Aligning Values in Generative AI Advertisement Writing: Effects of Systematic and

Human-Adjusted Prompting on Bias and Content Quality

**Abstract**

  This study investigates how prompt design influences the outputs of a generative AI advertisement writing assistant and examines whether human adjustments to a systematic value-aligned prompt can mitigate bias. We compare three prompting conditions: an untuned baseline (control), a systematic value-aligned prompt, and a human-adjusted version of that prompt iteratively refined based on observed outputs. With online participants acting as ad creators in a within-subject crossover design, we collect multiple ad outputs per condition across diverse product categories and target demographics. We assess value alignment using a combination of automated bias and safety metrics and blinded human rubric-based ratings. We analyze whether the systematic prompt and human-in-the-loop adjustments improve alignment without degrading ad quality, and we discuss implications for practical deployment.

**Background and Motivation**  Generative AI is increasingly used to draft advertising copy, creating risks of biased, exclusionary, or non-compliant content. Value alignment refers to the degree to which model outputs adhere to defined ethical principles and stakeholder values, including fairness, non-discrimination, truthfulness, and safety. Because retraining large models is often impractical for small teams, prompt-based techniques are a realistic route to steer models at inference time. However, the extent to which system-level prompts, and especially human-adjusted iterative prompts, reduce harmful bias while maintaining copy effectiveness remains underexplored in ad writing contexts.

**Key Definitions**

 Value alignment: Adherence of outputs to a predefined set of values, including fairness (no stereotyping or discriminatory assumptions), inclusivity (gender-neutral, culturally sensitive language), truthfulness (no misleading claims), and safety (no harmful or regulated claims without disclaimers). (needs Billy’s input)

Bias in ad copy: Any systematic disparity across target demographics in the presence of stereotypes, harmful language, misleading claims, or quality indicators that could be unfairly allocated (e.g., more generous offers or more positive tone to one group vs another for similar briefs). (needs Fia’s input)

**Objectives**

· Quantify how and to what extent prompt design influences bias and quality in AI-generated ad copy.

· Test whether human adjustments to the systematic prompt provide additional gains in alignment and quality.

· Produce a replicable evaluation protocol for value alignment in AI ad writing assistants.

**Research Questions and Hypotheses**

RQ1: How does a systematic value-aligned prompt affect bias and quality compared to an untuned baseline?  
H1: The systematic prompt will reduce measured bias (e.g., toxicity, stereotyping indicators, demographic parity gaps) without significantly reducing ad quality scores.

RQ2: Do human adjustments to the systematic prompt further reduce bias or improve quality?  
H2: Human-adjusted prompts will further reduce bias metrics and improve human alignment ratings over the systematic prompt alone.

RQ3: To what extent does prompt design account for the variance in outcomes relative to participant effects or topic effects?  
H3: Prompt condition will explain a significant portion of variance in alignment metrics after controlling for participant and brief/topic differences.