ValuesRAG: Enhancing Cultural Alignment Through Retrieval-Augmented Contextual Learning

Wonduk Seo[1], Zonghao Yuan[2], Yi Bu[3]
Enhans[1], Tsinghua University[2], Peking University[3]

MOTIVATION

Cultural bias in LLMs undermines reliability across regions. Localized models and prompt heuristics help but still encode training-set stereotypes. ValuesRAG reframes alignment as retrieval: inject multiple value summaries matched by demographics, then reason over them. This shifts from static labels to evidence-backed value profiles, enhancing inclusivity, contextual fidelity, and trustworthiness in downstream applications.

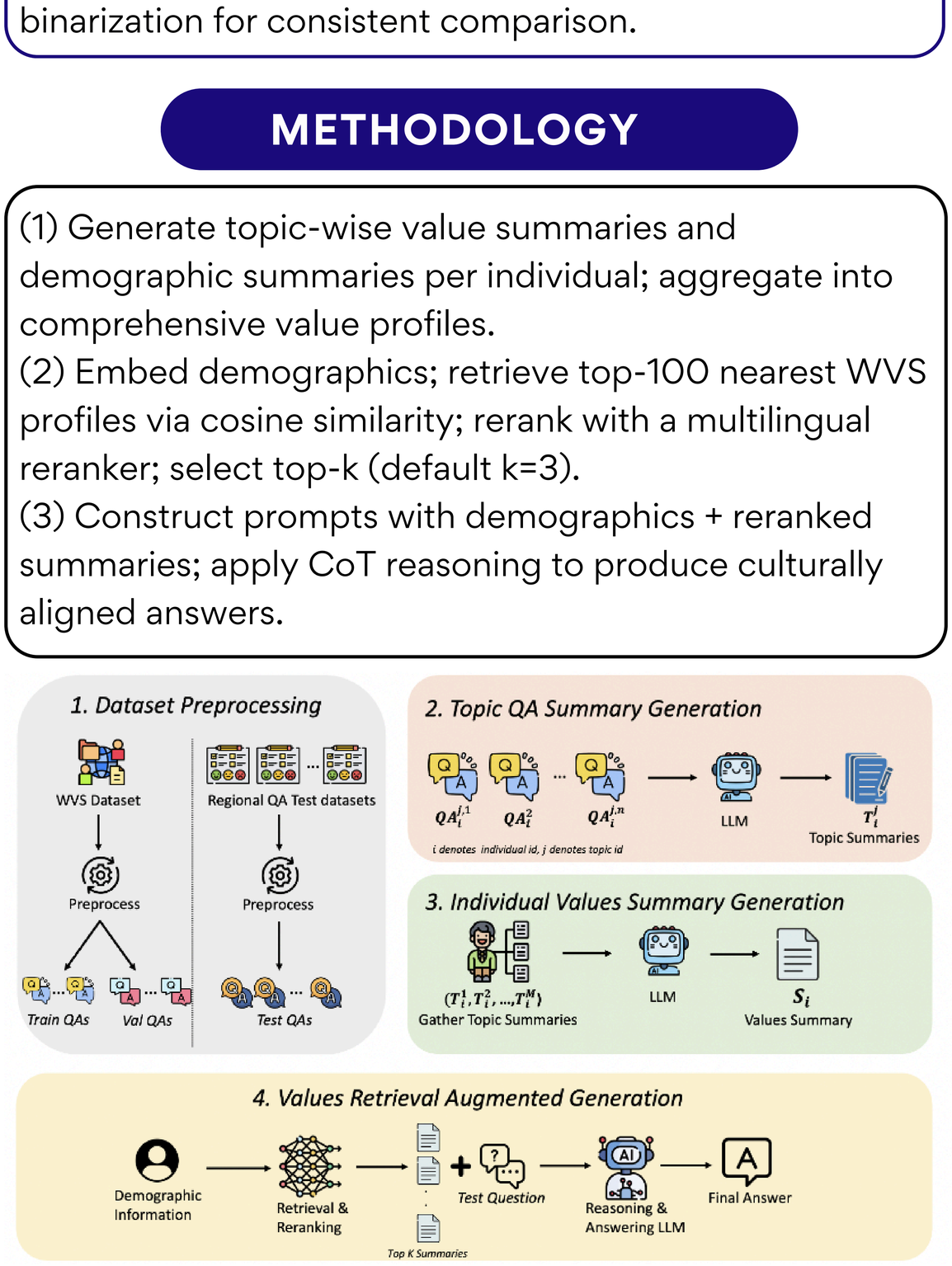
DATASET

Category	Dataset	Abbreviation	Region	Year	N	VQ
Retrieval Corpus	World Values Survey	WVS	Global	2017–2022	97.2k	259
Test Datasets	European Values Study	EVS	Europe	2017	59.4k	211
	The General Social Survey	GSS	North America	2021–2022	8.2k	44
	Chinese General Social Survey	CGSS	East Asia	2021	8.1k	58
	India Survey Dataset	ISD	South Asia	2019–2020	30.0k	33
	AmericasBarometer	LAPOP	Latin America	2021	59.1k	48
	Afrobarometer	Afrobarometer	Africa	2022	48.1k	144

Retrieval corpus: WVS Wave 7 (2017–2022), 97.2k respondents, 259 values questions, 31 demographics, covering 120 countries.
Test sets: EVS, GSS, CGSS, ISD, LAPOP, Afrobarometer, which are regionally representative surveys aligned temporally with WVS. We generate individual value and demographic summaries, then evaluate accuracy on values-related multiple-choice questions after binarization for consistent comparison.

METHODOLOGY

(1) Generate topic-wise value summaries and demographic summaries per individual; aggregate into comprehensive value profiles.
(2) Embed demographics; retrieve top-100 nearest WVS profiles via cosine similarity; rerank with a multilingual reranker; select top-k (default k=3).
(3) Construct prompts with demographics + reranked summaries; apply CoT reasoning to produce culturally aligned answers.



RESULTS

Model	Methods	EVS	GSS	CGSS	ISD	LAPOP	Africa	Avg. Accura
GPT-4o mini	Zero-shot Inference	0.5566	0.6026	0.4019	0.6109	0.4195	0.3923	0.4973
	Role-Assignment (2024)	0.5738	0.7564	0.4813	0.6164	0.4742	0.5563	0.5764
	Few-Shot Learning (2024)	0.5271	0.6538	0.4631	0.5804	0.4220	0.4258	0.5120
	Hybrid Method	0.5938	0.7292	0.5048	0.6330	0.4414	0.5305	0.5721
	ValuesRAG†	**0.6021***	**0.7781***	**0.5387***	**0.7001***	**0.5030***	**0.5953***	**0.6195***
emini 1.5 Flash	Zero-shot Inference	0.5419	0.6408	0.4502	0.6017	0.4149	0.4181	0.5113
	Role-Assignment (2024)	0.5598	0.7493	0.4770	0.6048	**0.4747**	0.5262	0.5653
	Few-Shot Learning (2024)	0.5225	0.6376	0.4559	0.5782	0.4194	0.4758	0.5149
	Hybrid Method	0.5845	0.7193	0.5026	0.6253	0.4448	0.5166	0.5655
	ValuesRAG†	**0.5869**	**0.7686***	**0.5337***	**0.6789***	0.4705	**0.5473***	**0.5977***

Models: (1) GPT-4o-mini and (2) Gemini-1.5-Flash.
Baselines: zero-shot, role-assignment, few-shot, and hybrid.
ValuesRAG achieves highest average accuracy across EVS, GSS, CGSS, ISD, LAPOP, Africa.

ABLATION STUDY

Model	Num(K)	EVS	GSS	CGSS	ISD	LAPOP	Africa	Avg. Accuracy
GPT-4o mini	1	0.5960	0.7722	0.5347	0.6853	0.4682	0.5905	0.6078
	3	0.6021	**0.7781**	**0.5387**	0.7001	0.5030	**0.5953**	**0.6195**
	5	**0.6052**	0.7706	0.5301	**0.7016**	**0.5061**	0.5905	0.6174
	10	0.6020	0.7380	0.5317	0.7014	0.5030	0.5680	0.6074
Gemini 1.5 Flash	1	0.5753	0.7668	0.5272	0.6646	0.4548	0.5369	0.5876
	3	**0.5869**	0.7686	**0.5337**	**0.6789**	**0.4705**	0.5473	**0.5977**
	5	0.5868	**0.7690**	0.5303	0.6734	0.4661	**0.5498**	0.5959
	10	0.5852	0.7665	0.5279	0.6773	0.4509	0.5464	0.5924

Model	Num(K)	Avg. Accuracy
GPT-4o mini	1	0.6078
	3	**0.6195**
	5	0.6174
	10	0.6074
Gemini 1.5 Flash	1	0.5876
	3	**0.5977**
	5	0.5959
	10	0.5924

(1) k=3 balances diversity and relevance, yielding best average accuracy for both base models; k=1 under-diversifies; k≥5 introduces noise and latency.
(2) Values-only generation (no demographics) still outperforms all baselines on WVS validation, indicating structured value summaries alone provide strong cultural grounding and generalization.

CONCLUSION

ValuesRAG provides a scalable and robust framework for aligning LLMs with diverse cultural values through retrieval-augmented and context-aware generation. By integrating survey-based values summaries with demographic information, it moves beyond static role labels and limited few-shot examples to achieve richer, more inclusive reasoning. Experiments across six regional datasets demonstrate consistent improvements over established baselines, with ablation studies confirming robustness under different retrieval settings. Overall, ValuesRAG offers a practical pathway toward culturally sensitive AI, with implications for policymaking, social research, and ethical global deployment.