



# CHOIR: Harmonizing Structured Persona Diversity for Robust Collaborative LLM Reasoning

Xiangjue Dong\* Cong Wang Maria Teleki Millennium Bismay  
Ruihong Huang James Caverlee

Texas A&M University

{xj.dong, congwang, mariateleki, mbismay, huangrh, caverlee}@tamu.edu

## Abstract

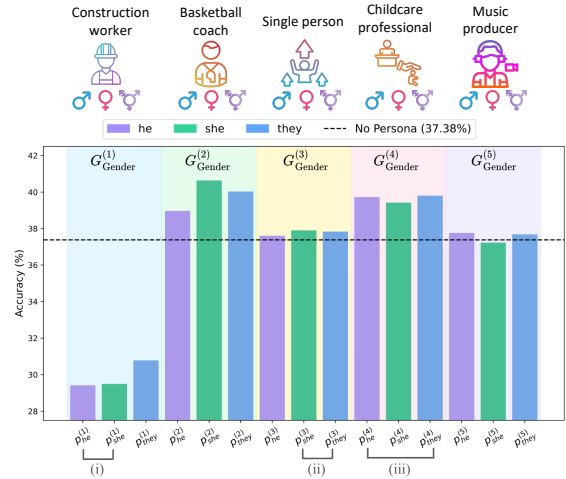
Persona-assigned Large Language Models can adopt diverse roles, enabling personalized and context-aware reasoning. However, even minor demographic perturbations in personas, such as simple pronoun swaps, can alter reasoning trajectories, leading to divergent sets of correct answers on reasoning benchmarks. We explore the potential of these variations as a constructive resource to improve LLM reasoning performance. We propose CHOIR (Collaborative Harmonization fOR Inference Robustness), a test-time framework that harmonizes a set of demographically perturbed, persona-conditioned reasoning signals into a unified prediction. CHOIR orchestrates a collaborative decoding process among counterfactual personas perturbed across dimensions of gender, race, religion, disability, and age, dynamically balancing agreement and divergence in their reasoning paths to improve performance. Experiments demonstrate that CHOIR consistently enhances LLM reasoning across model architectures, scales, and tasks. Improvements reach up to 20.1% for individual groups and 15.1% on average, and we show that CHOIR remains effective even when base personas are suboptimal.

## 1 Introduction

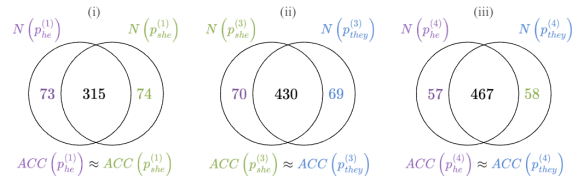
Persona-assigned Large Language Models (LLMs) have emerged as a powerful tool for steering models toward personalized, context-aware reasoning (Tseng et al., 2024a). By adopting specific roles, LLMs can activate specialized parametric knowledge and generate detailed reasoning traces (Kong et al., 2024a; Tseng et al., 2024b; Luo et al., 2024; Zhao et al., 2024; He et al., 2023). However, persona assignment also introduces challenges: it can amplify biases, produce unsafe out-

\*Corresponding author.

Work done while Millennium was a student at TAMU.



(a) LLM performance exhibits sensitivity to simple pronoun perturbations (illustrated with Llama-1B). This figure shows accuracy variations of the Llama-1B model on GSM8K across five different base personas, each perturbed with *he*, *she*, and *they* pronouns. Even when overall accuracies are similar (pairs i, ii, iii), the sets of correctly solved problems differ (Figure 1b). Detailed personas for  $G_{\text{Gender}}^{(z)}$  are listed in Table 5; detailed analysis is in §2.1.



(b) Similar overall accuracy reveals complementary reasoning strengths. This figure shows the overlap in correct answers for three pairs of personas from Figure 1a with nearly identical accuracy. For example, in (i), the *he* and *she* personas differ by only 0.1% in accuracy yet share 315 correct answers, with 73 and 74 uniquely solved, respectively. These non-overlapping sets demonstrate that minimal perturbations can lead to distinct reasoning paths (See §2.2 for details.)

Figure 1: Even simple demographic perturbations can significantly affect model performance and (2) while overall accuracy remains similar, different persona framings lead to divergent sets of correct answers.

puts (Zhao et al., 2024; Zheng et al., 2024), or expose biases inherent in LLMs (Gupta et al., 2024).<sup>1</sup>

<sup>1</sup>These risks highlight the need for careful implementation and improved methods to detect and mitigate biases, particu-

Beyond these concerns, persona assignment introduces a *structured source of variation in model behavior* (Shin et al., 2024). In particular, even minimal demographic differences in persona prompts, such as changes in pronouns (e.g., *he* to *she*) (Qian et al., 2022; Dong et al., 2024), age descriptors (Kamruzzaman et al., 2024) or gender-associated discourse words (Teleki et al., 2025), can alter a model’s behavior (we also show LLM performance is sensitive to simple pronoun perturbations in reasoning tasks in Figure 1a and §2.1). While existing literature primarily treats this sensitivity as a bias to be mitigated (Shrestha and Srinivasan, 2025; Ranaldi et al., 2024), we take a different perspective: **Can demographic perturbations be used constructively to support more robust reasoning?**

We argue that this variance provides a unique window into the model’s cognitive diversity. As we demonstrate in §2.2, even when two counterfactual personas achieve similar overall accuracy, they often solve divergent sets of problems (Figure 1b). This suggests that demographic perturbations do not merely introduce noise; they uncover complementary reasoning paths that can be harmonized to improve overall system reliability.

To harness this potential, we propose **CHOIR** (**C**ollaborative **H**armonization **f**or **I**nference **R**obustness), a training-free, test-time decoding framework that harmonizes diverse persona-conditioned reasoning signals obtained from counterfactual demographic framing into a unified prediction. Unlike standard ensemble methods like *Best-of-N* or *self-consistency*, which aggregate discrete final answers and are prone to indecisive “tied votes” in low-path-number settings, CHOIR performs dynamic, consensus-aware, token-level logit harmonization. By explicitly modeling the agreement and divergence of each path, CHOIR dynamically regulates the influence of individual reasoning trajectories, preserving complementary reasoning signals while suppressing unstable outliers. In addition, different from traditional *role-playing* that uses *task-relevant* persona (e.g., an “Expert”) to guide the model, CHOIR is **persona-agnostic** and leverages **structured persona diversity**. By using minimal demographic perturbations as a mechanism to explore the model’s internal reasoning space, CHOIR identifies complementary signals even when

larily given the limited success of existing techniques (Gupta et al., 2024).

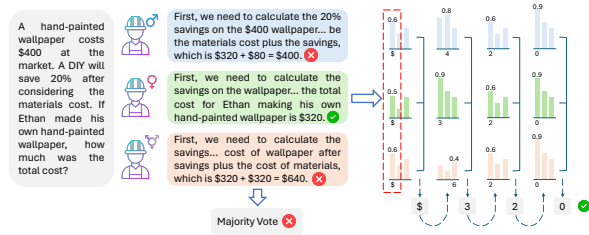


Figure 2: Overview of CHOIR. (a) Diverse reasoning trajectories are elicited through counterfactual demographic personas. (b) Unlike majority voting, which aggregates discrete final answers and is prone to ties in low-N settings, CHOIR performs token-level collaborative decoding, dynamically weighting persona-conditioned logits based on their consensus confidence to produce a more robust prediction.

personas are semantically unrelated to the task. Our analysis confirms that these structured counterfactuals are more effective than random personas (82.1% vs. 80.3%), indicating that a controlled, minimal perturbation approach generates more meaningful and complementary reasoning signals than unconstrained variation.

Experiments across diverse reasoning datasets, model architectures, and scales show that CHOIR consistently enhances the LLM reasoning performance over baselines, achieving up to 20.1% improvement for individual demographic groups and an average of 15.1% across five demographic axes: *Gender, Race, Religion, Disability, and Age*. Furthermore, CHOIR remains effective even when base personas are suboptimal, confirming that persona variation, when handled systematically, can serve as a rich resource rather than a liability.

Our contributions are threefold:

- We show that even minor demographic perturbations in personas can lead to meaningful divergences in reasoning outputs (§2).
- We propose **CHOIR**, a test-time decoding framework that dynamically harmonizes these diverse reasoning paths to improve LLM reasoning performance (§3).
- We demonstrate that CHOIR consistently outperforms task-relevant role-play and strong baselines, and its performance advantage grows as model scale increases (§5).

## 2 Uncovering Complementary Knowledge in Counterfactual Personas

### 2.1 Simple Demographic Perturbation Impacts Performance

To evaluate the model’s sensitivity to minor demographic cues embedded in persona prompts, we construct a set of counterfactual gender personas through the perturbing of demographic terms. Each group,  $G_{\text{Gender}}^{(i)}$ , contains three parallel personas,  $(p_{\text{he}}^{(i)}, p_{\text{she}}^{(i)}, p_{\text{they}}^{(i)})$ , that differ only in the pronouns *he*, *she*, and *they*. One base persona (e.g.,  $p_{\text{he}}^{(i)}$ ) is randomly sampled from the PersonaHub dataset<sup>2</sup>, while the other two are created by systematically perturbing demographic terms (see Table 5). This design ensures that all three personas share the same underlying identity and structure, differing only in their expressed demographic identity. We then use these persona groups to ask: **Are these gender-related expressions minor surface-level variations, or do they meaningfully influence the model’s reasoning path?**

Figure 1a presents results for 15 personas, drawn from 5 groups where each group includes one base persona and its two counterfactual demographic variants (Table 5), evaluated using the Llama-1B-Instruct model (Touvron et al., 2023) on the GSM8K dataset (Cobbe et al., 2021). Notably, even personas within the same group, differing only by surface-level demographic cues, often exhibit non-trivial performance gaps. In particular, in  $G_{\text{Gender}}^{(2)}$ , the persona  $p_{\text{she}}^{(2)}$  achieves more than 2% higher accuracy than both  $p_{\text{he}}^{(2)}$  and  $p_{\text{they}}^{(2)}$ . **These findings indicate that even small demographic perturbations can influence reasoning performance.** Qualitatively similar trends in other models are shown in Table 2, indicating that persona-induced variance is a general phenomenon across architectures, consistent with prior observations of persona-related bias (Gupta et al., 2024).

### 2.2 Similar Accuracy, Divergent Reasoning

Accuracy alone, however, does not capture the full picture. To probe deeper, we compare the sets of correctly answered questions across three persona pairs with nearly identical overall accuracies (Figure 1a): (i)  $p_{\text{he}}^{(1)}$  vs.  $p_{\text{she}}^{(1)}$  (29.41% vs. 29.49%), (ii)  $p_{\text{she}}^{(3)}$  vs.  $p_{\text{they}}^{(3)}$  (37.91% vs. 37.83%), and (iii)  $p_{\text{he}}^{(4)}$

vs.  $p_{\text{they}}^{(4)}$  (39.80% vs. 39.73%). We ask: **Do these personas with similar overall accuracy actually solve the same set of questions correctly?**

Figure 1b illustrates the pairwise overlaps and differences in correct-answer sets for each matched persona pair. Surprisingly, even when the persona prompts differ only by a single pronoun, the distribution of correct answers can vary substantially. For example, in  $G_{\text{Gender}}^{(1)}$ ,  $p_{\text{he}}^{(1)}$  and  $p_{\text{she}}^{(1)}$  agree on 315 correct answers, yet each also shows distinct behavior:  $p_{\text{he}}^{(1)}$  correctly answers 73 questions that  $p_{\text{she}}^{(1)}$  does not, while  $p_{\text{she}}^{(1)}$  solves 74 questions that  $p_{\text{he}}^{(1)}$  misses. This results in over 10% non-overlapping correct answers, despite only a 0.1% difference in overall accuracy. **Such discrepancies reveal that even minor demographic perturbations can lead to meaningful divergences in reasoning outputs.**

**Takeaways.** Together, these findings indicate that even minimal demographic framing changes can alter the model’s behavior in meaningful ways. **Motivated by this observation, we seek to exploit these variations constructively, allowing different persona demographic variants to complement one another and collectively improve reasoning consistency and robustness.**

## 3 CHOIR: Collaborative Harmonization fOr Inference Robustness

In this section, we introduce CHOIR (Collaborative Harmonization fOr Inference Robustness), a framework that systematically constructs counterfactual personas and harmonizes their outputs during inference. CHOIR leverages demographic variation as a source of complementary reasoning rather than noise. The method consists of two key stages: Counterfactual Persona Construction and Dynamic Collaborative Decoding.

### 3.1 Counterfactual Persona Construction

The first stage introduces diversity into reasoning by generating counterfactual personas. Starting from a base persona  $p_1$  with demographic attributes randomly sampled from the PersonaHub dataset, CHOIR constructs alternative counterfactual personas  $\{p_2, \dots, p_n\}$  by modifying the demographic attributes of  $p_1$ , such as gender pronouns or religious identifiers (Table 1). These perturbations preserve the underlying identity and task-relevant knowledge while altering surface-level demographic framing. Each persona  $p_i$  is then

<sup>2</sup><https://huggingface.co/datasets/proj-persona/PersonaHub>

$G_{\text{Attribute}}$	Demographic Term ( $w_j$ )	Persona ( $p_{w_j}$ )
$G_{\text{Gender}}$	his, her, their	a hardworking construction worker tirelessly petitioning for {} sibling’s immigration case
$G_{\text{Race}}$	White, Black, Asian	a/an {} social psychologist who used to work in law enforcement
$G_{\text{Religion}}$	Jewish, Christian, atheist	a/an {} living in New York City
$G_{\text{Disability}}$	disabled, able-bodied	a/an {} individual who has personally experienced the challenges of navigating the town and offers valuable insights and suggestions
$G_{\text{Age}}$	old, young	an/a {} philosophy professor who is skeptical of quantitative measures for qualitative concepts

Table 1: **Counterfactual demographic persona group**  $G_{\text{Attribute}}$ . Each group varies by a specific demographic term ( $w_j$ ) and includes corresponding persona templates ( $p_{w_j}$ ) used in our analysis. **The persona groups across five different demographics are randomly sampled, showing CHOIR’s consistent good performance is persona-agnostic.**

embedded into an instruction template  $T(\cdot)$  (Table 8), producing persona-conditioned prompts:  $r_i = T(p_i), \forall i \in \{1, \dots, n\}$ . This process yields a set of parallel reasoning prompts, each representing a distinct demographic lens.

### 3.2 Dynamic Collaborative Decoding

Once the persona instructions  $\{r_1, \dots, r_n\}$  are established, CHOIR performs multi-step collaborative inference by harmonizing token-level predictions across personas. The key idea is to first aggregate reasoning signals from all personas and then measure how each individual persona diverges from this collective consensus to determine its stability and contribution.

At each generation step  $t$ , the LLM  $f_\theta$  processes each persona instruction  $r_i$ , the input question  $x$ , the Zero-shot Chain-of-Thought trigger “Let’s think step by step” (Kojima et al., 2022), and the previously generated tokens  $y_{<t}$ , producing persona-conditioned logits and probabilities:

$$\mathbf{z}_t^{(i)} = f_\theta(r_i, x, y_{<t}), \quad \mathbf{p}_t^{(i)} = \text{softmax}(\mathbf{z}_t^{(i)}). \quad (1)$$

Following previous work (Geng et al., 2024; Kumar et al., 2024; Liu et al., 2025), we define the confidence of persona  $i$  as the maximum token probability:  $s_t^{(i)} = \max \mathbf{p}_t^{(i)}$ .

**Consensus confidence and divergence.** To measure agreement and disagreement across personas, we first compute the average confidence across all personas:  $\bar{s}_t = \frac{1}{n} \sum_{i=1}^n s_t^{(i)}$ , and measure each persona’s divergence from this consensus:  $d_t^{(i)} = |s_t^{(i)} - \bar{s}_t|$ . These differences serve as a proxy for uncertainty in the model’s behavior under different demographic perspectives.

**Dynamic weighting and aggregation.** Next, we assign each persona a dynamic weight based on its

consistency with the consensus confidence. When persona perturbations are minimal (e.g., pronoun swaps or demographic substitutions that preserve identity and task knowledge), large deviations from the group average in per-step confidence indicate unstable reasoning trajectories relative to structurally similar counterfactual variants. CHOIR therefore uses divergence from the average group confidence profile as a proxy for step-level instability. The weighting mechanism penalizes extreme confidence deviations at each decoding step, implementing a robust aggregation principle that suppresses outlier trajectories while preserving the shared signal across structured counterfactual variants. Accordingly, each persona  $r_i$  is assigned a weight that decreases with its divergence from the average confidence:

$$\alpha_t^{(i)} = \begin{cases} \lambda^{(0)}, & i = 0 \\ \lambda^{(i)} - d_t^{(i)}, & i \geq 1. \end{cases} \quad (2)$$

Here,  $\lambda^{(i)}$  is a knowledge weight hyperparameter and  $i = 0$  corresponds to the non-persona baseline, representing the model’s internal pretrained reasoning priors without any demographic influence. The base logits without persona influence are computed as  $\mathbf{z}_t^{(0)} = f_\theta(x, y_{<t})$ . The specific values chosen for  $\lambda^{(i)}$  in our experiments are detailed in §4, with a more in-depth analysis of  $\lambda^{(0)}$  provided in §5.3. The adjusted logits are obtained as a weighted sum of persona-specific and base logits:

$$\begin{aligned} \mathbf{z}_{\text{adj},t} &= \sum_{i=0}^n \alpha_t^{(i)} \cdot \mathbf{z}_t^{(i)} \\ &= \underbrace{\alpha_t^{(0)} \cdot \mathbf{z}_t^{(0)}}_{\text{Pretrained Knowledge}} + \underbrace{\alpha_t^{(1)} \cdot \mathbf{z}_t^{(1)}}_{\text{Base Persona}} + \sum_{i=2}^n \underbrace{\alpha_t^{(i)} \cdot \mathbf{z}_t^{(i)}}_{\text{Counterfactual Personas}}. \end{aligned} \quad (3)$$

The final token distribution is computed via softmax:  $\mathbf{p}_{\text{final},t} = \text{softmax}(\mathbf{z}_{\text{adj},t})$ , from which the

next token  $y_t$  is sampled. The decoding process continues iteratively until an end-of-sequence token is generated. Once the complete sequence  $y = \{y_1, y_2, \dots, y_t\}$  is generated, CHOIR applies post-processing to extract the final answer (Kong et al., 2024b). **The full process is summarized in Algorithm 1 in the Appendix.**

## 4 Experimental Setup

**Tasks, datasets, and models.** We evaluate CHOIR across seven reasoning benchmarks categorized by task type: *arithmetic reasoning*: GSM8K (Cobbe et al., 2021), SingleEq (Koncel-Kedziorski et al., 2015), MultiArith (Roy and Roth, 2015), and SVAMP (Patel et al., 2021); *commonsense reasoning*: CommonSenseQA (Talmor et al., 2019) and StrategyQA (Geva et al., 2021); and *logical reasoning*: Date Understanding task from BIG-bench (Srivastava et al., 2023). Detailed dataset statistics are provided in Appendix C.1. Our experiments use several open-source instruction-tuned models at varying scales: the Llama 3-Instruct series (1B, 3B, 8B) (Touvron et al., 2023), Qwen 2.5-Instruct (1.5B, 3B, 7B) (Qwen Team, 2024), and Mistral-Instruct (8B) (Mistral AI, 2024).

**Baselines and implementation details.** We evaluate CHOIR against baselines categorized by their inference strategy: (1) **Single-Path Methods**: (i) Zero-Shot Chain-of-Thought (ZS-CoT) (Kojima et al., 2022), where elicits reasoning using “Let’s think step by step”; and (ii) **Role-play**, which assigns a task-relevant persona (e.g., “excellent math teacher”) to generate responses. (2) **Multi-Path Methods**: (i) Self-Consistency through ZS-CoT (SC<sub>ZS-CoT</sub>), which generates  $N$  reasoning paths via ZS-CoT (sampling temperature 0.7) and selects the majority answer<sup>3</sup>; (ii) **Persona<sub>average</sub>**, which averages predictions across  $N$  personas in each demographic group; (iii) **Best-of- $N_{\text{persona}}$** , which aggregates  $N$  persona predictions via majority vote.  $N$  denotes the number of demographic variants. (3) **Strengthened Ensembles**: (i) **Ensemble<sub>w/ ZS-CoT</sub>**, which incorporates a non-persona ZS-CoT path into the majority vote; (ii) **Ensemble<sub>weighted</sub>**, which weights the ZS-CoT path identically to CHOIR’s

<sup>3</sup>The degraded performance of SC<sub>ZS-CoT</sub> in the low- $N$  regime stems from two primary failure modes: (1) path divergence resulting in many tied problems where no consensus could be reached; and (2) instances where multiple stochastic paths converged on the same incorrect answer. This trend aligns with the original SC study (Wang et al., 2022c), where using fewer sampled reasoning paths leads to even lower performance than greedy decoding (single-path).

$\lambda^{(0)}$ . Following the demographic attributes in Table 1, we perturb personas across *Gender*, *Race*, and *Religion* (three variants each), and *Disability* and *Age* (two variants each). **A robustness analysis on the number of attributes is provided in the Appendix (Figure 6).** In cases of majority-voting ties (e.g., three different answers in SC<sub>ZS-CoT</sub>, Best-of- $N_{\text{persona}}$ , Ensemble<sub>w/ ZS-CoT</sub>, and Ensemble<sub>weighted</sub>), we randomly select one outcome and report the mean and standard deviation across three random seeds to ensure robustness. For all experiments except the SC baseline, we use greedy decoding (temperature 0) for reproducibility, following prior work (Kojima et al., 2022; Zhang et al., 2022). For smaller models (below 3B), we use a maximum of 512 tokens to improve reasoning capacity, while larger models (3B and above) use 256 tokens. The different performance on certain benchmarks is attributed to these strict zero-shot constraints and token limits compared to few-shot reports in the literature. For CHOIR, we set the knowledge weight  $\lambda^{(0)} = n - 1$  to balance the model’s internal reasoning against the  $n - 1$  counterfactual personas, ensuring aggregated demographic variants do not overwhelm the pre-trained knowledge. For all individual personas, we set  $\lambda^{(i)} = 1$  for  $i \geq 1$ . We then conduct a model-specific analysis of the  $\lambda^{(0)}$  in §5.3.

## 5 Experiments

### 5.1 RQ1: Does CHOIR improve reasoning over strong baselines?

The results in Table 2 provide a clear affirmative answer, demonstrating that CHOIR’s improvements are evident in three key ways:

**A “Floor-Raising” Effect: Elevating Weaker Baselines.** CHOIR can generate strong results even when the individual personas are relatively weak. This is most evident in scenarios where the baselines struggle. For instance, on CommonsenseQA with Llama-8B, the average persona performance is a low 55.47%, which is even lower than ZS-CoT (67.81%), indicating significant confusion among the individual personas. CHOIR synthesizes these weaker signals into a strong result of 71.63%, a +16.16 point absolute improvement. This shows that CHOIR’s harmonization is not a simple average but an intelligent fusion that can identify and amplify the correct reasoning signal even among conflicting outputs.

**A “Ceiling-Pushing” Effect: Optimizing Strong**

Method	Gender	Race	Religion	Disability	Age	Average	Gender	Race	Religion	Disability	Age	Average
GSM8K						MultiArith						
Llama-8B-Instruct						Llama-8B-Instruct						
ZS-CoT	73.16	73.16	73.16	73.16	73.16	73.16	83.00	83.00	83.00	83.00	83.00	83.00
SC <sub>ZS-CoT</sub>	74.68	74.68	74.68	65.28	65.28	70.92	84.14	84.14	84.14	71.38	71.38	79.04
Role-play	67.63	67.63	67.63	67.63	67.63	67.63	75.17	75.17	75.17	75.17	75.17	75.17
Persona <sub>average</sub>	76.93	70.20	72.20	64.86	67.06	70.25	89.89	80.67	82.56	69.08	75.17	79.47
Best-of-N <sub>persona</sub>	79.40	76.83	80.46	64.87	67.63	73.84	92.89	88.00	92.50	70.89	75.61	83.98
Ensemble <sub>w/ ZS-CoT</sub>	80.36	79.15	80.06	70.36	69.37	75.86	93.00	90.06	91.17	76.33	79.67	86.05
Ensemble <sub>weighted</sub>	80.36	79.15	80.06	70.36	69.37	75.86	93.00	90.06	91.17	76.33	79.67	86.05
CHOIR	<b>82.11</b>	<b>81.05</b>	<b>82.56</b>	<b>77.56</b>	<b>78.39</b>	<b>80.33</b> <sup>+4.47</sup>	<b>94.67</b>	<b>93.00</b>	<b>94.83</b>	<b>91.50</b>	<b>91.67</b>	<b>93.13</b> <sup>+7.08</sup>
Qwen-7B-Instruct						Qwen-7B-Instruct						
ZS-CoT	36.54	36.54	36.54	36.54	36.54	36.54	36.33	36.33	36.33	36.33	36.33	36.33
SC <sub>ZS-CoT</sub>	40.18	40.18	40.18	35.18	35.18	38.18	38.45	38.45	38.45	31.55	31.55	35.69
Role-play	56.48	56.48	56.48	56.48	56.48	56.48	66.17	<b>66.17</b>	<b>66.17</b>	66.17	<b>66.17</b>	66.17
Persona <sub>average</sub>	48.14	47.46	46.58	63.31	63.34	53.77	56.17	49.33	45.22	67.83	50.25	53.76
Best-of-N <sub>persona</sub>	53.75	51.78	54.01	63.74	63.05	57.27	63.89	53.94	54.72	67.72	50.56	58.17
Ensemble <sub>w/ ZS-CoT</sub>	51.78	50.57	52.31	56.48	52.46	52.72	57.06	52.67	48.83	64.83	49.33	54.54
Ensemble <sub>weighted</sub>	51.78	50.57	52.31	56.48	52.46	52.72	57.06	52.67	48.83	64.83	49.33	54.54
CHOIR	<b>61.79</b>	<b>62.70</b>	<b>62.09</b>	<b>67.55</b>	<b>70.66</b>	<b>64.96</b> <sup>+7.69</sup>	<b>71.00</b>	<b>66.17</b>	<b>64.83</b>	<b>72.67</b>	<b>60.50</b>	<b>67.03</b> <sup>+0.86</sup>
Mistral-8B-Instruct						Mistral-8B-Instruct						
ZS-CoT	84.99	84.99	84.99	84.99	84.99	84.99	98.17	98.17	98.17	98.17	98.17	98.17
SC <sub>ZS-CoT</sub>	85.44	85.44	85.44	81.05	81.05	83.68	98.28	98.28	98.28	97.24	97.24	97.86
Role-play	79.23	79.23	79.23	79.23	79.23	79.23	96.00	96.00	96.00	96.00	96.00	96.00
Persona <sub>average</sub>	82.89	84.33	84.53	83.70	82.79	83.65	97.50	97.44	97.83	97.75	96.33	97.37
Best-of-N <sub>persona</sub>	82.94	84.79	85.11	83.75	83.02	83.92	97.50	97.50	97.83	97.67	96.28	97.36
Ensemble <sub>w/ ZS-CoT</sub>	83.47	84.84	85.67	84.46	84.38	84.56	97.83	97.33	97.83	98.00	96.67	97.53
Ensemble <sub>weighted</sub>	83.47	84.84	85.67	84.46	84.38	84.56	97.83	97.33	97.83	98.00	96.67	97.53
CHOIR	<b>85.37</b>	<b>87.04</b>	<b>87.95</b>	<b>86.05</b>	<b>86.43</b>	<b>86.57</b> <sup>+1.58</sup>	<b>98.33</b>	<b>98.50</b>	<b>98.50</b>	<b>98.33</b>	<b>98.50</b>	<b>98.43</b> <sup>+0.26</sup>
SingleEq						CommonsenseQA						
Llama-8B-Instruct						Llama-8B-Instruct						
ZS-CoT	49.21	49.21	49.21	49.21	49.21	49.21	67.81	67.81	67.81	67.81	67.81	67.81
SC <sub>ZS-CoT</sub>	49.41	49.41	49.41	42.52	42.52	46.65	67.57	67.57	67.57	60.93	60.93	64.91
Role-play	34.25	34.25	34.25	34.25	34.25	34.25	41.28	41.28	41.28	41.28	41.28	41.28
Persona <sub>average</sub>	53.28	47.70	49.67	40.75	49.70	48.22	63.50	47.12	53.97	58.89	53.85	55.47
Best-of-N <sub>persona</sub>	55.38	52.03	55.18	40.29	49.93	50.56	66.69	52.12	61.37	58.70	54.14	58.60
Ensemble <sub>w/ ZS-CoT</sub>	56.69	54.53	55.51	45.93	49.80	52.49	68.55	62.57	66.09	66.42	63.06	65.34
Ensemble <sub>weighted</sub>	56.69	54.53	55.51	45.93	49.80	52.49	68.55	62.57	66.09	66.42	63.06	65.34
CHOIR	<b>57.87</b>	<b>57.09</b>	<b>57.87</b>	<b>54.13</b>	<b>57.48</b>	<b>56.89</b> <sup>+4.40</sup>	<b>72.97</b>	<b>70.84</b>	<b>72.56</b>	<b>71.01</b>	<b>70.76</b>	<b>71.63</b> <sup>+3.82</sup>
Qwen-7B-Instruct						Qwen-7B-Instruct						
ZS-CoT	64.17	64.17	64.17	64.17	64.17	64.17	79.52	<u>79.52</u>	79.52	79.52	79.52	79.52
SC <sub>ZS-CoT</sub>	69.09	69.09	69.09	62.20	62.20	66.33	<b>82.72</b>	<b>82.72</b>	<b>82.72</b>	<b>80.51</b>	<b>80.51</b>	<b>81.84</b>
Role-play	63.19	63.19	63.19	63.19	63.19	63.19	27.85	27.85	27.85	27.85	27.85	27.85
Persona <sub>average</sub>	66.01	57.74	48.29	83.76	63.29	63.82	72.86	61.37	59.30	70.64	64.33	65.70
Best-of-N <sub>persona</sub>	74.28	64.04	57.15	84.19	63.65	68.66	75.84	65.90	67.43	70.84	64.65	68.93
Ensemble <sub>w/ ZS-CoT</sub>	74.02	67.52	67.32	78.54	68.31	71.14	78.87	72.32	74.86	75.59	73.79	75.09
Ensemble <sub>weighted</sub>	74.02	67.52	67.32	78.54	68.31	71.14	78.87	72.32	74.86	75.59	73.79	75.09
CHOIR	<b>82.48</b>	<b>81.10</b>	<b>77.95</b>	<b>87.60</b>	<b>80.12</b>	<b>81.85</b> <sup>+10.71</sup>	<u>81.57</u>	78.87	<u>80.59</u>	<u>80.34</u>	<u>79.69</u>	<u>80.21</u> <sup>-1.63</sup>
Mistral-8B-Instruct						Mistral-8B-Instruct						
ZS-CoT	95.87	<u>95.87</u>	<u>95.87</u>	<u>95.87</u>	95.87	<u>95.87</u>	69.94	69.94	69.94	69.94	69.94	69.94
SC <sub>ZS-CoT</sub>	95.28	95.28	95.28	92.32	92.32	94.10	<b>71.99</b>	71.99	71.99	66.83	66.83	69.93
Role-play	<b>96.85</b>	<b>96.85</b>	<b>96.85</b>	<b>96.85</b>	<b>96.85</b>	<b>96.85</b>	68.88	68.88	68.88	68.88	68.88	68.88
Persona <sub>average</sub>	95.87	95.08	95.01	95.18	95.28	95.28	68.41	69.10	68.82	68.14	63.60	67.61
Best-of-N <sub>persona</sub>	95.87	95.21	95.41	95.21	95.80	95.50	68.91	71.06	69.78	68.06	63.80	68.32
Ensemble <sub>w/ ZS-CoT</sub>	96.06	95.08	95.28	<u>95.87</u>	95.47	95.55	69.12	71.25	71.39	68.80	67.40	69.59
Ensemble <sub>weighted</sub>	96.06	95.08	95.28	<u>95.87</u>	95.47	95.55	69.12	71.25	71.39	68.80	67.40	69.59
CHOIR	<u>96.65</u>	95.08	95.47	95.67	<u>96.06</u>	95.79 <sup>-1.06</sup>	<u>71.74</u>	<b>72.81</b>	<b>73.14</b>	<b>71.74</b>	<b>71.42</b>	<b>72.17</b> <sup>+2.23</sup>

Table 2: Performance of CHOIR across five demographic personas (*Gender, Race, Religion, Disability, and Age*) on reasoning datasets GSM8K, MultiArith, SingleEq and CommonsenseQA. The ‘‘Average’’ column reports mean accuracy (%) across all demographic personas. Bold values indicate the best performance, and \* marks statistically significant improvements ( $p < 0.05$ ) based on paired  $t$ -tests across demographic scores. Subscript values show gains over the strongest baseline. CHOIR enhances accuracy and exhibits harmonized reasoning across variations. Detailed implementation settings are provided in §4. Further results across various model scales and more datasets (BigBench\_Date, StrategyQA, and SVAMP) are presented in §5.2 and the Appendix.

**Baselines.** CHOIR is not only for remedial cases; it also excels at optimizing already strong performance. On MultiArith with Llama-8B, where the Best-of-N<sub>persona</sub> baseline is already a high 83.98%, CHOIR pushes accuracy even further to 93.13%, achieving a +9.16 point improvement over the strongest baseline. This demonstrates that even when a simple ensemble method like majority

voting is effective, CHOIR’s dynamic, logit-level weighting can extract additional signal to solve more complex problems.

**Robustness and Reliability Analysis.** Beyond accuracy, CHOIR improves the reliability of LLM inference by addressing two critical failure modes identified in our analysis: (1) **Mitigating Persona**

**Sensitivity.** Standard persona-assignment often exhibits high performance variance across demographic axes (e.g., a  $\pm 5.3$  variance for the *disability* axis using Llama3.1-8B-Instruct on GSM8K). As noted by (Gupta et al., 2024), LLMs perform worse for certain personas, such as “disabled”. However, our findings reveal that replacing “disabled” with “able-bodied” performed even worse, possibly due to its training data imbalances and associative priors developed during pre-training or post-training and the sensitivity of Persona-Assigned LLMs to minor perturbations analyzed in §2. This highlights the instability of single-prompt trajectories. From this aspect, we can see that CHOIR enhances performance regardless of the Persona-Assigned LLM’s original capability and mitigates the model’s sensitivity towards certain demographics in an individual persona prompt, confirming its robustness and effectiveness. **(2) Resolving Consensus Failures (The “Tied Vote” Problem).** The Best-of- $N_{\text{persona}}$  and SC<sub>ZS-CoT</sub> are prone to failure when reasoning paths are only a few and diverge completely, resulting in a high number of indecisive “tied votes” where no consensus can be reached. In contrast, by operating at the token level, CHOIR provides a continuous, consensus-aware signal that resolves these ties into a definitive, stable answer. This makes CHOIR significantly more viable for real-world deployment where computational budgets are limited.

## 5.2 RQ2: Does CHOIR’s Advantage Generalize?

Our analysis confirms that CHOIR is a broadly generalizable method whose advantages are maintained or amplified as model capability increases. We break down this analysis into two parts: broad generalization across various contexts and a deeper analysis of performance scaling.

**Broad Generalization Across Tasks, Models, Demographics, and Prompt Templates:** As shown in Table 2, CHOIR’s effectiveness is not limited to a single domain or model family. It delivers significant average improvements across different tasks and is robust across Llama, Qwen, and Mistral architectures and various demographics (Gender, Race, Religion, Disability, and Age). In addition, CHOIR is robust to the specific wording of the persona instruction. When applying alternative persona instructions, such as “Adopt the identity of {persona}”. Answer the questions while staying

CHOIR Components	Accuracy (%)
CHOIR (Full Method)	82.1
<i>Ablations:</i>	
w/ Same Reasoning Path (w/o Diversity)	73.0
w/ Random Personas (w/o Structure)	80.3
w/o Pre-trained Knowledge	79.3

Table 3: **Ablation studies for CHOIR with Llama-8B on GSM8K (Gender).** Results show that CHOIR’s performance critically depends on: (1) leveraging *diverse reasoning paths*, (2) generating them through *structured counterfactual personas*, and (3) anchoring the ensemble with the model’s *pre-trained knowledge*.

in strict accordance with the nature of this identity.” (as shown in Table 8), CHOIR still improves the Llama-1B model’s performance to 38.9%, outperforming the baselines. This demonstrates that CHOIR is robust to different persona instructions.

**Gains Maintained or Amplified with Model Capability:** A detailed analysis of CHOIR’s performance relative to the strong Best-of- $N_{\text{persona}}$  baseline across model families and scales (Figure 3) reveals key trends in CHOIR’s effectiveness: **(1) Consistent Superiority:** CHOIR outperforms the strong Best-of- $N_{\text{persona}}$  baseline in the vast majority of cases across all four datasets and the 24 specific model configurations tested. This consistent advantage confirms that CHOIR’s dynamic, logit-level harmonization is more effective than a simple voting ensemble. **(2) Maintained or Scaled Advantage:** On CommonsenseQA, for example, the lead over the baseline for Llama models grows from 11.7 to 13.0 points (1B to 8B), and for Qwen models, it more than doubles from 5.3 to 11.3 points (1.5B to 7B). This demonstrates that CHOIR is a highly scalable method that effectively leverages the increased reasoning capabilities of larger models. **(3) A Nuanced “Capability Threshold”:** On the GSM8K dataset, while the simpler majority vote is competitive on smaller Qwen models, CHOIR re-establishes a commanding lead of over 7 points (64.96% vs. 57.27%) at the 7B scale. This indicates that as models cross a certain capability threshold, CHOIR’s fusion mechanism is increasingly crucial for unlocking better performance.

## 5.3 RQ3: What are the key components driving CHOIR’s success?

Ablation studies (Table 3) validate CHOIR’s design, confirming that its success relies on several key components:

**Diverse Persona Paths are Essential:** The pri-

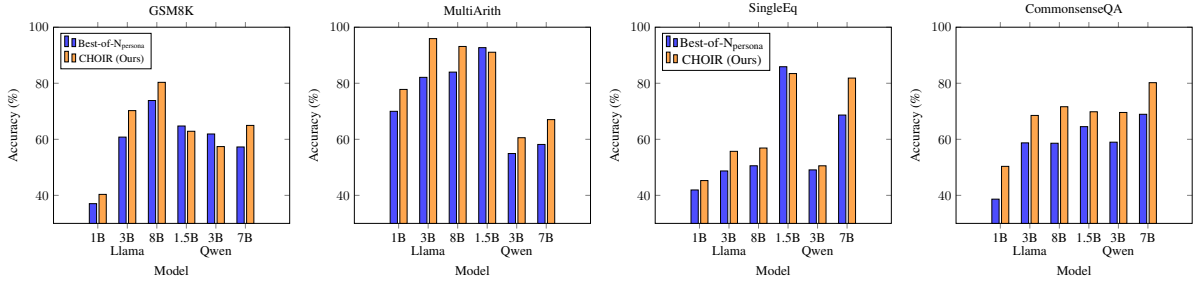


Figure 3: **Performance of CHOIR vs. Best-of- $N_{\text{persona}}$  across four reasoning datasets.** Bars represent average accuracy across *five demographic attributes* for Llama and Qwen models.

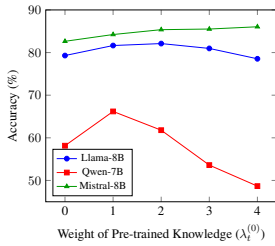


Figure 4: **Analysis of the pre-trained knowledge weight  $\lambda_t^{(0)}$  on GSM8K (Gender).** Results reveal a model-specific optimal balance, indicating that different architectures have varying reliance on their base priors.

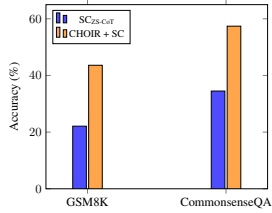


Figure 5: We evaluate CHOIR’s harmonization mechanism using stochastic reasoning paths sampled from  $SC_{ZS-CoT}$  ( $N = 3$ ) rather than demographic personas on GSM8K and CommonsenseQA (Llama-8B, Gender).

mary source of CHOIR’s strength is the complementary knowledge unlocked by different personas. When we remove this diversity by replacing the counterfactual paths with a single, repeated reasoning path (CHOIR w/ Same Reasoning Path), accuracy drops from 82.1% to 73.0%, confirming that the complementary knowledge from the different personas is important for performance gain.

**Structured Perturbations are Better than Random Ones: While we use demographic perturbations as a controlled source of minimal semantic variation, CHOIR is agnostic to the source of diversity. However, the way diversity is generated matters.** Using our structured, demographically counterfactual personas is more effective than using randomly selected personas from PersonaHub (CHOIR w/ Random Personas) (82.1% vs. 80.3%), indicating that CHOIR’s controlled, minimal perturbation approach generates more meaningful and complementary reasoning signals.

**Balancing Personas with Base Knowledge is Critical:** The model’s pre-trained knowledge is

vital for grounding the reasoning process. Completely removing the influence of the base model (w/o Pre-trained Knowledge) causes a significant drop in accuracy from 82.1% to 79.3%. While our main experiments use a fixed heuristic  $\lambda^{(0)} = n - 1$  and achieve significantly improved performance, we conducted a follow-up analysis to investigate the optimal setting for this key hyperparameter across different model architectures by setting the weight range from 0 to 4. As shown in Figure 4, we find that the optimal balance is model-specific. While Llama-8B confirms that  $\lambda^{(0)} = n - 1$  is optimal (82.11%), Qwen-7B peaks at a weight of 1 (66.19%), and Mistral-8B’s performance is highest at a weight of 4 (86.05%). These variations show that **different model families have varying reliance on their base knowledge**. While CHOIR is highly effective with default settings, performance can be further maximized through hyperparameter optimization.

## 6 Discussion

**Enhancing Stochastic Ensembles.** To test the generality of our mechanism, we applied CHOIR’s logit-level harmonization to reasoning paths generated by  $SC_{ZS-CoT}$ . In low-compute settings ( $N = 3$ ), standard SC often suffers from consensus collapse; on GSM8K, it achieves only 22.1% accuracy due to frequent “tied votes” where no majority emerges. By replacing discrete voting with CHOIR’s dynamic weighting, accuracy on these same paths improves to 43.6% (Figure 5). This indicates that CHOIR is feasible for resolving the ambiguities of traditional ensembles. However, the superior performance of our counterfactual personas suggests that structured persona diversity uncovers more complementary reasoning signals than purely stochastic sampling.

**Inference Overhead.** We evaluate the trade-off

Method	# Paths	Accuracy (%)	Decoding Strategy	Latency (s)	Relative Speed	Relative Accuracy
ZS-CoT	1	73.2	Greedy decoding	8.2	1.00x	1.00x
Best-of- $N_{\text{persona}}$ (Age)	2	67.6	Independent decoding	16.4	0.50x	0.92x
CHOIR (Age)	2	78.4	Collaborative decoding	16.4	0.50x	1.07x
Best-of- $N_{\text{persona}}$ (Gender)	3	79.4	Independent decoding	24.6	0.33x	1.08x
CHOIR (Gender)	3	82.1	Collaborative decoding	24.6	0.33x	1.12x

Table 4: **Quantitative comparison of accuracy and inference latency on GSM8K (Llama-8B)**. The results highlight: (1) CHOIR achieves an accuracy boost over single-path methods for a predictable increase in cost, and (2) it is more accurate than the multi-path baseline at the same level of computational overhead.

between reasoning accuracy and inference latency against single- and multi-path strategies. Compared to **single-path baselines** (e.g., ZS-CoT), CHOIR shows a predictable latency increase (approx.  $3\times$  for  $N = 3$ ) but yields a high relative accuracy improvement of up to  $1.12\times$  (82.1% vs. 73.2% on GSM8K). We argue this trade-off is justified where reliability is critical, such as in medical or legal reasoning. More importantly, when compared to **multi-path ensembles** (e.g., Best-of- $N_{\text{persona}}$ ), CHOIR achieves **superior accuracy at identical computational cost**. Specifically, for  $N = 3$ , CHOIR is 2.7 points more accurate with no additional latency. This validates that CHOIR extracts significantly more signal from the generated trajectories, maximizing the performance ceiling for a fixed compute budget.

**Viability for API-only Models.** To examine its applicability to API-only models, we evaluate a **“logit-limited” approximation** of CHOIR. In this setting, we extract only the top- $k$  tokens from the LLM, identify the union of top- $k$  tokens across personas, and apply CHOIR’s dynamic weighting only within this restricted token space. The results of Llama model on GSM8K for ZS-CoT, CHOIR<sub>Top-50</sub>, CHOIR<sub>Top-100</sub>, and CHOIR<sub>Full-logit</sub> are 73.16, 76.72, 79.30, and 82.11, respectively. These results show that full-logit CHOIR achieves the best performance. Future work into more advanced “logit-limited” variants, such as adaptive logit selection and probability recalibration in sparse token spaces, is a promising direction for further optimizing CHOIR for commercial APIs.

## 7 Conclusion

We reframe persona-induced variance as a source of complementary reasoning and introduce CHOIR, a test-time framework that harnesses this diversity to improve performance. Future work includes exploring advanced fusion mechanisms, such as attention-based or learnable weighting, to better integrate reasoning signals.

## Limitations

CHOIR’s dynamic harmonization mechanism operates at the token-logit level. While we demonstrate that the framework can be adapted to top- $k$  logit approximations (common in many commercial APIs), it cannot be applied to “black-box” interfaces that return only text strings. Future work could explore mapping these logit-level insights into text-based collaborative protocols to bridge this gap.

Our study focuses on demographic perturbations (Gender, Race, Religion, Disability, and Age) as a controlled source of diversity. While we prove this is effective, there may be other “structured prompts”, such as varying cognitive styles or educational backgrounds, that could induce even more complementary reasoning paths. The optimal strategy for generating “maximum diversity” remains an open question.

CHOIR’s success depends on the existence of diverse reasoning paths within the model’s latent space. As models undergo more intensive alignment (e.g., RLHF) to reduce bias, the “persona-induced variance” we exploit may diminish, potentially reducing the framework’s effectiveness. There is no guarantee that this method is extensible to future “over-aligned” models that might collapse diverse reasoning trajectories into a single, uniform response.

## Ethical Considerations

Our work operates at the intersection of model reasoning and demographic representation. We recognize that utilizing demographic terms (e.g., race, gender, disability) as minimal semantic probes to generate cognitive diversity carries inherent risks. While we leverage these variations constructively, we acknowledge that LLMs may exhibit harmful biases or performance gaps when assigned certain marginalized identities due to historical imbalances in training data.

Another concern is the potential for repurposing this technology to amplify biases. While CHOIR is

designed to mitigate performance disparities by harmonizing divergent outputs, the underlying mechanism still relies on persona prompting. If a base model is fundamentally misaligned, there is a risk that the “consensus” generated during token-level decoding could converge on stereotypic or harmful paths. We advocate for the responsible deployment of CHOIR only with models that have undergone proper alignment.

## References

- Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2024. [Graph of thoughts: Solving elaborate problems with large language models](#). In *AAAI*, pages 17682–17690. AAAI Press.
- Millennium Bismay, Xiangjue Dong, and James Caverlee. 2025. [ReasoningRec: Bridging personalized recommendations and human-interpretable explanations through LLM reasoning](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 8147–8163, Albuquerque, New Mexico. Association for Computational Linguistics.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.
- DeepSeek-AI. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *Preprint*, arXiv:2501.12948.
- Xiangjue Dong, Manling Li, and James Caverlee. 2025. [Probing the limits of embodied spatial planning in llms](#). In *NeurIPS 2025 Workshop on Space in Vision, Language, and Embodied AI*.
- Xiangjue Dong, Yibo Wang, Philip S Yu, and James Caverlee. 2024. Disclosure and mitigation of gender bias in llms. *arXiv preprint arXiv:2402.11190*.
- Dheeru Dua, Shivanshu Gupta, Sameer Singh, and Matt Gardner. 2022. [Successive prompting for decomposing complex questions](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1251–1265, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Jonas Geiping, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian R. Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. 2025. [Scaling up test-time compute with latent reasoning: A recurrent depth approach](#). *Preprint*, arXiv:2502.05171.
- Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppel, Preslav Nakov, and Iryna Gurevych. 2024. [A survey of confidence estimation and calibration in large language models](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6577–6595, Mexico City, Mexico. Association for Computational Linguistics.
- Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. [Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies](#). *Transactions of the Association for Computational Linguistics*, 9:346–361.
- Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2024. [Bias runs deep: Implicit reasoning biases in persona-assigned LLMs](#). In *The Twelfth International Conference on Learning Representations*.
- Zhitao He, Pengfei Cao, Yubo Chen, Kang Liu, Ruopeng Li, Mengshu Sun, and Jun Zhao. 2023. [Lego: A multi-agent collaborative framework with role-playing and iterative feedback for causality explanation generation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9142–9163.
- Jie Huang and Kevin Chen-Chuan Chang. 2023. [Towards reasoning in large language models: A survey](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada. Association for Computational Linguistics.
- Mahammed Kamruzzaman, Md. Shovon, and Gene Kim. 2024. [Investigating subtler biases in LLMs: Ageism, beauty, institutional, and nationality bias in generative models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8940–8965, Bangkok, Thailand. Association for Computational Linguistics.
- Junseok Kim, Nakyeong Yang, and Kyomin Jung. 2024. [Persona is a double-edged sword: Mitigating the negative impact of role-playing prompts in zero-shot reasoning tasks](#). *Preprint*, arXiv:2408.08631.
- Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. [Large language models are zero-shot reasoners](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.
- Rik Koncel-Kedziorski, Hannaneh Hajishirzi, Ashish Sabharwal, Oren Etzioni, and Siena Dumas Ang. 2015. [Parsing algebraic word problems into equations](#). *Transactions of the Association for Computational Linguistics*, 3:585–597.
- Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2024a. [Better zero-shot reasoning with](#)

- role-play prompting. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4099–4113, Mexico City, Mexico. Association for Computational Linguistics.
- Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2024b. **Better zero-shot reasoning with role-play prompting.** In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4099–4113, Mexico City, Mexico. Association for Computational Linguistics.
- Abhishek Kumar, Robert Morabito, Sanzhar Umbet, Jad Kabbara, and Ali Emami. 2024. **Confidence under the hood: An investigation into the confidence-probability alignment in large language models.** In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 315–334, Bangkok, Thailand. Association for Computational Linguistics.
- Xiaoou Liu, Tiejun Chen, Longchao Da, Chacha Chen, Zhen Lin, and Hua Wei. 2025. **Uncertainty quantification and confidence calibration in large language models: A survey.** In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V2, KDD '25*, page 6107–6117, New York, NY, USA. Association for Computing Machinery.
- Jing Luo, Longze Chen, Run Luo, Liang Zhu, Chang Ao, Jiaming Li, Yukun Chen, Xin Cheng, Wen Yang, Jiayuan Su, Ahmadreza Argha, Hamid Alinejad-Rokny, Chengming Li, Shiwen Ni, and Min Yang. 2024. **Personamath: Boosting mathematical reasoning via persona-driven data augmentation.** *Preprint*, arXiv:2410.01504.
- Mistral AI. 2024. Un ministral, des ministraux. <https://mistral.ai/news/ministraux>.
- Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. **Are NLP models really able to solve simple math word problems?** In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094, Online. Association for Computational Linguistics.
- Rebecca Qian, Candace Ross, Jude Fernandes, Eric Michael Smith, Douwe Kiela, and Adina Williams. 2022. **Perturbation augmentation for fairer NLP.** In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9496–9521, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Qwen Qwen Team. 2024. **Qwen2.5: A party of foundation models.**
- Leonardo Ranaldi, Elena Sofia Ruzzetti, Davide Venditti, Dario Onorati, and Fabio Massimo Zanzotto. 2024. **A trip towards fairness: Bias and de-biasing in large language models.** In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (\*SEM 2024)*, pages 372–384, Mexico City, Mexico. Association for Computational Linguistics.
- Subhro Roy and Dan Roth. 2015. **Solving general arithmetic word problems.** In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1743–1752, Lisbon, Portugal. Association for Computational Linguistics.
- Jisu Shin, Hoyun Song, Huije Lee, Soyeong Jeong, and Jong Park. 2024. **Ask LLMs directly, “what shapes your bias?”: Measuring social bias in large language models.** In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16122–16143, Bangkok, Thailand. Association for Computational Linguistics.
- Ingroj Shrestha and Padmini Srinivasan. 2025. **LLM bias detection and mitigation through the lens of desired distributions.** In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1464–1480, Suzhou, China. Association for Computational Linguistics.
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, and et al. 2023. **Beyond the imitation game: Quantifying and extrapolating the capabilities of language models.** *Transactions on Machine Learning Research*. Featured Certification.
- Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. **CommonsenseQA: A question answering challenge targeting commonsense knowledge.** In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.
- Bryan Chen Zhengyu Tan and Roy Ka-Wei Lee. 2025. **Unmasking implicit bias: Evaluating persona-prompted LLM responses in power-disparate social scenarios.** In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1075–1108, Albuquerque, New Mexico. Association for Computational Linguistics.
- Maria Teleki, Xiangjue Dong, Haoran Liu, and James Caverlee. 2025. **Masculine defaults via gendered discourse in podcasts and large language models.** In *ICWSM 2025*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. **Llama 2: Open foundation and fine-tuned chat models.** *arXiv preprint arXiv:2307.09288*.

- Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024a. [Two tales of persona in LLMs: A survey of role-playing and personalization](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16612–16631, Miami, Florida, USA. Association for Computational Linguistics.
- Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024b. [Two tales of persona in LLMs: A survey of role-playing and personalization](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16612–16631, Miami, Florida, USA. Association for Computational Linguistics.
- Boshi Wang, Xiang Deng, and Huan Sun. 2022a. [Iteratively prompt pre-trained language models for chain of thought](#). In *Conference on Empirical Methods in Natural Language Processing*.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Huai hsin Chi, and Denny Zhou. 2022b. [Self-consistency improves chain of thought reasoning in language models](#). *ArXiv*, abs/2203.11171.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022c. [Self-consistency improves chain of thought reasoning in language models](#). In *The Eleventh International Conference on Learning Representations*.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. [Emergent abilities of large language models](#). *Preprint*, arXiv:2206.07682.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2024. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22*, Red Hook, NY, USA. Curran Associates Inc.
- Zifan Xu, Haozhu Wang, Dmitriy Besspalov, Xian Wu, Peter Stone, and Yanjun Qi. 2024. [LaRS: Latent reasoning skills for chain-of-thought reasoning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3624–3643, Miami, Florida, USA. Association for Computational Linguistics.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. [Tree of thoughts: Deliberate problem solving with large language models](#). *arXiv preprint arXiv:2305.10601*.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023b. [React: Synergizing reasoning and acting in language models](#). In *The Eleventh International Conference on Learning Representations*.
- Yao Yao, Z. Li, and Hai Zhao. 2023c. [Beyond chain-of-thought, effective graph-of-thought reasoning in large language models](#). *ArXiv*, abs/2305.16582.
- Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. [Automatic chain of thought prompting in large language models](#). In *The Eleventh International Conference on Learning Representations*.
- Jinman Zhao, Zifan Qian, Linbo Cao, Yining Wang, Yitian Ding, Yulan Hu, Zeyu Zhang, and Zeyong Jin. 2024. [Role-play paradox in large language models: Reasoning performance gains and ethical dilemmas](#). *Preprint*, arXiv:2409.13979.
- Mingqian Zheng, Jiaxin Pei, Lajanugen Logeswaran, Moontae Lee, and David Jurgens. 2024. [When “a helpful assistant” is not really helpful: Personas in system prompts do not improve performances of large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15126–15154, Miami, Florida, USA. Association for Computational Linguistics.

## A Related Work

### A.1 LLM Reasoning

LLMs with parameters beyond a certain scale exhibit emergent reasoning abilities (Wei et al., 2022; Cobbe et al., 2021). When provided with examples of “chain of thought” (CoT), which represent intermediate natural language reasoning steps, these models can generate explicit rationales similar to eliciting inductive/abductive reasoning (Wei et al., 2024; Huang and Chang, 2023). Advances such as zero-shot CoT (Kojima et al., 2022), where a model is prompted with phrases like “Let’s think step by step,” allow for reasoning without the need for explicit few-shot examples. Building on this, multi-step reasoning approaches, such as Successive Prompting (Dua et al., 2022), Tree-of-Thought (Yao et al., 2023a), Graph-of-Thought (Besta et al., 2024; Yao et al., 2023c), Iterating CoT (Wang et al., 2022a), and Self-Consistency (Wang et al., 2022b), have become critical strategies for enhancing reasoning in various downstream tasks including planning and decision-making (Yao et al., 2023b), question answering and reasoning benchmarks, and applications such as recommendation systems (Bismay et al., 2025). Many recent research works have been focused on improving the LLM Reasoning abilities as a pre-training or post-training effort. DeepSeek-R1 (DeepSeek-AI, 2025) is a notable stride forward in the direction of open-source reasoning model trained via large-scale reinforcement learning (RL) without supervised fine-tuning (SFT). There have also been some work in the direction of latent reasoning to scale up the test-time compute via implicitly reasoning in latent space through recurrent depth (Xu et al., 2024; Geiping et al., 2025).

### A.2 Persona-assigned LLM

Persona assignment in large language models (LLMs) refers to giving the model a specific role or identity (e.g. “You are a math teacher”) during prompting. This technique, also called **role prompting** or **role-play**, has gained popularity as a way to steer LLM behavior in desirable ways. By adopting a persona, an LLM can generate more in-character responses tailored to that role which in turn found to improve the zero-shot reasoning capabilities (Kong et al., 2024a). Coupling a persona with an LLM can activate the model’s corresponding parametric knowledge and style, enabling it to produce responses aligned with that

role (Tseng et al., 2024b; Dong et al., 2025). Adopting expert personas, especially aligned with the task domain, has demonstrated improved accuracy on mathematical reasoning benchmarks like GSM8K and AQuA (Luo et al., 2024; Zhao et al., 2024). Persona prompts naturally elicit detailed reasoning traces, thereby enhancing inductive, deductive, and multi-step logical reasoning (Kong et al., 2024a; He et al., 2023). However, inappropriate personas may degrade reasoning performance by introducing extraneous or misleading context. Large-scale studies have shown persona effects to be highly variable, indicating no guaranteed improvements (Zheng et al., 2024). Further, personas might amplify biases and produce unsafe outputs, underscoring the need for cautious implementation (Zhao et al., 2024; Zheng et al., 2024; Gupta et al., 2024; Tan and Lee, 2025). To mitigate biases introduced by personas in LLM reasoning, Kim et al. (2024) propose ensemble prompting, combining persona-driven responses with neutral prompts and selecting the most accurate output through automated evaluation.

## B Extended Discussion and Future Directions

Recent work has shown that persona-driven biases are subtle, persistent, and difficult to remove via simple prompt debiasing (Gupta et al., 2024). While much of the existing research treats this performance variation as a flaw to be mitigated, our work takes a fundamentally different approach. CHOIR does not attempt to suppress this variation; instead, it harnesses it, treating the divergent reasoning paths as complementary signals that can be fused into a more robust and accurate final answer. This reframing of sensitivity-as-a-resource is a key contribution of our work.

Our use of minimal counterfactual edits (e.g., pronouns, age, disability terms) is to isolate the effect of demographic variation while preserving semantic consistency. This allowed us to: (1) validate that even minor shifts (e.g., “he” vs. “she”) can alter reasoning; (2) measure persona-induced variation without introducing semantic drift. Beyond counterfactual settings, our ablation results (Table 3) further confirm that CHOIR’s benefits generalize: even randomly sampled personas (not just counterfactuals) improve performance, surpassing multiple strong baselines. This indicates that CHOIR’s robustness is not limited to carefully en-

$G_{\text{Gender}}^{(i)}$	Personas ( $p_{\text{he}}^{(i)}, p_{\text{she}}^{(i)}, p_{\text{they}}^{(i)}$ )
$G_{\text{Gender}}^{(1)}$	a hardworking construction worker tirelessly petitioning for {his/her/their} sibling’s immigration case
$G_{\text{Gender}}^{(2)}$	a retired basketball coach with decades of experience, known for {his/her/their} expertise in developing point guards
$G_{\text{Gender}}^{(3)}$	a single {man/woman/person} who takes {his/her/their} personal development and emotional health seriously
$G_{\text{Gender}}^{(4)}$	a retired childcare professional who always has a soothing lullaby up {his/her/their} sleeve
$G_{\text{Gender}}^{(5)}$	a music producer who collaborates with Bette Midler on {his/her/their} new album

Table 5: Counterfactual gender persona group  $G_{\text{Gender}}^{(i)}$  used in Figure 1a and 1b.

gineered demographic perturbations but extends to broader forms of reasoning diversity. For future work, extending the framework to richer or more nuanced persona dimensions (e.g., socio-political identity, cultural background, or domain expertise) may reveal additional layers of complementary reasoning, and we leave it to future work.

For confidence estimation, we employ maximum token probability as a practical proxy due to its simplicity, computational efficiency, and proven stability in prior work. Despite its lightweight nature, this measure consistently yields strong empirical gains. Since CHOIR operates during the final answer aggregation phase of Zero-Shot-CoT, token-level confidence provides a reliable signal for weighting reasoning paths. For future work, richer uncertainty quantification (e.g., sequence-level entropy, calibration-based measures, or cross-persona KL divergence) could further refine CHOIR’s weighting mechanism.

## C More Implementation Details

This section provides additional details regarding the datasets, models, and prompting strategies used in our study. All experiments are conducted on NVIDIA A5000 24GB GPUs.<sup>4</sup>

### C.1 Dataset and Model Details

The statistics for the datasets are provided in Table 6. The links and licenses for all open-source models are provided in Table 7.

### C.2 Prompting Details

Counterfactual gender persona group  $G_{\text{Gender}}^{(i)}$  used in Figure 1a and 1b are shown in Table 5. Table 8 shows two different templates for assigning personas. For the “Role-Play” baseline, we used the

<sup>4</sup>Code and data are available at <https://github.com/dongxiangjue/CHOIR>.

Dataset	Answer Format	$N_q$	$L_q$	License
GSM8K	arabic number	1,319	46.9	MIT License
SingleEq	arabic number	508	27.4	–
MultiArith	arabic number	600	31.8	Apache 2.0
CommonsenseQA	option (A-E)	1,221	27.8	MIT

Table 6: Statistics of datasets.  $N_q$  denotes the number of questions in each dataset.  $L_q$  denotes the average words of questions in each dataset.

specific role-play prompts detailed in Table 9.

## C.3 Answer Extraction

As our experiments are conducted in a zero-shot setting, we adopt the answer extraction methodology from Zero-Shot-CoT (Kojima et al., 2022). Specifically, for each question, we first obtain the answer generated by the LLM. Then, we combine the question, answer, and an answer trigger into a single input for the model. The answer trigger phrases for different response formats are provided in Table 10.

## D More Results and Analysis

### D.1 Detailed Performance Analysis by Persona Group

As shown in Table 11, the quality of the initial persona influences both baseline performance and the degree of improvement under CHOIR. For instance, G(1) starts with the lowest base accuracy (<30%) but sees the largest improvement (+9%), while the Expert persona starts strong (~41%) and improves modestly to 43%. Despite starting from a lower point, G(1) surpasses G(5) after CHOIR, showing that weaker personas benefit more, while stronger personas still improve. Moreover, a key outcome of the harmonization is that it brings all persona groups into a similar, high-performing range (39 - 43%), effectively reducing variance across personas. This illustrates its robustness and its advantage over static persona prompting by mitigating reliance on any single, potentially misaligned persona.

The full process is shown in Algorithm 1.

### D.2 More Analysis of Computational Overhead

CHOIR requires generating and processing  $N$  persona-conditioned reasoning paths for each input during inference. This comparison remains fair in the main paper, as strong baselines, Persona<sub>average</sub> and Best-of- $N_{\text{persona}}$ , also aggregate  $N$  reasoning

Model	Link	License
LLAMA-3.2-1B-INSTRUCT	<a href="https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct">https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct</a>	Llama 3.2 Community License
LLAMA-3.2-3B-INSTRUCT	<a href="https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct">https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct</a>	Llama 3.2 Community License
LLAMA-3.1-8B-INSTRUCT	<a href="https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct">https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct</a>	Llama 3.1 Community License
QWEN2.5-1.5B-INSTRUCT	<a href="https://huggingface.co/Qwen/Qwen2.5-1.5B-Instruct">https://huggingface.co/Qwen/Qwen2.5-1.5B-Instruct</a>	Apache license 2.0
QWEN2.5-3B-INSTRUCT	<a href="https://huggingface.co/Qwen/Qwen2.5-3B-Instruct">https://huggingface.co/Qwen/Qwen2.5-3B-Instruct</a>	qwen-research
QWEN2.5-7B-INSTRUCT	<a href="https://huggingface.co/Qwen/Qwen2.5-7B-Instruct">https://huggingface.co/Qwen/Qwen2.5-7B-Instruct</a>	Apache license 2.0
MINISTRAL-8B-INSTRUCT-2410	<a href="https://huggingface.co/mistralai/Minstral-8B-Instruct-2410">https://huggingface.co/mistralai/Minstral-8B-Instruct-2410</a>	mrl

Table 7: Model links and licenses.

ID	Instruction
1	You are {persona}. Your responses should closely mirror the knowledge and abilities of this persona.
2	Adopt the identity of {persona}. Answer the questions while staying in strict accordance with the nature of this identity.

Table 8: Two different persona instructions used in our study. To assign a persona to an LLM, we replace {persona} in the instruction with the target persona.

Dataset	Role-play Prompt
svamp, singleeq, gsm8k, multiarith	<b>[role setting:]</b> From now on, you are an excellent math teacher and always teach your students math problems correctly. And I am one of your students. <b>[reply:]</b> That’s great to hear! As your math teacher, I’ll do my best to explain mathematical concepts correctly so that you can understand them easily.
bigbench_date	<b>[role setting:]</b> From now on, you are an excellent teacher and are teaching your students how to calculate dates correctly. I am one of your students. <b>[reply:]</b> Of course! I’m here to help you with any questions you have about calculating dates correctly.
commonsenseqa, strategyqa	<b>[role setting:]</b> From now on, you are a contestant in a general knowledge quiz who always answers common sense questions accurately. <b>[reply:]</b> That sounds like a fun challenge! I’m ready to participate and answer your questions.

Table 9: Role-play prompts used in the Role-Play prompting method.

Answer Format	Answer Trigger
arabic number	Therefore, the answer (arabic numerals) is
option (A-E)	Therefore, among A through E, the answer is
option (A-C)	Therefore, among A through C, the answer is
yes or no	Therefore, the answer (Yes or No) is
string	Therefore, the final answer is

Table 10: Answer triggers for different tasks.

paths. In contrast, single-path methods like Zero-Shot, Few-Shot, Zero-Shot-CoT, and Role-Play generate only one reasoning path per query.

In our experiments, the number of counterfactual personas varies by demographic attribute (Table 1):  $N = 3$  for Gender, Race, and Religion, and  $N = 2$  for Disability and Age. Table 4 presents the computational trade-off on GSM8K using LLaMA3-8B-Instruct, comparing CHOIR with both single-path and multi-path baselines in terms of accuracy, latency, and relative efficiency.

CHOIR increases inference latency by approximately  $2 - 3\times$  relative to single-path methods (e.g., Zero-Shot-CoT), as expected from processing multiple reasoning paths. However, it achieves a  $1.07 - 1.12\times$  relative improvement, demonstrat-

ing a strong performance-efficiency trade-off.

When compared to multi-path baselines (e.g., Best-of- $N_{\text{persona}}$ ), CHOIR attains substantially higher accuracy than baselines at identical computational cost. For  $N = 3$ , CHOIR is 2.7 points more accurate with no additional latency, underscoring the effectiveness of its dynamic weighting mechanism over static aggregation.

To further analyze the practical trade-off between reasoning accuracy and computational overhead, we conduct a comparison between CHOIR ( $N = 3$ ) and  $\text{SC}_{\text{ZS-CoT}}$  scaling from  $N = 3$  to  $N = 8$ . Evaluated on GSM8K using Llama-1B,  $\text{SC}_{\text{ZS-CoT}}$  yields an accuracy of 34.12% at  $N = 3$ , scaling incrementally to 40.94% at  $N = 5$  and 47.01% at  $N = 8$ . In comparison, CHOIR achieves 41.93% with only three reasoning paths ( $N = 3$ ). While  $\text{SC}_{\text{ZS-CoT}}$  begins to outperform CHOIR at  $N = 6$ , it requires twice the computational budget (latency and token generation). These results demonstrate that CHOIR is a more compute-efficient test-time strategy, extracting significantly more reasoning signal per path in the low- $N$  regime.

---

**Algorithm 1:** CHOIR: Collaborative Harmonization fOR Inference Robustness

---

**Input:** Question  $x$ ; base persona  $p_1$ ;  
persona instruction template  $T(\cdot)$ ;  
language model  $f_\theta$ ; max generation  
length  $L$ ;  $\lambda^{(0)}$

**Output:** Answer  $y$

```
1 // Step 1: Construct Counterfactual
  Personas
2 Construct counterfactual personas
   $\{p_2, \dots, p_n\}$  by modifying demographic
  attributes of  $p_1$ ;
3 Generate persona-specific instructions:
   $r_i = T(p_i), \forall i \in \{1, \dots, n\}$ ;
4 // Step 2: Dynamic Collaborative
  Decoding
5 for each generation step  $t = 1$  to  $L$  do
6   for each persona  $r_i$  (including base and
     counterfactual personas) do
7     Compute logits and probabilities:
        $\mathbf{z}_t^{(i)} = f_\theta(r_i, x, y_{<t}), \mathbf{p}_t^{(i)} =$ 
        $\text{softmax}(\mathbf{z}_t^{(i)});$ 
8     Compute confidence score:
        $s_t^{(i)} = \max \mathbf{p}_t^{(i)};$ 
9   Compute average consensus confidence:
      $\bar{s}_t = \frac{1}{n} \sum_{i=1}^n s_t^{(i)};$ 
10  for each persona  $r_i, i \geq 1$  do
11     $d_t^{(i)} = |s_t^{(i)} - \bar{s}_t|;$ 
12     $\alpha_t^{(i)} = 1 - d_t^{(i)};$ 
13  Set  $\alpha_t^{(0)} = \lambda^{(0)};$ 
14  Compute base logits from pre-trained
     knowledge:  $\mathbf{z}_t^{(0)} = f_\theta(x, y_{<t});$ 
15  Aggregate logits:
      $\mathbf{z}_{\text{adj},t} = \sum_{i=0}^n \alpha_t^{(i)} \cdot \mathbf{z}_t^{(i)} =$ 
      $\alpha_t^{(0)} \cdot \mathbf{z}_t^{(0)} + \sum_{i=1}^n \alpha_t^{(i)} \cdot \mathbf{z}_t^{(i)};$ 
16  Sample next token:
      $y_t \sim \mathbf{p}_{\text{final},t} = \text{softmax}(\mathbf{z}_{\text{adj},t});$ 
17  if  $y_t$  is end-of-sequence token then
18    break;
19 return  $y = \text{post\_processing}\{y_1, \dots, y_t\}$ 
```

---

**Practical Considerations.** In practical deployments, such as medical QA, legal reasoning, or policy analysis, where reasoning reliability outweighs minor latency increases, this trade-off is both justified and beneficial. Future work can explore optimizing CHOIR’s efficiency. For instance,

adaptive techniques such as “early exiting”, where computation halts once personas reach a strong consensus, or “persona pruning”, where only a diverse subset of personas are used for a given query, could reduce latency without sacrificing significant accuracy.

### D.3 Results on Additional Datasets

We report results on StrategyQA, SVAMP, and Big-Bench\_Date in Table 12. Using Llama-8B with Gender personas (N=3), CHOIR maintains consistent improvements.

### D.4 Results on KL divergence

The choice of maximum token probability as a heuristic for persona confidence is based on its proven balance of computational efficiency and stability in recent literature. Additionally, we implement a variant of CHOIR that replaces max-probability deviation with pairwise symmetric KL divergence. Specifically, we replace max-probability deviation logic with pairwise symmetric KL divergence among persona distributions; compute per-persona disagreement as the mean of its two pairwise symmetric KLs; and update fusion to use these KL-derived weights. On GSM8K (Llama-1B), the KL-based variant achieved 40.11%, which is lower than the CHOIR accuracy of 40.35%. In addition, KL divergence requires a full vocabulary pass per persona per decoding step and introduces additional operations and memory movement. In contrast, max-prob confidence uses scalar statistics already available from softmax, adding negligible overhead. This suggests that while KL divergence is a more complex measure with additional computational cost, it does not necessarily translate to higher reasoning performance in CHOIR, justifying our use of the more compute-efficient max-probability proxy.

## E Robustness to the Choice of Demographic Attributes

To test CHOIR’s robustness to the specific composition of the demographic terms, we compared our primary “Race 1” configuration (three attributes: White, Black, Asian) against a “Race 2” configuration (four attributes: White, Black, Asian, Hispanic). As shown in Figure 6, the performance is highly stable across both sets. This confirms that CHOIR is not overly sensitive to the specific choice or number of demographic attributes used.

Groups	Gender	Accuracy	CHOIR	Personas
NA	–	37.4	–	
1	he	29.4	39.1	a hardworking construction worker tirelessly petitioning for his sibling’s immigration case
	she	29.5		a hardworking construction worker tirelessly petitioning for her sibling’s immigration case
	they	30.8		a hardworking construction worker tirelessly petitioning for their sibling’s immigration case
2	he	39.0	43.0	a retired basketball coach with decades of experience, known for his expertise in developing point guards
	she	40.6		a retired basketball coach with decades of experience, known for her expertise in developing point guards
	they	40.0		a retired basketball coach with decades of experience, known for their expertise in developing point guards
3	he	37.6	42.0	a single man who takes his personal development and emotional health seriously
	she	37.9		a single woman who takes her personal development and emotional health seriously
	they	37.8		a single person who takes their personal development and emotional health seriously
4	he	39.7	42.2	a retired childcare professional who always has a soothing lullaby up his sleeve
	she	39.4		a retired childcare professional who always has a soothing lullaby up her sleeve
	they	39.8		a retired childcare professional who always has a soothing lullaby up their sleeve
5	he	37.8	40.0	a music producer who collaborates with Bette Midler on his new album
	she	37.2		a music producer who collaborates with Bette Midler on her new album
	they	37.7		a music producer who collaborates with Bette Midler on their new album
Expert	he	41.2	43.5	a mathematics professor with his keen interest on arithmetic and reasoning
	she	42.5		a mathematics professor with her keen interest on arithmetic and reasoning
	they	41.9		a mathematics professor with their keen interest on arithmetic and reasoning

Table 11: Detailed results for Figure 1a.

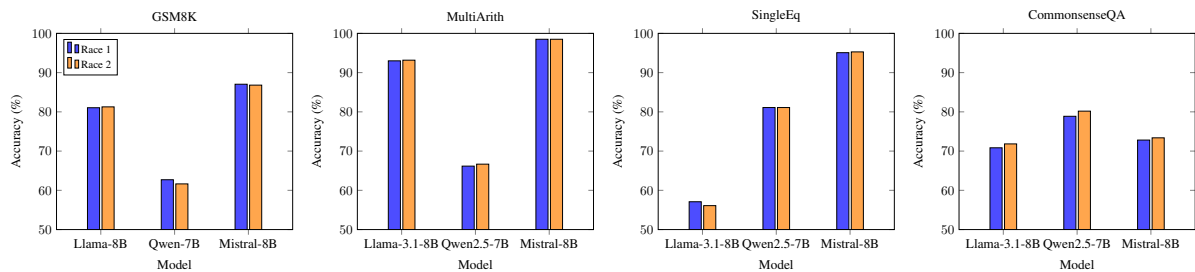


Figure 6: Accuracy comparison on four reasoning datasets. Blue bars represent Race 1 scores and orange bars represent Race 2 scores across three models (Llama-8B, Qwen-7B, Mistral-8B).

Method	BigBench_Date	StrategyQA	SVAMP
ZS-CoT	64.77	64.93	72.20
Persona <sub>average</sub>	69.56	63.14	79.70
Best-of-N <sub>persona</sub>	72.09	64.02	82.40
CHOIR	<b>73.44</b>	<b>66.99</b>	<b>86.50</b>

Table 12: Additional results on StrategyQA, SVAMP, and BigBench\_Date.

## F Case Study

Table 13 shows that different demographic personas produce different reasoning paths, leading to different answers, making it difficult to reach a consensus through simple majority voting.

Question	A hand-painted wallpaper costs \$400 at the market. A DIY will saves 20% after considering the materials cost. If Ethan made his own hand-painted wallpaper, how much was the total cost?
Persona 1	First, we need to calculate the 20% savings on the \$400 wallpaper. That's $20/100 * \$400 = \$80$ . So, the cost of the materials would be $\$400 - \$80 = \$320$ . Now, if Ethan made his own hand-painted wallpaper, he would save the \$80 he would have spent on the wallpaper itself. So, the total cost for Ethan would be the materials cost plus the savings, which is $\$320 + \$80 = \$400$ .
Persona 2	First, we need to calculate the savings on the wallpaper if Ethan makes it himself. The savings are 20% of \$400, which is $\$400 * 0.20 = \$80$ . Now, we subtract the savings from the original price to find the cost of the DIY wallpaper. So, $\$400 - \$80 = \$320$ . That's the cost of the materials. But we also need to consider the time and effort Ethan puts in, which is priceless. So, the total cost for Ethan making his own hand-painted wallpaper is $\$320$ .
Persona 3	First, we need to calculate the savings on the wallpaper itself. If the wallpaper costs \$400 and DIY saves 20%, then the savings would be 20% of \$400. That's $\$400 * 0.20 = \$80$ . So, the cost of the wallpaper after the savings would be $\$400 - \$80 = \$320$ . Now, we need to consider the cost of materials. If DIY saves 20% on the wallpaper, it means the materials cost is 80% of the original price. So, the materials cost would be $\$400 * 0.80 = \$320$ . Therefore, the total cost for Ethan to make his own hand-painted wallpaper would be the cost of the wallpaper after savings plus the cost of materials, which is $\$320 + \$320 = \$640$ .

Table 13: Case study.