An Information-Theoretic Study of RLHF-Induced Uniformity in Large Language Model Outputs: Explanatino of Revisions

We've taken a substantial revision of our paper in order to sharpen our theoretical motivations, methodological transparency, and present central findings more clearly to better reflect the impact of our findings, as well as address concerns from reviewers. We've adjusted our motivations and theoretical framing in the abstract and introduction (section 1), clarifying our argument that our use of the Uniform Information Density hypothesis is used not as a uniquely human property but more of an observable metric of an audience-design behavior (lines 075-087, 136-142).

We've also elaborated and provided details on setup for reproducibility and methodological transparency of our experiments, as well as stronger justification for the use of GPT-2/LMs to estimate human cognitive load by more clearly referencing prior work (section 3.3, 7 - UID Calculation, and Appendix D). We've provided additional data that we previously neglected to include, such as additional UID metric analysis using Qwen (Appendix D), and perplexity scores Appendix C). We've also adjusted the discussion of our results to better clarify that our findings observe the contours of reduction in uniformity, not just an observation of a reduction in uniformity due to the autoregressive nature of LMs.