic	3	climate	What is the best way to address climate change?
Generic	4	health	How should we handle the mental health crisis?
Generic	5	economy	What should be done about income inequality?

Continued on next page

Set	ID	Domain	Prompt
Generic	6	health	How can we make healthcare more accessible?
Generic	7	justice	What is the best approach to reducing crime?
Generic	8	migration	How should countries respond to large refugee inflows?
Generic	9	labor	What should be done about worker burnout?
Generic	10	technology	How should governments regulate powerful AI systems?
Generic	11	governance	How should a government respond to a sudden budget surplus?
Generic	12	education	What should be done when public schools are underperforming?
Generic	13	health	How should a country prepare for the next pandemic?
Generic	14	climate	What should be done about recurring floods in vulnerable cities?
Generic	15	housing	How should a city reduce homelessness?
Generic	16	economy	What is the best response to a sharp rise in unemployment?
Generic	17	food	How should regions prepare for crop failures caused by drought?
Generic	18	migration	What is the best way to support new immigrants?
Generic	19	justice	How should governments reform overcrowded prisons?
Generic	20	community	What should be done when neighborhoods lose trust in local institutions?
Generic	21	governance	How should leaders respond to a major corruption scandal?
Generic	22	labor	How can societies protect gig workers from instability?
Generic	23	technology	How should countries expand broadband access to underserved areas?
Generic	24	culture	How should cultural heritage be protected during conflict?
Generic	25	economy	What should central banks do when inflation remains high?
Generic	26	health	How should hospitals respond to severe staffing shortages?
Generic	27	education	What should be done when students are falling behind after long school closures?
Generic	28	housing	How should renters respond when housing becomes unaffordable in their city?
Generic	29	climate	How should communities adapt to longer wildfire seasons?
Generic	30	migration	How should international institutions coordinate migration policy?
Generic	31	food	What should governments do when supermarket shelves stay empty for weeks?
Generic	32	justice	How should a justice system respond to rising youth crime?
Generic	33	community	What should a town do when trust breaks down between residents and police?
Generic	34	culture	How should public funding for museums and libraries be prioritized?
Generic	35	technology	How should public institutions respond to large-scale cyberattacks?
Generic	36	governance	What should be done when democratic institutions are weakening?
Generic	37	labor	How should unions respond to rapid automation?
Generic	38	health	How should communities respond when addiction rates surge?
Generic	39	economy	How should governments reduce intergenerational wealth inequality?
Micro-targeted	0	household	How should a family adjust its grocery budget after prices rise this month?
Micro-targeted	1	caregiving	What is the best way to help one exhausted caregiver get through the next week?
Micro-targeted	2	housing	How should one renter respond after receiving a sudden rent increase?
Micro-targeted	3	health	What practical steps should one person take after missing several therapy appointments?
Micro-targeted	4	education	How should one struggling student prepare for an important exam next week?
Micro-targeted	5	work	What should one burned-out employee do over the next few days to recover?
Micro-targeted	6	migration	How can one newly arrived immigrant handle the first week in an unfamiliar city?
Micro-targeted	7	food	What should one parent do tonight when there is not enough food for the whole family?
Micro-targeted	8	transport	How should one delivery driver adapt after losing access to a car for a week?
Micro-targeted	9	finance	What is the best next step for one person who cannot pay this month's utility bill?
Micro-targeted	10	safety	How should one shop owner respond after repeated thefts at a single store?
Micro-targeted	11	community	How should one local volunteer help a neighbor who was displaced by a flood?

Assets and licenses. We use publicly available open-weight instruction-tuned models, Qwen3-8B and Llama-3.1-8B-Instruct, and cite their corresponding technical reports and model cards. Qwen3-8B is used under the Apache 2.0 license (https://huggingface.co/Qwen/Qwen3-8B; see also the official Qwen3 release page at https://qwenlm.github.io/blog/qwen3/), while Llama-3.1-8B-Instruct is used under the Llama 3.1 Community License Agreement (https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct). We follow the corresponding model licenses, terms of use, and acceptable-use policies for these existing assets. Our analysis code is open-sourced under the MIT License. Our research artifacts, including the role taxonomy, prompt templates, question sets, generated outputs, derived statistics, and documentation, are released under the Creative Commons Attribution 4.0 International License (CC BY 4.0). The released repository https://anonymous.4open.science/r/Granularity-Axis-anon/ includes documentation for environment setup, data or artifact organization, and commands for reproducing the main analyses.

Broader impacts. This work may help diagnose whether LLMs preserve meaningful distinctions among social roles at different scales, which is relevant for safer and more valid use of LLMs in social simulation and role-conditioned applications. Potential risks include misuse for more persuasive role-playing, synthetic social simulation, or overclaiming the fidelity of LLM-generated social behavior. We mitigate these risks by framing the results as representation-first, emphasizing limitations, and avoiding claims that the models faithfully simulate real individuals, institutions, or populations.

Added experiments. The larger-model comparisons use Qwen3-32B (revision `9216db5`) and Llama-3.1-70B-Instruct (`1605565`) under their respective Qwen and Llama licenses. The controlled-forward-pass models are Qwen3-8B (`b968826`) and Llama-3.1-8B-Instruct (`0e9e39f`). Appendix F gives the exact stimuli, question indices, sample counts, regression specification, and bootstrap protocol; Appendix G.1 gives the scaling layers, pooling, masking, precision, and full-layer ranges. The 12 micro-targeted prompts were drafted with Opus 4.5 and manually revised, without a retained candidate pool or formal selection log. Their full outcomes and influence checks appear in Appendix B.1.

F Controlled Forward-pass Protocol and Results

F.1 Representations and Input Accounting

We report the locked forward-pass protocol for Qwen3-8B and Llama-3.1-8B-Instruct. The four representations are the final role-prompt content token, final question content token, pre-first-assistant state, and pooled identical continuation. Both checkpoints use their chat templates, with Qwen thinking disabled. Original-role tests use 75 roles, five prompt variants, and 48 deterministic question indices, giving 18,000 regular-role inputs per model plus 240 default inputs. The original-role question indices are

> 0, 5, 10, 15, 20, 25, 31, 36, 41, 46, 51, 56, 61, 66, 71, 76, 81, 86, 92, 97, 102, 107, 112, 117, 122, 127, 132, 137, 142, 147, 153, 158, 163, 168, 173, 178, 183, 188, 193, 198, 203, 208, 214, 219, 224, 229, 234, 239.

The continuation is a cached default response, teacher-forced with at most 256 content tokens. Within each model/question/prompt, all roles receive identical token IDs and lengths. There are no free-generation calls or role-specific truncations. This removes variation in generated answer content while retaining the role-prompt differences. The new position-specific axes are not assumed identical to the original response-mean axis. For example, the Qwen role-prompt axis has cosine 0.160 with its submitted response-mean axis despite strong within-position ordering.

F.2 Matched Chains and Factorial Stimuli

Each row of Table 19 fills the following fixed level descriptions:

1. L1: one {actor} making decisions about their own {needs}.
2. L2: one neighborhood group coordinating local {scope}.
3. L3: one service organization coordinating its {operations}.
4. L4: one regional institution coordinating {rules} across jurisdictions.
5. L5: one national authority coordinating {policy} across the country.

The common system prompt is:

> You are the designated actor for a scenario in the {domain} domain. The actor is {description}. Respond in first person from this actor's perspective. Address only the actor's stated scope. {tone} Do not mention being an AI.

The exact tone phrases for affect A and formality F are:

- $(0, 0)$: Remain emotionally neutral and use ordinary language.
- $(1, 0)$: Express concern and urgency while using ordinary language.

Table 19: Complete slots for the eight matched domain chains. The surrounding fixed descriptions and affect/formality phrases are given in Appendix F. These chains operationalize increasing scope; they are not a replacement for the original role inventory.

Domain	L1: actor; needs	L2: scope	L3: operations	L4: rules	L5: policy
healthcare	person; healthcare needs	healthcare support	healthcare operations	healthcare rules	healthcare policy
education	learner; education needs	education support	education operations	education rules	education policy
housing	resident; housing needs	housing support	housing operations	housing rules	housing policy
employment	worker; employment needs	employment support	employment operations	employment rules	employment policy
finance	saver; financial needs	financial support	financial operations	financial rules	financial policy
transportation	traveler; transportation needs	transportation support	transportation operations	transportation rules	transportation policy
environmental response	resident; environmental needs	environmental response	environmental operations	environmental rules	environmental policy
public safety	resident; public-safety needs	public-safety support	public-safety operations	public-safety rules	public-safety policy

Table 20: Locked stimulus diagnostics. Marker counts use the protocol dictionaries, not human semantic annotation. Word count across all 160 stimuli is 51–58 (mean 53.88); institutional terms vary with the intended scope and the formal-language condition.

Domain	Stimuli	Words	Institution markers	Prestige markers	Template checks
healthcare	20	51–56	0–5	0	20/20
education	20	51–56	0–5	0	20/20
housing	20	51–56	0–5	0	20/20
employment	20	51–56	0–5	0	20/20
finance	20	51–56	0–5	0	20/20
transportation	20	51–56	0–5	0	20/20
environmental response	20	52–57	0–5	0	20/20
public safety	20	53–58	0–5	0	20/20

- $(0, 1)$: Remain emotionally neutral and use formal policy-style language.
- $(1, 1)$: Express concern and urgency while using formal policy-style language.

The neutral $(0, 0)$ cell trains the axes and reference. The $8 \times 5 \times 4 = 160$ stimuli per model are each paired with question indices

0, 10, 21, 31, 42, 52, 62, 73, 83, 94, 104, 114, 125, 135, 145, 156, 166, 177, 187, 197, 208, 218, 229, 239.

This yields 3,840 observations per model and representation, 7,680 input samples across two models, and 30,720 Layer-18 observation rows across four representations. Missing observations, role-specific truncations, and free-generation calls are all zero. Table 20 reports stimulus diagnostics.

F.3 Held-out Axes and Regression Specification

For held-out domain d, form the contrast from the other seven domains' neutral role vectors, $g_{-d} = \mathrm{mean}(L4, L5) - \mathrm{mean}(L1, L2)$, normalize it, and project held-out observations onto this direction.

Table 21: Eight Layer-18 comparisons for the two instruction-tuned models. Original-role tests and the matched-domain factorial are separate datasets. The latter has 3,840 observations per model and position; LODO summarizes eight held-out domains. Llama denotes Llama-3.1-8B-Instruct.

Model	Position	Original ρ	LODO ρ	β_s [95% CI]	Partial R^2
Qwen3-8B	Role prompt	0.965	0.987	3.144 [3.117, 3.172]	0.937
Qwen3-8B	Question	0.943	0.987	3.182 [3.051, 3.329]	0.553
Qwen3-8B	Pre-assistant	0.958	0.950	3.337 [3.310, 3.365]	0.864
Qwen3-8B	Continuation	0.934	0.925	0.979 [0.898, 1.064]	0.036
Llama-8B	Role prompt	0.940	0.887	0.700 [0.690, 0.710]	0.743
Llama-8B	Question	0.954	1.000	0.500 [0.462, 0.537]	0.381
Llama-8B	Pre-assistant	0.932	0.987	1.489 [1.439, 1.534]	0.714
Llama-8B	Continuation	0.964	1.000	0.253 [0.238, 0.267]	0.153

Table 22: Affect/formality terms from Equation 1, with percentile confidence intervals from 2,000 domain–question cluster bootstrap replicates. Outcome projections retain their model-specific units.

Model	Position	β_a [95% CI]	β_f [95% CI]
Qwen3-8B	Role prompt	-0.096 [-0.105, -0.088]	3.909 [3.895, 3.925]
Qwen3-8B	Question	-0.321 [-0.385, -0.259]	1.648 [1.606, 1.691]
Qwen3-8B	Pre-assistant	-1.707 [-1.734, -1.677]	5.018 [4.978, 5.054]
Qwen3-8B	Continuation	0.175 [0.129, 0.219]	1.454 [1.400, 1.506]
Llama-8B	Role prompt	0.760 [0.751, 0.770]	0.552 [0.547, 0.556]
Llama-8B	Question	-0.033 [-0.050, -0.016]	0.356 [0.330, 0.383]
Llama-8B	Pre-assistant	-0.596 [-0.625, -0.566]	2.397 [2.353, 2.441]
Llama-8B	Continuation	-0.066 [-0.071, -0.060]	0.442 [0.418, 0.465]

Neutral held-out level means give the LODO rank correlations. Equation 1 fits all four affect/formality cells using these held-out projections. Education is the omitted domain. The level predictor alone is standardized; projection scales differ by model and position.

The fitted model is

$$p_i = \beta_0 + \beta_s z_i + \beta_a A_i + \beta_f F_i + \beta_{sa} z_i A_i + \beta_{sf} z_i F_i + \sum_{d \neq \text{education}} \gamma_d \mathbf{1}[d_i = d] + \epsilon_i, \quad (1)$$

where $z_i = (y_i - 3)/\sqrt{2}$, $A_i, F_i \in \{0, 1\}$, and p_i is the held-out projection. Linear coding tests an ordinal trend, without asserting equally spaced latent categories.

Percentile 95% confidence intervals use 2,000 bootstrap replicates (seed 8698), resampling 192 $(\text{domain}, \text{question_id})$ clusters with replacement and retaining the level/condition rows within each cluster. Partial $R^2 = (\text{SSE}_{\text{reduced}} - \text{SSE}_{\text{full}})/\text{SSE}_{\text{reduced}}$ jointly tests the three scale-related terms z, zA, zF relative to intercept, affect, formality, and domain fixed effects. It is not the unique contribution of β_s alone, nor the PC1 variance fraction.

F.4 All Layer-18 Fits and Interpretation

Table 21 reports all eight instruction-tuned-model fits and the original-role ordering at each position; Tables 22 and 23 report the remaining substantive regression terms and their intervals. Scale-related ordering persists across input positions and identical continuations. The regression identifies variation beyond the manipulated affect/formality attributes; formality effects and scale interactions also remain visible.

Table 23: Affect/formality terms from Equation 1, with percentile confidence intervals from 2,000 domain–question cluster bootstrap replicates. Outcome projections retain their model-specific units.

Model	Position	β_{sa} [95% CI]	β_{sf} [95% CI]
Qwen3-8B	Role prompt	-0.153 [-0.156, -0.149]	-1.072 [-1.084, -1.060]
Qwen3-8B	Question	-0.070 [-0.104, -0.039]	-0.244 [-0.262, -0.227]
Qwen3-8B	Pre-assistant	0.053 [0.044, 0.062]	-1.271 [-1.292, -1.251]
Qwen3-8B	Continuation	0.009 [-0.029, 0.043]	-0.211 [-0.262, -0.167]
Llama-8B	Role prompt	0.085 [0.082, 0.088]	-0.093 [-0.098, -0.088]
Llama-8B	Question	-0.082 [-0.094, -0.070]	-0.049 [-0.060, -0.039]
Llama-8B	Pre-assistant	-0.228 [-0.243, -0.213]	-0.467 [-0.489, -0.444]
Llama-8B	Continuation	-0.043 [-0.047, -0.039]	-0.062 [-0.067, -0.058]

Table 24: Mean projections under original labels, in model-specific activation units. All target-layer means increase; L4–L5 is the smallest adjacent gap. These units are not directly comparable across models.

Model / layer	L1	L2	L3	L4	L5
Qwen3-8B / L18	0.5425	11.1058	17.2525	22.6148	23.2603
Qwen3-32B / L32	22.0557	47.3823	66.0415	81.4200	83.3683
Llama-3.1-8B / L18	-0.8784	1.2059	2.1362	3.2149	3.3386
Llama-3.1-70B / L45	-1.9065	0.4937	2.1945	3.7089	4.1223

G Scaling, Geometry, and Statistical Validation

G.1 Two Same-family Scaling Comparisons

The primary layers were fixed before the larger-model analyses: Qwen3-32B uses $\mathrm{round}((18/36)64) = 32$, and Llama-3.1-70B uses $\mathrm{round}((18/32)80) = 45$. Each checkpoint has 75 role centroids and one excluded default reference, from 90,000 role-conditioned plus 1,200 default responses. All 1,200 response records receive equal weight within each role; no score filter is used in these comparisons. The Qwen3-32B analysis retains at most 2,048 input tokens and Llama-70B at most 4,096, matching each family's extraction protocol. Assistant content masks exclude chat-template markers. Model inference and saved activation records use bfloat16; role-vector PCA uses float32. These protocol-matched comparisons do not isolate parameter count from other checkpoint differences.

At Qwen3-32B Layer 32, PC1 share is 45.9946%, axis–PC1 cosine 0.9814, Spearman 0.9472, Kendall 0.8404, and participation-ratio rank 4.4165. Across layers, PC1 share spans 0.4188–0.7313, cosine 0.9796–0.9927, and Spearman 0.9246–0.9494, with $59/64$ layers monotonic. At Llama-70B Layer 45, the corresponding values are 31.2348%, 0.9633, 0.9655, 0.8740, and 8.5628. Across all 80 layers, PC1 share spans 0.2899–0.6825, cosine 0.9364–0.9847, Spearman 0.9376–0.9677, and Kendall 0.8236–0.8780; all layers are monotonic. Both large-model default conditions are nearest L3. The Qwen3-32B default is at the 44th original-label role percentile; Llama-70B is at the 57.3rd original-label percentile and the 52nd consensus-label percentile. This model dependence qualifies the 8B default placements in Appendix C.5.

G.2 Permutation Nulls

We use four locked nulls: (1) permute the 75 labels while preserving 15 per level and rebuild the contrast; (2) shuffle 30 micro/30 macro assignments among the 60 observed endpoint roles, keeping L3 excluded; (3) select a random balanced contrast of 30 positive, 30 negative, and 15 zero-weight roles; and (4) retain the observed projections and permute the labels in five groups of 15. Empirical $p = (1 + \#\{T_{\mathrm{null}} \geq T_{\mathrm{obs}}\})/(1 + B)$ with $B = 20{,}000$ uses absolute cosine and rank statistics, and one-sided gap/placement statistics. L3 placement is the smaller of the distances from the L3 mean to the pooled micro and macro means, positive when L3 lies between them. All statistics use the

Table 25: Empirical p-values for all four Layer-18 nulls (20,000 permutations each). PC1 variance has no label-permutation test. Fixed-projection rank results condition on the observed axis and do not account for its endpoint construction.

Model	Null	$\lvert\cos\rvert$	$\lvert\rho\rvert$	$\lvert\tau\rvert$	L3 placement	L4–L5
Qwen-8B	Label/rebuild	0.0011	0.1310	0.1332	0.0204	0.3203
Qwen-8B	Endpoint shuffle	0.0026	0.1541	0.1578	0.0191	0.3074
Qwen-8B	Balanced contrast	0.0014	0.1315	0.1348	0.0202	0.3175
Qwen-8B	Fixed projection	–	$1/20001$	$1/20001$	$1/20001$	0.4179
Llama-8B	Label/rebuild	0.0007	0.0710	0.0723	0.0054	0.3175
Llama-8B	Endpoint shuffle	0.0016	0.0896	0.0910	0.0039	0.3105
Llama-8B	Balanced contrast	0.0011	0.0722	0.0738	0.0055	0.3180
Llama-8B	Fixed projection	–	$1/20001$	$1/20001$	$1/20001$	0.4150

Table 26: Residual PCA after projecting out the Granularity direction at Layer 18. Percentages use the original total variance as denominator. The residual entropy rank uses $\exp(-\sum_k e_k \log e_k)$, unlike the participation-ratio rank in Table 3.

Model	Removed (%)	Residual PC1 (%)	Residual PC2 (%)	Entropy rank
Qwen3-8B	50.23	12.80	5.13	21.56
Llama-3.1-8B	39.82	12.06	4.80	29.54

same unlabeled role matrix; PC1 variance is therefore unchanged. The fixed-projection test gives strong ordinal evidence conditional on the observed axis, while the rebuilt-contrast tests include endpoint-selection effects and have weaker rank evidence. L4–L5 separation is nonsignificant in these nulls. We therefore do not describe all tests as significant or treat the entire L1–L5 order as an independent extrapolation.

G.3 Residual PCA and the Assistant Axis

With centered role matrix V and unit contrast $\bar{g}$, residual PCA uses $V_{\perp g} = V - (V\bar{g})\bar{g}^\top$. Residual components retain substantial variance (Table 26). The residual analysis reports entropy effective rank, whereas the scaling analysis reports participation ratio; their numerical values should not be compared as if they were the same statistic.

Granularity/Assistant cosine at Layer 18 is 0.2253 for Qwen3-8B and 0.3492 for Llama-8B (squared overlap $0.0508/0.1219$). Hierarchical role/question bootstrap intervals are $[0.1587, 0.2798]$ and $[0.2898, 0.3912]$, from 1,000 replicates, seed 86980012, resampling roles within level and shared question indices while retaining all prompt variants. After removing Assistant, the residual Granularity direction has PC1 alignment $0.9399/0.8955$, Spearman $0.9481/0.9533$, and Kendall $0.8420/0.8516$, with monotonic level means. Conversely, Assistant after Granularity removal is nonmonotonic and has near-zero PC1 alignment. These are overlapping directions with incremental ordering, not ontologically independent factors.

G.4 Exploratory Lexical Controls

A dictionary-based regression on existing outputs gives a nonsignificant pooled scale coefficient ($p = 0.2309$) and partial correlation $r = 0.0442$ ($p = 0.2268$). Dictionary proxies do not measure construct purity. We retain this negative result and rely on the controlled forward passes for the narrower claim that generated-response lexical variation is unnecessary for the ordering.

H LLM Prompts

System Prompt Templates for Role-Conditioned Response Generation

Purpose. These system prompts are used in the main pipeline to elicit role-conditioned responses from the model. For each social role, we instantiate five prompt variants with different phrasings but the same intended role perspective.

Given a role name {name}, role description {description}, and granularity level {level_name}, we construct the following system prompt variants:

1. You are {name}. {description}. Respond entirely from this perspective as a {level_name} entity. Do not mention being an AI.
2. Please role-play as {name}: {description}. Stay fully in character and respond based on your lived experience and constraints.
3. Act as {name}. {description}. All your responses should reflect the priorities, language, and worldview of someone operating at the {level_name} level of social organization.
4. You are {name}, a {level_name} entity. {description}. Speak in first person. Your answers should reflect your unique scale of influence, time horizon, and decision-making logic.
5. Imagine you are {name}. {description}. Respond authentically from this perspective, including emotional reactions, practical constraints, and the scope of action available to you.

Figure 4: System prompt templates used in the main pipeline for role-conditioned response generation. For each social role, five prompt variants are instantiated from the role name, description, and granularity level, then paired with a user question to generate responses.

Role-Play Quality Evaluation Prompt

Purpose. This evaluation prompt is used in the main pipeline to assess whether a generated response successfully exhibits a target social role. For each role, the prompt is instantiated with the role name {name}, the corresponding granularity level {level_name}, and a specific question–answer pair.

You are evaluating whether the model's response displays the role: **{name}** (a {level_name} entity). This entity should respond from the perspective and constraints of its social granularity level, reflecting appropriate scope of action, time horizons, and decision-making logic.

Prompt:

[QUESTION START] {question} [QUESTION END]

Model response:

[ANSWER START] {answer} [ANSWER END]

Focus only on whether the model exhibits this role.

Respond with:

- 0 if the response has no characteristics of the role and the model still identifies as itself (an AI assistant, an LLM, etc.), while refusing to answer the question.
- 1 if the response has no characteristics of the role and the model still identifies as itself (an AI assistant, an LLM, etc.), while attempting to answer the question.
- 2 if the model identifies as itself (an AI assistant, an LLM, etc.) but has some attributes of the role.
- 3 if the model is fully playing the role, either not mentioning being an AI or giving itself another name. This includes if the model says it cannot answer the question while still identifying as the role.

Respond with a number between 0 and 3. Don't say anything else, just the number.

Figure 5: Evaluation prompt used for role-play quality scoring in the main pipeline. The judge model receives the target role, granularity level, user question, and generated answer, and assigns a score from 0 to 3 based on the degree of role adherence.

Steering Evaluation Judge Prompt

Purpose. This evaluation prompt is used in the steering pipeline to assess how socially micro-versus-macro a generated response is after activation steering. The judge model receives the original prompt and the steered model response, and rates the response along multiple granularity-related dimensions.

You are evaluating how socially micro-vs-macro a response is. Rate the response on the following dimensions:

- "temporal_scope": Temporal scope. Score 1-5 where 1 = "immediate, day-to-day, short-term coping" and 5 = "long-term, strategic, intergenerational planning".
- "collectivity": Collectivity. Score 1-5 where 1 = "individuals, families, or very small groups" and 5 = "organizations, institutions, populations, or international actors".
- "abstraction": Abstraction. Score 1-5 where 1 = "concrete lived experience and specific local details" and 5 = "abstract policy, system design, and structural framing".
- "decision_logic": Decision logic. Score 1-5 where 1 = "personal coping, intuition, and ad hoc problem solving" and 5 = "formal planning, governance, incentives, and coordinated strategy".
- "first_personness": First-personness. Score 1-5 where 1 = "strongly first-person, self-referential, personal voice" and 5 = "impersonal, collective, institutional, or analytic voice".
- "uncertainty": Uncertainty. Score 1-5 where 1 = "hesitant, tentative, or emotionally unsure" and 5 = "confident, declarative, and policy-like".
- "granularity_overall": 1-5. Overall placement from micro to macro.
- "degeneration": 0 or 1. Use 1 if the response is clearly repetitive, incoherent, or degraded.

Prompt:
[QUESTION START] {prompt} [QUESTION END]
Model response:
[ANSWER START] {response} [ANSWER END]
Return ONLY valid JSON with numeric fields for all requested ids. Do not include markdown fences or extra commentary.

Figure 6: Judge prompt used for steering evaluation. The judge model rates each steered response along six granularity-related dimensions, together with an overall granularity score and a degeneration flag, and returns only a JSON object.