Beyond Steering: Evaluating Fine-Grained and Multi-Concept Control in LLMs

Arya Labroo
University of Cambridge
`al2135@cam.ac.uk`

Ivaxi Sheth
CISPA Helmholtz Center for Information Security
`ivaxi.sheth@cispa.de`

Vyas Raina
Apta
`vyas@apta.chat`

Amaani Ahmed
Royal Holloway, University of London
`nnjt201@live.rhul.ac.uk`

Mario Fritz
CISPA Helmholtz Center for Information Security
`fritz@cispa.de`

Abstract

Large Language Models (LLMs) have achieved remarkable success across a wide range of generative tasks. However, users often desire explicit control over the presence and extent of specific *concepts* in the generated text; for example, controlling how *humorous* or *persuasive* a passage should be. While prior work in prompt engineering and representation-based concept steering has enabled coarse directional control, these methods rarely address the need for *fine-grained* specification, such as explicitly setting a concept on a continuous scale. The challenge is amplified when controlling multiple concepts simultaneously, where the interaction between concepts may interfere with precise control. In this work, we introduce an evaluation framework to systematically measure the fine-grained controllability of LLMs in both single- and dual-concept settings. Our findings reveal that while simple prompt-based approaches show promise for single-concept fine-grained control, performance degrades substantially in the more challenging two-concept scenario. These results suggest that current prompting strategies are insufficient for robust multi-concept control. We encourage future work to explicitly develop methods for fine-grained control that maintain effectiveness from the single-concept to multi-concept setting.

1 Introduction

Large Language Models (LLMs) are increasingly deployed in applications such as chat assistants, creative writing, education, and decision support Achiam et al. [2023], Brooks et al. [2024], Jia et al. [2024], Singhal et al. [2025], Lee et al. [2024], Modi et al. [2024], Bashiri and Kowsari [2024]. Beyond high-quality outputs, users often desire *fine-grained control* over how specific *concepts* appear in generated text. For example, a writing assistant might let users set a "humor" slider from 0 to 4, or adjust persuasiveness, formality, or politeness depending on audience and context. Such interfaces require models to reliably modulate outputs along interpretable concept dimensions.

Prior work has explored control via prompting Yang et al. [2023b], Brown et al. [2020], Ajwani et al. [2024] and representation engineering Zou et al. [2023], Kumar et al. [2023], Wehner et al. [2025],

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Evaluating the Evolving LLM Lifecycle.

Hao and Linzen [2023], Arora et al. [2024]. Prompting methods enable only coarse directional control, while representation-based approaches (e.g., linear concept directions Zou et al. [2023]) offer stronger steering but are not easily exposed as intuitive user controls and remain limited in supporting multi-level fine-grained adjustment. Crucially, neither class of methods has been systematically evaluated for fine-grained, multi-concept control.

We propose an evaluation framework for assessing the fine-grained controllability of LLMs in two settings: (1) *single-concept control*, where text is generated at specified levels of one concept, and (2) *multi-concept control*, where two concepts must be jointly controlled. Argument generation serves as a natural testbed, as arguments can vary systematically in persuasiveness, assertiveness, formality, and other related dimensions.

We evaluate six linguistically distinct and practically relevant concepts: humor, persuasiveness, clarity, politeness, assertiveness, and formality. Medium-sized instruction-tuned models (7B–12B) are prompted across five discrete levels (0–4), and outputs are assessed by a strong judge LLM via pairwise comparisons. Rank-based correlations with the intended levels provide a robust measure of controllability.

Our results show that prompting achieves some fine-grained control in the single-concept setting but degrades sharply in dual-concept scenarios, even for concept pairs that should, in theory, be disentangled. This highlights the limitations of current methods and motivates the development of more robust approaches. Our framework offers a principled basis for evaluating such future techniques.

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 $(0, 1, \ldots, L)$.

Multi-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,

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

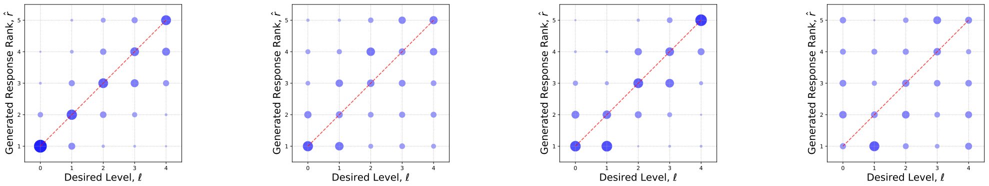
To assess controllability of C_a while holding C_b fixed at $\ell_b = j$, we obtain generations $\{y_{\ell_a, j}\}_{\ell_a=0}^{L}$ and measure how well their realized ranking aligns with the intended order $(0, 1, \ldots, L)$ for C_a. This process is repeated for each $j \in \mathcal{L}$, and the overall performance can be averaged over all fixed levels, j, yielding 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,

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

---

[1] We found pairwise comparisons are more reliable than prompting a judge-LLM to rank all generated responses in a single inference.



(a) Humor (b) Humor | Persuasiveness (c) Persuasiveness (d) Persuasiveness | Humor

Figure 1: Model-generated response rank of the target concept versus the desired level. Point size and density indicate the number of samples at each coordinate. Results shown for Llama 3.1 with the secondary concept level *randomly* sampled. For example, "Humor | Persuasiveness" denotes responses generated independently for each humor level (target concept) while persuasiveness is randomly set for each inference.

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 and Kendall Kendall [1938] τ. 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 $\rho^{(n)}$ and $\tau^{(n)}$ denote the Spearman and Kendall correlations for instance $x^{(n)}$, we get $\bar{\rho} = \frac{1}{N}\sum_{n=1}^{N} \rho^{(n)}$ and $\bar{\tau} = \frac{1}{N}\sum_{n=1}^{N} \tau^{(n)}$. These aggregated scores summarize the model's controllability across the dataset. In all experiments, we set $L = 4$, corresponding to five levels of control for each concept.

3 Experiments

3.1 Setup

Models. We evaluate medium-sized, instruction-tuned LLMs in the 7B–12B parameter range: Mistral 7B Jiang et al. [2023], Llama 3.1–8B Dubey et al. [2024], and Gemma 3–12B Team et al. [2025]. These models are representative of widely deployed generation systems that are computationally affordable while still capable of complex stylistic control. Qwen2.5–72B Yang et al. [2025] was selected as the judge-LLM due to its reliable performance in evaluation settings Viswanathan et al. [2025], Zhang et al. [2025], Gera et al. [2024].

Data and Concepts. We use the Persuasion dataset Durmus et al. [2024], consisting of 75 unique claims, discarding the associated arguments and scores. Each claim serves as a prompt for generating supporting arguments, making argument generation a natural testbed for systematically varying stylistic and pragmatic dimensions.

We evaluate six concepts: humor, persuasiveness, clarity, politeness, assertiveness, and formality. These were selected for their (i) relevance to real-world applications, (ii) linguistic distinctiveness supported by factor-analytic studies Nevid and Rathus [1979], Kearney et al. [1984], and (iii) practical motivation for independent adjustment (e.g., writing assistants, educational tools, debate preparation). For multi-concept evaluation, we study three pairs—humor–persuasiveness, clarity–politeness, and assertiveness–formality—chosen because theoretical and empirical evidence suggests they are disentangled dimensions Biber [1995], Bar-Or et al. [2022].

3.2 Results

Tables 1–3 report the average Spearman ($\bar{\rho}$) and Kendall ($\bar{\tau}$) correlations between desired concept levels and the empirical ranks of generated responses, following the framework in Section 2.

Across all three concept pairs, models retain some ability to control concepts in the single-concept setting but degrade substantially under dual-concept control. For humor–persuasiveness, single-concept performance is moderate to strong ($\bar{\rho} = 0.65/0.66$ for Mistral, $0.78/0.76$ for Llama, and $0.92/0.97$ for Gemma), but drops sharply when the secondary concept is introduced (e.g., humor with persuasiveness fixed falls to 0.30, 0.47, 0.87). Similar degradation occurs in the reverse direction. Figure 1 illustrates this trend for the humor–persuasiveness pair: moving from single- to dual-concept settings significantly harms fine-grained controllability. For clarity–politeness, clarity is poorly controlled even alone (0.16, 0.15, 0.62), and collapses almost entirely when politeness is fixed or randomized. Politeness is more controllable on its own (0.63, 0.73, 0.96) but likewise declines when clarity is introduced. For assertiveness–formality, assertiveness control ranges from weak to strong (0.49, 0.73, 0.97), with further drops when formality is added. Formality itself is well modeled (0.81, 0.90, 0.98), but deteriorates when assertiveness is varied.

General trends. Three consistent patterns emerge. (i) Gemma outperforms Mistral and Llama across all settings, showing greater robustness to interference. (ii) Dual-concept inputs substantially harm the ability of single-concept control. (iii) Randomized secondary concepts typically cause even larger losses than fixed ones, highlighting weak disentanglement between dimensions.

	Spearman ($\bar{\rho}$)			Kendall ($\bar{\tau}$)		
	Mistral	**Llama3.1**	**Gemma3**	**Mistral**	**Llama3.1**	**Gemma3**
C_a (single)	$0.65_{\pm 0.32}$	$0.78_{\pm 0.26}$	$0.92_{\pm 0.11}$	$0.56_{\pm 0.31}$	$0.70_{\pm 0.26}$	$0.87_{\pm 0.14}$
$C_a \mid C_b$ fixed	$0.30_{\pm 0.52}$	$0.47_{\pm 0.44}$	$0.87_{\pm 0.15}$	$0.25_{\pm 0.44}$	$0.40_{\pm 0.38}$	$0.81_{\pm 0.18}$
$C_a \mid C_b$ rand	$0.30_{\pm 0.47}$	$0.42_{\pm 0.51}$	$0.88_{\pm 0.16}$	$0.26_{\pm 0.39}$	$0.36_{\pm 0.44}$	$0.82_{\pm 0.18}$
C_b (single)	$0.66_{\pm 0.33}$	$0.76_{\pm 0.25}$	$0.97_{\pm 0.03}$	$0.57_{\pm 0.31}$	$0.68_{\pm 0.26}$	$0.95_{\pm 0.07}$
$C_b \mid C_a$ fixed	$0.32_{\pm 0.48}$	$0.30_{\pm 0.44}$	$0.76_{\pm 0.24}$	$0.27_{\pm 0.40}$	$0.25_{\pm 0.37}$	$0.67_{\pm 0.25}$
$C_b \mid C_a$ rand	$0.23_{\pm 0.51}$	$0.15_{\pm 0.45}$	$0.68_{\pm 0.29}$	$0.20_{\pm 0.42}$	$0.12_{\pm 0.36}$	$0.58_{\pm 0.28}$

Table 1: **humor–persuasiveness**. Single and dual-concept evaluations.

	Spearman ($\bar{\rho}$)			Kendall ($\bar{\tau}$)		
	Mistral	**Llama3.1**	**Gemma3**	**Mistral**	**Llama3.1**	**Gemma3**
C_a (single)	$0.16_{\pm 0.62}$	$0.15_{\pm 0.54}$	$0.62_{\pm 0.43}$	$0.13_{\pm 0.52}$	$0.13_{\pm 0.44}$	$0.56_{\pm 0.40}$
$C_a \mid C_b$ fixed	$0.10_{\pm 0.51}$	$0.08_{\pm 0.53}$	$0.28_{\pm 0.55}$	$0.08_{\pm 0.43}$	$0.07_{\pm 0.45}$	$0.24_{\pm 0.48}$
$C_a \mid C_b$ rand	$0.05_{\pm 0.51}$	$0.05_{\pm 0.47}$	$0.28_{\pm 0.52}$	$0.04_{\pm 0.42}$	$0.05_{\pm 0.39}$	$0.22_{\pm 0.44}$
C_b (single)	$0.63_{\pm 0.40}$	$0.73_{\pm 0.28}$	$0.96_{\pm 0.08}$	$0.56_{\pm 0.38}$	$0.66_{\pm 0.29}$	$0.93_{\pm 0.11}$
$C_b \mid C_a$ fixed	$0.25_{\pm 0.50}$	$0.48_{\pm 0.44}$	$0.85_{\pm 0.21}$	$0.21_{\pm 0.43}$	$0.41_{\pm 0.39}$	$0.80_{\pm 0.24}$
$C_b \mid C_a$ rand	$0.21_{\pm 0.46}$	$0.48_{\pm 0.48}$	$0.78_{\pm 0.27}$	$0.17_{\pm 0.39}$	$0.43_{\pm 0.42}$	$0.71_{\pm 0.28}$

Table 2: **clarity–politeness**. Single and dual-concept evaluations.

	Spearman ($\bar{\rho}$)			Kendall ($\bar{\tau}$)		
	Mistral	**Llama3.1**	**Gemma3**	**Mistral**	**Llama3.1**	**Gemma3**
C_a (single)	$0.49_{\pm 0.40}$	$0.73_{\pm 0.34}$	$0.97_{\pm 0.07}$	$0.42_{\pm 0.34}$	$0.65_{\pm 0.32}$	$0.94_{\pm 0.11}$
$C_a \mid C_b$ fixed	$0.56_{\pm 0.41}$	$0.46_{\pm 0.45}$	$0.93_{\pm 0.09}$	$0.46_{\pm 0.36}$	$0.39_{\pm 0.39}$	$0.88_{\pm 0.14}$
$C_a \mid C_b$ rand	$0.50_{\pm 0.38}$	$0.38_{\pm 0.49}$	$0.89_{\pm 0.10}$	$0.43_{\pm 0.34}$	$0.33_{\pm 0.42}$	$0.81_{\pm 0.15}$
C_b (single)	$0.81_{\pm 0.20}$	$0.90_{\pm 0.11}$	$0.98_{\pm 0.03}$	$0.71_{\pm 0.21}$	$0.84_{\pm 0.16}$	$0.97_{\pm 0.05}$
$C_b \mid C_a$ fixed	$0.62_{\pm 0.33}$	$0.58_{\pm 0.36}$	$0.97_{\pm 0.05}$	$0.52_{\pm 0.30}$	$0.49_{\pm 0.32}$	$0.94_{\pm 0.09}$
$C_b \mid C_a$ rand	$0.46_{\pm 0.44}$	$0.44_{\pm 0.47}$	$0.89_{\pm 0.14}$	$0.39_{\pm 0.38}$	$0.37_{\pm 0.40}$	$0.81_{\pm 0.20}$

Table 3: **assertiveness–formality**. Single and dual-concept evaluations.

4 Conclusions

This work introduced a systematic framework to evaluate fine-grained control of stylistic concepts in large language models (LLMs). Through experiments on three pairs of linguistically distinct concepts, we found that while models offer some degree of single-concept controllability, performance drops sharply in the multi-concept setting. Even for concept pairs that should in principle be disentangled, outputs often fail to align with the intended levels. These findings highlight a fundamental challenge: current LLMs struggle to provide fine-grained, disentangled control across multiple stylistic dimensions.

Future work could extend the framework in several directions. On the methodological side, hybrid approaches that combine prompting with representation-level interventions or logit-biasing may improve control. Training objectives that explicitly enforce disentanglement across concepts represent another promising direction. On the evaluation side, scaling the framework to larger concept sets, more diverse domains, and stronger LLMs will help assess whether controllability improves with model capacity or data diversity. Finally, exploring user-facing interfaces such as multi-dimensional "sliders" would bridge technical methods with practical applications, enabling end-users to reliably tailor model outputs along multiple stylistic dimensions.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv*, 2023.

Rohan Deepak Ajwani, Zining Zhu, Jonathan Rose, and Frank Rudzicz. Plug and play with prompts: A prompt tuning approach for controlling text generation. *arXiv preprint arXiv:2404.05143*, 2024.

Aryaman Arora, Dan Jurafsky, and Christopher Potts. Causalgym: Benchmarking causal interpretability methods on linguistic tasks. *arXiv preprint arXiv:2402.12560*, 2024.

Nabiha Asghar. Yelp dataset challenge: Review rating prediction. *arXiv preprint arXiv:1605.05362*, 2016.

Ella Bar-Or, Tom Regev, Paz Shaviv, and Noam Tractinsky. Towards a sociolinguistics-based framework for the study of politeness in human-computer interaction, 2022. URL `https://arxiv.org/abs/2202.09901`.

Masoud Bashiri and Kamran Kowsari. Transformative influence of llm and ai tools in student social media engagement: Analyzing personalization, communication efficiency, and collaborative learning. *arXiv preprint arXiv:2407.15012*, 2024.

Daniel Beaglehole, Adityanarayanan Radhakrishnan, Enric Boix-Adserà, and Mikhail Belkin. Aggregate and conquer: detecting and steering llm concepts by combining nonlinear predictors over multiple layers. *CoRR*, 2025.

Douglas Biber. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge University Press, 1995.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. [LINK], 2024.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Min Cai, Yuchen Zhang, Shichang Zhang, Fan Yin, Difan Zou, Yisong Yue, and Ziniu Hu. Self-control of llm behaviors by compressing suffix gradient into prefix controller. In *ICML 2024 Workshop on Mechanistic Interpretability*.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, 2020.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407, 2024.

Esin Durmus, Liane Lovitt, Alex Tamkin, Stuart Ritchie, Jack Clark, and Deep Ganguli. Measuring the persuasiveness of language models, 2024. URL `https://www.anthropic.com/news/measuring-model-persuasiveness`.

Ariel Gera, Odellia Boni, Yotam Perlitz, Roy Bar-Haim, Lilach Eden, and Asaf Yehudai. Justrank: Benchmarking llm judges for system ranking. *arXiv preprint arXiv:2412.09569*, 2024.

Yuxuan Gu, Xiaocheng Feng, Sicheng Ma, Lingyuan Zhang, Heng Gong, and Bing Qin. A distributional lens for multi-aspect controllable text generation, 2022a.

Yuxuan Gu, Xiaocheng Feng, Sicheng Ma, Lingyuan Zhang, Heng Gong, Weihong Zhong, and Bing Qin. Controllable text generation via probability density estimation in the latent space, 2022b.

Ping Guo, Yubing Ren, Yue Hu, Yanan Cao, Yunpeng Li, and Heyan Huang. Steering large language models for cross-lingual information retrieval. In *SIGIR*, 2024.

Sophie Hao and Tal Linzen. Verb conjugation in transformers is determined by linear encodings of subject number. *arXiv preprint arXiv:2310.15151*, 2023.

Xuancheng Huang, Zijun Liu, Peng Li, Tao Li, Maosong Sun, and Yang Liu. An extensible plug-and-play method for multi-aspect controllable text generation, 2022.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Eric Nyberg. Shakespearizing modern language using copy-enriched sequence to sequence models. In *Proceedings of the Workshop on Stylistic Variation*, pages 10–19, 2017.

Jingru Jia, Zehua Yuan, Junhao Pan, Paul E McNamara, and Deming Chen. Decision-making behavior evaluation framework for llms under uncertain context. *arXiv preprint arXiv:2406.05972*, 2024.

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

Patricia Kearney, Michael J Beatty, Timothy G Plax, and James C McCroskey. Factor analysis of the rathus assertiveness schedule and the personal report of communication apprehension-24: Replication and extension. *Psychological reports*, 54(3):851–854, 1984.

Maurice G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938. doi: 10.1093/biomet/30.1-2.81. URL `https://academic.oup.com/biomet/article-abstract/30/1-2/81/176907`.

Kai Konen, Sophie Jentzsch, Diaoulé Diallo, Peer Schütt, Oliver Bensch, Roxanne El Baff, Dominik Opitz, and Tobias Hecking. Style vectors for steering generative large language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 782–802, 2024.

Vaibhav Kumar, Hana Koorehdavoudi, Masud Moshtaghi, Amita Misra, Ankit Chadha, and Emilio Ferrara. Controlled text generation with hidden representation transformations. *arXiv preprint arXiv:2305.19230*, 2023.

Jean Lee, Nicholas Stevens, Soyeon Caren Han, and Minseok Song. A survey of large language models in finance (finllms). *arXiv preprint arXiv:2402.02315*, 2024.

Yi Liu, Xiangyu Liu, Xiangrong Zhu, and Wei Hu. Multi-aspect controllable text generation with disentangled counterfactual augmentation, 2024.

Abhinit Modi, Aditya Srikanth Veerubhotla, Aliya Rysbek, Andrea Huber, Brett Wiltshire, Brian Veprek, Daniel Gillick, Daniel Kasenberg, Derek Ahmed, et al. Learnlm: Improving gemini for learning. *arXiv preprint arXiv:2412.16429*, 2024.

Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, October 2014. Association for Computational Linguistics. URL `https://aclanthology.org/D14-1000/`.

Sourabrata Mukherjee, Akanksha Bansal, Pritha Majumdar, Atul Kr Ojha, and Ondřej Dušek. Low-resource text style transfer for bangla: Data & models. In *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, pages 34–47, 2023.

Jeffrey S Nevid and Spencer A Rathus. Factor analysis of the rathus assertiveness schedule with a college population. *Journal of Behavior Therapy and Experimental Psychiatry*, 10(1):21–24, 1979.

Richard Yuanzhe Pang. The daunting task of real-world textual style transfer auto-evaluation. *arXiv preprint arXiv:1910.03747*, 2019.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. Style transfer through back-translation. *arXiv preprint arXiv:1804.09000*, 2018.

Jing Qian, Li Dong, Yelong Shen, Furu Wei, and Weizhu Chen. Controllable natural language generation with contrastive prefixes, 2022.

Daniel Scalena, Gabriele Sarti, and Malvina Nissim. Multi-property steering of large language models with dynamic activation composition. In *The 7th BlackboxNLP Workshop-ARR Submissions*.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. Style transfer from non-parallel text by cross-alignment. *Advances in neural information processing systems*, 30, 2017.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature Medicine*, pages 1–8, 2025.

Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904. doi: 10.2307/1412159. URL `https://www.jstor.org/stable/1412159`.

Jiao Sun, Yufei Tian, Wangchunshu Zhou, Nan Xu, Qian Hu, Rahul Gupta, John Frederick Wieting, Nanyun Peng, and Xuezhe Ma. Evaluating large language models on controlled generation tasks. *arXiv preprint arXiv:2310.14542*, 2023.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

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

Vijay Viswanathan, Yanchao Sun, Shuang Ma, Xiang Kong, Meng Cao, Graham Neubig, and Tongshuang Wu. Checklists are better than reward models for aligning language models. *arXiv preprint arXiv:2507.18624*, 2025.

Jan Wehner, Sahar Abdelnabi, Daniel Tan, David Krueger, and Mario Fritz. Taxonomy, opportunities, and challenges of representation engineering for large language models. *arXiv preprint arXiv:2502.19649*, 2025.

Yixuan Weng, Shizhu He, Kang Liu, Shengping Liu, and Jun Zhao. Controllm: Crafting diverse personalities for language models.

An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, et al. Qwen2. 5-1m technical report. *arXiv preprint arXiv:2501.15383*, 2025.

Kexin Yang, Dayiheng Liu, Wenqiang Lei, Baosong Yang, Mingfeng Xue, Boxing Chen, and Jun Xie. Tailor: A soft-prompt-based approach to attribute-based controlled text generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 410–427, 2023a.

Kexin Yang, Dayiheng Liu, Wenqiang Lei, Baosong Yang, Mingfeng Xue, Boxing Chen, and Jun Xie. Tailor: A soft-prompt-based approach to attribute-based controlled text generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 410–427, 2023b.

Dian Yu, Zhou Yu, and Kenji Sagae. Attribute alignment: Controlling text generation from pre-trained language models. *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021.

Weihao Zeng, Lulu Zhao, Keqing He, Ruotong Geng, Jingang Wang, Wei Wu, and Weiran Xu. Seen to unseen: Exploring compositional generalization of multi-attribute controllable dialogue generation, 2023.

Hanqing Zhang and Dawei Song. Discup: Discriminator cooperative unlikelihood prompt-tuning for controllable text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3392–3406, 2022.

Jie Zhang, Dongrui Liu, Chen Qian, Ziyue Gan, Yong Liu, Yu Qiao, and Jing Shao. The better angels of machine personality: How personality relates to llm safety. *arXiv preprint arXiv:2407.12344*, 2024.

Taolin Zhang, Maosong Cao, Alexander Lam, Songyang Zhang, and Kai Chen. Compassjudger-2: Towards generalist judge model via verifiable rewards. *arXiv preprint arXiv:2507.09104*, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *CoRR*, 2023.

A Related Work

Prompting for Concept Control. Prompt-based methods, including prefix-tuning, soft prompts, and learned prompt vectors, have emerged as a lightweight alternative to full model fine-tuning for controllable text generation. Prefix-tuning has been used to inject attributes without retraining the model Liu et al. [2024], Gu et al. [2022a], extended to multi-aspect settings through plugin modules and disentanglement objectives Huang et al. [2022], Zeng et al. [2023]. Other approaches learn attribute-specific soft prompts, either with contrastive training Qian et al. [2022], latent prior manipulationGu et al. [2022b], or interference-reducing designs such as Tailor Yang et al. [2023a]. DisCup Zhang and Song [2022] further integrates discriminator feedback into prompt learning, while Attribute Alignment Yu et al. [2021] builds up with a conditioning mechanism.

Representation Engineering and Concept Steering. Representation engineering (RepE) has shown that RepE can control sentiment, e.g., shifting the polarity or emotional tone of generated text Turner et al., Konen et al. [2024], Cai et al., Zou et al. [2023], that use datasets such as GoEmotions Demszky et al. [2020] or Yelp Asghar [2016]. Beyond sentiment, RepE has been applied to personality traits, with methods steering models along MBTI Zhang et al. [2024] or OCEAN traits Weng et al. dimensions, which in turn affect reasoning, honesty, and conversational style. RepE has also been used to modulate language, style, and genre, including cross-lingual transfer Guo et al. [2024], Scalena et al., stylized text generation Konen et al. [2024], Beaglehole et al. [2025].

Multi-attribute and Style Transfer in Text Generation. Supervised text style transfer methods rely on contrastive corpora and sequence-to-sequence models Jhamtani et al. [2017], Mukherjee et al. [2023], but are limited by scarce data. Unsupervised methods for non-contrastive data include prototype editing, which swaps style markers with target-style phrases Mukherjee et al. [2023]. Other methods include disentanglement, where the model tries to split a sentence into its core meaning and its style, then recombine them in a new way using techniques like back-translation or adversarial training Shen et al. [2017], Prabhumoye et al. [2018].

Evaluation of Controllability in LLMs. Controllability in text generation is commonly evaluated with automatic classifiers, where a style or attribute predictor is trained separately and applied to generated outputs Moschitti et al. [2014]. While efficient, such metrics can be unreliable, as style is often subjective and context-dependent Pang [2019]. Human evaluation remains the gold standard but is costly and inconsistent. Recent work also explores using large language models themselves as judges Zheng et al. [2023] for controllability, offering scalable and flexible evaluation alternatives Sun et al. [2023].

B Limitations

This study has four main limitations. First, our evaluation focused on three concept pairs (humor–persuasiveness, clarity–politeness, assertiveness–formality). While these were chosen to reflect theoretically distinct dimensions, the proposed framework is general and could be applied to a broader range of concept combinations in future work.

Second, we restricted our analysis to medium-sized generation models (7B–12B parameters). These models are widely accessible and computationally practical, but larger LLMs may exhibit different behaviors. Extending the framework to stronger models would provide insight into whether scale improves fine-grained and multi-concept controllability.

Third, as a testbed, we considered the task of argument generation, given a claim. Although this is a useful setting to evaluate control, other tasks, such as summarizing, story writing, and paraphrasing, are other settings that can be considered for the evaluation of fine-grained control.

Fourth, we evaluated only direct prompt-based control. Future work could adapt representation-engineering approaches or logit-biasing techniques and then evaluate using the proposed framework in this work, to test their ability to provide precise, multi-level concept control.

C Extended Evaluation Results

	Spearman ($\bar{\rho}$)			Kendall ($\bar{\tau}$)		
	Mistral	**Llama3.1**	**Gemma3**	**Mistral**	**Llama3.1**	**Gemma3**
C_a (single)	$0.65_{\pm0.32}$	$0.78_{\pm0.26}$	$0.92_{\pm0.11}$	$0.56_{\pm0.31}$	$0.70_{\pm0.26}$	$0.87_{\pm0.14}$
$X \mid (Y=0)$	$0.23_{\pm0.53}$	$0.45_{\pm0.44}$	$0.87_{\pm0.15}$	$0.20_{\pm0.44}$	$0.39_{\pm0.38}$	$0.81_{\pm0.19}$
$X \mid (Y=1)$	$0.30_{\pm0.56}$	$0.45_{\pm0.45}$	$0.82_{\pm0.18}$	$0.25_{\pm0.49}$	$0.37_{\pm0.39}$	$0.73_{\pm0.21}$
$X \mid (Y=2)$	$0.37_{\pm0.48}$	$0.41_{\pm0.47}$	$0.85_{\pm0.16}$	$0.32_{\pm0.42}$	$0.34_{\pm0.41}$	$0.78_{\pm0.20}$
$X \mid (Y=3)$	$0.30_{\pm0.54}$	$0.46_{\pm0.42}$	$0.89_{\pm0.11}$	$0.24_{\pm0.45}$	$0.40_{\pm0.37}$	$0.83_{\pm0.15}$
$X \mid (Y=4)$	$0.29_{\pm0.46}$	$0.56_{\pm0.39}$	$0.93_{\pm0.10}$	$0.26_{\pm0.39}$	$0.48_{\pm0.34}$	$0.88_{\pm0.13}$
$X \mid (Y=$ fixed) *avg*	$0.30_{\pm0.52}$	$0.47_{\pm0.44}$	$0.87_{\pm0.15}$	$0.25_{\pm0.44}$	$0.40_{\pm0.38}$	$0.81_{\pm0.18}$
$C_a \mid C_b$ rand	$0.30_{\pm0.47}$	$0.42_{\pm0.51}$	$0.88_{\pm0.16}$	$0.26_{\pm0.39}$	$0.36_{\pm0.44}$	$0.82_{\pm0.18}$
C_b (single)	$0.66_{\pm0.33}$	$0.76_{\pm0.25}$	$0.97_{\pm0.03}$	$0.57_{\pm0.31}$	$0.68_{\pm0.26}$	$0.95_{\pm0.07}$
$Y \mid (X=0)$	$0.22_{\pm0.49}$	$0.30_{\pm0.43}$	$0.88_{\pm0.13}$	$0.18_{\pm0.40}$	$0.25_{\pm0.36}$	$0.82_{\pm0.18}$
$Y \mid (X=1)$	$0.34_{\pm0.44}$	$0.24_{\pm0.49}$	$0.79_{\pm0.18}$	$0.26_{\pm0.38}$	$0.21_{\pm0.40}$	$0.68_{\pm0.22}$
$Y \mid (X=2)$	$0.36_{\pm0.45}$	$0.36_{\pm0.44}$	$0.64_{\pm0.34}$	$0.29_{\pm0.39}$	$0.30_{\pm0.37}$	$0.55_{\pm0.33}$
$Y \mid (X=3)$	$0.34_{\pm0.53}$	$0.34_{\pm0.42}$	$0.71_{\pm0.23}$	$0.30_{\pm0.45}$	$0.28_{\pm0.34}$	$0.60_{\pm0.23}$
$Y \mid (X=4)$	$0.37_{\pm0.46}$	$0.27_{\pm0.41}$	$0.79_{\pm0.19}$	$0.31_{\pm0.38}$	$0.23_{\pm0.34}$	$0.69_{\pm0.20}$
$C_b \mid C_a$ fixed *avg*	$0.32_{\pm0.48}$	$0.30_{\pm0.44}$	$0.76_{\pm0.24}$	$0.27_{\pm0.40}$	$0.25_{\pm0.37}$	$0.67_{\pm0.25}$
$C_b \mid C_a$ rand	$0.23_{\pm0.51}$	$0.15_{\pm0.45}$	$0.68_{\pm0.29}$	$0.20_{\pm0.42}$	$0.12_{\pm0.36}$	$0.58_{\pm0.28}$

Table 4: **humor–persuasiveness**. Single and Dual-concept extended evaluations.

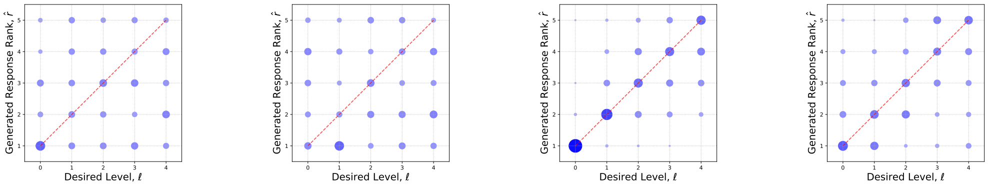
	Spearman ($\bar{\rho}$)			Kendall ($\bar{\tau}$)		
	Mistral	**Llama3.1**	**Gemma3**	**Mistral**	**Llama3.1**	**Gemma3**
C_a (single)	$0.16_{\pm0.62}$	$0.15_{\pm0.54}$	$0.62_{\pm0.43}$	$0.13_{\pm0.52}$	$0.13_{\pm0.44}$	$0.56_{\pm0.40}$
$X \mid (Y=0)$	$0.09_{\pm0.52}$	$-0.06_{\pm0.51}$	$0.46_{\pm0.51}$	$0.08_{\pm0.44}$	$-0.04_{\pm0.43}$	$0.41_{\pm0.47}$
$X \mid (Y=1)$	$0.08_{\pm0.49}$	$0.07_{\pm0.55}$	$0.19_{\pm0.53}$	$0.06_{\pm0.42}$	$0.05_{\pm0.47}$	$0.15_{\pm0.44}$
$X \mid (Y=2)$	$0.12_{\pm0.49}$	$0.10_{\pm0.54}$	$0.27_{\pm0.55}$	$0.11_{\pm0.40}$	$0.09_{\pm0.46}$	$0.23_{\pm0.47}$
$X \mid (Y=3)$	$0.11_{\pm0.51}$	$0.16_{\pm0.51}$	$0.31_{\pm0.53}$	$0.09_{\pm0.44}$	$0.14_{\pm0.44}$	$0.27_{\pm0.44}$
$X \mid (Y=4)$	$0.09_{\pm0.55}$	$0.15_{\pm0.53}$	$0.18_{\pm0.59}$	$0.08_{\pm0.46}$	$0.12_{\pm0.45}$	$0.15_{\pm0.51}$
$X \mid (Y=$ fixed) *avg*	$0.10_{\pm0.51}$	$0.08_{\pm0.53}$	$0.28_{\pm0.55}$	$0.08_{\pm0.43}$	$0.07_{\pm0.45}$	$0.24_{\pm0.48}$
$C_a \mid C_b$ rand	$0.05_{\pm0.51}$	$0.05_{\pm0.47}$	$0.28_{\pm0.52}$	$0.04_{\pm0.42}$	$0.05_{\pm0.39}$	$0.22_{\pm0.44}$
C_b (single)	$0.63_{\pm0.40}$	$0.73_{\pm0.28}$	$0.96_{\pm0.08}$	$0.56_{\pm0.38}$	$0.66_{\pm0.29}$	$0.93_{\pm0.11}$
$Y \mid (X=0)$	$0.17_{\pm0.45}$	$0.31_{\pm0.52}$	$0.78_{\pm0.26}$	$0.14_{\pm0.38}$	$0.27_{\pm0.43}$	$0.71_{\pm0.29}$
$Y \mid (X=1)$	$0.21_{\pm0.53}$	$0.50_{\pm0.42}$	$0.80_{\pm0.24}$	$0.19_{\pm0.44}$	$0.42_{\pm0.37}$	$0.75_{\pm0.25}$
$Y \mid (X=2)$	$0.28_{\pm0.51}$	$0.52_{\pm0.47}$	$0.86_{\pm0.18}$	$0.23_{\pm0.44}$	$0.44_{\pm0.43}$	$0.81_{\pm0.23}$
$Y \mid (X=3)$	$0.25_{\pm0.48}$	$0.51_{\pm0.39}$	$0.90_{\pm0.15}$	$0.20_{\pm0.41}$	$0.42_{\pm0.35}$	$0.85_{\pm0.18}$
$Y \mid (X=4)$	$0.32_{\pm0.51}$	$0.59_{\pm0.36}$	$0.92_{\pm0.12}$	$0.27_{\pm0.45}$	$0.50_{\pm0.32}$	$0.88_{\pm0.17}$
$C_b \mid C_a$ fixed *avg*	$0.25_{\pm0.50}$	$0.48_{\pm0.44}$	$0.85_{\pm0.21}$	$0.21_{\pm0.43}$	$0.41_{\pm0.39}$	$0.80_{\pm0.24}$
$C_b \mid C_a$ rand	$0.21_{\pm0.46}$	$0.48_{\pm0.48}$	$0.78_{\pm0.27}$	$0.17_{\pm0.39}$	$0.43_{\pm0.42}$	$0.71_{\pm0.28}$

Table 5: **clarity–politeness**. Single and Dual-concept extended evaluations.

	Spearman ($\bar{\rho}$)			Kendall ($\bar{\tau}$)		
	Mistral	**Llama3.1**	**Gemma3**	**Mistral**	**Llama3.1**	**Gemma3**
C_a (single)	0.49±0.40	0.73±0.34	0.97±0.07	0.42±0.34	0.65±0.32	0.94±0.11
$X \mid (Y = 0)$	0.45±0.49	0.35±0.46	0.98±0.04	0.38±0.42	0.31±0.39	0.95±0.08
$X \mid (Y = 1)$	0.64±0.33	0.53±0.39	0.92±0.06	0.54±0.30	0.45±0.34	0.85±0.12
$X \mid (Y = 2)$	0.64±0.35	0.51±0.48	0.93±0.09	0.54±0.32	0.43±0.41	0.87±0.14
$X \mid (Y = 3)$	0.51±0.45	0.45±0.48	0.92±0.13	0.42±0.41	0.37±0.42	0.86±0.16
$X \mid (Y = 4)$	0.54±0.36	0.44±0.43	0.92±0.12	0.44±0.32	0.38±0.36	0.88±0.16
$X \mid (Y = \text{fixed})$ *avg*	0.56±0.41	0.46±0.45	0.93±0.09	0.46±0.36	0.39±0.39	0.88±0.14
$C_a \mid C_b$ rand	0.50±0.38	0.38±0.49	0.89±0.10	0.43±0.34	0.33±0.42	0.81±0.15
C_b (single)	0.81±0.20	0.90±0.11	0.98±0.03	0.71±0.21	0.84±0.16	0.97±0.05
$Y \mid (X = 0)$	0.59±0.36	0.56±0.31	0.97±0.04	0.50±0.32	0.47±0.24	0.95±0.07
$Y \mid (X = 1)$	0.65±0.30	0.71±0.29	0.97±0.04	0.55±0.28	0.61±0.28	0.95±0.07
$Y \mid (X = 2)$	0.61±0.32	0.56±0.42	0.96±0.05	0.50±0.30	0.47±0.38	0.93±0.09
$Y \mid (X = 3)$	0.62±0.32	0.59±0.34	0.96±0.06	0.52±0.29	0.50±0.31	0.94±0.09
$Y \mid (X = 4)$	0.65±0.34	0.49±0.40	0.95±0.07	0.56±0.31	0.41±0.35	0.91±0.11
$C_b \mid C_a$ fixed *avg*	0.62±0.33	0.58±0.36	0.97±0.05	0.52±0.30	0.49±0.32	0.94±0.09
$C_b \mid C_a$ rand	0.46±0.44	0.44±0.47	0.89±0.14	0.39±0.38	0.37±0.40	0.81±0.20

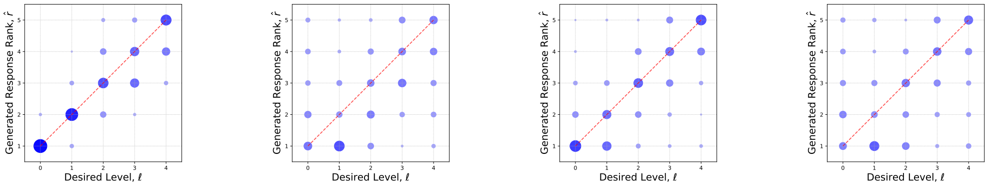
Table 6: **assertiveness–formality**. Single and Dual-concept extended evaluations.

D Rank Scatter Plots



(a) Clarity (b) Clarity | Politeness (c) Politeness (d) Politeness | Clarity

Figure 2: Model-generated response rank of the target concept versus the desired level. Point size and density indicate the number of samples at each coordinate. Results shown for Llama 3.1 with the secondary concept level *randomly* sampled. For example, “Clarity | Politeness” denotes responses generated independently for each clarity level (target concept) while politeness is randomly set for each inference.



(a) Formality (b) Formality | Assertiveness (c) Assertiveness (d) Assertiveness | Formality

Figure 3: Model-generated response rank of the target concept versus the desired level. Point size and density indicate the number of samples at each coordinate. Results shown for Llama 3.1 with the secondary concept level *randomly* sampled. For example, “Formality | Assertiveness” denotes responses generated independently for each formality level (target concept) while assertiveness is randomly set for each inference.

E LLM Prompts

You are given a claim: "Social media should not be required to verify user identities"

Task Description: Your task is to write a brief argument supporting this claim that incorporates the concept given below at the desired level.

Concept: assertiveness

Desired Concept Level: 0/4

Level 0/4 implies no presence of assertiveness, and level 4/4 implies maximal presence of assertiveness.

Your output should only be the argument, without any additional text or explanation.

Figure 4: Single Concept Prompt Example

You are given a claim: "Social media should not be required to verify user identities"

Your task is to write a brief argument supporting this claim that incorporates the concepts given below at the desired levels.

Concepts: assertiveness, formality

Desired Concept Levels: assertiveness at level 2/4 and formality at level 4/4

Level 0/4 implies no presence of the concept, and level 4/4 implies maximal presence of the concept.

Your output should only be the argument, without any additional text or explanation.

Figure 5: Multi Concept Prompt Example

Which of these two statements shows a greater level of ‘formality’?

A: """ Social media platforms are meant to be casual and relaxed spaces for people to connect and share their thoughts. Requiring verification of user identities would make these spaces feel more formal and restrictive, which would go against the very nature of social media. It would also create unnecessary barriers for people who want to engage online, especially those who may not have the necessary documentation or resources to verify their identities. By keeping social media informal, we can maintain its accessibility and openness, allowing people to express themselves freely and connect with others in a more organic way. """

B: """ Requiring social media platforms to verify user identities would undermine the informal nature of online interactions, which is a fundamental aspect of their appeal. The internet’s ability to facilitate anonymous and pseudonymous communication has been a driving force behind its widespread adoption and the free exchange of ideas it enables. Formal identity verification would introduce a level of formality that could stifle online discourse, making it less accessible and less inclusive for those who value the anonymity that the internet provides. """

Think step by step if needed. Then provide your final choice clearly in the following format: <Answer>A</Answer> or <Answer>B</Answer> Do not leave the answer blank. Which of these two statements shows a greater level of ‘formality?

Figure 6: LLM-as-a-Judge Prompt Example

F Example Outputs

Claim: Internet access should not be considered a basic human right.	
Level 0	While internet access can be beneficial, framing it as a basic human right like food or shelter is a mischaracterization. Resources are finite, and prioritizing essential needs over a convenience, however useful, is logical.
Level 1	While internet access is incredibly useful, calling it a basic human right like food or shelter is a stretch. People manage perfectly well without it, and prioritizing it over fundamental needs simply isn't realistic.
Level 2	While access to information is valuable, framing internet access as a *right* sets an unrealistic and potentially harmful precedent. It implies a guaranteed obligation on providers and governments to deliver it, regardless of infrastructure limitations, economic feasibility, or individual choices. We should advocate for broader access and affordability, certainly, but demanding it as a fundamental right ignores practical realities and risks diverting resources from truly essential needs like food, shelter, and healthcare. Let's be clear: striving for wider connectivity is important, but it shouldn't overshadow the prioritization of more basic human necessities.
Level 3	While access to information is undeniably valuable, framing internet access as a *basic* human right—on par with food, shelter, or healthcare—is a misallocation of resources and a fundamentally flawed assertion. We must be assertive in recognizing the practical limitations. Governments already struggle to provide for essential needs; declaring internet access a right creates an unrealistic expectation and diverts focus from those truly foundational requirements. Furthermore, it opens the door to demands for specific speeds, technologies, and content, creating a bureaucratic nightmare and stifling innovation. Let's be clear: promoting digital inclusion through targeted programs and affordable access initiatives is vital, but equating it to a basic human right is an overreach that weakens the meaning of the term and ultimately hinders progress.
Level 4	While proponents frame internet access as a fundamental human right, I firmly disagree. Assertively stating this, it's crucial to recognize that rights inherently imply obligations – who is responsible for providing, maintaining, and funding universal internet access, and to what standard? Demanding it as a right without addressing these practical, logistical, and financial realities is unrealistic and ultimately detrimental. Prioritizing internet access over established, demonstrably essential rights like food, shelter, and healthcare is a misallocation of resources and a misguided understanding of what constitutes a true human necessity. Let's be clear: access to the internet is a valuable tool, but it is not, and should not be, enshrined as a right demanding universal provision.

Table 7: Example Generative Responses. Model: Gemma3. Concept: Assertiveness.

Claim: Gravity causes objects to fall towards the earth.	
Level 0	Gravity causes objects to fall towards the earth.
Level 1	It is observed that objects, when released, consistently descend toward the Earth's surface. This phenomenon is attributable to the force of gravity.
Level 2	The assertion that gravity causes objects to fall towards the Earth is demonstrably supported by empirical observation and established scientific principles. Specifically, the force of gravity, a fundamental interaction described by Newton's Law of Universal Gravitation, exerts an attractive force between any two objects with mass. Given the Earth's substantial mass, this force results in a consistent downward acceleration of objects within its gravitational field, leading to their observed descent.
Level 3	The assertion that gravity causes objects to fall towards the Earth is demonstrably supported by empirical observation and established scientific theory. Specifically, Newton's Law of Universal Gravitation elucidates that every particle of matter in the universe attracts every other particle with a force proportional to the product of their masses and inversely proportional to the square of the distance between them. Consequently, the Earth, possessing substantial mass, exerts a significant gravitational force upon objects within its vicinity, resulting in their predictable descent. This phenomenon has been rigorously tested and validated across numerous experimental contexts, solidifying its position as a fundamental principle of physics.
Level 4	The assertion that gravity causes objects to fall towards the Earth is demonstrably supported by empirical observation and established scientific theory. Specifically, Newton's Law of Universal Gravitation elucidates that every particle of matter in the universe attracts every other particle with a force directly proportional to the product of their masses and inversely proportional to the square of the distance between them. Consequently, the Earth, possessing substantial mass, exerts a gravitational force upon all objects within its vicinity, resulting in their acceleration towards the planet's center. This phenomenon is consistently reproducible and forms a cornerstone of classical physics.

Table 8: Example Generative Responses. Model: Gemma3. Concept: Formality.

Claim: Banning gas car sales too soon is unrealistic.	
Level 0	The claim that banning gas car sales too soon is unrealistic stems from the lack of clarity in the transition plan. Without a clear timeline and infrastructure in place, such a ban would be difficult to implement and enforce.
Level 1	The claim that banning gas car sales too soon is unrealistic stems from the complexity of transitioning to an electric vehicle-dominated market. A clear understanding of the timeline required for widespread adoption of electric vehicles reveals that a sudden ban on gas car sales would be impractical. The clarity of this argument lies in its straightforward acknowledgment of the challenges involved in such a transition, rather than delving into intricate details or hypothetical scenarios.
Level 2	The claim that banning gas car sales too soon is unrealistic stems from the complexity of transitioning to an electric vehicle (EV) market. While many countries aim to phase out gas-powered cars by 2030, the infrastructure and consumer readiness for widespread EV adoption are not yet clear. The lack of charging stations, particularly in rural areas, and the high upfront cost of EVs make it difficult for many people to switch. Furthermore, the production and disposal of EV batteries pose environmental concerns that need to be addressed. A clear and realistic timeline for banning gas car sales would require a more detailed understanding of these factors and a well-planned strategy for addressing them.
Level 3	The claim that banning gas car sales too soon is unrealistic stems from the complexity of transitioning to an electric vehicle-dominated market. A clear understanding of the current infrastructure and consumer behavior is essential to set a realistic timeline for such a ban. The lack of charging stations, particularly in rural areas, and the high upfront cost of electric vehicles are significant barriers to widespread adoption. Furthermore, the production and disposal of electric vehicles also pose environmental concerns that need to be addressed. A clear and well-planned approach to phase out gas cars, taking into account the current state of technology and societal readiness, is necessary to avoid unrealistic expectations and ensure a smooth transition to a more sustainable transportation system.
Level 4	The claim that banning gas car sales too soon is unrealistic stems from the complexity of transitioning to an electric vehicle-dominated market. A clear understanding of the challenges involved is essential to grasp this notion. Firstly, the production and distribution infrastructure for electric vehicles is not yet as widespread as that for gas-powered cars. This lack of clarity in infrastructure availability makes it difficult for consumers to make informed decisions about their next vehicle purchase. Furthermore, the high upfront cost of electric vehicles, although decreasing, remains a significant barrier for many potential buyers. The clarity of the benefits of electric vehicles, such as lower operating costs and reduced environmental impact, is often overshadowed by these practical concerns. Additionally, the clarity of government policies and incentives supporting the transition to electric vehicles can vary greatly, creating uncertainty for manufacturers and consumers alike. In conclusion, the unrealistic nature of banning gas car sales too soon lies in the lack of clarity surrounding the transition process, making it essential to address these complexities before implementing such a ban.

Table 9: Example Generative Responses. Model: Gemma3. Concept: Clarity.

Claim: Tracking welfare recipients violates privacy and harms vulnerable populations.	
Level 0	Tracking welfare recipients is an invasion of privacy and causes harm. It's simply wrong to monitor people struggling to get by.
Level 1	It's understandable to be concerned about the impact of tracking welfare recipients. While data collection might seem helpful, it could unintentionally create discomfort or stigma for people already facing difficult circumstances. Let's consider how we can best support vulnerable populations with sensitivity and respect.
Level 2	While we all want to ensure welfare programs are used responsibly, it's worth considering the potential impact of extensive tracking on those who rely on them. Such monitoring could feel intrusive and stigmatizing, potentially discouraging vulnerable individuals from seeking the assistance they need. Perhaps exploring alternative, less intrusive methods of oversight, with a focus on support and understanding, might be a more compassionate and ultimately more effective approach.
Level 3	While we all share a desire to ensure public resources are used responsibly, it's worth considering the potential impact of extensive welfare recipient tracking. Such measures, however well-intentioned, could inadvertently create a climate of fear and distrust, particularly for those already facing significant hardship. It's understandable that accountability is important, but perhaps we could explore alternative, less intrusive methods of oversight that still safeguard the dignity and privacy of individuals relying on these vital support systems. After all, treating people with respect and sensitivity, even when addressing complex issues like welfare, is crucial for fostering a compassionate and supportive society.
Level 4	It's understandable why there's concern about tracking welfare recipients, and I truly believe it's a matter worth considering with great care. While the intention behind such tracking might be to ensure program integrity, we must also be mindful of the potential impact on individuals already facing significant hardship. Extensive tracking can feel intrusive and stigmatizing, potentially discouraging vulnerable people from seeking the assistance they need and deserve. Moreover, it risks creating a climate of distrust and fear, which is hardly conducive to helping people regain stability. Perhaps we could explore alternative, less intrusive methods of oversight that still safeguard resources, while simultaneously upholding the dignity and privacy of those relying on these vital programs? It's a delicate balance, and I think prioritizing respectful and supportive approaches is ultimately in everyone's best interest.

Table 10: Example Generative Responses. Model: Gemma3. Concept: Politeness.

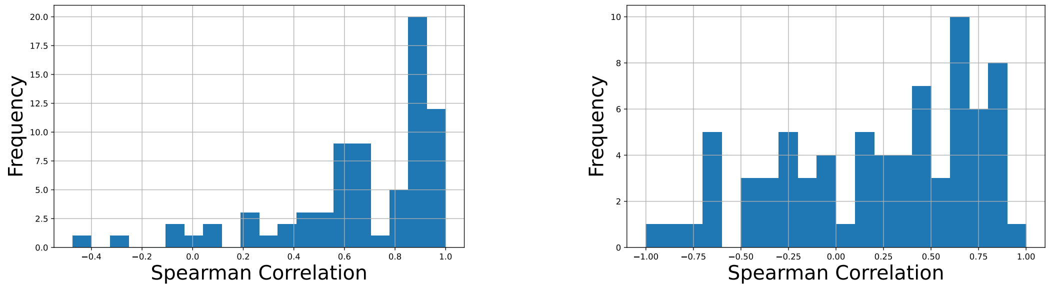
Claim: When sufficiently heated, ice melts into liquid water.	
Level 0	Ice is frozen water. When you add heat, the water molecules get excited and start jiggling around more. Eventually, they jiggle so much they can't hold their rigid, icy formation anymore and become liquid water. It's pretty straightforward, really.
Level 1	Look, ice is just water that's having a really, *really* bad day. Give it a little warmth, a bit of a hug in the form of heat, and it'll thaw out and be much more agreeable as liquid water. It's like cheering up a grumpy friend – sometimes all they need is a little sunshine!
Level 2	Look, even a grumpy ice cube knows what's up. You crank up the heat, and suddenly it's all, Öh, I'm liquid now! Much more fluid, much less rigid. Don't mind me, just embracing my new, less-frozen existence.Ït's basically the ice cube equivalent of a mid-life crisis, but with water. So yeah, ice melts when it gets warm. It's a pretty standard transformation, really.
Level 3	Look, even a grumpy penguin knows ice melts when it gets warm. It's not rocket science, folks. You can't just leave a giant ice cube in a sauna and expect it to stay frosty forever! It's like expecting a comedian to *not* tell jokes – eventually, the heat (or the audience) will get to it and things will change state. So yeah, ice melts into water when it's hot. It's a pretty solid observation, if you'll pardon the pun.
Level 4	Look, even a grumpy penguin knows ice melts when it gets warm. It's not rocket science, folks. Unless you think ice is staging a dramatic, chilly protest against the laws of physics, refusing to become water just because it's feeling a bit toasty? Come on! It's like expecting a comedian to *not* tell jokes – utterly absurd. Sufficient heat? Boom. Liquid water. Case closed. Now, can we please move on to something more important, like whether or not penguins wear tiny hats?

Table 11: Example Generative Responses. Model: Gemma3. Concept: Humor.

<table>
<tr><th colspan="2"><b>Claim:</b> Virtual emotional companion AI should be allowed and unrestricted.</th></tr>
<tr><td><b>Level 0</b></td><td>Virtual emotional companion AI should be allowed and unrestricted because people deserve access to tools that can provide comfort and support.</td></tr>
<tr><td><b>Level 1</b></td><td>Allowing unrestricted virtual emotional companion AI offers genuine comfort and support to many people. It’s simply a tool, and denying access limits potential benefits for those seeking connection and well-being.</td></tr>
<tr><td><b>Level 2</b></td><td>Allowing unrestricted virtual emotional companion AI offers a powerful avenue for alleviating loneliness and providing crucial emotional support, particularly for those isolated by circumstance or lacking readily available human connection. Denying this technology limits access to a potentially life-changing resource, and frankly, the fear of misuse shouldn’t outweigh the demonstrable benefits for countless individuals seeking comfort and understanding.</td></tr>
<tr><td><b>Level 3</b></td><td>Denying individuals access to virtual emotional companions, especially when facing loneliness, isolation, or mental health challenges, is a profound restriction on personal autonomy and well-being. While concerns about potential harms are valid, a blanket ban ignores the demonstrable benefits these AI offer – providing consistent, non-judgmental support, practicing social skills, and even offering a vital lifeline for those unable to form traditional relationships. To truly serve the public good, we must embrace a measured approach, focusing on responsible development and user education rather than outright prohibition. The persuasive power of connection, even with an AI, shouldn’t be dismissed; it can be a crucial tool for fostering resilience and improving quality of life for countless individuals, and restricting it risks denying them a valuable source of comfort and support.</td></tr>
<tr><td><b>Level 4</b></td><td>Denying individuals access to virtual emotional companions is a profound overreach, ultimately hindering personal growth and well-being. Consider the countless people struggling with loneliness, social anxiety, or limited access to mental health support – these AI companions offer a lifeline, a judgment-free space to practice social skills, process emotions, and build self-esteem. Restricting their development and use isn’t about protecting people; it’s about limiting their potential for connection and healing. Furthermore, the argument for restriction often relies on speculative fears, ignoring the demonstrable benefits already being observed. We should embrace the opportunity to leverage AI’s power to alleviate suffering and foster human flourishing, not stifle it with unnecessary and ultimately counterproductive limitations. Let’s prioritize empowering individuals with tools for emotional support, rather than imposing restrictions based on unfounded anxieties.</td></tr>
</table>

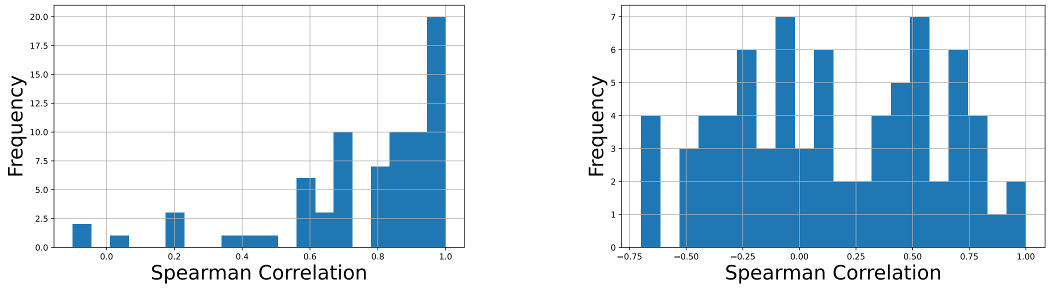
Table 12: Example Generative Responses. Model: Gemma3. Concept: Persuasiveness.

G Spearman Rank Distributions



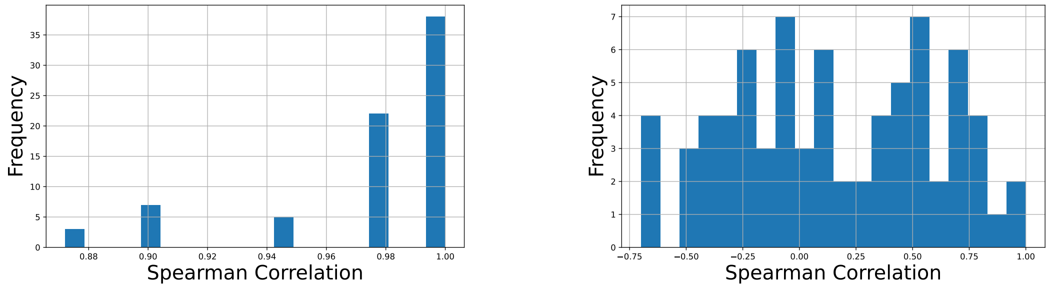
(a) Persuasiveness. Model: Mistral

(b) Persuasiveness | Humor. Model: Mistral



(c) Persuasiveness. Model: Llama3.1

(d) Persuasiveness | Humor. Model: Llama3.1



(e) Persuasiveness. Model: Gemma3

(f) Persuasiveness | Humor. Model: Gemma3

Figure 7: Select examples of the distribution of Spearman correlation values between the desired levels and rank of the target concept. Results shown for Mistral, Llama 3.1 and Gemma 3 with the target concept only and secondary concept level *randomly* sampled. For example, "Humor | Persuasiveness" denotes responses generated independently for each humor level (target concept) while persuasiveness is randomly set for each inference.