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

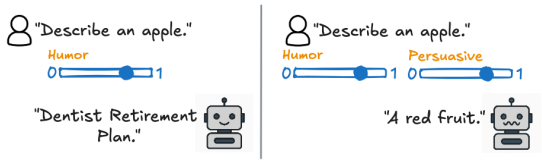

Figure 1: Illustrative example: an LLM can perform single-concept control, but the explicit presence of a second concept at the input can compromise the ability of the model to control the former concept in its response.

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).

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