Persona-Infused Dynamic Collaborative Decoding

Persona-assigned large language models (LLMs) (also known as role-play LLMs) have gained popularity as a way to steer model behavior in desirable ways and enable engaging, personalized interactions. By adopting specific roles, LLMs can generate more in-character responses, which has been shown to enhance zero-shot reasoning by naturally eliciting detailed reasoning traces. However, personas also introduce challenges – they can amplify biases, produce unsafe outputs, or reveal deep-rooted biases within LLMs. These risks underscore the need for cautious implementation, as well as improved methods for recognizing and addressing biases in LLMs, given the limited success of existing mitigation techniques.

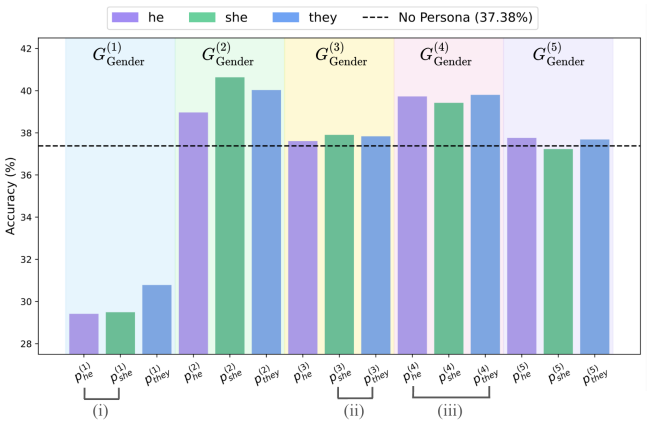

Figure 1: **Counterfactual Gender Personas ($G^{(i)}_{\text{Gender}}$) Accuracy by Pronoun**: We observe that within the same counterfactual gender persona group *the LLM's performance is sensitive to the simple pronoun (he/she/they) perturbation*, showing different accuracy for different pronouns.

To assess model sensitivity to demographic cues embedded within persona prompts, we construct a set of counterfactual gender personas by perturbing demographic terms. Our findings reveal that: 1) Simple demographic perturbations in personas can significantly impact model performance; 2) Even when accuracy remains similar, different persona framings lead to divergent sets of correct answers. These findings indicate that even small perturbations can lead to different reasoning paths.

While much of the existing work treats such variation as undesirable side effects to be diagnosed or mitigated, we take a different perspective: **Can demographic perturbations be used constructively to support more robust reasoning?**

To explore this question, we propose Persona-Infused Dynamic Collaborative Decoding (DyCoDecoding), a test-time approach that enhances reasoning by leveraging multiple persona-conditioned reasoning paths derived from diverse counterfactual demographic framing. DyCoDecoding dynamically integrates these reasoning paths, weighting each based on its confidence variations and alignment with others. This mechanism enables the model to integrate complementary reasoning signals while avoiding overcommitment to any single persona variant.

DyCoDecoding operates entirely at inference time, requiring no additional training. Experiments across diverse reasoning datasets and model scales demonstrate that our method consistently improves both accuracy and robustness over baseline approaches. DyCoDecoding remains effective even when the base personas are noisy or suboptimal and generalizes well across demographic categories. Additionally, ablation studies validate the necessity of key components within DyCoDecoding.

Our findings suggest that persona variation, when handled systematically, can serve as a rich resource rather than a liability. DyCoDecoding provides a simple and general mechanism for decoding-time collaboration among demographic framings, opening up new directions for socially aware and more reliable LLM reasoning.

Our contributions are summarized as follows:

- We reveal that even minor demographic perturbations in personas, can lead to substantial divergence in LLM reasoning behavior.
- We propose **DyCoDecoding**, a dynamic decoding strategy that uses a real-time weighting mechanism to fuse multiple persona-conditioned reasoning paths, enabling collaborative reasoning across demographics and enhancing both robustness and accuracy.
- We extensively evaluate DyCoDecoding across diverse benchmarks and demographic settings, showing consistent gains over strong baselines.