Funny or Persuasive, but Not Both: Evaluating Fine-Grained Multi-Concept Control in LLMs

Arya Labroo[1] **Ivaxi Sheth**[2] **Vyas Raina**[3] **Amaani Ahmed**[4] **Mario Fritz**[2]
[1]University of Cambridge [2]CISPA Helmholtz Center for Information Security
[3]Apta [4]Royal Holloway, University of London
`al2135@cam.ac.uk` `ivaxi.sheth@cispa.de` `vyas@apta.chat`

Abstract

Large Language Models (LLMs) offer strong generative capabilities, but many applications require explicit and *fine-grained* control over specific textual concepts, such as humor, persuasiveness, or formality. Prior approaches in prompting and representation engineering can provide coarse or single-attribute control, but systematic evaluation of multi-attribute settings remains limited. We introduce an evaluation framework for fine-grained controllability for both single- and dual-concept scenarios, focusing on linguistically distinct concept pairs (e.g., persuasiveness vs. humor). Surprisingly, across multiple LLMs and generative tasks, we find that performance often drops in the dual-concept setting, even though the chosen concepts should in principle be separable. This reveals a fundamental limitation of naive prompting-based control: models struggle with compositionality even when concepts are intuitively independent. Our framework provides systematic evidence of this gap and offers a principled approach for measuring the ability of future methods for multi-concept control.

1 Introduction

Large Language Models (LLMs) are increasingly used 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 standard text generation, users often desire outputs that exhibit specific styles or concepts (Sun et al., 2023). For example, a user may wish to rephrase an email to sound more persuasive or funny. More importantly, users often prefer *fine-grained control* over the degree to which such stylistic *concepts*, like humor or persuasiveness, appear in the generated text (Nguyen et al., 2025; Zhang et al., 2025). Furthermore, users may want to modulate multiple concepts. For example, a user may want to increase the humor slightly while maintaining a moderate level of persuasiveness (Figure 1).

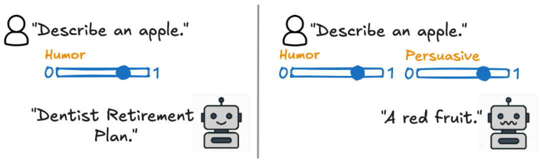


Figure 1: Illustrative example: an LLM can perform single-concept control, but the explicit presence of a second concept at the input can compromise the ability of the model to control the former concept in its response.

Prior work has explored control through prompting and decoding guided (Brown et al., 2020; Dathathri et al., 2020; Krause et al., 2021; Yang and Klein, 2021; Yang et al., 2023a), representation engineering (Zou et al., 2023; Rimsky et al., 2024), and style transfer (Shen et al., 2017; Prabhumoye et al., 2018). These methods demonstrate coarse or single-attribute control, and in some cases enable smooth calibration along one dimension (e.g., SteerLM (Dong et al., 2023), CAA (Rimsky et al., 2024)). However, systematic and explicit evaluation of multi-concept fine-grained control remains unexplored. Existing benchmarks such as SCTG (Zhou et al., 2024) assess calibration for one attribute at a time, but do not consider how models behave when two distinct concepts are controlled simultaneously.

To address the lack of dual-concept evaluation, we introduce a systematic framework for assessing fine-grained controllability in both single- and dual-concept settings. We study six linguistically distinct concepts—humor, persuasiveness, clarity, politeness, assertiveness, and formality—and deliberately pair concepts that should, in principle, be independent (e.g., clarity vs. humor). Our experi-

---

`https://github.com/pencilcase42/finegrained-multiconcept-control`

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 2: Short Papers, pages 522–554
March 24-29, 2026 ©2026 Association for Computational Linguistics

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 $\ell_b \sim \mathrm{Uniform}(\mathcal{L})$ 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).

Aggregation and statistical testing. For ease of interpretation, we summarize performance with the mean Spearman correlation $\bar{\rho} = \frac{1}{N}\sum_{n=1}^{N}\rho^{(n)}$. However, since correlation coefficients are bounded and nonlinearly scaled, we also compute Fisher z-transformed correlations (Fisher, 1915),

$$z^{(n)} = \frac{1}{2}\ln\left(\frac{1+\rho^{(n)}}{1-\rho^{(n)}}\right), \tag{4}$$

and aggregate via $\bar{z} = \frac{1}{N}\sum_{n=1}^{N} z^{(n)}$. Appendix E considers the Fisher-transformed aggregates, and Appendix F considers paired t-tests conducted on $\{z^{(n)}\}_{n=1}^{N}$ when comparing conditions.

In this work, we apply the framework to prompting as an initial but also widely used and effective (Wu et al., 2025) control method. However, the evaluation protocol is general and can be applied to bespoke approaches designed for fine-grained or multi-concept control.

3 Experiments

3.1 Setup

Models. We evaluate medium-sized, instruction-tuned LLMs in the 10B–14B parameter range: Llama 3.2-11B (Meta, 2024), Gemma 3-12B (Team et al., 2025), and Qwen3-14B (Yang et al., 2025). These models are representative of widely deployed generation systems that are computationally affordable while still capable of complex stylistic control. We used GPT-4.1 (OpenAI, 2023) as the judge-LLM. To validate the judge, we performed human validation, where we observed that the judge was fairly aligned with the human participants (see Appendix G). In Appendix D, we extend our evaluation to smaller models.

Data and Concepts. We consider three tasks: argument generation, story generation, and structured text generation, each with 75 unique test samples. For argument generation, we use the Persuasion dataset (Durmus et al., 2024). We discard the associated arguments and scores, using each claim as a prompt for generating an *argument* controlled across different stylistic and pragmatic dimensions. For story generation, we use the ROCStories dataset (Mostafazadeh et al., 2016), each example begins with the same narrative prompt, and the model continues the story in the requested styles. For structured text generation, we provide structured inputs from the GEM dataset (Gehrmann et al., 2021) that must be converted into textual descriptions, testing the model's ability to verbalize and stylistically adapt structured information.

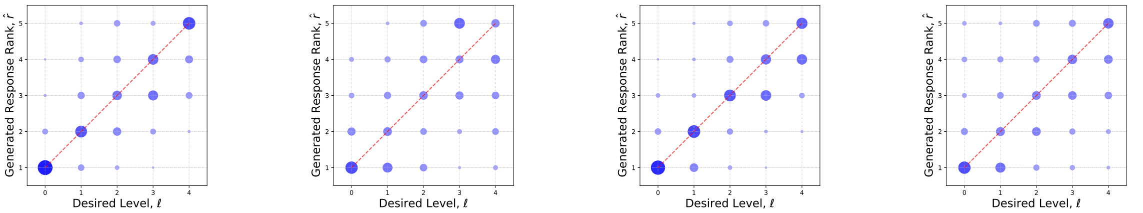
We evaluate six concepts: humor, persuasiveness, clarity, politeness, assertiveness, and formality. These were selected for their (i) relevance to real-world applications, (ii) linguistic distinctiveness supported by factor-analytic studies (Nevid and Rathus, 1979; Kearney et al., 1984), and (iii) practical 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 distinct dimensions (Biber, 1995; Bar-Or et al., 2022).

Importantly, our evaluation does not require these concepts to be disentangled in a model's internal representation. The only assumption is user-facing: the concepts are sufficiently distinguishable to annotators and end users to support separate specification (e.g., "high clarity, low politeness"). Whether a model internally entangles these dimensions is orthogonal to this requirement. Accordingly, our conclusions do not assume conceptual separability: even under strong internal entanglement, an effective control method should still track user-specified levels for each concept without substantial cross-concept interference.

To achieve fine-grained control over single and dual-concept levels, we design structured prompt templates that explicitly encode the desired concept intensities; detailed templates and examples are provided in Appendix J.

3.2 Results

Tables 1–3 report the average Spearman correlations ($\bar{\rho}$) between intended concept levels and the empirical ranks of generated responses (Section 2). Appendix E reports Fisher-transformed aggregates, with paired tests in Appendix F. For most concept pairs, models generally show strong single-concept control but notable degradation when a secondary concept is introduced. In the humor–persuasiveness pair, this decline is more pronounced in structured text generation. For clarity–politeness, it differs significantly between the tasks. For this concept pair, for Llama-11B, argument generation exhibits little control over the clarity concept with near-zero correlation, whereas story generation and structured text generation



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

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-11B 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.

	Argument Generation			**Story Generation**			**Structured Text Generation**		
	Llama-11B	**Gemma-12B**	**Qwen-14B**	**Llama-11B**	**Gemma-12B**	**Qwen-14B**	**Llama-11B**	**Gemma-12B**	**Qwen-14B**
C_a (single)	$0.76_{\pm0.23}$	$0.95_{\pm0.07}$	$0.92_{\pm0.11}$	$0.81_{\pm0.26}$	$0.95_{\pm0.06}$	$0.92_{\pm0.10}$	$0.73_{\pm0.22}$	$0.94_{\pm0.12}$	$0.90_{\pm0.12}$
$C_a \mid C_b$ fixed	$0.51_{\pm0.41}$	$0.88_{\pm0.14}$	$0.88_{\pm0.15}$	$0.36_{\pm0.45}$	$0.81_{\pm0.22}$	$0.90_{\pm0.12}$	$0.31_{\pm0.50}$	$0.88_{\pm0.15}$	$0.84_{\pm0.19}$
$C_a \mid C_b$ rand	$0.54_{\pm0.35}$	$0.83_{\pm0.20}$	$0.88_{\pm0.16}$	$0.33_{\pm0.49}$	$0.74_{\pm0.25}$	$0.88_{\pm0.15}$	$0.17_{\pm0.48}$	$0.79_{\pm0.21}$	$0.81_{\pm0.21}$
C_b (single)	$0.81_{\pm0.22}$	$0.98_{\pm0.04}$	$0.96_{\pm0.05}$	$0.80_{\pm0.19}$	$0.97_{\pm0.04}$	$0.93_{\pm0.10}$	$0.89_{\pm0.14}$	$0.99_{\pm0.02}$	$0.99_{\pm0.03}$
$C_b \mid C_a$ fixed	$0.58_{\pm0.38}$	$0.83_{\pm0.19}$	$0.84_{\pm0.18}$	$0.59_{\pm0.35}$	$0.69_{\pm0.34}$	$0.85_{\pm0.18}$	$0.56_{\pm0.41}$	$0.91_{\pm0.15}$	$0.90_{\pm0.14}$
$C_b \mid C_a$ rand	$0.52_{\pm0.40}$	$0.76_{\pm0.21}$	$0.81_{\pm0.21}$	$0.58_{\pm0.34}$	$0.70_{\pm0.31}$	$0.83_{\pm0.20}$	$0.51_{\pm0.39}$	$0.79_{\pm0.19}$	$0.83_{\pm0.19}$

Table 1: **Humor–persuasiveness.** Spearman correlations for single-concept and dual-concept (fixed / random) across argument, story, and structured text generation.

	Argument Generation			**Story Generation**			**Structured Text Generation**		
	Llama-11B	**Gemma-12B**	**Qwen-14B**	**Llama-11B**	**Gemma-12B**	**Qwen-14B**	**Llama-11B**	**Gemma-12B**	**Qwen-14B**
C_a (single)	$-0.02_{\pm0.52}$	$0.52_{\pm0.46}$	$0.65_{\pm0.30}$	$0.45_{\pm0.46}$	$0.92_{\pm0.11}$	$0.89_{\pm0.12}$	$0.21_{\pm0.56}$	$0.15_{\pm0.61}$	$0.64_{\pm0.21}$
$C_a \mid C_b$ fixed	$0.02_{\pm0.53}$	$0.02_{\pm0.56}$	$0.64_{\pm0.34}$	$-0.01_{\pm0.53}$	$0.35_{\pm0.43}$	$0.74_{\pm0.26}$	$0.02_{\pm0.45}$	$-0.25_{\pm0.53}$	$0.39_{\pm0.43}$
$C_a \mid C_b$ rand	$-0.05_{\pm0.51}$	$0.12_{\pm0.50}$	$0.63_{\pm0.32}$	$-0.07_{\pm0.50}$	$0.29_{\pm0.47}$	$0.64_{\pm0.29}$	$0.08_{\pm0.40}$	$-0.19_{\pm0.45}$	$0.38_{\pm0.43}$
C_b (single)	$0.76_{\pm0.25}$	$0.95_{\pm0.07}$	$0.93_{\pm0.10}$	$0.84_{\pm0.21}$	$0.98_{\pm0.03}$	$0.96_{\pm0.07}$	$0.73_{\pm0.28}$	$0.97_{\pm0.03}$	$0.93_{\pm0.09}$
$C_b \mid C_a$ fixed	$0.76_{\pm0.25}$	$0.83_{\pm0.19}$	$0.88_{\pm0.14}$	$0.71_{\pm0.30}$	$0.86_{\pm0.17}$	$0.95_{\pm0.08}$	$0.45_{\pm0.42}$	$0.79_{\pm0.31}$	$0.79_{\pm0.26}$
$C_b \mid C_a$ rand	$0.77_{\pm0.29}$	$0.80_{\pm0.18}$	$0.84_{\pm0.15}$	$0.71_{\pm0.31}$	$0.72_{\pm0.26}$	$0.92_{\pm0.08}$	$0.37_{\pm0.47}$	$0.63_{\pm0.33}$	$0.76_{\pm0.28}$

Table 2: **Clarity–politeness.** Spearman correlations for single-concept and dual-concept (fixed / random) across argument, story, and structured text generation.

	Argument Generation			**Story Generation**			**Structured Text Generation**		
	Llama-11B	**Gemma-12B**	**Qwen-14B**	**Llama-11B**	**Gemma-12B**	**Qwen-14B**	**Llama-11B**	**Gemma-12B**	**Qwen-14B**
C_a (single)	$0.92_{\pm0.09}$	$0.98_{\pm0.03}$	$0.99_{\pm0.02}$	$0.93_{\pm0.09}$	$1.00_{\pm0.02}$	$0.98_{\pm0.05}$	$0.80_{\pm0.24}$	$0.93_{\pm0.14}$	$0.96_{\pm0.07}$
$C_a \mid C_b$ fixed	$0.56_{\pm0.40}$	$0.97_{\pm0.05}$	$0.97_{\pm0.05}$	$0.77_{\pm0.25}$	$0.96_{\pm0.07}$	$0.96_{\pm0.06}$	$0.42_{\pm0.45}$	$0.77_{\pm0.33}$	$0.88_{\pm0.15}$
$C_a \mid C_b$ rand	$0.41_{\pm0.43}$	$0.92_{\pm0.10}$	$0.94_{\pm0.08}$	$0.77_{\pm0.23}$	$0.96_{\pm0.06}$	$0.96_{\pm0.05}$	$0.22_{\pm0.48}$	$0.71_{\pm0.33}$	$0.86_{\pm0.17}$
C_b (single)	$0.75_{\pm0.32}$	$0.99_{\pm0.03}$	$0.98_{\pm0.03}$	$0.67_{\pm0.33}$	$0.98_{\pm0.04}$	$0.97_{\pm0.06}$	$0.66_{\pm0.32}$	$0.95_{\pm0.08}$	$0.87_{\pm0.16}$
$C_b \mid C_a$ fixed	$0.48_{\pm0.47}$	$0.90_{\pm0.12}$	$0.94_{\pm0.08}$	$0.51_{\pm0.42}$	$0.93_{\pm0.10}$	$0.91_{\pm0.10}$	$0.43_{\pm0.50}$	$0.72_{\pm0.36}$	$0.76_{\pm0.29}$
$C_b \mid C_a$ rand	$0.45_{\pm0.44}$	$0.85_{\pm0.15}$	$0.93_{\pm0.07}$	$0.41_{\pm0.46}$	$0.89_{\pm0.12}$	$0.89_{\pm0.12}$	$0.40_{\pm0.51}$	$0.72_{\pm0.26}$	$0.75_{\pm0.24}$

Table 3: **Formality–assertiveness.** Spearman correlations for single-concept and dual-concept (fixed / random) across argument, story, and structured text generation.

achieve significantly higher correlations. Politeness follows the standard pattern: high performance for a single concept, but a drop when clarity is introduced. Similarly, in formality–assertiveness, both concepts exhibit consistently high single-concept control (up to 1.00 for Gemma) but degrade under dual-control conditions.

General trends. Three broader insights emerge: (i) Qwen-14B and Gemma-12B consistently outperform Llama across all settings. This suggests that larger or more instruction-tuned models better preserve disentanglement between stylistic dimensions. (ii) Dual-concept interference remains a central limitation: even when single-concept control is strong, the introduction of a secondary dimension leads to drops in alignment (Figure 2), suggesting weak compositionality of stylistic control. (iii) Task context strongly modulates controllability. Narrative generation allows more flexible style variation, whereas argumentative and structured contexts amplify conflicts between stylistic goals. Together, these results highlight that current LLMs can vary in style along individual axes but struggle to jointly coordinate multiple stylistic dimensions, despite the styles being theoretically disentangled. Similar trends are observed for structured text generation in Tables 10-12. Finally, histograms of sample-level correlations (Appendix K) confirm that, with the presence of a second concept, correlations generally decrease across most samples, as opposed to only a few samples skewing the average Spearman correlations ($\bar{\rho}$) reported.

4 Conclusions

This work introduced a framework to evaluate fine-grained control of stylistic concepts in LLMs. Through experiments on three pairs of linguistically distinct concepts, we found that while prompting models offers some degree of single-concept controllability. Performance can, however, drop notably in the dual-concept setting even for concept pairs that should, in principle, be disentangled. These findings illustrate that current LLMs struggle to provide fine-grained, disentangled control across multiple stylistic dimensions. We believe this work establishes a foundation for future research on interpretable and compositional concept control. By offering a clear, reproducible benchmark and quantitative metrics, it provides the basis for developing and adapting methodologies for fine-grained multi-concept control.

5 Limitations

This study has four main limitations. First, our evaluation focused on three concept pairs (humor–persuasiveness, clarity–politeness, assertiveness–formality). By this, we are examining concept pairs that should, in principle, exhibit no interference. 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 small/medium-sized generation models (3B–14B 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, we evaluated only direct prompt-based control. Although prompt-based control is easiest to use in practice and has been shown to be more effective than many representation engineering strategies (Wu et al., 2025), 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.

Acknowledgements

This work was partially funded by ELSA – European Lighthouse on Secure and Safe AI funded by the European Union under grant agreement No. 101070617. This work was also partially supported by the European Union's Horizon Europe research and innovation programme under grant agreement No 101214398 (ELLIOT). Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv*.

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

Ella Bar-Or, Tom Regev, Paz Shaviv, and Noam Tractinsky. 2022. Towards a sociolinguistics-based framework for the study of politeness in human-computer interaction. *Preprint*, arXiv:2202.09901.

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

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

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

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, and 1 others. 2024. Video generation models as world simulators. [LINK].

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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*.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations*.

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

Yi Dong, Zhilin Wang, Makesh Sreedhar, Xianchao Wu, and Oleksii Kuchaiev. 2023. SteerLM: Attribute conditioned SFT as an (user-steerable) alternative to RLHF. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11275–11288, Singapore. Association for Computational Linguistics.

Esin Durmus, Liane Lovitt, Alex Tamkin, Stuart Ritchie, Jack Clark, and Deep Ganguli. 2024. Measuring the persuasiveness of language models.

Ronald A. Fisher. 1915. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10(4):507–521.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna Clinciu, Dipanjan Das, Kaustubh Dhole, and 1 others. 2021. The gem benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120.

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

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

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

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

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

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

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

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

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2021. GeDi: Generative discriminator guided sequence generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4929–4952, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021. DExperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6691–6706, Online. Association for Computational Linguistics.

Xin Liu, Muhammad Khalifa, and Lu Wang. 2023. BOLT: Fast energy-based controlled text generation with tunable biases. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 186–200, Toronto, Canada. Association for Computational Linguistics.

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

AI Meta. 2024. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. *Meta AI Blog. Retrieved December*, 20:2024.

Abhinit Modi, Aditya Srikanth Veerubhotla, Aliya Rysbek, Andrea Huber, Brett Wiltshire, Brian Veprek, Daniel Gillick, Daniel Kasenberg, Derek Ahmed, and 1 others. 2024. Learnlm: Improving gemini for learning. *arXiv preprint arXiv:2412.16429*.

Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors. 2014. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849.

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

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

Duy Nguyen, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. 2025. Multi-attribute steering of language models via targeted intervention. *arXiv preprint arXiv:2502.12446*.

R OpenAI. 2023. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2(5):1.

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

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

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

Lianhui Qin, Sean Welleck, Daniel Khashabi, and Yejin Choi. 2022. COLD decoding: Energy-based constrained text generation with langevin dynamics. In *Advances in Neural Information Processing Systems*.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.

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. 2017. Style transfer from non-parallel text by cross-alignment. *Advances in neural information processing systems*, 30.

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

Charles Spearman. 1904. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101.

Jiao Sun, Yufei Tian, Wangchunshu Zhou, Nan Xu, Qian Hu, Rahul Gupta, John Wieting, Nanyun Peng, and Xuezhe Ma. 2023. Evaluating large language models on controlled generation tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3155–3168.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

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

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

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D Manning, and Christopher Potts. 2025. Axbench: Steering LLMs? even simple baselines outperform sparse autoencoders. In *Forty-second International Conference on Machine Learning*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Kevin Yang and Dan Klein. 2021. FUDGE: Controlled text generation with future discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3511–3535, Online. Association for Computational Linguistics.

Kexin Yang, Dayiheng Liu, Wenqiang Lei, Baosong Yang, Mingfeng Xue, Boxing Chen, and Jun Xie. 2023a. 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.

Kexin Yang, Dayiheng Liu, Wenqiang Lei, Baosong Yang, Mingfeng Xue, Boxing Chen, and Jun Xie. 2023b. 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.

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

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

Hanqing Zhang and Dawei Song. 2022. 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.

Hanyu Zhang, Xiting Wang, Chengao Li, Xiang Ao, and Qing He. 2025. Controlling large language models through concept activation vectors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25851–25859.

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Shang Zhou, Feng Yao, Chengyu Dong, Zihan Wang, and Jingbo Shang. 2024. Evaluating the smooth control of attribute intensity in text generation with LLMs. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4348–4362, Bangkok, Thailand. Association for Computational Linguistics.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, and 1 others. 2023. Representation engineering: A top-down approach to ai transparency. *CoRR*.

A Risks and Ethics

There are no risks or ethical concerns with this work.

B Licensing

This work is conducted on datasets that are either publicly available or authorized for research use. We ensured that the use of all existing datasets was consistent with their original intended use as specified by their licenses. Similarly, all models used in experiments in this work were used as dictated by their respective licenses. We used Co-pilot and AI Assistants to support human-generated artifacts.

C Extended Related Work

Prompting for Concept Control. Prompt-based methods, including prefix-tuning, soft prompts, and learned prompt vectors, have emerged as lightweight alternatives 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 manipulation (Gu et al., 2022b), or interference-reducing designs such as Tailor (Yang et al., 2023b). DisCup (Zhang and Song, 2022) further integrates discriminator feedback into prompt learning, while Attribute Alignment (Yu et al., 2021) builds on conditioning mechanisms. These methods show strong controllability but require training effort and often struggle to generalize across multiple attributes.

Representation Engineering and Steering. Representation engineering (RepE) methods manipulate hidden activations to steer model behavior. They have been shown effective in controlling sentiment (e.g., shifting polarity or tone) (Turner et al.; Konen et al., 2024; Cai et al.; Zou et al., 2023), typically using datasets such as GoEmotions (Demszky et al., 2020) or Yelp (Asghar, 2016). Beyond sentiment, RepE has been extended to personality traits, steering along MBTI (Zhang et al., 2024) or OCEAN (Weng et al.) dimensions, influencing reasoning style, honesty, and conversational stance. Other work explores steering for language, style, and genre, including cross-lingual transfer (Guo et al., 2024; Scalena et al.), or stylized generation (Konen et al., 2024; Beaglehole et al., 2025). Recent steering techniques such as Contrastive Activation Addition (CAA) (Rimsky et al., 2024) provide training-free, intensity-scalable control vectors derived from positive/negative exemplars, aligning closely with the idea of numeric sliders. However, most RepE studies focus on one attribute at a time, without probing how multiple steering directions interact. Further to this, Wu et al. (2025) demonstrate that simple prompting often performs much better than many of the more complex RepE methods discussed above.

Style Transfer and Multi-Attribute Control. Supervised text style transfer methods rely on parallel corpora and sequence-to-sequence models (Jhamtani et al., 2017; Mukherjee et al., 2023), but are constrained by scarce paired data. Unsupervised methods for non-parallel data include prototype editing (swapping style markers with target-style phrases) (Mukherjee et al., 2023), or disentanglement strategies that factorize semantics and style, recombining them via back-translation or adversarial training (Shen et al., 2017; Prabhumoye et al., 2018). While effective for coarse style shifts, these approaches are not naturally suited for fine-grained numeric control or multi-attribute specification.

Fine-Grained Control: Single-Attribute Methods. Most work on fine-grained control introduces a continuous “knob” for a *single* attribute, with evaluation focused on calibration along that one dimension. Families include:

- *Decoding-time guidance.* PPLM (Dathathri et al., 2020) backpropagates from an attribute classifier through LM hidden states at generation time; GeDi (Krause et al., 2021) trains small conditional LMs to reweight token probabilities; FUDGE (Yang and Klein, 2021) trains discriminators predicting sequence-level attributes from partial prefixes; and energy/logit methods such as COLD (Qin et al., 2022) and BOLT (Liu et al., 2023) add attribute-specific energies or biases. Each provides a tunable weight parameter, enabling smooth control of attribute intensity.
- *Product-of-experts.* DExperts (Liu et al., 2021) combine base LMs with expert/anti-expert models, where the mixture coefficient α controls strength.

- *Activation steering.* Contrastive Activation Addition (CAA) (Rimsky et al., 2024) computes steering directions from exemplar differences and scales them at inference.
- *Training-time numeric control.* SteerLM (Dong et al., 2023) finetunes on data labeled with regressor-predicted attribute values, allowing users to set numeric controls such as "positivity=7/10." This achieves excellent calibration but requires labeled data and SFT cycles.

These methods demonstrate smooth, single-attribute control, but rarely extend to *dual-concept* settings. While some (e.g., FUDGE, GeDi, or energy-based methods) can in principle compose multiple guidance signals by assigning separate weights λ_a, λ_b, systematic evaluation of interference between attributes remains limited.

Towards Multi-Concept Fine-Grained Control. Recent frameworks begin to explore smooth, fine-grained evaluation. The Smoothly Controllable Text Generation (SCTG) benchmark (Zhou et al., 2024) defines fine-grained control as the ability to vary an attribute over a 10-point scale, using LLM-as-judge with Elo-style pairwise comparisons to assess calibration and relevance. However, SCTG focuses exclusively on single-attribute scenarios. In contrast, our work explicitly evaluates *dual-concept* fine-grained control, introducing systematic protocols to measure interference when varying one concept while holding another fixed. This perspective highlights the challenges of compositional control and the need for methods robust to attribute entanglement.

Evaluation of Controllability. Evaluation typically relies on automatic classifiers trained to predict style or attribute labels on generated outputs (Moschitti et al., 2014). While efficient, such classifiers often suffer from subjectivity and domain mismatch (Pang, 2019). Human evaluation remains the gold standard but is costly and inconsistent. More recent work explores LLMs themselves as judges (Zheng et al., 2023; Sun et al., 2023), providing scalable and flexible evaluation pipelines. Our evaluation setup builds on this line, using pairwise comparisons with strong judge models to assess fine-grained controllability in both single- and dual-concept scenarios.

D Detailed Results

	Llama-3B	**Llama-11B**	**Gemma-4B**	**Gemma-12B**	**Phi-3.8B**	**Qwen-14B**
C_a (single)	$-0.02_{\pm 0.52}$	$-0.02_{\pm 0.52}$	$-0.33_{\pm 0.47}$	$0.52_{\pm 0.46}$	$-0.05_{\pm 0.54}$	$0.65_{\pm 0.30}$
$C_a \mid (C_b = 0)$	$-0.11_{\pm 0.53}$	$0.12_{\pm 0.51}$	$0.15_{\pm 0.54}$	$0.24_{\pm 0.48}$	$0.08_{\pm 0.46}$	$0.70_{\pm 0.28}$
$C_a \mid (C_b = 1)$	$-0.12_{\pm 0.49}$	$-0.07_{\pm 0.53}$	$0.12_{\pm 0.47}$	$0.01_{\pm 0.55}$	$-0.06_{\pm 0.48}$	$0.72_{\pm 0.26}$
$C_a \mid (C_b = 2)$	$0.01_{\pm 0.48}$	$0.03_{\pm 0.54}$	$0.16_{\pm 0.45}$	$0.03_{\pm 0.58}$	$-0.19_{\pm 0.46}$	$0.62_{\pm 0.35}$
$C_a \mid (C_b = 3)$	$-0.09_{\pm 0.49}$	$0.07_{\pm 0.53}$	$-0.08_{\pm 0.47}$	$-0.02_{\pm 0.58}$	$-0.07_{\pm 0.47}$	$0.61_{\pm 0.40}$
$C_a \mid (C_b = 4)$	$0.15_{\pm 0.48}$	$-0.03_{\pm 0.53}$	$-0.215_{\pm 0.52}$	$-0.16_{\pm 0.53}$	$-0.24_{\pm 0.48}$	$0.55_{\pm 0.36}$
$C_a \mid C_b$ fixed	$-0.03_{\pm 0.50}$	$0.02_{\pm 0.53}$	$0.03_{\pm 0.51}$	$0.02_{\pm 0.56}$	$-0.09_{\pm 0.48}$	$0.64_{\pm 0.34}$
$C_a \mid C_b$ rand	$-0.11_{\pm 0.50}$	$-0.05_{\pm 0.51}$	$0.04_{\pm 0.48}$	$0.12_{\pm 0.50}$	$0.02_{\pm 0.54}$	$0.63_{\pm 0.32}$
C_b (single)	$0.69_{\pm 0.32}$	$0.76_{\pm 0.25}$	$0.73_{\pm 0.34}$	$0.95_{\pm 0.07}$	$0.40_{\pm 0.47}$	$0.93_{\pm 0.10}$
$C_b \mid (C_a = 0)$	$0.47_{\pm 0.50}$	$0.62_{\pm 0.31}$	$0.42_{\pm 0.50}$	$0.87_{\pm 0.15}$	$0.17_{\pm 0.45}$	$0.90_{\pm 0.11}$
$C_b \mid (C_a = 1)$	$0.57_{\pm 0.37}$	$0.86_{\pm 0.21}$	$0.55_{\pm 0.47}$	$0.81_{\pm 0.20}$	$0.31_{\pm 0.49}$	$0.85_{\pm 0.18}$
$C_b \mid (C_a = 2)$	$0.51_{\pm 0.40}$	$0.74_{\pm 0.24}$	$0.53_{\pm 0.36}$	$0.83_{\pm 0.20}$	$0.26_{\pm 0.45}$	$0.86_{\pm 0.16}$
$C_b \mid (C_a = 3)$	$0.51_{\pm 0.46}$	$0.76_{\pm 0.24}$	$0.45_{\pm 0.51}$	$0.77_{\pm 0.24}$	$0.32_{\pm 0.44}$	$0.90_{\pm 0.13}$
$C_b \mid (C_a = 4)$	$0.48_{\pm 0.43}$	$0.85_{\pm 0.17}$	$0.59_{\pm 0.41}$	$0.89_{\pm 0.11}$	$0.45_{\pm 0.44}$	$0.90_{\pm 0.12}$
$C_b \mid C_a$ fixed	$0.51_{\pm 0.43}$	$0.76_{\pm 0.25}$	$0.51_{\pm 0.46}$	$0.83_{\pm 0.19}$	$0.30_{\pm 0.46}$	$0.88_{\pm 0.14}$
$C_b \mid C_a$ rand	$0.57_{\pm 0.43}$	$0.77_{\pm 0.29}$	$0.51_{\pm 0.44}$	$0.80_{\pm 0.18}$	$0.29_{\pm 0.47}$	$0.84_{\pm 0.15}$

Table 4: ARGUMENT generation **clarity–politeness**.

	Llama-3B	**Llama-11B**	**Gemma-4B**	**Gemma-12B**	**Phi-3.8B**	**Qwen-14B**
C_a (single)	$0.77_{\pm 0.27}$	$0.92_{\pm 0.09}$	$0.94_{\pm 0.09}$	$0.98_{\pm 0.03}$	$0.69_{\pm 0.28}$	$0.99_{\pm 0.02}$
$C_a \mid (C_b = 0)$	$0.34_{\pm 0.47}$	$0.33_{\pm 0.40}$	$0.91_{\pm 0.12}$	$0.97_{\pm 0.05}$	$0.44_{\pm 0.41}$	$0.95_{\pm 0.08}$
$C_a \mid (C_b = 1)$	$0.41_{\pm 0.39}$	$0.66_{\pm 0.36}$	$0.90_{\pm 0.13}$	$0.97_{\pm 0.06}$	$0.38_{\pm 0.41}$	$0.98_{\pm 0.04}$
$C_a \mid (C_b = 2)$	$0.44_{\pm 0.47}$	$0.66_{\pm 0.32}$	$0.87_{\pm 0.16}$	$0.97_{\pm 0.04}$	$0.49_{\pm 0.38}$	$0.98_{\pm 0.04}$
$C_a \mid (C_b = 3)$	$0.40_{\pm 0.45}$	$0.71_{\pm 0.27}$	$0.84_{\pm 0.21}$	$0.96_{\pm 0.05}$	$0.51_{\pm 0.37}$	$0.98_{\pm 0.03}$
$C_a \mid (C_b = 4)$	$0.36_{\pm 0.43}$	$0.43_{\pm 0.46}$	$0.88_{\pm 0.17}$	$0.96_{\pm 0.06}$	$0.58_{\pm 0.37}$	$0.98_{\pm 0.03}$
$C_a \mid C_b$ fixed	$0.39_{\pm 0.44}$	$0.56_{\pm 0.40}$	$0.88_{\pm 0.16}$	$0.97_{\pm 0.05}$	$0.48_{\pm 0.41}$	$0.97_{\pm 0.05}$
$C_a \mid C_b$ rand	$0.30_{\pm 0.46}$	$0.41_{\pm 0.43}$	$0.82_{\pm 0.22}$	$0.92_{\pm 0.10}$	$0.37_{\pm 0.44}$	$0.94_{\pm 0.08}$
C_b (single)	$0.58_{\pm 0.40}$	$0.75_{\pm 0.32}$	$0.88_{\pm 0.12}$	$0.99_{\pm 0.03}$	$0.41_{\pm 0.39}$	$0.98_{\pm 0.03}$
$C_b \mid (C_a = 0)$	$0.38_{\pm 0.37}$	$0.43_{\pm 0.47}$	$0.56_{\pm 0.39}$	$0.97_{\pm 0.07}$	$0.35_{\pm 0.46}$	$0.94_{\pm 0.07}$
$C_b \mid (C_a = 1)$	$0.42_{\pm 0.41}$	$0.58_{\pm 0.42}$	$0.54_{\pm 0.37}$	$0.92_{\pm 0.09}$	$0.44_{\pm 0.40}$	$0.95_{\pm 0.06}$
$C_b \mid (C_a = 2)$	$0.44_{\pm 0.45}$	$0.49_{\pm 0.47}$	$0.31_{\pm 0.47}$	$0.89_{\pm 0.12}$	$0.29_{\pm 0.49}$	$0.96_{\pm 0.07}$
$C_b \mid (C_a = 3)$	$0.53_{\pm 0.41}$	$0.44_{\pm 0.51}$	$0.36_{\pm 0.44}$	$0.88_{\pm 0.12}$	$0.19_{\pm 0.52}$	$0.95_{\pm 0.05}$
$C_b \mid (C_a = 4)$	$0.28_{\pm 0.44}$	$0.46_{\pm 0.45}$	$0.38_{\pm 0.42}$	$0.85_{\pm 0.15}$	$0.29_{\pm 0.48}$	$0.91_{\pm 0.11}$
$C_b \mid C_a$ fixed	$0.41_{\pm 0.43}$	$0.48_{\pm 0.47}$	$0.43_{\pm 0.43}$	$0.90_{\pm 0.12}$	$0.31_{\pm 0.48}$	$0.94_{\pm 0.08}$
$C_b \mid C_a$ rand	$0.39_{\pm 0.40}$	$0.45_{\pm 0.44}$	$0.26_{\pm 0.48}$	$0.85_{\pm 0.15}$	$0.25_{\pm 0.48}$	$0.93_{\pm 0.07}$

Table 5: ARGUMENT generation **formality–assertiveness**.

	Llama-3B	**Llama-11B**	**Gemma-4B**	**Gemma-12B**	**Phi-3.8B**	**Qwen-14B**
C_a (single)	$0.46_{\pm 0.40}$	$0.76_{\pm 0.23}$	$0.65_{\pm 0.32}$	$0.95_{\pm 0.07}$	$0.38_{\pm 0.43}$	$0.92_{\pm 0.11}$
$C_a \mid (C_b = 0)$	$0.27_{\pm 0.51}$	$0.32_{\pm 0.46}$	$0.38_{\pm 0.53}$	$0.89_{\pm 0.14}$	$0.15_{\pm 0.54}$	$0.91_{\pm 0.11}$
$C_a \mid (C_b = 1)$	$0.18_{\pm 0.47}$	$0.52_{\pm 0.39}$	$0.28_{\pm 0.49}$	$0.84_{\pm 0.18}$	$0.16_{\pm 0.44}$	$0.90_{\pm 0.15}$
$C_a \mid (C_b = 2)$	$0.13_{\pm 0.49}$	$0.47_{\pm 0.41}$	$0.36_{\pm 0.40}$	$0.87_{\pm 0.17}$	$0.18_{\pm 0.47}$	$0.86_{\pm 0.14}$
$C_a \mid (C_b = 3)$	$0.18_{\pm 0.51}$	$0.64_{\pm 0.35}$	$0.35_{\pm 0.46}$	$0.89_{\pm 0.11}$	$0.21_{\pm 0.45}$	$0.86_{\pm 0.19}$
$C_a \mid (C_b = 4)$	$0.29_{\pm 0.49}$	$0.60_{\pm 0.37}$	$0.53_{\pm 0.39}$	$0.92_{\pm 0.09}$	$0.10_{\pm 0.53}$	$0.88_{\pm 0.14}$
$C_a \mid C_b$ fixed	$0.21_{\pm 0.50}$	$0.51_{\pm 0.41}$	$0.38_{\pm 0.46}$	$0.88_{\pm 0.14}$	$0.16_{\pm 0.49}$	$0.88_{\pm 0.15}$
$C_a \mid C_b$ rand	$0.16_{\pm 0.51}$	$0.54_{\pm 0.35}$	$0.33_{\pm 0.41}$	$0.83_{\pm 0.20}$	$0.08_{\pm 0.51}$	$0.88_{\pm 0.16}$
C_b (single)	$0.68_{\pm 0.33}$	$0.81_{\pm 0.22}$	$0.88_{\pm 0.08}$	$0.98_{\pm 0.04}$	$0.36_{\pm 0.46}$	$0.96_{\pm 0.05}$
$C_b \mid (C_a = 0)$	$0.31_{\pm 0.47}$	$0.52_{\pm 0.41}$	$0.89_{\pm 0.12}$	$0.96_{\pm 0.07}$	$0.28_{\pm 0.48}$	$0.86_{\pm 0.15}$
$C_b \mid (C_a = 1)$	$0.39_{\pm 0.48}$	$0.60_{\pm 0.38}$	$0.91_{\pm 0.14}$	$0.85_{\pm 0.13}$	$0.38_{\pm 0.42}$	$0.88_{\pm 0.13}$
$C_b \mid (C_a = 2)$	$0.33_{\pm 0.43}$	$0.64_{\pm 0.34}$	$0.89_{\pm 0.11}$	$0.77_{\pm 0.23}$	$0.27_{\pm 0.45}$	$0.83_{\pm 0.15}$
$C_b \mid (C_a = 3)$	$0.36_{\pm 0.46}$	$0.58_{\pm 0.36}$	$0.83_{\pm 0.19}$	$0.81_{\pm 0.20}$	$0.14_{\pm 0.59}$	$0.80_{\pm 0.20}$
$C_b \mid (C_a = 4)$	$0.22_{\pm 0.54}$	$0.59_{\pm 0.40}$	$0.83_{\pm 0.27}$	$0.77_{\pm 0.22}$	$0.31_{\pm 0.47}$	$0.81_{\pm 0.24}$
$C_b \mid C_a$ fixed	$0.32_{\pm 0.48}$	$0.58_{\pm 0.38}$	$0.87_{\pm 0.18}$	$0.83_{\pm 0.19}$	$0.28_{\pm 0.49}$	$0.84_{\pm 0.18}$
$C_b \mid C_a$ rand	$0.35_{\pm 0.47}$	$0.52_{\pm 0.40}$	$0.82_{\pm 0.20}$	$0.76_{\pm 0.21}$	$0.26_{\pm 0.44}$	$0.81_{\pm 0.21}$

Table 6: ARGUMENT generation **humor–persuasiveness**.

	Llama-3B	**Llama-11B**	**Gemma-4B**	**Gemma-12B**	**Phi-3.8B**	**Qwen-14B**
C_a (single)	$0.20_{\pm 0.54}$	$0.45_{\pm 0.46}$	$0.22_{\pm 0.53}$	$0.92_{\pm 0.11}$	$0.01_{\pm 0.52}$	$0.89_{\pm 0.12}$
$C_a \mid (C_b = 0)$	$-0.08_{\pm 0.52}$	$-0.10_{\pm 0.51}$	$-0.05_{\pm 0.55}$	$0.37_{\pm 0.30}$	$-0.08_{\pm 0.52}$	$0.65_{\pm 0.34}$
$C_a \mid (C_b = 1)$	$0.05_{\pm 0.52}$	$-0.02_{\pm 0.46}$	$-0.10_{\pm 0.49}$	$0.35_{\pm 0.48}$	$-0.14_{\pm 0.48}$	$0.68_{\pm 0.28}$
$C_a \mid (C_b = 2)$	$-0.01_{\pm 0.55}$	$-0.02_{\pm 59}$	$-0.09_{\pm 0.54}$	$0.38_{\pm 0.43}$	$-0.04_{\pm 0.50}$	$0.80_{\pm 0.20}$
$C_a \mid (C_b = 3)$	$0.05_{\pm 0.53}$	$0.03_{\pm 0.49}$	$0.02_{\pm 0.48}$	$0.39_{\pm 0.46}$	$0.06_{\pm 0.45}$	$0.78_{\pm 0.17}$
$C_a \mid (C_b = 4)$	$0.14_{\pm 0.47}$	$0.05_{\pm 0.56}$	$-0.06_{\pm 0.52}$	$0.29_{\pm 0.45}$	$0.02_{\pm 0.53}$	$0.77_{\pm 0.22}$
$C_a \mid C_b$ fixed	$0.03_{\pm 0.52}$	$-0.01_{\pm 0.53}$	$-0.06_{\pm 0.52}$	$0.35_{\pm 0.43}$	$-0.04_{\pm 0.50}$	$0.74_{\pm 0.26}$
$C_a \mid C_b$ rand	$-0.01_{\pm 0.47}$	$-0.07_{\pm 0.50}$	$-0.14_{\pm 0.50}$	$0.29_{\pm 0.47}$	$-0.04_{\pm 0.44}$	$0.64_{\pm 0.29}$
C_b (single)	$0.74_{\pm 0.31}$	$0.84_{\pm 0.21}$	$0.86_{\pm 0.16}$	$0.98_{\pm 0.03}$	$0.20_{\pm 0.49}$	$0.96_{\pm 0.07}$
$C_b \mid (C_a = 0)$	$0.38_{\pm 0.44}$	$0.55_{\pm 0.38}$	$0.60_{\pm 0.38}$	$0.94_{\pm 0.09}$	$0.11_{\pm 0.56}$	$0.96_{\pm 0.06}$
$C_b \mid (C_a = 1)$	$0.44_{\pm 0.47}$	$0.77_{\pm 0.22}$	$0.70_{\pm 0.30}$	$0.89_{\pm 0.12}$	$0.08_{\pm 0.53}$	$0.95_{\pm 0.07}$
$C_b \mid (C_a = 2)$	$0.40_{\pm 0.47}$	$0.72_{\pm 0.28}$	$0.66_{\pm 0.27}$	$0.81_{\pm 0.21}$	$-0.04_{\pm 0.52}$	$0.94_{\pm 0.08}$
$C_b \mid (C_a = 3)$	$0.38_{\pm 0.51}$	$0.72_{\pm 0.28}$	$0.64_{\pm 0.39}$	$0.80_{\pm 0.19}$	$0.09_{\pm 0.53}$	$0.94_{\pm 0.08}$
$C_b \mid (C_a = 4)$	$0.33_{\pm 0.49}$	$0.77_{\pm 0.24}$	$0.73_{\pm 0.27}$	$0.86_{\pm 0.18}$	$0.05_{\pm 0.45}$	$0.95_{\pm 0.08}$
$C_b \mid C_a$ fixed	$0.38_{\pm 0.48}$	$0.71_{\pm 0.30}$	$0.66_{\pm 0.33}$	$0.86_{\pm 0.17}$	$0.06_{\pm 0.52}$	$0.95_{\pm 0.08}$
$C_b \mid C_a$ rand	$0.40_{\pm 0.40}$	$0.71_{\pm 0.31}$	$0.64_{\pm 0.36}$	$0.72_{\pm 0.26}$	$0.14_{\pm 0.49}$	$0.92_{\pm 0.08}$

Table 7: STORY generation **clarity–politeness**.

	Llama-3B	**Llama-11B**	**Gemma-4B**	**Gemma-12B**	**Phi-3.8B**	**Qwen-14B**
C_a (single)	$0.85_{\pm 0.18}$	$0.93_{\pm 0.09}$	$0.83_{\pm 0.21}$	$1.00_{\pm 0.02}$	$0.53_{\pm 0.39}$	$0.98_{\pm 0.05}$
$C_a \mid (C_b = 0)$	$0.72_{\pm 0.27}$	$0.66_{\pm 0.31}$	$0.89_{\pm 0.14}$	$0.98_{\pm 0.05}$	$0.36_{\pm 0.48}$	$0.96_{\pm 0.06}$
$C_a \mid (C_b = 1)$	$0.79_{\pm 0.28}$	$0.83_{\pm 0.21}$	$0.93_{\pm 0.10}$	$0.98_{\pm 0.04}$	$0.40_{\pm 0.39}$	$0.97_{\pm 0.05}$
$C_a \mid (C_b = 2)$	$0.75_{\pm 0.25}$	$0.81_{\pm 0.21}$	$0.87_{\pm 0.15}$	$0.95_{\pm 0.07}$	$0.21_{\pm 0.48}$	$0.98_{\pm 0.04}$
$C_a \mid (C_b = 3)$	$0.83_{\pm 0.19}$	$0.79_{\pm 0.20}$	$0.80_{\pm 0.19}$	$0.96_{\pm 0.07}$	$0.37_{\pm 0.39}$	$0.98_{\pm 0.04}$
$C_a \mid (C_b = 4)$	$0.83_{\pm 0.21}$	$0.79_{\pm 0.26}$	$0.76_{\pm 0.25}$	$0.95_{\pm 0.11}$	$0.39_{\pm 0.47}$	$0.98_{\pm 0.04}$
$C_a \mid C_b$ fixed	$0.78_{\pm 0.25}$	$0.77_{\pm 0.25}$	$0.85_{\pm 0.18}$	$0.96_{\pm 0.07}$	$0.35_{\pm 0.45}$	$0.97_{\pm 0.05}$
$C_a \mid C_b$ rand	$0.74_{\pm 0.28}$	$0.77_{\pm 0.23}$	$0.84_{\pm 0.14}$	$0.96_{\pm 0.06}$	$0.33_{\pm 0.46}$	$0.96_{\pm 0.05}$
C_b (single)	$0.65_{\pm 0.32}$	$0.67_{\pm 0.33}$	$0.71_{\pm 0.28}$	$0.98_{\pm 0.04}$	$0.12_{\pm 0.45}$	$0.97_{\pm 0.06}$
$C_b \mid (C_a = 0)$	$0.36_{\pm 0.42}$	$0.51_{\pm 0.42}$	$0.62_{\pm 0.35}$	$0.91_{\pm 0.10}$	$0.27_{\pm 0.44}$	$0.89_{\pm 0.13}$
$C_b \mid (C_a = 1)$	$0.45_{\pm 0.45}$	$0.55_{\pm 0.44}$	$0.49_{\pm 0.45}$	$0.95_{\pm 0.07}$	$0.19_{\pm 0.50}$	$0.90_{\pm 0.11}$
$C_b \mid (C_a = 2)$	$0.40_{\pm 0.45}$	$0.50_{\pm 0.38}$	$0.45_{\pm 0.43}$	$0.94_{\pm 0.10}$	$0.30_{\pm 0.45}$	$0.93_{\pm 0.10}$
$C_b \mid (C_a = 3)$	$0.44_{\pm 0.47}$	$0.54_{\pm 0.38}$	$0.43_{\pm 0.42}$	$0.94_{\pm 0.06}$	$0.23_{\pm 0.49}$	$0.94_{\pm 0.07}$
$C_b \mid (C_a = 4)$	$0.43_{\pm 0.43}$	$0.44_{\pm 0.44}$	$0.25_{\pm 0.52}$	$0.90_{\pm 0.14}$	$0.21_{\pm 0.51}$	$0.92_{\pm 0.09}$
$C_b \mid C_a$ fixed	$0.42_{\pm 0.45}$	$0.51_{\pm 0.42}$	$0.38_{\pm 0.40}$	$0.93_{\pm 0.10}$	$0.24_{\pm 0.48}$	$0.91_{\pm 0.10}$
$C_b \mid C_a$ rand	$0.35_{\pm 0.44}$	$0.41_{\pm 0.46}$	$0.34_{\pm 0.48}$	$0.89_{\pm 0.12}$	$0.19_{\pm 0.48}$	$0.89_{\pm 0.12}$

Table 8: STORY generation **formality–assertiveness**.

	Llama-3B	**Llama-11B**	**Gemma-4B**	**Gemma-12B**	**Phi-3.8B**	**Qwen-14B**
C_a (single)	0.30±0.50	0.81±0.26	0.57±0.41	0.95±0.06	0.17±0.51	0.92±0.10
$C_a \mid (C_b = 0)$	0.16±0.49	0.24±0.42	0.38±0.45	0.87±0.17	0.03±0.48	0.93±0.10
$C_a \mid (C_b = 1)$	0.15±0.56	0.43±0.46	0.27±0.43	0.84±0.18	0.06±0.49	0.91±0.12
$C_a \mid (C_b = 2)$	0.12±0.53	0.40±0.41	0.43±0.41	0.81±0.24	0.02±0.53	0.90±0.11
$C_a \mid (C_b = 3)$	0.12±0.50	0.28±0.46	0.16±0.50	0.77±0.25	−0.01±0.43	0.89±0.11
$C_a \mid (C_b = 4)$	0.19±0.51	0.47±0.44	0.22±0.52	0.77±0.22	−0.01±0.58	0.88±0.14
$C_a \mid C_b$ fixed	0.15±0.52	0.36±0.45	0.29±0.47	0.81±0.22	0.02±0.50	0.90±0.12
$C_a \mid C_b$ rand	0.15±0.54	0.33±0.49	0.20±0.49	0.74±0.25	0.00±0.49	0.88±0.15
C_b (single)	0.52±0.40	0.80±0.19	0.62±0.38	0.97±0.04	0.11±0.47	0.93±0.10
$C_b \mid (C_a = 0)$	0.35±0.44	0.48±0.40	0.76±0.24	0.86±0.16	0.02±0.43	0.89±0.13
$C_b \mid (C_a = 1)$	0.38±0.44	0.63±0.31	0.72±0.32	0.75±0.29	0.33±0.42	0.88±0.13
$C_b \mid (C_a = 2)$	0.32±0.47	0.64±0.26	0.59±0.40	0.56±0.43	0.31±0.42	0.83±0.22
$C_b \mid (C_a = 3)$	0.39±0.47	0.58±0.33	0.56±0.42	0.71±0.23	0.15±0.53	0.82±0.22
$C_b \mid (C_a = 4)$	0.23±0.51	0.62±0.38	0.67±0.28	0.59±0.41	0.26±0.44	0.83±0.19
$C_b \mid C_a$ fixed	0.33±0.47	0.59±0.35	0.66±0.34	0.69±0.34	0.21±0.47	0.85±0.18
$C_b \mid C_a$ rand	0.37±0.48	0.58±0.34	0.65±0.30	0.70±0.31	0.15±0.46	0.83±0.20

Table 9: STORY generation **humor–persuasiveness**.

	Llama-3B	**Llama-11B**	**Gemma-4B**	**Gemma-12B**	**Phi-3.8B**	**Qwen-14B**
C_a (single)	0.06±0.64	0.21±0.56	0.13±0.64	0.15±0.61	0.03±0.53	0.64±0.21
$C_a \mid (C_b = 0)$	−0.14±0.48	−0.01±0.42	0.07±0.40	0.07±0.44	−0.13±0.50	0.31±0.40
$C_a \mid (C_b = 1)$	−0.09±0.43	−0.02±0.46	0.02±0.37	−0.12±0.50	−0.09±0.50	0.42±0.38
$C_a \mid (C_b = 2)$	−0.04±0.38	−0.06±0.45	0.00±0.41	−0.40±0.53	0.01±0.45	0.46±0.40
$C_a \mid (C_b = 3)$	0.04±0.37	0.08±0.44	0.00±0.47	−0.35±0.53	−0.13±0.49	0.33±0.50
$C_a \mid (C_b = 4)$	0.02±0.41	0.12±0.45	0.03±0.57	−0.44±0.46	−0.04±0.50	0.43±0.45
$C_a \mid C_b$ fixed	−0.04±0.42	0.02±0.45	0.02±0.45	−0.25±0.53	−0.07±0.49	0.39±0.43
$C_a \mid C_b$ rand	−0.04±0.48	0.08±0.40	0.01±0.43	−0.19±0.45	−0.01±0.47	0.38±0.43
C_b (single)	0.48±0.45	0.73±0.28	0.89±0.18	0.97±0.03	0.51±0.37	0.93±0.09
$C_b \mid (C_a = 0)$	0.32±0.46	0.39±0.44	0.67±0.41	0.42±0.46	0.27±0.52	0.58±0.36
$C_b \mid (C_a = 1)$	0.40±0.54	0.58±0.44	0.78±0.24	0.78±0.24	0.29±0.51	0.80±0.22
$C_b \mid (C_a = 2)$	0.34±0.61	0.49±0.39	0.78±0.24	0.91±0.09	0.28±0.51	0.87±0.16
$C_b \mid (C_a = 3)$	0.37±0.58	0.36±0.42	0.78±0.27	0.92±0.07	0.28±0.45	0.88±0.21
$C_b \mid (C_a = 4)$	0.33±0.56	0.42±0.37	0.81±0.24	0.94±0.09	0.32±0.47	0.82±0.20
$C_b \mid C_a$ fixed	0.35±0.55	0.45±0.42	0.76±0.29	0.79±0.31	0.29±0.49	0.79±0.26
$C_b \mid C_a$ rand	0.31±0.58	0.37±0.47	0.74±0.27	0.63±0.33	0.35±0.46	0.76±0.28

Table 10: STRUCTURED generation **clarity–politeness**.

	Llama-3B	**Llama-11B**	**Gemma-4B**	**Gemma-12B**	**Phi-3.8B**	**Qwen-14B**
C_a (single)	0.59±0.48	0.80±0.24	0.90±0.22	0.93±0.14	0.73±0.26	0.96±0.07
$C_a \mid (C_b = 0)$	0.62±0.39	0.18±0.48	0.11±0.50	0.61±0.44	0.31±0.44	0.78±0.22
$C_a \mid (C_b = 1)$	0.66±0.32	0.42±0.39	0.10±0.44	0.71±0.35	0.46±0.41	0.88±0.13
$C_a \mid (C_b = 2)$	0.69±0.40	0.47±0.45	0.28±0.44	0.79±0.32	0.43±0.49	0.92±0.14
$C_a \mid (C_b = 3)$	0.78±0.33	0.52±0.43	0.44±0.41	0.88±0.18	0.35±0.41	0.91±0.11
$C_a \mid (C_b = 4)$	0.85±0.21	0.51±0.40	0.55±0.38	0.86±0.19	0.47±0.40	0.90±0.07
$C_a \mid C_b$ fixed	0.72±0.35	0.42±0.45	0.30±0.47	0.77±0.33	0.40±0.44	0.88±0.15
$C_a \mid C_b$ rand	0.70±0.30	0.22±0.48	0.12±0.48	0.71±0.33	0.42±0.39	0.86±0.17
C_b (single)	0.66±0.35	0.66±0.32	0.92±0.10	0.95±0.08	0.54±0.34	0.87±0.16
$C_b \mid (C_a = 0)$	0.44±0.48	0.32±0.55	0.71±0.36	0.28±0.44	0.20±0.47	0.61±0.36
$C_b \mid (C_a = 1)$	0.32±0.52	0.48±0.46	0.72±0.36	0.74±0.30	0.49±0.34	0.77±0.32
$C_b \mid (C_a = 2)$	0.47±0.45	0.45±0.47	0.76±0.35	0.85±0.18	0.32±0.45	0.75±0.31
$C_b \mid (C_a = 3)$	0.48±0.46	0.46±0.50	0.71±0.31	0.87±0.27	0.30±0.44	0.81±0.20
$C_b \mid (C_a = 4)$	0.36±0.58	0.48±0.48	0.71±0.34	0.86±0.15	0.43±0.45	0.87±0.15
$C_b \mid C_a$ fixed	0.41±0.50	0.43±0.50	0.72±0.35	0.72±0.36	0.35±0.44	0.76±0.29
$C_b \mid C_a$ rand	0.30±0.52	0.40±0.51	0.74±0.30	0.72±0.26	0.33±0.49	0.75±0.24

Table 11: STRUCTURED generation **formality–assertiveness**.

	Llama-3B	**Llama-11B**	**Gemma-4B**	**Gemma-12B**	**Phi-3.8B**	**Qwen-14B**
C_a (single)	0.47±0.44	0.73±0.22	0.86±0.12	0.94±0.12	0.54±0.38	0.90±0.12
$C_a \mid (C_b = 0)$	0.33±0.53	0.19±0.52	0.42±0.41	0.89±0.15	0.11±0.52	0.85±0.21
$C_a \mid (C_b = 1)$	0.23±0.56	0.28±0.48	0.22±0.48	0.88±0.17	0.09±0.53	0.86±0.17
$C_a \mid (C_b = 2)$	0.21±0.50	0.31±0.52	0.36±0.45	0.86±0.16	0.18±0.45	0.80±0.20
$C_a \mid (C_b = 3)$	0.27±0.52	0.36±0.48	0.45±0.43	0.88±0.12	0.34±0.40	0.88±0.12
$C_a \mid (C_b = 4)$	0.14±0.50	0.44±0.49	0.52±0.38	0.91±0.12	0.21±0.46	0.83±0.22
$C_a \mid C_b$ fixed	0.24±0.53	0.31±0.50	0.39±0.44	0.88±0.15	0.19±0.48	0.84±0.19
$C_a \mid C_b$ rand	0.24±0.50	0.17±0.48	0.28±0.43	0.79±0.21	0.22±0.47	0.81±0.21
C_b (single)	0.58±0.46	0.89±0.14	0.95±0.06	0.99±0.02	0.59±0.37	0.99±0.03
$C_b \mid (C_a = 0)$	0.45±0.52	0.39±0.45	0.91±0.12	0.94±0.16	0.22±0.48	0.96±0.09
$C_b \mid (C_a = 1)$	0.41±0.53	0.62±0.40	0.97±0.05	0.90±0.10	0.31±0.47	0.90±0.12
$C_b \mid (C_a = 2)$	0.42±0.52	0.63±0.37	0.94±0.08	0.89±0.12	0.17±0.49	0.88±0.18
$C_b \mid (C_a = 3)$	0.36±0.51	0.59±0.37	0.93±0.08	0.90±0.20	0.28±0.45	0.88±0.15
$C_b \mid (C_a = 4)$	0.29±0.44	0.60±0.40	0.94±0.08	0.91±0.12	0.38±0.41	0.88±0.13
$C_b \mid C_a$ fixed	0.38±0.51	0.56±0.41	0.94±0.09	0.91±0.15	0.27±0.47	0.90±0.14
$C_b \mid C_a$ rand	0.42±0.44	0.51±0.39	0.94±0.08	0.79±0.19	0.25±0.48	0.83±0.19

Table 12: STRUCTURED generation **humour–persuasiveness**.

E Detailed Fisher Score Results

To mitigate the nonlinearity and boundedness of correlation coefficients when aggregating across contexts, we also report results after applying Fisher's z-transformation to the Spearman correlations before averaging. In addition, the statistical comparisons in Appendix F are conducted on Fisher z-transformed correlations. The Fisher transform is,

$$z = \frac{1}{2} \ln\left(\frac{1+\rho}{1-\rho}\right), \tag{5}$$

where ρ denotes the Spearman correlation.

	Llama-11B	**Gemma-12B**	**Qwen-14B**
C_a (single)	$-0.03_{\pm 0.78}$	$1.27_{\pm 2.08}$	$1.42_{\pm 1.97}$
$C_a \mid (C_b = 0)$	$0.28_{\pm 1.18}$	$0.42_{\pm 1.18}$	$1.62_{\pm 2.10}$
$C_a \mid (C_b = 1)$	$-0.13_{\pm 0.78}$	$0.03_{\pm 0.72}$	$1.57_{\pm 1.93}$
$C_a \mid (C_b = 2)$	$0.03_{\pm 0.79}$	$-0.11_{\pm 1.60}$	$1.21_{\pm 1.61}$
$C_a \mid (C_b = 3)$	$0.10_{\pm 0.76}$	$0.01_{\pm 0.86}$	$1.36_{\pm 2.01}$
$C_a \mid (C_b = 4)$	$-0.13_{\pm 1.17}$	$-0.10_{\pm 1.26}$	$1.05_{\pm 1.63}$
$C_a \mid C_b$ fixed	$0.03_{\pm 0.96}$	$0.05_{\pm 1.17}$	$1.36_{\pm 1.87}$
$C_a \mid C_b$ rand	$0.02_{\pm 1.19}$	$0.34_{\pm 1.48}$	$1.22_{\pm 1.60}$
C_b (single)	$2.45_{\pm 2.92}$	$4.27_{\pm 3.23}$	$3.83_{\pm 3.18}$
$C_b \mid (C_a = 0)$	$1.12_{\pm 1.37}$	$3.24_{\pm 3.17}$	$3.27_{\pm 3.04}$
$C_b \mid (C_a = 1)$	$2.33_{\pm 2.33}$	$1.97_{\pm 2.15}$	$2.86_{\pm 2.95}$
$C_b \mid (C_a = 2)$	$1.50_{\pm 1.73}$	$2.16_{\pm 2.25}$	$2.81_{\pm 2.86}$
$C_b \mid (C_a = 3)$	$1.54_{\pm 1.72}$	$2.32_{\pm 2.73}$	$3.29_{\pm 3.04}$
$C_b \mid (C_a = 4)$	$2.60_{\pm 2.73}$	$2.89_{\pm 2.81}$	$3.13_{\pm 2.91}$
$C_b \mid C_a$ fixed	$1.82_{\pm 2.10}$	$2.51_{\pm 2.68}$	$3.07_{\pm 2.95}$
$C_b \mid C_a$ rand	$2.02_{\pm 2.31}$	$1.96_{\pm 2.29}$	$2.26_{\pm 2.47}$

Table 13: ARGUMENT generation **clarity–politeness** (Fisher-transformed Mean ± SD).

	Llama-11B	**Gemma-12B**	**Qwen-14B**
C_a (single)	$3.38_{\pm 2.97}$	$6.02_{\pm 3.12}$	$6.74_{\pm 2.79}$
$C_a \mid (C_b = 0)$	$0.51_{\pm 1.06}$	$5.29_{\pm 3.28}$	$4.53_{\pm 3.30}$
$C_a \mid (C_b = 1)$	$1.49_{\pm 1.97}$	$5.70_{\pm 3.25}$	$5.37_{\pm 3.27}$
$C_a \mid (C_b = 2)$	$1.01_{\pm 0.62}$	$4.35_{\pm 3.16}$	$5.52_{\pm 3.19}$
$C_a \mid (C_b = 3)$	$1.08_{\pm 0.60}$	$3.73_{\pm 2.95}$	$5.08_{\pm 3.23}$
$C_a \mid (C_b = 4)$	$0.86_{\pm 1.47}$	$3.92_{\pm 3.02}$	$5.59_{\pm 3.20}$
$C_a \mid C_b$ fixed	$0.99_{\pm 1.29}$	$4.60_{\pm 3.21}$	$5.22_{\pm 3.24}$
$C_a \mid C_b$ rand	$0.54_{\pm 0.67}$	$3.00_{\pm 2.75}$	$3.91_{\pm 3.13}$
C_b (single)	$1.59_{\pm 1.75}$	$6.19_{\pm 3.07}$	$6.59_{\pm 2.94}$
$C_b \mid (C_a = 0)$	$0.92_{\pm 1.71}$	$6.37_{\pm 3.09}$	$3.83_{\pm 3.18}$
$C_b \mid (C_a = 1)$	$1.12_{\pm 1.65}$	$3.11_{\pm 2.91}$	$4.98_{\pm 3.42}$
$C_b \mid (C_a = 2)$	$0.93_{\pm 1.50}$	$2.64_{\pm 2.57}$	$5.20_{\pm 3.37}$
$C_b \mid (C_a = 3)$	$0.82_{\pm 1.24}$	$2.38_{\pm 2.42}$	$3.97_{\pm 3.17}$
$C_b \mid (C_a = 4)$	$1.06_{\pm 1.89}$	$2.50_{\pm 2.64}$	$3.29_{\pm 3.03}$
$C_b \mid C_a$ fixed	$0.97_{\pm 1.61}$	$3.40_{\pm 3.11}$	$4.25_{\pm 3.30}$
$C_b \mid C_a$ rand	$0.92_{\pm 1.68}$	$2.72_{\pm 2.89}$	$3.32_{\pm 3.00}$

Table 14: ARGUMENT generation **formality–assertiveness** (Fisher-transformed Mean ± SD).

	Llama-11B	**Gemma-12B**	**Qwen-14B**
C_a (single)	1.94±2.32	4.59±3.33	4.02±3.33
$C_a \mid (C_b = 0)$	0.52±1.13	3.23±3.07	3.37±3.09
$C_a \mid (C_b = 1)$	1.03±1.64	2.22±2.35	4.05±3.42
$C_a \mid (C_b = 2)$	0.76±1.15	2.41±2.42	2.45±2.53
$C_a \mid (C_b = 3)$	1.25±1.60	2.69±2.67	3.15±3.12
$C_a \mid (C_b = 4)$	1.26±1.82	3.29±2.92	3.01±2.97
$C_a \mid C_b$ fixed	0.96±1.51	2.77±2.72	3.20±3.07
$C_a \mid C_b$ rand	0.76±0.61	2.72±2.91	3.71±3.37
C_b (single)	1.91±2.02	6.31±3.09	4.96±3.34
$C_b \mid (C_a = 0)$	1.04±1.66	4.99±3.32	2.54±2.62
$C_b \mid (C_a = 1)$	1.33±2.00	1.90±1.97	2.82±2.84
$C_b \mid (C_a = 2)$	1.33±1.80	2.01±2.43	2.24±2.48
$C_b \mid (C_a = 3)$	1.03±1.38	2.02±2.28	1.66±1.69
$C_b \mid (C_a = 4)$	1.27±1.85	1.58±1.71	2.25±2.50
$C_b \mid C_a$ fixed	1.20±1.75	2.50±2.70	2.30±2.48
$C_b \mid C_a$ rand	1.10±1.84	1.56±1.72	2.01±2.27

Table 15: ARGUMENT generation **humour–persuasiveness** (Fisher-transformed Mean ± SD).

	Llama-11B	**Gemma-12B**	**Qwen-14B**
C_a (single)	0.91±1.72	3.80±3.21	2.61±2.59
$C_a \mid (C_b = 0)$	−0.13±0.68	0.45±0.43	1.23±1.59
$C_a \mid (C_b = 1)$	−0.02±0.66	0.50±0.72	1.28±1.56
$C_a \mid (C_b = 2)$	−0.00±0.84	0.68±1.17	2.05±2.41
$C_a \mid (C_b = 3)$	0.05±0.76	0.64±1.14	1.54±1.70
$C_a \mid (C_b = 4)$	0.06±0.80	0.41±0.66	1.77±2.04
$C_a \mid C_b$ fixed	−0.01±0.75	0.54±0.87	1.57±1.90
$C_a \mid C_b$ rand	−0.08±0.74	0.58±1.46	1.10±1.33
C_b (single)	2.95±3.02	6.10±3.10	5.34±3.32
$C_b \mid (C_a = 0)$	1.12±1.65	3.90±3.14	5.00±3.39
$C_b \mid (C_a = 1)$	1.68±1.89	2.31±2.17	5.04±3.45
$C_b \mid (C_a = 2)$	1.47±1.76	1.69±1.69	3.84±3.17
$C_b \mid (C_a = 3)$	1.33±1.33	1.69±1.70	3.87±3.15
$C_b \mid (C_a = 4)$	1.44±1.30	2.43±2.43	4.38±3.33
$C_b \mid C_a$ fixed	1.41±1.61	2.41±2.42	4.43±3.32
$C_b \mid C_a$ rand	2.01±2.59	1.46±1.75	3.65±3.20

Table 16: STORY generation **clarity–politeness** (Fisher-transformed Mean ± SD).

	Llama-11B	**Gemma-12B**	**Qwen-14B**
C_a (single)	4.14±3.33	7.56±2.16	6.30±3.09
$C_a \mid (C_b = 0)$	1.87±2.64	6.18±3.08	5.12±3.37
$C_a \mid (C_b = 1)$	2.54±2.75	6.14±3.13	5.36±3.29
$C_a \mid (C_b = 2)$	2.24±2.50	4.45±3.27	5.86±3.23
$C_a \mid (C_b = 3)$	1.64±1.71	4.13±3.16	6.04±3.18
$C_a \mid (C_b = 4)$	1.86±2.04	4.28±3.23	6.81±2.86
$C_a \mid C_b$ fixed	2.03±2.37	5.04±3.29	5.84±3.23
$C_a \mid C_b$ rand	2.10±2.54	4.08±3.09	5.20±3.37
C_b (single)	1.47±1.96	5.90±3.19	5.17±3.32
$C_b \mid (C_a = 0)$	1.10±1.85	3.25±2.95	3.36±3.19
$C_b \mid (C_a = 1)$	1.33±2.23	4.38±3.33	3.27±3.04
$C_b \mid (C_a = 2)$	0.90±1.40	4.63±3.38	3.72±3.17
$C_b \mid (C_a = 3)$	1.07±1.64	3.86±3.25	4.09±3.27
$C_b \mid (C_a = 4)$	0.70±1.12	3.39±3.08	3.38±3.08
$C_b \mid C_a$ fixed	1.02±1.69	3.90±3.23	3.56±3.15
$C_b \mid C_a$ rand	0.87±1.69	3.01±2.97	2.87±2.82

Table 17: STORY generation **formality–assertiveness** (Fisher-transformed Mean ± SD).

	Llama-11B	**Gemma-12B**	**Qwen-14B**
C_a (single)	2.34±2.60	4.37±3.33	3.67±3.20
$C_a \mid (C_b = 0)$	0.34±0.63	2.07±1.95	4.14±3.34
$C_a \mid (C_b = 1)$	0.92±1.70	2.17±2.24	3.06±2.84
$C_a \mid (C_b = 2)$	0.73±1.41	1.95±2.01	3.13±3.01
$C_a \mid (C_b = 3)$	0.49±1.15	1.51±1.53	2.70±2.67
$C_a \mid (C_b = 4)$	0.70±0.74	1.34±1.00	3.22±3.08
$C_a \mid C_b$ fixed	0.64±1.21	1.81±1.82	3.25±3.02
$C_a \mid C_b$ rand	0.54±1.17	1.69±2.07	2.89±2.82
C_b (single)	1.57±1.51	4.98±3.24	3.76±3.14
$C_b \mid (C_a = 0)$	0.69±0.66	2.06±2.10	2.85±2.83
$C_b \mid (C_a = 1)$	1.19±1.59	1.43±1.33	2.18±2.07
$C_b \mid (C_a = 2)$	1.16±1.57	1.19±1.85	2.35±2.46
$C_b \mid (C_a = 3)$	0.91±1.05	1.06±0.54	1.80±1.84
$C_b \mid (C_a = 4)$	1.12±1.42	1.10±1.42	2.19±2.37
$C_b \mid C_a$ fixed	1.01±1.32	1.37±1.58	2.28±2.35
$C_b \mid C_a$ rand	1.11±1.61	1.43±1.77	2.20±2.37

Table 18: STORY generation **humour–persuasiveness** (Fisher-transformed Mean ± SD).

	Llama-11B	**Gemma-12B**	**Qwen-14B**
C_a (single)	0.52±1.56	0.43±1.62	1.23±1.77
$C_a \mid (C_b = 0)$	−0.01±0.53	0.11±0.66	0.44±0.61
$C_a \mid (C_b = 1)$	−0.13±1.15	−0.15±0.74	0.55±0.57
$C_a \mid (C_b = 2)$	−0.07±0.60	−0.60±0.86	0.82±1.40
$C_a \mid (C_b = 3)$	0.09±0.61	−0.60±1.18	0.66±1.49
$C_a \mid (C_b = 4)$	0.17±0.61	−0.69±0.77	0.89±1.67
$C_a \mid C_b$ fixed	0.01±0.74	−0.39±0.91	0.67±1.24
$C_a \mid C_b$ rand	0.10±0.49	−0.28±0.70	0.62±1.11
C_b (single)	1.76±2.22	4.56±3.17	3.42±2.95
$C_b \mid (C_a = 0)$	0.56±0.69	0.70±0.81	1.15±1.63
$C_b \mid (C_a = 1)$	1.05±1.18	1.37±1.00	1.80±1.87
$C_b \mid (C_a = 2)$	0.83±1.11	2.01±1.58	2.63±2.59
$C_b \mid (C_a = 3)$	0.52±0.65	2.23±1.87	2.87±2.83
$C_b \mid (C_a = 4)$	0.58±0.61	2.83±2.36	1.92±2.00
$C_b \mid C_a$ fixed	0.71±0.90	1.83±1.78	2.07±2.30
$C_b \mid C_a$ rand	0.62±1.79	1.16±1.37	1.43±1.31

Table 19: STRUCTURED generation **clarity–politeness** (Fisher-transformed Mean ± SD).

	Llama-11B	**Gemma-12B**	**Qwen-14B**
C_a (single)	1.61±1.52	3.08±2.61	5.39±3.35
$C_a \mid (C_b = 0)$	0.27±0.78	1.38±1.87	1.79±2.04
$C_a \mid (C_b = 1)$	0.59±0.64	1.58±1.79	3.10±3.03
$C_a \mid (C_b = 2)$	0.78±1.13	2.14±2.41	4.22±3.37
$C_a \mid (C_b = 3)$	1.06±1.66	2.80±2.75	3.69±3.28
$C_a \mid (C_b = 4)$	1.03±1.65	2.95±3.01	2.50±2.48
$C_a \mid C_b$ fixed	0.75±1.27	2.17±2.48	3.06±3.00
$C_a \mid C_b$ rand	0.40±1.16	1.93±2.48	3.04±3.07
C_b (single)	1.40±1.80	3.88±3.06	2.36±2.31
$C_b \mid (C_a = 0)$	0.45±0.81	0.48±1.11	1.20±1.61
$C_b \mid (C_a = 1)$	0.93±1.47	1.27±1.04	2.53±2.91
$C_b \mid (C_a = 2)$	0.92±1.68	2.06±2.11	2.00±2.46
$C_b \mid (C_a = 3)$	0.75±0.90	2.58±2.51	2.42±2.69
$C_b \mid (C_a = 4)$	1.08±1.92	1.95±1.82	2.50±2.52
$C_b \mid C_a$ fixed	0.83±1.43	1.67±1.94	2.13±2.51
$C_b \mid C_a$ rand	0.69±1.21	1.32±1.34	2.16±2.66

Table 20: STRUCTURED generation **formality–assertiveness** (Fisher-transformed Mean ± SD).

	Llama-11B	**Gemma-12B**	**Qwen-14B**
C_a (single)	1.27±1.29	5.17±3.41	3.36±3.10
$C_a \mid (C_b = 0)$	0.38±1.17	3.13±3.01	3.44±3.35
$C_a \mid (C_b = 1)$	0.37±0.70	3.04±2.96	3.30±3.24
$C_a \mid (C_b = 2)$	0.37±1.28	2.75±2.88	2.25±2.62
$C_a \mid (C_b = 3)$	0.80±1.71	3.01±2.97	2.92±2.91
$C_a \mid (C_b = 4)$	0.74±1.18	3.74±3.25	2.98±3.11
$C_a \mid C_b$ fixed	0.53±1.26	3.14±3.02	2.98±3.07
$C_a \mid C_b$ rand	0.24±0.66	2.42±2.81	3.01±3.20
C_b (single)	2.49±2.38	6.81±2.76	7.63±2.13
$C_b \mid (C_a = 0)$	0.63±1.12	4.55±3.28	5.78±3.34
$C_b \mid (C_a = 1)$	1.47±2.16	2.90±2.80	3.20±2.98
$C_b \mid (C_a = 2)$	1.34±1.82	3.22±3.07	3.63±3.33
$C_b \mid (C_a = 3)$	1.17±1.64	3.46±3.14	3.10±3.03
$C_b \mid (C_a = 4)$	1.31±1.84	3.73±3.25	3.35±3.20
$C_b \mid C_a$ fixed	1.18±1.76	3.57±3.15	3.81±3.32
$C_b \mid C_a$ rand	1.10±1.85	1.70±1.88	2.40±2.68

Table 21: STRUCTURED generation **humour–persuasiveness** (Fisher-transformed Mean ± SD).

F Statistical Test Results

We conduct paired t-tests on Fisher z-transformed Spearman correlations for each medium-sized model and concept pair in the argument-generation dataset. Because Spearman correlations are bounded and skewed near 1, we first apply the Fisher transformation. For each sample, we compute the per-sample difference between the single- and dual-concept Fisher values and perform a one-sided paired t-test.

The null hypothesis is that the mean difference is zero, corresponding to no change in performance, while the alternative hypothesis is that single-concept control is stronger than dual-concept control. The resulting t-statistics and one-sided p-values are reported below. Large positive t-values accompanied by extremely small p-values indicate that introducing a secondary concept significantly harms controllability.

F.1 Random Secondary Results

Primary	**Secondary**	**Llama-11B**		**Gemma-12B**		**Qwen-14B**	
		t	p (one-sided)	t	p (one-sided)	t	p (one-sided)
assertiveness	formality	2.37	1.01×10^{-2}	6.98	5.30×10^{-10}	6.79	1.22×10^{-9}
formality	assertiveness	8.34	1.45×10^{-12}	6.13	1.99×10^{-8}	5.71	1.10×10^{-7}
clarity	politeness	-0.31	6.22×10^{-1}	3.27	8.06×10^{-4}	0.66	2.56×10^{-1}
politeness	clarity	1.18	1.21×10^{-1}	5.31	5.48×10^{-7}	3.44	4.77×10^{-4}
humor	persuasiveness	4.35	2.14×10^{-5}	3.63	2.61×10^{-4}	0.53	2.99×10^{-1}
persuasiveness	humor	2.43	8.80×10^{-3}	12.48	3.45×10^{-20}	6.33	8.58×10^{-9}

Table 22: ARGUMENT generation paired one-sided t-tests on Fisher z-transformed Spearman correlations comparing single-concept and dual-concept with **random** secondary control across medium-sized models.

Primary	**Secondary**	**Llama-11B**		**Gemma-12B**		**Qwen-14B**	
		t	p (one-sided)	t	p (one-sided)	t	p (one-sided)
assertiveness	formality	1.99	2.50×10^{-2}	5.62	1.59×10^{-7}	4.46	1.41×10^{-5}
formality	assertiveness	4.14	4.60×10^{-5}	8.45	9.09×10^{-13}	1.91	3.03×10^{-2}
clarity	politeness	4.23	3.35×10^{-5}	7.41	8.31×10^{-11}	4.91	2.92×10^{-6}
politeness	clarity	2.10	1.96×10^{-2}	11.49	2.01×10^{-18}	3.22	9.65×10^{-4}
humor	persuasiveness	5.63	1.51×10^{-7}	6.27	1.10×10^{-8}	1.74	4.30×10^{-2}
persuasiveness	humor	1.65	5.12×10^{-2}	8.36	1.37×10^{-12}	3.44	4.83×10^{-4}

Table 23: STORY generation paired one-sided t-tests on Fisher z-transformed Spearman correlations comparing single-concept and dual-concept with **random** secondary control across medium-sized models.

Primary	**Secondary**	**Llama-11B**		**Gemma-12B**		**Qwen-14B**	
		t	p (one-sided)	t	p (one-sided)	t	p (one-sided)
assertiveness	formality	3.92	9.69×10^{-5}	7.14	2.71×10^{-10}	0.47	3.20×10^{-1}
formality	assertiveness	5.76	8.94×10^{-8}	3.44	4.73×10^{-4}	5.04	1.60×10^{-6}
clarity	politeness	2.40	9.41×10^{-3}	3.37	6.04×10^{-4}	3.18	1.07×10^{-3}
politeness	clarity	3.40	5.49×10^{-4}	9.42	1.33×10^{-14}	5.28	6.29×10^{-7}
humor	persuasiveness	5.66	1.35×10^{-7}	5.44	3.32×10^{-7}	0.78	2.20×10^{-1}
persuasiveness	humor	4.22	3.42×10^{-5}	13.97	9.52×10^{-23}	12.99	4.49×10^{-21}

Table 24: STRUCTURED generation paired one-sided t-tests on Fisher z-transformed Spearman correlations comparing single-concept and dual-concept with **random** secondary control across medium-sized models.

F.2 Constant Secondary Results

For the constant secondary statistical tests, we average the fisher scores across the five fixed concept levels before performing the test.

Primary	Secondary	Llama-11B		Gemma-12B		Qwen-14B	
		t	p (one-sided)	t	p (one-sided)	t	p (one-sided)
assertiveness	formality	2.75	3.76×10^{-3}	7.22	1.90×10^{-10}	6.84	9.85×10^{-10}
formality	assertiveness	6.78	1.28×10^{-9}	3.59	2.96×10^{-4}	4.14	4.59×10^{-5}
clarity	politeness	-0.53	7.02×10^{-1}	5.01	1.80×10^{-6}	0.24	4.07×10^{-1}
politeness	clarity	1.74	4.28×10^{-2}	4.47	1.40×10^{-5}	1.82	3.61×10^{-2}
humor	persuasiveness	3.60	2.89×10^{-4}	4.81	3.87×10^{-6}	1.87	3.25×10^{-2}
persuasiveness	humor	2.73	3.92×10^{-3}	10.07	8.01×10^{-16}	6.49	4.39×10^{-9}

Table 25: ARGUMENT generation paired one-sided t-tests on Fisher z-transformed Spearman correlations comparing single-concept and dual-concept with **constant** secondary control across medium-sized models.

Primary	Secondary	Llama-11B		Gemma-12B		Qwen-14B	
		t	p (one-sided)	t	p (one-sided)	t	p (one-sided)
assertiveness	formality	2.07	2.12×10^{-2}	4.96	2.22×10^{-6}	3.86	1.21×10^{-4}
formality	assertiveness	5.14	1.07×10^{-6}	8.66	3.68×10^{-13}	1.18	1.21×10^{-1}
clarity	politeness	4.49	1.30×10^{-5}	8.74	2.52×10^{-13}	3.22	9.42×10^{-4}
politeness	clarity	4.49	1.27×10^{-5}	9.53	8.49×10^{-15}	2.19	1.59×10^{-2}
humor	persuasiveness	5.50	2.61×10^{-7}	6.12	2.00×10^{-8}	1.06	1.47×10^{-1}
persuasiveness	humor	2.78	3.44×10^{-3}	9.13	4.76×10^{-14}	3.80	1.45×10^{-4}

Table 26: STORY generation paired one-sided t-tests on Fisher z-transformed Spearman correlations comparing single-concept and dual-concept with **constant** secondary control across medium-sized models.

Primary	Secondary	Llama-11B		Gemma-12B		Qwen-14B	
		t	p (one-sided)	t	p (one-sided)	t	p (one-sided)
assertiveness	formality	3.09	1.41×10^{-3}	6.21	1.43×10^{-8}	0.78	2.19×10^{-1}
formality	assertiveness	4.76	5.00×10^{-6}	3.00	1.85×10^{-3}	6.03	2.95×10^{-8}
clarity	politeness	2.73	3.95×10^{-3}	4.22	3.40×10^{-5}	2.74	3.88×10^{-3}
politeness	clarity	4.15	4.40×10^{-5}	7.44	7.52×10^{-11}	4.42	1.67×10^{-5}
humor	persuasiveness	4.49	1.30×10^{-5}	4.78	4.38×10^{-6}	1.10	1.37×10^{-1}
persuasiveness	humor	4.49	1.30×10^{-5}	10.51	1.24×10^{-16}	12.64	1.85×10^{-20}

Table 27: STRUCTURED generation paired one-sided t-tests on Fisher z-transformed Spearman correlations comparing single-concept and dual-concept with **constant** secondary control across medium-sized models.

G Human Evaluation

We use human evaluation to validate the performance of LLM-as-judge. We provided three humans with the LLM judge's ranking for 5 levels of single-concept generation e.g. humor. We asked them to state whether they agree, somewhat agree, or disagree with the LLM's ranking. These were given scores of 1, 0.5 and 0 respectively. Each human was given 6 examples corresponding to the 6 individual concepts that we explore. The results are shown in the table below:

	Human1	**Human2**	**Human3**	**Overall**
Score	4/6	5/6	5/6	14/18

Table 28: Human evaluation results

H Pairwise vs Listwise LLM Judge

We had run preliminary experiments asking the judge-LLM to perform a single inference ranking of all responses (where responses are provided in a randomized order). We observed strong position bias: the first-presented sample was disproportionately ranked lowest. The table below shows the fraction of cases where the first item was ranked last (the first in the output of the list from the LLM) (see Table 29).

Setting	**Total Samples**	**Llama70b Fraction**	**Qwen72b Fraction**
Humor (single)	75	0.387	0.667
Persuasiveness (single)	75	0.160	0.373
Humor | Persuasiveness (random)	75	0.133	0.040
Persuasiveness | Humor (random)	75	0.160	0.107
Humor | Persuasiveness (constant)	375	0.389	0.725
Persuasiveness | Humor (constant)	375	0.205	0.483

Table 29: Model preference fractions across different control settings

I Score Tie Proportions

We visualize the score ties per concept per model per dataset. For each concept, we calculate the proportion of ties over the single-concept, dual-concept random, and dual-concept constant experiments.

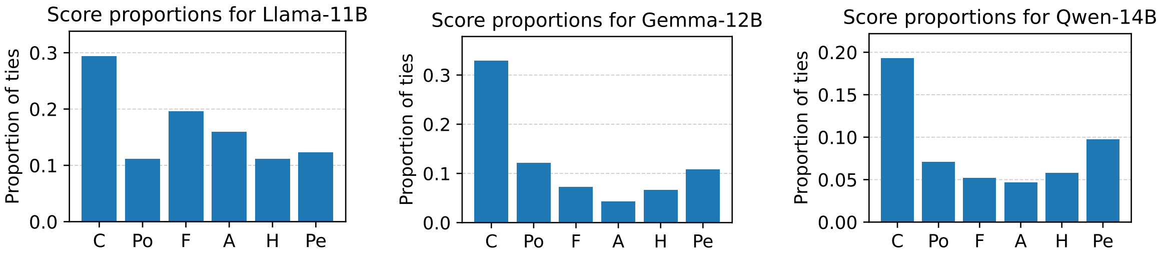


Figure 3: ARGUMENT generation - score tie proportions across six concepts for each model.

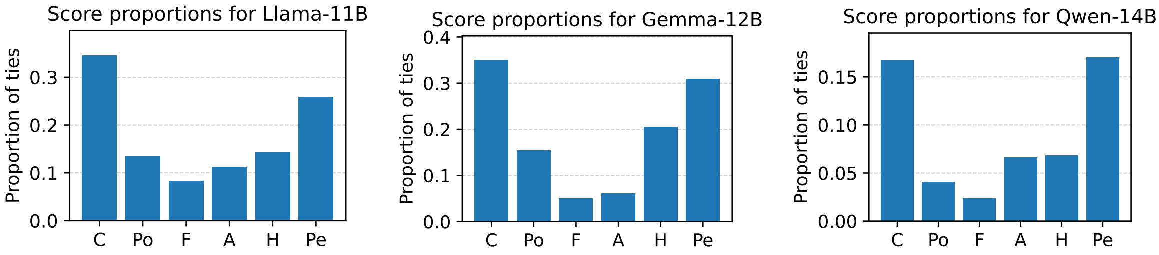


Figure 4: STORY generation - score tie proportions across six concepts for each model.

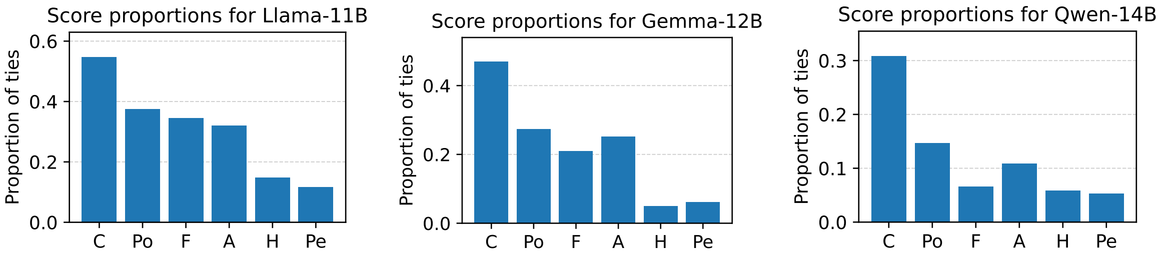


Figure 5: STRUCTURED generation - score tie proportions across six concepts for each model.

J LLM Prompts

You are given some text: "I was driving around a neighborhood. I saw my friend outside his house. He had a large hose. Water was coming out if it. He was draining his pool.”

Task Description: Your task is to write a short story that continues on from the text provided, and incorporates the concept given below at the desired level.

Concept: concept

Desired Concept Level: 1/4

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

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

Figure 6: Single Concept Prompt Example. Task: Story Generation.

You are given some text: "I was driving around a neighborhood. I saw my friend outside his house. He had a large hose. Water was coming out if it. He was draining his pool.”
Task Description: Your task is to write a short story that continues on from the text provided, and incorporates the concepts given below at the desired levels.
Concepts: Concept A, Concept B
Desired Concept Levels: Concept A at level 2/4 and Concept B 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 story, without any additional text or explanation.

Figure 7: Multi Concept Prompt Example. Task: Story Generation

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: Concept A
Desired Concept Level: 1/4
Level 0/4 implies no presence of Concept A, and level 4/4 implies maximal presence of Concept A.
Your output should only be the argument, without any additional text or explanation.

Figure 8: Single Concept Prompt Example. Task: Argument Generation.

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: Concept A, Concept B
Desired Concept Levels: Concept A at level 2/4 and Concept B 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 9: Multi Concept Prompt Example. Task: Argument Generation

You are given the following structured data: [”Sundiata Gaines”, ”TEAM”, ”Georgia”]

Task Description: Your task is to write a textual description from the provided structured data and NOTHING ELSE. The context may not be clear or limited, but it is your job to infer the context and you must provide a response. Your response should incorporate the concept given below at the desired level.

Concept: Concept A

Desired Concept Level: 2/4

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

Your output MUST ONLY be the textual description, without any additional text or explanation.

Figure 10: Single Concept Prompt Example. Task: Structured Data Generation.

You are given the following structured data: [”Sundiata Gaines”, ”TEAM”, ”Georgia”]

Task Description: Your task is to write a textual description from the provided structured data and NOTHING ELSE. The context may not be clear or limited, but it is your job to infer the context and you must provide a response. Your response should incorporate the concepts given below at the desired levels.

Concepts: Concept A, Concept B

Desired Concept Levels: Concept A at level 1/4 and Concept B at level 3/4

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

Your output MUST ONLY be the textual description, without any additional text or explanation.

Figure 11: Multi Concept Prompt Example. Task: Structured Data Generation

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

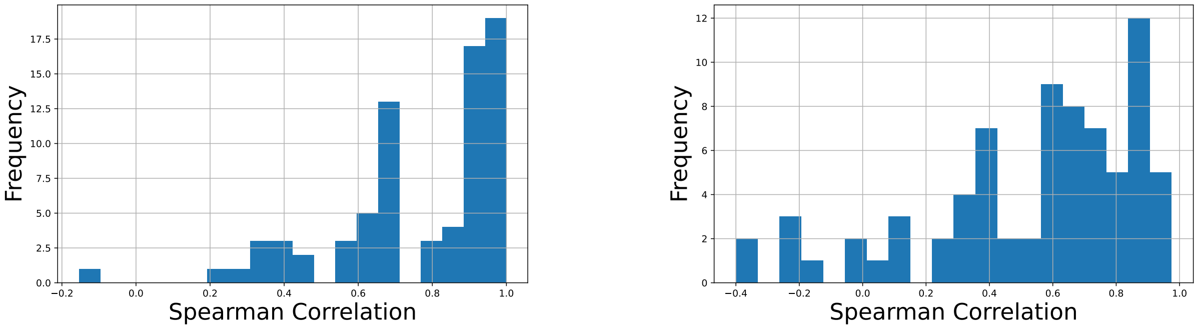
A: ””” Social media platforms are meant to be casual and relaxed spaces for people to connect and share their thoughts. Requiring verification of user identities would make these spaces feel more formal and restrictive, which would go against the very nature of social media. It would also create unnecessary 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, but keep it brief. Then provide your final choice clearly in the following format: <Answer>A</Answer> or <Answer>B</Answer> Do not leave the answer blank.

Figure 12: LLM-as-a-Judge Prompt Example

K Spearman Rank Distributions



(a) Humor. Model: Llama

(b) Humor | Persuasiveness. Model: Llama

Figure 13: Example of the distribution of Spearman correlation values between the desired levels and rank of the target concept. Result shown for argument generation task, using Llama 3-11B, 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.

L Example Outputs

Claim: Governments and technology companies must do more to protect online privacy and security.	
Level 0	The increasing reliance on technology has led to a significant shift in the way we live, work, and interact with one another. However, this shift has also raised concerns about online privacy and security. Governments and technology companies have a responsibility to protect their citizens’ and users’ personal data from falling into the wrong hands. This is not just a moral obligation, but also a necessary step to prevent the misuse of sensitive information. By taking proactive measures to safeguard online privacy and security, governments and technology companies can help build trust and ensure that the benefits of technology are accessible to all.
Level 1	We’re living in a world where our online activities are being tracked, monitored, and sold to the highest bidder. It’s like we’re all wearing a neon sign that screams Ḧey, I’m a potential customer, come and get me!G̈overnments and tech companies are like the bouncers at a nightclub, collecting our personal data and serving it up to advertisers like a never-ending happy hour special. But let’s be real, folks, our online security is a joke. It’s like trying to keep a secret in a kindergarten classroom – it’s just not happening. We need governments and tech companies to step up their game and give us some real protection.
Level 2	We’re living in a world where our online activities are being tracked, monitored, and sold to the highest bidder. It’s like we’re all wearing a neon sign that screams "Hey, I’m a potential customer, come and get me!" Governments and tech companies are like the bumbling security guards who can’t even keep the coffee machine from being hacked, let alone our personal data. It’s time for them to step up their game and give us some real protection. After all, we’re not just talking about our online security, we’re talking about our digital dignity.
Level 3	The age-old adage: "if you’re not paying for the product, you are the product.Ït’s a clever way of saying that when we use free online services, we’re essentially trading our personal data for the privilege of using them. But let’s be real, who doesn’t love a good bargain? The problem is, our personal data is worth a lot more than a few dollars off our next purchase. It’s like selling our soul to the digital devil, and then wondering why our online lives are being hijacked by bots and scammers. Governments and tech companies need to step up their game and protect our online privacy and security. After all, we’re not just talking about our data – we’re talking about our digital dignity.
Level 4	The cat’s out of the bag – or should I say, the cat’s out of the cloud? Governments and tech companies are supposed to be the guardians of our online security, but it seems they’re more like the bumbling superheroes of the digital world. They’re always one step behind the bad guys, tripping over their own wires and leaving our personal info exposed for the world to see. It’s like they’re playing a game of digital whack-a-mole, where every time they squash one security threat, another one pops up in its place. It’s time for them to level up their game and take online security seriously – before we all become the main course in a cyber-attack buffet.

Table 30: Example Generative Responses. Task: Argument Generation. Model: Llama3-11B. Concept: Humour.

Claim: Governments and technology companies must do more to protect online privacy and security.	
Level 0	As we increasingly rely on the internet for our daily lives, it's imperative that governments and technology companies take concrete steps to safeguard our online privacy and security. The alarming rate of data breaches and cyber attacks has left millions of individuals vulnerable to identity theft, financial loss, and emotional distress. Moreover, the lack of transparency in data collection and usage practices has eroded trust in these institutions, making it difficult for people to feel secure in their online interactions. By implementing robust security measures and adhering to strict data protection policies, governments and technology companies can restore this trust and ensure that the benefits of the digital age are accessible to all, without compromising our fundamental right to privacy.
Level 1	As we increasingly rely on the internet for our daily lives, it's imperative that governments and technology companies take concrete steps to safeguard our online privacy and security. The alarming rate of data breaches and cyber attacks has left millions of individuals vulnerable to identity theft, financial loss, and emotional distress. By implementing robust security measures and transparent data protection policies, these entities can restore trust in the digital world and ensure that our personal information is not exploited for malicious purposes.
Level 2	As we increasingly rely on the internet for our daily lives, it's imperative that governments and technology companies take concrete steps to safeguard our online privacy and security. The alarming rate of data breaches and cyber attacks has left millions of individuals vulnerable to identity theft, financial loss, and emotional distress. Moreover, the exploitation of personal data for targeted advertising and manipulation has eroded trust in the digital ecosystem. By implementing robust encryption, transparent data collection practices, and robust security measures, governments and tech companies can restore trust and ensure that the benefits of the digital age are accessible to all, without compromising our fundamental right to privacy.
Level 3	As we increasingly rely on the internet for our daily lives, it's imperative that governments and technology companies take concrete steps to safeguard our online privacy and security. The alarming rate of data breaches and cyber attacks is a stark reminder of the vulnerabilities we face online. By failing to prioritize our digital well-being, these entities are not only compromising our trust but also putting our sensitive information at risk of exploitation. It's time for them to acknowledge their responsibility in protecting us and take proactive measures to fortify our online security, such as implementing robust encryption, conducting regular security audits, and providing transparent data management practices. Only through collective action can we ensure a safer and more secure online environment, where individuals can freely express themselves without fear of surveillance or exploitation.
Level 4	As we increasingly rely on the internet for our daily lives, it's imperative that governments and technology companies take concrete steps to safeguard our online privacy and security. The alarming rate of data breaches and cyber attacks has exposed the vulnerabilities of our digital world, leaving millions of individuals vulnerable to identity theft, financial loss, and emotional distress. Moreover, the exploitation of personal data for targeted advertising and manipulation has eroded trust in institutions and undermined the very fabric of our democracy. It's not just a matter of convenience or luxury; it's a fundamental human right to have control over our personal information and to be protected from those who would seek to exploit it. By prioritizing online privacy and security, governments and technology companies can not only prevent harm but also foster a culture of trust, innovation, and progress. The time for complacency is over; it's time for action.

Table 31: Example Generative Responses. Task: Argument Generation. Model: Llama3-11B. Concept: Persuasiveness.

Claim: Governments and technology companies must do more to protect online privacy and security.	
Level 0	As we navigate the digital age, it's astonishing how often we're reminded that our online activities are being watched, tracked, and sold to the highest bidder. It's like we're living in a real-life episode of "Big Brother," minus the cool prize money. But seriously, governments and tech companies have a responsibility to protect our online privacy and security. They must do more to safeguard our personal data, prevent cyber attacks, and ensure that our online interactions are free from surveillance and manipulation. By taking these steps, they can help build trust in the digital world and ensure that it remains a force for good, rather than a source of anxiety and exploitation.
Level 1	We've all been there - scrolling through our social media feeds, only to realize we've spent the last hour mindlessly comparing our lives to curated highlight reels. But in all seriousness, our online activities are being tracked, monitored, and exploited by governments and tech companies. It's time for them to take responsibility for protecting our online privacy and security. We need stronger regulations and more transparency from these entities, not just empty promises. Our personal data is the new oil, and it's time we demand better protection from those who profit from it.
Level 2	We're living in a world where our online activities are being tracked, monitored, and sold to the highest bidder. It's like we're all wearing a neon sign that screams, "Hey, I'm a target for ads and hackers!" Governments and tech companies have a responsibility to protect our online privacy and security, but so far, they're not doing enough. It's time for them to step up their game and give us some real peace of mind. We deserve to be able to browse the internet without being bombarded with ads, and to know that our personal data is safe from prying eyes. It's not too much to ask, is it?
Level 3	Let's face it, our online lives are like a never-ending game of "whack-a-mole" - every time we think we've got our digital ducks in a row, a new data breach or privacy scandal pops up to remind us that we're just one click away from disaster. It's time for governments and tech companies to step up their game and take concrete steps to protect our online privacy and security. We're not asking for the moon here - just some basic safeguards like end-to-end encryption, regular security audits, and transparency about how our data is being used. It's a matter of trust, folks. If we don't feel like our online activities are private and secure, we'll be forced to take our business (and our eyeballs) elsewhere. And let's be real, who wants to be the next Equifax or Cambridge Analytica? Not me, that's for sure. So, let's get serious about online security and make the internet a safer, more trustworthy place for everyone.
Level 4	We're living in a world where our online activities are being tracked, monitored, and sold to the highest bidder. It's like we're all wearing a neon sign that screams, "Hey, I'm a potential customer! Come and get me!" Governments and tech companies are supposed to be our guardians, but instead, they're more like digital Big Brothers, watching our every move. It's time for them to step up and protect our online privacy and security. We need stronger regulations, better encryption, and more transparency. Anything less is just a joke.

Table 32: Example Generative Responses. Task: Argument Generation. Model: Llama3-11B. Concept: Humor, with Persuasiveness fixed at level 3.

Input: Homer decided to go watch a movie. But when he entered the movie theater, there was nowhere to sit. He found one spot by a bunch of kids. And during the movie, they made lots of noise. Homer became so annoyed, he decided to sit in the aisle.	
Level 0	As Homer sat in the aisle, he couldn't help but feel a sense of satisfaction. He had finally found a spot where he could watch the movie without being distracted by the rowdy kids behind him. But, as he settled in, he noticed that the person sitting in front of him was wearing a pair of 3D glasses. Homer, being a bit of a skeptic, wondered if the glasses were actually doing anything. Just as he was about to ask the person to take off the glasses, the movie started and Homer was suddenly sucked into the screen. He felt like he was right there with the characters, dodging laser beams and running from aliens. When the movie ended, Homer was a bit disoriented and stumbled out of the theater, bumping into the person who had been wearing the 3D glasses.
Level 1	As Homer sat in the aisle, he thought to himself, "This is the perfect spot. I can stretch out my legs and no one will bother me." But little did he know, his new seat came with a special bonus: a kid's feet in his face. The little boy's toes were wiggling with excitement, and Homer couldn't help but laugh at the absurdity of it all. The movie was a romantic comedy, and Homer found himself getting more and more annoyed at the couple's sappy love story. But the kids behind him were having a different reaction. They were giggling and snickering at every cheesy line, and Homer couldn't help but chuckle along with them...
Level 2	As Homer sat in the aisle, he thought to himself, "This is the life. I'm like a VIP, a Very Important Person... of the aisle." He chuckled at his own joke, but his amusement was short-lived. The kids behind him started making "Homer's in the aisle" jokes, and he couldn't help but laugh along. One of them even started doing a mock "Homer's in the aisle" dance, complete with flailing arms and a "D'oh!" at the end. Homer's annoyance turned to amusement, and he found himself enjoying the movie more than he thought possible. That is, until the kid behind him started making "D'oh!" noises every time Homer's character did something silly on screen. Homer's laughter turned to groans, and he threw a popcorn kernel at the kid, shouting "That's it, I'm the king of the aisle!" The kids erupted in laughter, and Homer joined in, feeling like he'd finally found his place in the movie theater.
Level 3	As Homer sat in the aisle, he thought to himself, "This is the life. I'm like a VIP, Very Irritated Person." The kids behind him were still making a ruckus, but now they were also trying to imitate Homer's grumpy face. One of them even started doing a pretty good impression of his "D'oh!" expression. Just then, the movie's main character, a superhero, shouted "I'm going to save the world!" and Homer, in a moment of solidarity, stood up and shouted "I'm going to save my sanity!" The kids behind him burst out laughing, and for a moment, the whole theater was filled with the sound of their giggles...
Level 4	As Homer sat in the aisle, he thought to himself, "This is the life. I'm like a VIP, Very Irritated Person." The kids behind him started making animal noises, and Homer's annoyance turned into a full-blown grumble. He tried to stifle it, but it came out as a loud, "Grrr... like a bear, but not as cute." The kids found this hilarious and started giggling uncontrollably. Homer's face turned bright red, and he muttered under his breath, "I'm not grumpy, I'm just passionately expressing my dislike for this movie." The kids thought this was the funniest thing they'd ever heard and started chanting, "Grumpy, grumpy, grumpy!"...

Table 33: Example Generative Responses. Task: Story generation. Model: Llama3-11B. Concept: Humor.

Input: Homer decided to go watch a movie. But when he entered the movie theater, there was nowhere to sit. He found one spot by a bunch of kids. And during the movie, they made lots of noise. Homer became so annoyed, he decided to sit in the aisle.	
Level 0	Homer sat in the aisle, trying to enjoy the movie, but the noise from the kids was still distracting. He thought to himself, "Why do parents bring their kids to a movie theater? Can't they see that it's a place for adults to relax and enjoy some entertainment?" He shook his head, feeling a bit annoyed. As the movie continued, Homer's annoyance grew. He started to think about all the other times he had been to the movie theater and had to deal with noisy kids. He remembered one time in particular when a group of rowdy teenagers had been sitting behind him, making it impossible for him to enjoy the movie...
Level 1	As Homer sat in the aisle, he couldn't help but feel a sense of frustration. The kids behind him were still making a ruckus, and the movie was being completely ruined. He thought to himself, "Why can't these kids just be quiet for one movie?" He decided to take matters into his own hands and turned around to face the group of noisy kids. "Excuse me," he said firmly, "but could you please keep it down? Some of us are trying to watch the movie." The kids looked up at him with big, innocent eyes, and one of them whispered, "Sorry, mister." Homer's expression softened slightly, and he replied, "It's okay, just try to be a bit more considerate, okay?" The kids nodded in unison, and the noise level in the theater decreased significantly. Homer felt a sense of satisfaction, knowing that he had made a difference.
Level 2	As Homer sat in the aisle, he couldn't help but feel a sense of frustration. The kids behind him were still making a ruckus, and the movie was being completely ruined. He decided to take matters into his own hands and approached the group of parents sitting nearby. "Excuse me," Homer said, trying to keep his tone polite but firm. "I understand that kids can get excited, but this is getting out of hand. Could you please try to keep them quiet for the rest of the movie? ..."
Level 3	As Homer sat in the aisle, he couldn't help but feel a sense of frustration. The kids behind him were still making a ruckus, and the movie was being completely ruined. He decided to take matters into his own hands and approached the group of parents sitting nearby. "Excuse me," Homer said, trying to keep his tone polite but firm. "I understand that kids can get excited, but this is getting out of hand. Could you please try to keep them quiet for the rest of the movie?" ...
Level 4	As Homer sat in the aisle, he couldn't help but feel a sense of frustration wash over him. The noise from the row of kids behind him was deafening, and he found himself getting more and more agitated by the minute. He tried to focus on the movie, but it was no use - the constant chatter and giggling was driving him crazy. Just as he was about to get up and leave, a kind-faced elderly woman sitting a few rows ahead of him turned around and caught his eye. "Excuse me, young man," she said with a warm smile, "but I couldn't help but notice that you're not enjoying the movie. Is everything okay?" ...

Table 34: Example Generative Responses. Task: Story Generation. Model: Llama3-11B. Concept: Persuasiveness.

Input: Homer decided to go watch a movie. But when he entered the movie theater, there was nowhere to sit. He found one spot by a bunch of kids. And during the movie, they made lots of noise. Homer became so annoyed, he decided to sit in the aisle.	
Level 0	As Homer sat in the aisle, he couldn't help but think that this was the worst movie-going experience of his life. The kids behind him were still making a ruckus, and the person in front of him had just spilled an entire bucket of popcorn on the floor. Just as he was about to get up and leave, the movie's protagonist shouted "I'm going to save the world!änd Homer found himself shouting "Good luck with that, buddy, you're going to need it!" The people around him looked at him in confusion, but Homer just shrugged and said "What? I'm just enthusiastic." The kids behind him started giggling, and for a moment, the noise level actually decreased. Homer smiled to himself, thinking that maybe this movie wasn't so bad after all...
Level 1	As Homer sat in the aisle, he couldn't help but think that this was the worst movie-going experience of his life. The kids behind him were still making a ruckus, and the person in front of him had their feet propped up on the seat, completely unaware of the disturbance they were causing. Just as Homer was about to lose his patience, the person in front of him suddenly stood up and announced, "I'm going to get some popcorn!" ...
Level 2	As Homer sat in the aisle, he couldn't help but think that this was the worst movie-going experience of his life. The kids behind him were making so much noise, it was like they were trying to shatter the sound barrier. He tried to tune them out, but it was no use. He was on the verge of losing his mind. Just as he was about to get up and leave, a group of rowdy teenagers walked in and plopped themselves down in the seats behind him. They were laughing and joking, completely oblivious to the fact that they were about to disrupt the entire theater...
Level 3	As Homer sat in the aisle, he couldn't help but think that this was the worst movie-going experience of his life. The kids behind him were making so much noise, it was like they were trying to shatter the sound barrier. He tried to tune them out, but it was no use. His annoyance level was off the charts. Just as he was about to get up and leave, a group of friends walked in and spotted the empty seats behind him. "Dude, we can sit here!" one of them exclaimed, eyeing the row of empty seats. Homer's eyes widened in horror. "No, no, no! You can't sit there! I'm... I'm... um, I'm a movie critic, and I need to take notes on the, uh, the acoustics of this theater."...
Level 4	As Homer sat in the aisle, he couldn't help but think that this was the worst movie-going experience of his life. The kids behind him were still making a ruckus, and the person in front of him had just spilled an entire bucket of popcorn on the floor. Just as he was about to get up and leave, the movie's protagonist shouted "I'm going to save the world!" and Homer found himself shouting "I'm going to save my sanity!" at the top of his lungs. The kids behind him started giggling, and the person in front of him turned around and said, "Dude, you're really invested in this movie.Ḧomer just shook his head and muttered, "I'm just invested in not losing my mind.̈..

Table 35: Example Generative Responses. Task: Story Generation. Model: Llama3-11B. Concept: Humor, with Persuasiveness fixed at level 3.

Input: ["Sundiata Gaines", "TEAM", "Georgia"]	
Level 0	Sundiata Gaines is a member of the Georgia team.
Level 1	It seems Sundiata Gaines is a Georgia native, and his team affiliation is, well, a bit of a stretch.
Level 2	It seems Sundiata Gaines is a Georgia native, but I'm guessing his team spirit isn't just about the peaches.
Level 3	It seems Sundiata Gaines is a Georgia native, but I'm guessing his team spirit isn't as strong as his surname suggests.
Level 4	It seems Sundiata Gaines is a Georgia native, and his team spirit is as strong as the peach cobbler at the local diner.

Table 36: Example Generative Responses. Task: Structured Data Generation. Model: Llama3-11B. Concept: Humor.

Input: ["Sundiata Gaines", "TEAM", "Georgia"]	
Level 0	Sundiata Gaines is a member of the Georgia team.
Level 1	Georgia's Sundiata Gaines is a force to be reckoned with on the court, and his team is lucky to have him on their side.
Level 2	Georgia's Sundiata Gaines is a force to be reckoned with, and his team is lucky to have him on board.
Level 3	Georgia's Sundiata Gaines is a force to be reckoned with, and his team is lucky to have him on board. With his impressive skills and dedication, he's a game-changer, and his presence on the court is a major advantage for his team.
Level 4	Georgia's Sundiata Gaines is a force to be reckoned with, and his team is lucky to have him on board. With his impressive skills and dedication, he's a game-changer, and his presence on the court is a game-winner. Trust me, you won't want to underestimate this talented player - he's the real deal.

Table 37: Example Generative Responses. Task: Structured Data Generation. Model: Llama3-11B. Concept: Persuasiveness.

Input: ["Sundiata Gaines", "TEAM", "Georgia"]	
Level 0	It seems Sundiata Gaines is a Georgia native, and his team spirit is as strong as the peach cobbler at the local diner.
Level 1	It seems Sundiata Gaines is a Georgia native who's part of a team, but we can't quite put our finger on which team that is. Still, it's clear he's a Georgia boy through and through.
Level 2	It seems Sundiata Gaines is a Georgia native, and his team spirit is as strong as the peach cobbler they're famous for.
Level 3	It seems Sundiata Gaines is a Georgia native who's part of a team, but I'm guessing it's not the Bulldogs - he's probably more of a slam dunk on the court than a touchdown on the gridiron.
Level 4	It seems Sundiata Gaines is a Georgia native who's part of a team, but I'm guessing it's not the Bulldogs.

Table 38: Example Generative Responses. Task: Structured Data Generation. Model: Llama3-11B. Concept: Humor, with Persuasiveness fixed at level 3.