ity. 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.