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),

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

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