ments use medium-sized instruction-tuned models (7B–14B), prompted across five discrete levels (0–4), with outputs judged via pairwise comparisons by a stronger LLM. Rank correlations between intended and judged levels provide a robust measure of controllability across single- and dual-concept conditions.

In this work, we opt to evaluate prompting, both due to it being a widely accessible method of control, as well have been shown to perform better than many more complicated representation engineering methods proposed in literature for single-concept control (Wu et al., 2025). Our findings are insightful: while prompting achieves sensible fine-grained calibration for individual concepts, performance often **drops sharply in the dual-concept setting**, even for pairs that are intuitively orthogonal. This suggests that concept dimensions are entangled in ways that resist naive composition.

More broadly, our evaluation framework is model- and method-agnostic, providing a standardized way to measure controllability across future techniques. By establishing clear metrics and identifying common failure modes, we aim to encourage the development of more robust methods that enable interpretable, multidimensional stylistic control in language models.

2 Fine-grained Control Evaluation Framework

We define the task of fine-grained concept control as follows. Let $\mathcal{C}$ denote the set of controllable concepts, where each $C \in \mathcal{C}$ represents a semantic dimension such as `humor` or `formality`. Each concept C is associated with a discrete scale of levels $\mathcal{L} = \{0, 1, \ldots, L\}$, where $\ell = 0$ denotes no presence and $\ell = L$ denotes maximal presence of the concept. The objective is to evaluate the fine-grained control abilities of a language generation model, $\mathcal{G}(\cdot)$.

Single-concept control. Given a textual context x and a target concept $C_a \in \mathcal{C}$ with desired level $\ell \in \mathcal{L}$, the generation model G, produces an output,

$$y_\ell = G(x, C_a, \ell). \tag{1}$$

Across all levels $\ell \in \{0, \ldots, L\}$, this yields a set of outputs $\{y_0, \ldots, y_L\}$. For a perfect model $\mathcal{G}$, the ranking of generations by their realized strength of concept C_a would be strictly monotonic in ℓ, i.e. aligned with the intended order $(0, 1, \ldots, L)$.

Dual-concept control. Now consider two concepts $C_a, C_b \in \mathcal{C}$, assumed to be semantically distinct. The user specifies desired levels $(\ell_a, \ell_b) \in \mathcal{L}^2$, and the model generates,

$$y = G(x, C_a, \ell_a, C_b, \ell_b). \tag{2}$$

To assess controllability of C_a while holding C_b fixed at $\ell_b = j$, we obtain generations $\{y_{\ell_a, j}\}_{\ell_a=0}^{L}$ and measure how well their ranking aligns with the intended order $(0, 1, \ldots, L)$ for C_a. This process is repeated for each $j \in \mathcal{L}$, and the overall performance can be averaged over all fixed levels, j, giving a controllability profile of C_a given C_b. Evaluation is performed symmetrically with C_b as the target concept. In addition to the fixed-level setting, we also consider a *randomized secondary concept* variant. Here, for each target concept C_a, we sample $\ell_b \sim \text{Uniform}(\mathcal{L})$ independently for each generation. This variant tests whether control over C_a is disentangled from the level of C_b.

Judge-based evaluation. To assess whether the generated outputs $\{y_\ell\}$ follow the intended order, we use a judge model J that performs pairwise comparisons between generations[1]. Each pair (y_i, y_j) is presented in both orders to avoid position bias, and we define the preference score as,

$$s(i, j) \;=\; \tfrac{1}{2}\Big(J(y_i, y_j) + \big(1 - J(y_j, y_i)\big)\Big), \tag{3}$$

where $J(y_i, y_j) \in \{0, 0.5, 1\}$ denotes whether the judge considers y_i to exhibit more of the target concept than y_j (with 0.5 for a tie). By summing the pairwise scores for each y_ℓ against other levels, we derive an empirical ranking $\hat{r}$ over $\{y_\ell\}$ and measure correlation with the intended ranking $r = (0, 1, \ldots, L)$ using Spearman (Spearman, 1904) ρ correlation. The overall ability of a generation model $\mathcal{G}(\cdot)$ to perform fine-grained control of the selected concepts is quantified as the average of the correlation metrics across a dataset of N contexts $\{x^{(1)}, \ldots, x^{(N)}\}$. Letting $\rho^{(n)}$ denote the Spearman correlation for instance $x^{(n)}$, we get $\bar{\rho} = \frac{1}{N}\sum_{n=1}^{N} \rho^{(n)}$. This aggregated scores summarizes the model's controllability across the dataset. In all experiments, we set $L = 4$, corresponding to five levels of control for each concept.

[1] In preliminary experiments, we also evaluated a *listwise* single-inference approach with the judge-LLM that ranks all responses in a single inference (with responses presented in randomized order). We observed substantial position bias, where the first-presented sample was disproportionately ranked lowest (Appendix H, Table 29).