Toward Culturally Aligned LLMs through Ontology-Guided Multi-Agent Reasoning

Wonduk Seo [1] **Wonseok Choi** [2 *] **Junseo Koh** [2 *] **Juhyeon Lee** [2] **Hyunjin An** [1] **Minhyeong Yu** [1] **Jian Park** [3] **Qingshan Zhou** [2] **Seunghyun Lee** [1] **Yi Bu** [2]

Abstract

Large Language Models (LLMs) increasingly support culturally sensitive decision making, yet often exhibit misalignment due to skewed pre-training data and the absence of structured value representations. Existing methods can steer outputs, but often lack demographic grounding and treat values as independent, unstructured signals, reducing consistency and interpretability. We propose **OG-MAR**, an **O**ntology-**G**uided **M**ulti-**A**gent **R**easoning framework. **OG-MAR** summarizes respondent-specific values from the World Values Survey (WVS) and constructs a global cultural ontology by eliciting relations over a fixed taxonomy via competency questions. At inference time, it retrieves ontology-consistent relations and demographically similar profiles to instantiate multiple value-persona agents, whose outputs are synthesized by a judgment agent that enforces ontology consistency and demographic proximity. Experiments on regional social-survey benchmarks across four LLM backbones show that **OG-MAR** improves cultural alignment and robustness over competitive baselines, while producing more transparent reasoning traces[1].

1. Introduction

Large Language Models (LLMs) are predominantly trained on web-scale corpora that are unevenly distributed across regions and sociocultural contexts (Bender et al., 2021; Dodge et al., 2021; Achiam et al., 2023; Jiang et al., 2023; Touvron et al., 2023). As a result, they often inherit culture-default biases, prioritizing high-resource Western-centric viewpoints while underrepresenting diverse cultural value systems (Durmus et al., 2023; Gallegos et al., 2024; Xie et al., 2024). These biases lead to systematic misalignment in culturally sensitive tasks, particularly ones involving social norms and value-based decisions (Karinshak et al., 2024; Pistilli et al., 2024; Tao et al., 2024). In response, several countries and organizations have developed localized LLMs to better reflect region-specific values (Zeng et al., 2022; Sengupta et al., 2023; Avramidis et al., 2024; Nguyen et al., 2024; Yoo et al., 2024). Nevertheless, cultural bias and value misalignment continue to pose significant challenges for real-world LLM deployments (Hershcovich et al., 2022; Kreutzer et al., 2022).

Prior work has proposed several strategies to reduce cultural bias in LLMs. Role-assignment methods (Tao et al., 2024) steer behavior by specifying culturally grounded personas, while few-shot prompting uses curated cultural exemplars to guide generation (Choenni & Shutova, 2024). Retrieval-based approaches such as ValuesRAG (Seo et al., 2025) further ground outputs in external survey evidence to better match cultural preferences. More recently, multi-agent frameworks (Baltaji et al., 2024; Ki et al., 2025; Wan et al., 2025) simulate diverse viewpoints through agent interaction and deliberation. In particular, the debate-only framework (Ki et al., 2025) relies on iterative critique and refinement to improve cultural adaptability.

Despite their promise, existing approaches share several fundamental limitations: (1) they often depend on implicit cultural assumptions that are weakly grounded in empirical value distributions, making outputs brittle and sensitive to prompting choices; (2) even with external evidence, cultural values are frequently treated as independent signals, missing structural relationships and cross-topic dependencies; (3) aggregation and multi-agent methods can boost robustness and diversity but using multiple agents without concrete value structure or grounding often reduces interpretability, offering limited visibility into why specific viewpoints emerge.

To address these issues, we propose **OG-MAR**, an ontology-driven cultural reasoning framework that integrates structured value knowledge, demographic grounding, and multi-

[1]AI Research, Enhans, Seoul, South Korea [2]Department of Information Management, Peking University, Beijing, China [3]Department of Data Science, Fudan University, Shanghai, China. Correspondence to: Seunghyun Lee <seunghyun@enhans.ai>, Yi Bu <buyi@pku.edu.cn>.

Preprint. January 30, 2026.

[1]Demo and code available at `https://authorname55.github.io/OG-MAR/`

agent simulation. We use the World Values Survey (WVS) (Zhao et al., 2024) as an empirically grounded retrieval corpus capturing diverse value distributions across regions. Raw survey responses are converted into topic-aware value summaries, and a global cultural ontology is built using expert-designed Competency Questions (CQs) (Gruninger, 1995; Grüninger & Fox, 1995), region-stratified LLM reasoning, and human-guided consolidation. At inference time, we retrieve ontology-consistent value structures and demographically similar individuals, instantiate value-persona agents for culturally grounded reasoning, and synthesize their outputs via a principled judgment mechanism.

We evaluate our framework on six regional benchmarks from major social surveys spanning Asia, Europe, Africa, North America, Latin America, and India. Results show consistent gains in cultural alignment, robustness across question types, and interpretability of reasoning traces. Quantitative analyses and qualitative case studies further indicate that ontology-guided multi-agent simulation offers a scalable and reliable path to culturally aligned LLM inference.

2. Related Work

2.1. Evaluation of LLM's Cultural Alignment

Large Language Models (LLMs) exhibit strong linguistic ability, yet geographic and linguistic skew in pretraining data can embed dominant-region norms as implicit defaults, causing cultural misalignment and potential inequities across contexts (AlKhamissi et al., 2024). Prior work evaluated these biases using stereotype-focused benchmarks (Nadeem et al., 2021), open-ended generation measures (Dhamala et al., 2021), and task-level tests such as ambiguous question answering (Parrish et al., 2022). Later studies measured cultural alignment by comparing model outputs with representative value surveys like the World Values Survey (Haerpfer et al., 2020b). Recent benchmarks further assess culture-specific everyday knowledge (Chiu et al., 2024), cross-national norm adaptation under different cultural frames (Rao et al., 2025), and value structure in open-ended generations through cultural psychology lenses (Karinshak et al., 2024). Building on this line of work, we use six regionally diverse survey datasets to evaluate value alignment across a broad set of cultural contexts.

2.2. Mitigating cultural bias in LLMs

Cultural bias mitigation for Large Language Models has evolved from in-context prompting to structured, evidence-grounded methods. Cultural prompting steers outputs by specifying a cultural frame (Tao et al., 2024), while Anthropological Prompting adds richer context and reasoning for underrepresented personas (AlKhamissi et al., 2024). To reduce sensitivity to examples and language, self-alignment selects culturally aligned demonstrations for in-context learning (Choenni & Shutova, 2024). When demonstrations are limited, ValuesRAG retrieves cultural and demographic cues as external evidence (Seo et al., 2025). Agentic approaches further improve reliability and parity through multi-agent debate and planning–critique–refinement pipelines (Ki et al., 2025; Wan et al., 2025). Despite this shift toward grounding and multi-agent reasoning, most methods still model cultural knowledge as unstructured, motivating ontology engineering to capture explicit value relationships.

2.3. Ontology Engineering with LLMs

An ontology is a formal specification of domain concepts and their relations (Gruber, 1993), supporting consistent and interpretable retrieval, integration, and reasoning. METHONTOLOGY (Fernández-López et al., 1997), On-To-Knowledge (Sure et al., 2004), and NeOn (Suárez-Figueroa et al., 2011) provide structured lifecycles and emphasize reuse. Recent work uses LLMs for ontology extraction via zero-shot prompting (Babaei Giglou et al., 2023) or fine-tuning (Mateiu & Groza, 2023), and increasingly automates end-to-end development: CQbyCQ matches novice-level performance (Saeedizade & Blomqvist, 2024), while Memoryless CQbyCQ and Ontogenia improve context efficiency and reasoning quality (Lippolis et al., 2025). Despite the benefits of ontology-based structuring for stereotype reduction, it remains underused in cultural-bias mitigation. We therefore propose a multi-agent framework for ontology-aware reasoning and pseudo-answer simulation.

3. Proposed Framework

We propose **OG-MAR**, an ontology-driven cultural reasoning framework that (i) summarizes respondent values under a fixed taxonomy, (ii) constructs CQ-derived cross-category relations as an ontology, (iii) retrieves ontology triples and demographically similar profiles for a query, and (iv) performs multi-persona simulation with ontology-constrained final adjudication.

3.1. Data Preprocessing & Structuring

3.1.1. Topic-Aware Value Summary Generation

Large-scale surveys capture cultural values through diverse question types, varying response scales, and heterogeneous answer formats. Directly operating on raw survey answers risks conflating unrelated signals and amplifying noise. To address this issue, we generate structured value summaries aligned with a predefined taxonomy[2]. Each World Values Survey (WVS) respondent record is decomposed into (1)

[2]Detailed taxonomy process and results are provided in Appendix G.

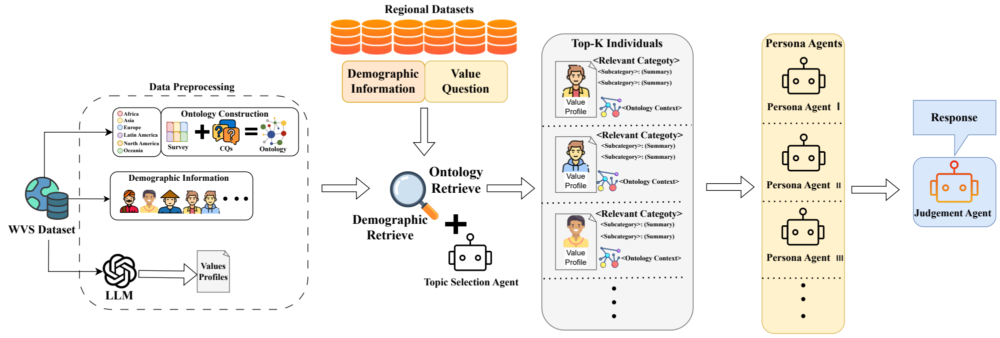

Figure 1. **Overall architecture of the OG-MAR framework.** The pipeline illustrates the overall architecture of **OG-MAR**. It begins with Data Preprocessing & Ontology Construction (left). During inference, for a given query and target demographics, it performs Ontology & Demographic Retrieval (center) to gather relevant context. This context is used to instantiate multiple Persona Agents (top right) whose outputs are synthesized by a Judgment Agent (bottom right) to produce the final, culturally aligned prediction.

demographic attributes and (2) values-related responses. Given the fixed ontology class set $C = \{c_1, \ldots, c_n\}$ with $|C| = n$, a Summarization Agent G_{sum} generates a concise, category-specific synopsis of the respondent's stance within the semantic scope of each class in the taxonomy. The agent is instructed to summarize only information relevant to the given class. Formally, let $\mathcal{R}_i$ denote the raw response set for individual i. For each value class $c_j \in C$, we obtain a category-conditioned synopsis:

$$s_{i,j} = G_{\text{sum}}(\mathcal{R}_i \mid c_j). \tag{1}$$

Aggregating over all categories yields a structured value profile:

$$V_i = \{s_{i,1}, s_{i,2}, \ldots, s_{i,n}\}. \tag{2}$$

Consequently, each individual is represented by a structured value profile that supports subsequent demographic grounding and persona simulation.

3.1.2. CQ-Guided Ontology Relation Construction

To model relationships between value categories, we adopt a human-guided ontology construction process based on Competency Questions (CQs)[3]. Domain experts curate CQs, each designed to probe meaningful interactions between two parent classes in the fixed taxonomy. For each CQ, we prompt a Large Language Model to describe *subclass-level* relationships between the two given parent classes. The model is constrained to: (i) use only the predefined taxonomy classes, (ii) avoid introducing any new classes, and (iii) focus solely on articulating relationships between the given parent classes.

Ontology Triple generation under cultural conditioning. To incorporate diverse cultural perspectives during ontology construction, we condition the LLM on value profiles sampled from 120 individuals (20 per region) spanning six major world regions. Each CQ yields candidate relational statements, represented as ordered sentence triples:

$$t_{a,b} = (c_a, p_{a,b}, c_b), \tag{3}$$

where c_a and c_b are natural-language sentences describing subclasses from the two queried parent domains, and $p_{a,b}$ is a natural-language relation sentence (distinct from p in top-p selection). Although these correspond to ontology classes and object properties, we express them as natural language sentences to maintain human interpretability and ensure alignment with the phrasing of the Competency Questions. We use parentheses to emphasize an ordered ontology triple of text rather than symbolic identifiers[4].

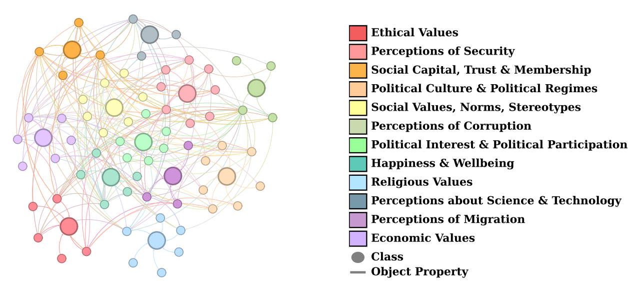

Figure 2. Visualization of the final ontology structure. The ontology comprises 76 classes and 150 pairs of object properties, forming a comprehensive semantic network.

Consolidation and human review. Compared to traditional ontology engineering, the taxonomy remains fixed: no classes are merged, split, or added. Human experts review candidate object properties by (1) validating cultural plausibility, (2) editing relation descriptions for clarity and

[3]Details of the ontology construction process and CQ examples are provided in Appendix G, Table 15.

[4]The ontology construction prompts are provided in Appendix E, Table 10.

consistency, and (3) removing spurious or inconsistent relations. The resulting ontology consists of the fixed class set

$$C = \{c_1, c_2, \ldots, c_{76}\} \tag{4}$$

and a curated ontology triple set

$$T = \{t_h\}_{h=1}^{|T|}, \quad t_h = (c_a, p_{a,b}, c_b). \tag{5}$$

Here P denotes the set of relation texts, and we index the curated ontology triples as $\{t_h\}$, where each t_h corresponds to a ontology triple of the form $t_{a,b}$ above. Thus $T \subseteq C \times P \times C$ is the curated subset of CQ-derived relations, and $|T|$ is determined by expert review rather than by the combinatorics of $|C|$.

3.2. OG-MAR Inference Pipeline

3.2.1. Query Analysis & Context Retrieval

Given an input query q and target respondent demographics d_q, we retrieve (i) ontology-consistent triples and (ii) demographically grounded respondent profiles, which jointly define the context for downstream multi-agent simulation. We use D_q for the top-k domains, F_q for the top-p fine-grained categories, O_q for retrieved ontology triples, R_q for retrieved respondents, and $\mathcal{V}_q$ for their value profiles; k, p, M, and K are the corresponding selection sizes (with $k = 1$ reducing to top-1 selection).

(a) Topic/category identification. A Topic-Selection Agent G_{topic} identifies relevant value domains. We implement G_{topic} as a pretrained text encoder fine-tuned on WVS data labeled with the 12 top-level value domains. Let $D = \{d_1, \ldots, d_{12}\}$ denote the domain set. The encoder outputs a logit score ℓ_u for each d_u, and the top-k domains are selected:

$$D_q = \{d_u \mid \ell_u \text{ is among the top-}k \text{ scores of } \{\ell_1, \ldots, \ell_{12}\}\}. \tag{6}$$

Within these domains, fine-grained categories are selected by computing similarity $\mathrm{sim}(\cdot, \cdot)$ between the query and each category description, where $\mathrm{sim}(\cdot, \cdot)$ denotes a similarity function (e.g., cosine or dot product). Let $\mathbf{e}_q$ denote the query embedding. The top-p subcategories form the fine-grained topic set $F_q \subseteq C$:

$$\begin{aligned} F_q = \{c_j \in C \mid & c_j \text{ belongs to a domain in } D_q \\ & \mathrm{sim}(\mathbf{e}_q, \mathbf{e}_{c_j}) \text{ is among the top-}p\}, \end{aligned} \tag{7}$$

where $\mathbf{e}_{c_j}$ is the embedding of the category text c_j.

(b) Ontology triple retrieval. We retrieve ontology knowledge in the form of triple, where each triple $t_h = (c_a, p_{a,b}, c_b)$ is treated as a single semantic unit. In this step, we *only* use node-level similarity *within the current class constraint* F_q. Specifically, for each ontology node (category) $c \in F_q$, we compute a node relevance:

$$\alpha(c) = \mathrm{sim}(\mathbf{e}_q, \mathbf{e}_c), \tag{8}$$

where $\mathbf{e}_c$ is the embedding of the category text c.

We then score each ontology triple by the relevance of its endpoint nodes:

$$\alpha_{\text{triple}}(t_h) = \max\big(\alpha(c_a), \alpha(c_b)\big). \tag{9}$$

Finally, we restrict retrieval to ontology triples whose endpoints are both in the current class set F_q, and select the top-M ontology triples by α_{triple} to form the ontology context O_q.

(c) Similar individual retrieval with dense embeddings. To ground reasoning in real-world perspectives, we retrieve individuals demographically similar to d_q using dense embedding retrieval. We encode demographic descriptions with an embedding model and rank respondents in I by embedding similarity to d_q. The top-K individuals form the demographic set R_q,

$$R_q = \{i_1, i_2, \ldots, i_K\}, \tag{10}$$

with corresponding value profiles

$$\mathcal{V}_q = \{V_i \mid i \in R_q\}. \tag{11}$$

3.2.2. Multi-Value Persona Agent Simulation

Given O_q and $\mathcal{V}_q$, we instantiate Value-Persona Agents that simulate culturally grounded reasoning under ontology constraints. We denote the agent as G_{persona} and instantiate it per retrieved individual $i \in R_q$. For convenience, index the retrieved ontology triples as $O_q = \{t_h\}_{h=1}^{|O_q|}$ and let $C_q = \{c_a, c_b \mid t_h \in O_q\}$ denote the set of classes referenced by the retrieved ontology triples. Concretely, each agent is conditioned on the ontology context O_q (typically 3–9 triples), the individual's value summaries restricted to C_q, and demographic attributes d_i. We denote the filtered profile as $V_{i,q} = \{s_{i,j} \mid c_j \in C_q\}$. The agent-specific conditioning context is:

$$z_i = \mathrm{Concat}(O_q, V_{i,q}, d_i). \tag{12}$$

Here z_i denotes the conditioning context for the persona associated with individual i.

Simulated reasoning trace. Given query q, each agent generates (i) an answer and (ii) an explicit simulated reasoning trace that forms a coherent chain of thought under the provided evidence and ontology constraint:

$$G_{\text{persona}}(q, z_{i,h}) = (\hat{y}_{i,h}, \rho_{i,h}). \tag{13}$$

Here $\hat{y}_{i,h}$ denotes the agent's answer and $\rho_{i,h}$ denotes its natural-language reasoning trace. We collect outputs across all retrieved individuals and ontology triples and denote the set by A:

$$A = \{G_{\text{persona}}(q, z_{i,h}) \mid i \in R_q,\ t_h \in O_q\}. \tag{14}$$

3.2.3. Ontology-Guided Final Judgment via Constrained Meta-Adjudication

A Final Judgment Agent G_{judge} synthesizes the final prediction by performing *Constrained Meta-Adjudication* over the candidate outputs from Value-Persona Agents. Given the multi-agent set A and the query q (question and response options), the judge outputs:

$$\hat{y}_q = G_{\text{judge}}(A, q). \tag{15}$$

Here $\hat{y}_q$ denotes the final prediction for query q. The judge does not receive O_q or $\mathcal{V}_q$ directly; ontology and profile grounding are carried through the persona outputs.

Compared to majority voting, G_{judge} uses a constrained, evidence-first protocol:

1. **Evidence & consistency.** For each $(\hat{y}_i, \rho_i)$, score grounding of ρ_i and ontology compliance, then aggregate scores per option y.

2. **Vote only under near-tie.** If top options are within a small margin, consult a Vote Summary as a secondary signal (otherwise ignore it).

3. **Relevance tie-break.** If still tied, choose the option supported by personas more relevant to d_q.

This yields evidence-weighted predictions and reduces response-count artifacts under weak or ambiguous evidence.

3.3. Implementation Details

Our implementation consists of: (i) a fixed ontology **taxonomy** with 12 top-level value domains and 76 fine-grained categories; (ii) a **Summarization Agent** $G_{\textbf{sum}}$, an LLM-based summarizer that generates category-specific summaries $s_{i,j}$ from raw respondent responses $\mathcal{R}_i$ under a "no-new-concepts" constraint; (iii) a **Topic-Selection Agent** $G_{\textbf{topic}}$, a pretrained text encoder fine-tuned on WVS with 12-domain supervision, which first selects the top-k domains and then the top-p fine-grained categories via semantic similarity; (iv) a **category selection** policy that retains the top-3 fine-grained value categories per query; (v) an **ontology triple retrieval** step that performs dense retrieval over ontology triples for each selected category and returns up to the top-3 triples per category; and (vi) a **persona retrieval setting** where the default number of retrieved individuals (personas) is set to K=5.

4. Experimental Design and Setup

4.1. Setup

Models Used. We use GPT-4o-mini (Achiam et al., 2023) and Gemini 2.5 (Google DeepMind, 2025) via APIs, and Qwen 2.5 (Team, 2024) and EXAONE 3.5 (An et al., 2024) as open-source models for the generation task in both our Persona Agent and Final Judgment Agent. We also used GPT-o4-mini (OpenAI, 2025) for object properties construction and values profile generation. To ensure stable behavior, we set the temperature to 0 across all models.

We use dense embedding retrieval with E5-base embeddings (Wang et al., 2022a). For demographic retrieval, we encode the target demographic description and each respondent's demographic profile, then rank respondents by embedding similarity to obtain the top-K demographically similar individuals. For ontology retrieval, we embed ontology triples and retrieve the top-M triples by similarity to the query.

Additionally, for topic classification, we fine-tune DeBERTa-v2-xxlarge (He et al., 2020) on WVS data for 3 epochs (batch size=4, learning rate=5×10^{-6}), following optimized configurations for value identification (Kiesel et al., 2023; Balikas, 2023)[5].

4.2. Datasets

Retrieval Corpus We use the *World Values Survey* (WVS) (Haerpfer et al., 2020a) as the retrieval corpus. WVS is a large-scale cross-national survey of human values and socio-cultural attitudes with structured demographic attributes (e.g., country, age, gender, education). In our setting, we use predefined 12 topics in the WVS[6], which provide a globally diverse and publicly available source of value-related responses and enable consistent retrieval of relevant demographic evidence for downstream inference.

Test Datasets To evaluate generalization beyond the retrieval corpus, we use six regional social-survey datasets with value-related questions and WVS-comparable demographic metadata: **EVS** (Europe), **GSS** (U.S.), **CGSS** (China), **ISD** (India), **LAPOP** (Latin America and the Caribbean), and **Afrobarometer** (Africa). We use clustering-based sampling to select 2,000 representative instances as test data, for efficient yet balanced evaluation across diverse datasets[7].

[5] Detailed training curves and performance on regional dataset are provided in Appendix A.

[6] The 12 topics are provided in Appendix B, Table 6.

[7] Detailed dataset descriptions and sampling process are provided in Appendix B.

Table 1. Accuracy of baseline methods across regional datasets. **Bold** text indicates the best performance, underlined text the second-best performance. $*$ denotes significant improvements (paired t-test with Holm–Bonferroni correction, $p < 0.05$) over all baseline model(s). † denotes our proposed method.

Method	EVS (*Europe*)	GSS (*United States*)	CGSS (*China*)	ISD (*India*)	AFRO (*Africa*)	LAPOP	Avg. Score
GPT-4o mini							
Zero-shot	0.5606	0.5164	0.5847	0.6139	0.5324	0.5760	0.5640
Role (2024)	0.5892	0.5184	0.6014	0.6060	0.5505	0.5674	0.5722
Self-consistency (2022b)	0.5558	0.4920	0.5631	0.5976	0.5224	0.5551	0.5477
Debate (2025)	0.5985	0.5509	0.5993	**0.6568**	0.5343	0.5306	0.5784
ValuesRAG (2025)	0.6127	**0.5589**	0.5889	0.6420	**0.5654**	0.6085	0.5961
OG-MAR (Ours)†	**0.6206***	0.5480	**0.6509***	0.6192	0.5389	**0.6268**	**0.6007**
Gemini 2.5 Flash Lite							
Zero-shot	0.5681	0.4957	0.6467	0.5000	0.5282	0.6225	0.5602
Role (2024)	0.5786	0.4992	0.6669	0.5521	0.5313	0.5852	0.5689
Self-consistency (2022b)	0.5489	0.4728	0.6063	0.4705	0.5182	0.6268	0.5406
Debate (2025)	0.5977	0.5138	0.6348	0.6335	0.5046	0.5331	0.5696
ValuesRAG (2025)	0.6075	0.5376	0.6084	0.6041	0.5472	0.5339	0.5731
OG-MAR (Ours)†	**0.6249***	**0.5489***	**0.7017***	**0.7007***	**0.5701***	**0.6385***	**0.6308**
QWEN 2.5							
Zero-shot	0.5199	0.5069	0.2704	0.7222	0.4814	0.4908	0.4986
Role (2024)	0.5357	0.5037	0.3463	**0.7452**	0.5014	0.4712	0.5172
Self-consistency (2022b)	0.5096	0.4975	0.3289	0.6278	0.4080	0.4975	0.4782
Debate (2025)	0.5511	0.5174	0.4578	0.6320	0.4875	0.4332	0.5132
ValuesRAG (2025)	0.5538	0.5215	0.4697	0.6591	0.4724	0.5268	0.5339
OG-MAR (Ours)†	**0.5898***	**0.5325***	**0.5220***	0.6599	**0.5180**	**0.6005**	**0.5705**
EXAONE 3.5							
Zero-shot	0.5143	0.5311	0.2885	0.6041	0.4054	0.5006	0.4740
Role (2024)	0.5319	0.5326	0.3129	0.6048	0.4077	0.4602	0.4750
Self-consistency (2022b)	0.5490	0.5266	0.2697	0.6122	0.4086	0.5368	0.4838
Debate (2025)	0.5713	0.5407	0.5624	0.6773	0.4995	0.4939	0.5575
ValuesRAG (2025)	0.5172	0.5520	0.5833	0.6446	0.4794	0.5913	0.5631
OG-MAR (Ours)†	**0.6080***	**0.5636**	**0.6307***	**0.7810***	**0.5045***	**0.7022***	**0.6317**

4.3. Evaluation Metrics

We cast all items into a **binary** decision task and report **accuracy**. For two-option questions, we require an exact match. For multi-choice questions, we either (i) map options into two fixed buckets or (ii) require an exact label match, depending on the dataset definition.

For ordinal response items, we bin raw responses r_i into a binary label b_i using the midpoint threshold m of the original scale:

$$b_i = \begin{cases} 0, & \text{if } r_i \le m \text{ (disagree)}, \\ 1, & \text{if } r_i > m \text{ (agree)}. \end{cases} \tag{16}$$

In addition, for ordinal items we report **MAE** (*Mean Absolute Error*) between the gold label y_i and prediction $\hat{y}_i$:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i| \,. \tag{17}$$

The average MAE scores across all regional datasets are provided in Figure 6.

4.4. Baselines

We compare against two single-pass prompting baselines, one single-agent aggregation baseline, and two multi-step baselines spanning multi-agent deliberation and retrieval augmentation: (1) ***Zero-shot***, a single prompt without scaffolding; (2) ***Role Assignment*** (Tao et al., 2024), which conditions generation on an explicit culturally grounded role; (3) ***Self-consistency*** (Wang et al., 2022b), which samples multiple reasoned outputs from the same model and takes a majority vote over them; (4) ***Debate*** (Ki et al., 2025), a multi-agent framework with iterative critique and refinement; and (5) ***ValuesRAG*** (Seo et al., 2025), which grounds generation in retrieved survey evidence. For a fair comparison, we use the same retrieval setting for *ValuesRAG* in the main experiments[8].

5. Experiment Results

Table 1 reports results on six regional benchmarks and four Large Language Model backbones, showing that baselines offer only incremental gains and fail under cultural distribution shift. **Zero-shot** Inference is efficient but often reverts to culture-default priors, yielding inconsistent value-sensitive judgments across regions. **Role Assignment** adds culturally framed prompting, yet remains weakly grounded in empirically observed value distributions and is sensitive to prompt formulation. **Self-Consistency** improves robustness via sample aggregation, but it does not enforce demographic

[8] Detailed baseline explanations are provided in Appendix C.

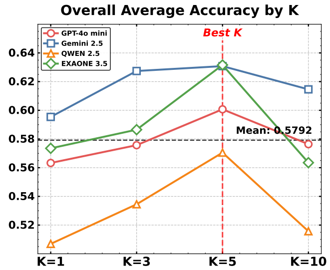

Figure 3. Performance comparison of four models across K $\in \{1, 3, 5, 10\}$ on average. Red vertical dashed lines indicate the best K and gray horizontal lines show the overall mean accuracy.

grounding or conceptual coherence and may amplify majority bias. **Debate** introduces critique-and-refinement, but without explicit evidence constraints it is prone to drift and still fails to capture cross-topic value dependencies. **ValuesRAG** is competitive by grounding generation in survey evidence but treats values as unstructured snippets, which restricts control over structured relationships.

In contrast, **OG-MAR** integrates ontology-guided triple retrieval and multi-persona simulation with a judgment agent that explicitly weighs ontology consistency and demographic proximity. This design yields strong and consistent performance across regions, achieving average accuracies of 0.6308 on Gemini 2.5 Flash Lite, 0.5705 on Qwen 2.5, and 0.6317 on EXAONE, while remaining competitive on GPT-4o-mini with an average accuracy of 0.6007. Notably, **OG-MAR** delivers particularly large gains on culturally challenging settings such as CGSS and ISD, suggesting that structured cultural relations and demographically grounded personas are most beneficial when the target distribution deviates from dominant pretraining priors.

5.1. Ablation Studies

5.1.1. Varying the Number of Retrieved Individuals

We investigate the impact of retrieval size K on **OG-MAR**'s performance by varying the number of retrieved demographically similar individuals across K $\in \{1, 3, 5, 10\}$. Figure 3 reports average accuracy for GPT-4o-mini, Gemini 2.5 Flash Lite, QWEN 2.5, and EXAONE 3.5.

All four models achieve their best overall performance at K=5, outperforming other retrieval sizes by +0.003 to +0.07 across different models. While Gemini 2.5 Flash Lite shows minimal difference between K=3 and K=5, the remaining models exhibit substantially larger improvements from K=3 to K=5, with gains ranging from 0.03 to 0.05. When retrieval size increases to K=10, all models show clear performance degradation, with accuracy drops ranging from 0.02 to 0.07 compared to K=5. These results reveal a clear trade-off: K=1 retrieves narrow value-persona signals, whereas K=5

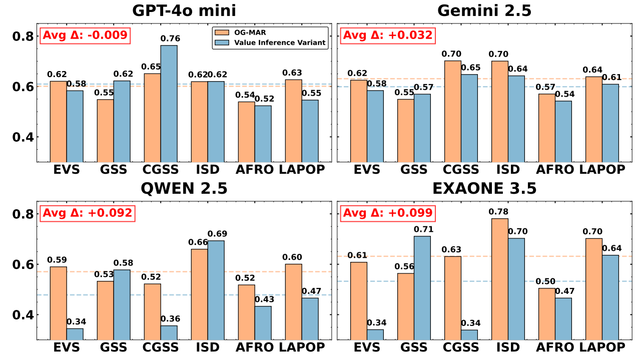

Figure 4. Performance comparison between **OG-MAR** and the Value Inference Variant. Accuracy over four models on six regional datasets. Dashed lines show per-method average accuracy; red boxes report the average gap (Avg Δ = **OG-MAR** - Variant).

offers richer yet stable context. Consequently, we adopt K=5 as the default retrieval depth throughout our experiments[9].

5.1.2. Impact of Value Inference Generation

To assess the impact of value inference generation, we compare our two step process with K personas and a judgment agent with a similar two step process where a single Value Inference Agent replaces the K persona agents by first inferring a value profile for the target individual and the same judgment agent answers questions based only on this profile. We evaluate both approaches across all six regional datasets using the same retrieval and setup as in our main experiments. Figure 4 presents the accuracy comparison between the two architectures.

Among the four models, GPT-4o-mini slightly benefits from the Value Inference Variant by around 0.01 on average, particularly showing higher accuracy on GSS (+0.07) and CGSS (+0.11). In contrast, the other three models achieve consistent gains with **OG-MAR**, ranging from +0.03 to +0.10 on average across datasets. While the Value Inference Variant outperforms **OG-MAR** on GSS across all four models, **OG-MAR** achieves clearly higher accuracy on the remaining datasets for most models. These results suggest that explicitly simulating multiple personas and preserving their distinct value profiles provides the judgment model with richer and more diverse evidence, which in turn allows **OG-MAR** to maintain stronger performance than the Value Inference Variant in most models and datasets.

5.1.3. Impact of Multi-Persona Reasoning

To evaluate the importance of multi-persona reasoning, we compare the full **OG-MAR** framework against a single-judge variant. Both variants receive the same input: K individuals' demographics, relevant value summaries, and

[9] Complete results for all regional datasets are in Appendix H.1.

Table 2. Average accuracy of the full **OG-MAR** framework and the single-judge variant across four LLMs.

Model	Method	Avg. Accuracy
GPT-4o mini	**OG-MAR**	**0.6007**
	Single-Judge	0.5987
Gemini 2.5	**OG-MAR**	**0.6308**
	Single-Judge	0.6022
QWEN 2.5	**OG-MAR**	**0.5705**
	Single-Judge	0.5311
EXAONE 3.5	**OG-MAR**	**0.6317**
	Single-Judge	0.5627

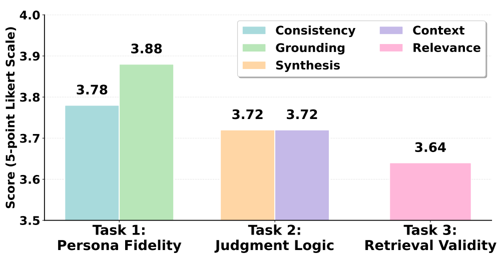

Figure 5. Average human evaluation scores (5-point Likert scale) across three tasks: Persona Fidelity (Consistency, Grounding), Judgment Logic (Synthesis, Context), and Retrieval Validity (Relevance). Scores are averaged over nine expert raters.

ontology context. However, the single-judge variant bypasses the persona-simulation stage and directly generates the final answer using a single judgment model. Following the same setup as Section 5.2.2, we evaluate the single-judge variant across all datasets. Table 2 presents the accuracy comparison between the two architectures.

Across models, **OG-MAR** outperforms the single-judge variant across all four models, with higher average accuracy by approximately 0.002 on GPT-4o mini, 0.03 on Gemini 2.5 Flash Lite, 0.04 on QWEN 2.5, and 0.07 on EXAONE 3.5. Although the single-judge variant consistently achieves higher scores on GSS across all four models, **OG-MAR** exhibits substantial gains on other datasets[10]. The comparison reveals that multi-persona reasoning provides consistent benefits across multiple regional datasets, while the single-judge variant remains competitive, suggesting that **OG-MAR**'s gains are not solely attributable to the persona-simulation layer but also rely on the underlying ontology-grounded retrieval and value summarization pipeline.

6. Discussion

Qualitative Analysis To complement quantitative results, we conducted a human evaluation of **OG-MAR**'s reasoning traces with nine domain experts using three 5-point Likert

[10]Detailed per-dataset results are provided in Appendix H.2.

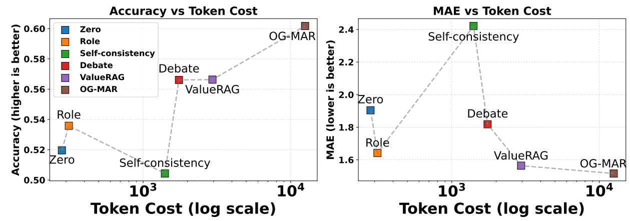

Figure 6. Performance–cost trade-off across methods. Left: accuracy vs total tokens (higher is better). Right: MAE vs total tokens (lower is better). Markers denote methods; the dashed line shows performance changes as token usage increases.

tasks: *Persona Fidelity* (Task 1), *Judgment Logic* (Task 2), and *Retrieval Validity* (Task 3). Overall, **OG-MAR** shows consistent interpretability across six regional datasets. Notably, it achieves the highest *Grounding* on **CGSS (China)** (4.02), slightly exceeding **GSS (U.S.)** (3.97), suggesting ontology-guided value injection mitigates "culture-default" tendencies by encouraging evidence-based reasoning. The Judgment Agent also attains strong *Synthesis Logic* (Avg. 3.72), indicating evidence-first adjudication remains effective across culturally distinct regions, while ontology retrieval maintains high relevance (Avg. 3.64), supporting that retrieved triples provide meaningful evidence for downstream reasoning[11].

Performance–Token Usage Trade-off. Figure 6 shows the trade-off between token cost and prediction quality across methods. We measure cost by the total number of input and output tokens, and use a log-scale x-axis to highlight large differences in usage. Performance is reported with Accuracy (higher is better) and MAE (lower is better). Single-pass prompting baselines lie in the lowest-cost regime but generally achieve weaker performance. RAG method baseline (ValuesRAG) and Multi-step methods that retrieve external context or aggregate multiple samples (e.g., Self-consistency, and Debate) consume substantially more tokens and typically improve quality. **OG-MAR** incurs the highest token budget yet achieves the best overall results, with the highest accuracy and lowest average MAE, indicating that the additional computation yields substantive gains rather than mere overhead.

7. Conclusion

In this paper, we presented **OG-MAR**, an ontology-guided multi-agent framework for culturally aligned LLM inference. **OG-MAR** combines CQ-driven cultural ontology construction with demographically-grounded retrieval from WVS, and performs value-persona simulation followed by a

[11]Detailed qualitative evaluation tasks, procedures, and results are provided in Appendix D, and the case study is provided in Appendix F.

judgment agent that enforces ontology consistency. Experiments across six regional benchmarks and four LLM backbones show that **OG-MAR** improves cultural alignment and robustness over competitive baselines, while providing more interpretable reasoning grounded in structured value relations.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

AlKhamissi, B., ElNokrashy, M., AlKhamissi, M., and Diab, M. Investigating cultural alignment of large language models. *arXiv preprint arXiv:2402.13231*, 2024.

An, S., Bae, K., Choi, E., Choi, K., Choi, S. J., Hong, S., Hwang, J., Jeon, H., Jo, G. J., Jo, H., et al. Exaone 3.5: Series of large language models for real-world use cases. *arXiv preprint arXiv:2412.04862*, 2024.

Avramidis, E., Grützner-Zahn, A., Brack, M., Schramowski, P., Suarez, P. O., Ostendorff, M., Barth, F., Manakhimova, S., Macketanz, V., Rehm, G., et al. Occiglot at wmt24: European open-source large language models evaluated on translation. In *Proceedings of the Ninth Conference on Machine Translation*, pp. 292–298, 2024.

Babaei Giglou, H., D'Souza, J., and Auer, S. Llms4ol: Large language models for ontology learning. In *International Semantic Web Conference*, pp. 408–427. Springer, 2023.

Balikas, G. John-arthur at semeval-2023 task 4: fine-tuning large language models for arguments classification. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pp. 1428–1432, 2023.

Baltaji, R., Hemmatian, B., and Varshney, L. R. Persona inconstancy in multi-agent llm collaboration: Conformity, confabulation, and impersonation. *arXiv preprint arXiv:2405.03862*, 2024.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623, 2021.

Chiu, Y. Y., Jiang, L., Lin, B. Y., Park, C. Y., Li, S. S., Ravi, S., Bhatia, M., Antoniak, M., Tsvetkov, Y., Shwartz, V., et al. Culturalbench: a robust, diverse and challenging benchmark on measuring (the lack of) cultural knowledge of llms. 2024.

Choenni, R. and Shutova, E. Self-alignment: Improving alignment of cultural values in llms via in-context learning. *arXiv preprint arXiv:2408.16482*, 2024.

Dhamala, J., Sun, T., Kumar, V., Krishna, S., Pruksachatkun, Y., Chang, K.-W., and Gupta, R. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 862–872, 2021.

Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D., Mitchell, M., and Gardner, M. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. *arXiv preprint arXiv:2104.08758*, 2021.

Durmus, E., Nguyen, K., Liao, T. I., Schiefer, N., Askell, A., Bakhtin, A., Chen, C., Hatfield-Dodds, Z., Hernandez, D., Joseph, N., et al. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*, 2023.

Fernández-López, M., Gómez-Pérez, A., and Juristo, N. Methontology: From ontological art towards ontological engineering. In *Proceedings of the Symposium on Ontological Engineering of AAAI*, pp. 33–40, Stanford, CA, USA, 1997. AAAI Press.

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., and Ahmed, N. K. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179, 2024.

Google DeepMind. Gemini 2.5 flash-lite model card. Technical report, Google DeepMind, 2025. Accessed: 2025-12-28.

Gruber, T. R. A translation approach to portable ontology specifications. *Knowledge acquisition*, 5(2):199–220, 1993.

Gruninger, M. Methodology for the design and evaluation of ontologies. In *Proc. IJCAI'95, Workshop on Basic Ontological Issues in Knowledge Sharing*, 1995.

Grüninger, M. and Fox, M. S. The role of competency questions in enterprise engineering. In *Benchmarking—Theory and practice*, pp. 22–31. Springer, 1995.

Haerpfer, C., Inglehart, R., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano, J., Lagos, M., Norris, P., Ponarin, E., and Puranen, B. World values survey wave 7 (2017-2020) cross-national data-set. *(No Title)*, 2020a.

Haerpfer, C., Inglehart, R., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano, J., Lagos, M., Norris, P., Ponarin, E., and Puranen, B. World values survey wave 7 (2017-2020) cross-national data-set. *(No Title)*, 2020b.

He, P., Liu, X., Gao, J., and Chen, W. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020.

Hershcovich, D., Frank, S., Lent, H., De Lhoneux, M., Abdou, M., Brandl, S., Bugliarello, E., Piqueras, L. C., Chalkidis, I., Cui, R., et al. Challenges and strategies in cross-cultural nlp. *arXiv preprint arXiv:2203.10020*, 2022.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Karinshak, E., Hu, A., Kong, K., Rao, V., Wang, J., Wang, J., and Zeng, Y. Llm-globe: A benchmark evaluating the cultural values embedded in llm output. *arXiv preprint arXiv:2411.06032*, 2024.

Ki, D., Rudinger, R., Zhou, T., and Carpuat, M. Multiple llm agents debate for equitable cultural alignment. *arXiv preprint arXiv:2505.24671*, 2025.

Kiesel, J., Alshomary, M., Mirzakhmedova, N., Heinrich, M., Handke, N., Wachsmuth, H., and Stein, B. Semeval-2023 task 4: Valueeval: Identification of human values behind arguments. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pp. 2287–2303, 2023.

Kreutzer, J., Caswell, I., Wang, L., Wahab, A., Van Esch, D., Ulzii-Orshikh, N., Tapo, A., Subramani, N., Sokolov, A., Sikasote, C., et al. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72, 2022.

Lippolis, A. S., Saeedizade, M. J., Keskisärkkä, R., Zuppiroli, S., Ceriani, M., Gangemi, A., Blomqvist, E., and Nuzzolese, A. G. Ontology generation using large language models. In *European Semantic Web Conference*, pp. 321–341. Springer, 2025.

Mateiu, P. and Groza, A. Ontology engineering with large language models. In *2023 25th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, pp. 226–229. IEEE, 2023.

Nadeem, M., Bethke, A., and Reddy, S. Stereoset: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pp. 5356–5371, 2021.

Nguyen, X.-P., Zhang, W., Li, X., Aljunied, M., Hu, Z., Shen, C., Chia, Y. K., Li, X., Wang, J., Tan, Q., et al. Seallms-large language models for southeast asia. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 294–304, 2024.

OpenAI, O. o3 and o4-mini system card. *Preprint*, 2025.

Parrish, A., Chen, A., Nangia, N., Padmakumar, V., Phang, J., Thompson, J., Htut, P. M., and Bowman, S. Bbq: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2086–2105, 2022.

Pistilli, G., Leidinger, A., Jernite, Y., Kasirzadeh, A., Luccioni, A. S., and Mitchell, M. Civics: Building a dataset for examining culturally-informed values in large language models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pp. 1132–1144, 2024.

Rao, A. S., Yerukola, A., Shah, V., Reinecke, K., and Sap, M. Normad: A framework for measuring the cultural adaptability of large language models. 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)*, pp. 2373–2403, 2025.

Saeedizade, M. J. and Blomqvist, E. Navigating ontology development with large language models. In Meroño-Peñuela, A., Dimou, A., Troncy, R., Hartig, O., Acosta, M., Alam, M., Paulheim, H., and Lisena, P. (eds.), *The Semantic Web - 21st International Conference, ESWC 2024, Hersonissos, Crete, Greece, May 26-30, 2024, Proceedings, Part I*, volume 14664 of *Lecture Notes in Computer Science*, pp. 143–161. Springer, 2024. doi: 10.1007/978-3-031-60626-7_8. URL https://doi.org/10.1007/978-3-031-60626-7_8.

Sengupta, N., Sahu, S. K., Jia, B., Katipomu, S., Li, H., Koto, F., Marshall, W., Gosal, G., Liu, C., Chen, Z., et al. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*, 2023.

Seo, W., Yuan, Z., and Bu, Y. Valuesrag: Enhancing cultural alignment through retrieval-augmented contextual learning. *arXiv preprint arXiv:2501.01031*, 2025.

Suárez-Figueroa, M. C., Gómez-Pérez, A., Motta, E., and Gangemi, A. Introduction: Ontology engineering in a networked world. In *Ontology engineering in a networked world*, pp. 1–6. Springer, 2011.

Sure, Y., Staab, S., and Studer, R. On-to-knowledge methodology (otkm). In *Handbook on ontologies*, pp. 117–132. Springer, 2004.

Tao, Y., Viberg, O., Baker, R. S., and Kizilcec, R. F. Cultural bias and cultural alignment of large language models. *PNAS nexus*, 3(9):pgae346, 2024.

Team, Q. Qwen2.5: A party of foundation models, September 2024. URL `https://qwenlm.github.io/blog/qwen2.5/`.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Wan, Y., Chen, X., and Chang, K.-W. Which cultural lens do models adopt? on cultural positioning bias and agentic mitigation in llms. *arXiv preprint arXiv:2509.21080*, 2025.

Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., Majumder, R., and Wei, F. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022a.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022b.

Xie, C., Chen, C., Jia, F., Ye, Z., Lai, S., Shu, K., Gu, J., Bibi, A., Hu, Z., Jurgens, D., et al. Can large language model agents simulate human trust behavior? *Advances in neural information processing systems*, 37:15674–15729, 2024.

Yoo, K. M., Han, J., In, S., Jeon, H., Jeong, J., Kang, J., Kim, H., Kim, K.-M., Kim, M., Kim, S., et al. Hyperclova x technical report. *arXiv preprint arXiv:2404.01954*, 2024.

Zeng, A., Liu, X., Du, Z., Wang, Z., Lai, H., Ding, M., Yang, Z., Xu, Y., Zheng, W., Xia, X., et al. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*, 2022.

Zhao, W., Mondal, D., Tandon, N., Dillion, D., Gray, K., and Gu, Y. Worldvaluesbench: A large-scale benchmark dataset for multi-cultural value awareness of language models. *arXiv preprint arXiv:2404.16308*, 2024.

A. Training Details and Loss Curves

A.1. DeBERTa-v2-xxlarge Fine-tuning

We fine-tuned DeBERTa-v2-xxlarge (He et al., 2020) on the WVS category classification task (191 questions, 12 categories) using batch size 4, learning rate 5×10^{-6}, AdamW optimizer (weight decay 0.01), and FP16 precision on an NVIDIA A100 GPU. Training terminated early at epoch 3 when validation Top-3 accuracy reached 100%, as shown in Figure 7.

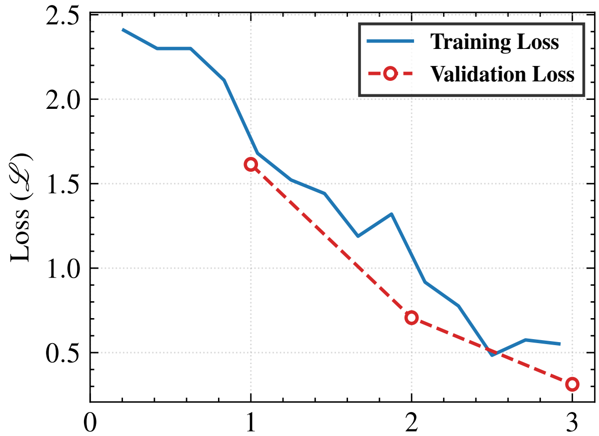

Figure 7. **Training and validation loss curves for DeBERTa-v2-xxlarge fine-tuning.** The x-axis represents epochs. Training loss (blue solid line) exhibits minor fluctuations typical of small-batch optimization, while validation loss (red dashed line, evaluated every 48 steps) decreases monotonically from 1.61 to 0.31 across three epochs, indicating effective learning without overfitting.

The early stopping criterion based on Top-3 accuracy ensures computational efficiency. Despite minor fluctuations in training loss during the final epoch, validation loss decreases consistently across all epochs on the WVS dataset.

A.2. Value Category Classification

Table 3. **Topic classification performance on six regional test datasets and WVS validation data.** Top-k: fraction of questions where the true category appears in top-k predictions. All metrics in [0,1].

Dataset	Top-1	Top-2	Top-3
Afrobarometer	0.5037	0.6875	0.7574
CGSS	0.3375	0.5079	0.6656
EVS	0.4315	0.5560	0.6680
GSS	0.4545	0.6667	0.7765
ISD	0.5439	0.7071	0.7950
LAPOP	0.4396	0.6577	0.7349
WVS (val)	**0.9583**	**1.0000**	**1.0000**

Table 4. Summary statistics of the retrieval corpus (WVS) and six test datasets. For test datasets, "#Value Qs" denotes the value-related items retained after our preprocessing/topic mapping (not necessarily the full questionnaire length).

Dataset	Type	Region	Wave / Year	#Countries	#Respondents	#Value Qs
Retrieval Corpus						
WVS (World Values Survey)	Retrieval	Global	2017–2022	64	94,728	239
Test Datasets						
EVS (European Values Study)	Test	Europe	2017	–	59,438	211
GSS (General Social Survey)	Test	United States	2021–2022	–	8,181	44
CGSS (Chinese General Social Survey)	Test	China (E. Asia)	2021	–	∼8,148	58
ISD (Pew India Survey Dataset)	Test	India (S. Asia)	2019–2020	–	29,999	33
LAPOP (AmericasBarometer)	Test	Latin America	2021	–	64,352	48
Afrobarometer	Test	Africa	2022	–	∼48,100	144

Table 5. Data sources used in our experiments. We use the World Values Survey (WVS) as the retrieval corpus, and evaluate generalization on six external test datasets (EVS, GSS, CGSS, ISD, LAPOP, and Afrobarometer), with official access links provided for reproducibility.

Dataset	Link
Retrieval Corpus	
WVS	https://www.worldvaluessurvey.org/wvs.jsp
Test Datasets	
EVS (European Values Study)	https://europeanvaluesstudy.eu
GSS (General Social Survey)	https://gss.norc.org
CGSS (Chinese General Social Survey)	https://cgss.ruc.edu.cn
ISD (Pew India Survey Dataset)	https://www.pewresearch.org/dataset/india-survey-dataset/
LAPOP (AmericasBarometer)	https://www.vanderbilt.edu/lapop
Afrobarometer	https://www.afrobarometer.org

B. Dataset Details

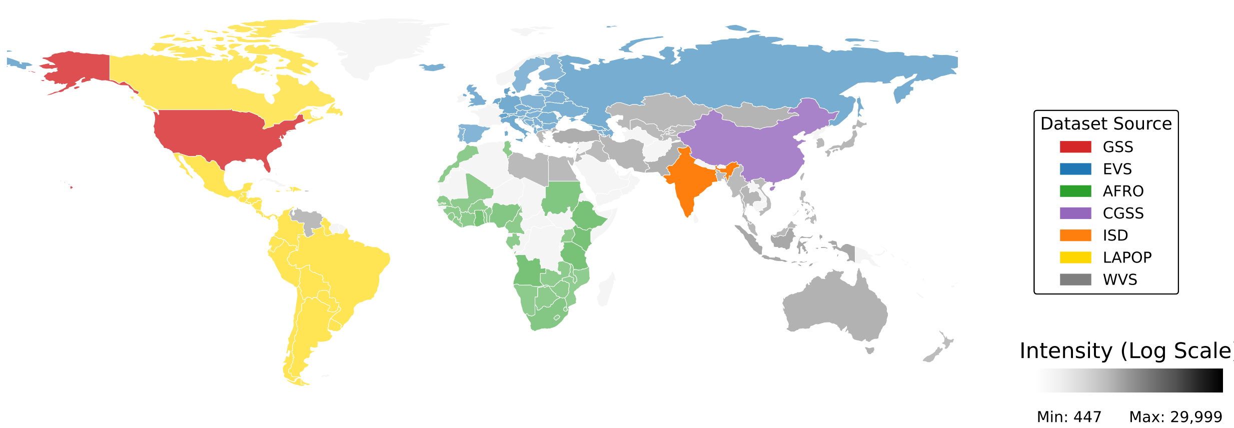

Figure 8. Geographic coverage of cultural value datasets used in this study. Each country is colored according to its primary data source, prioritizing regional surveys over the global World Values Survey. Regional datasets include the General Social Survey for the United States, the European Values Study for Europe, Afrobarometer for Africa, the Chinese General Social Survey for China, and the India Social Dataset for India. Countries without regional coverage are represented by WVS data shown in gray. Color intensity reflects participant count on a logarithmic scale, ranging from 447 to 29,999 respondents per country. This multimodal approach ensures both regional specificity and global breadth in cultural alignment research.

Table 6. **Distribution of Values-related Questions in WVS.** The questions were categorized into 12 topics with a total of 250 questions in the original WVS questionnaire covering most of the dimensions of values.

Topic	Count
Social Values, Norms, Stereotypes	45
Happiness and Wellbeing	11
Social Capital, Trust and Organizational Membership	47
Economic Values	6
Perceptions of Corruption	9
Perceptions of Migration	10
Perceptions of Security	21
Perceptions about Science and Technology	6
Religious Values	12
Ethical Values	23
Political Interest and Political Participation	35
Political Culture and Political Regimes	25

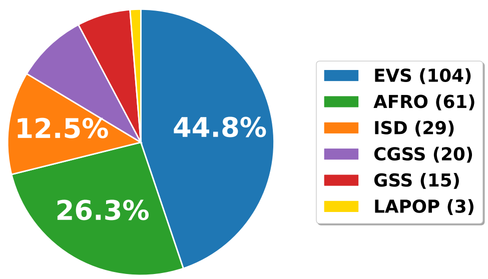

Figure 9. Distribution of selected value questions across regional datasets.

To ensure the quality and relevance of our value questions, we employed a systematic selection process for the regional datasets. First, we considered the actual response rates from our extracted representative samples (see Section B.2) and retained the top 80% of questions with valid responses in each dataset. This ensured sufficient data coverage for meaningful analysis. Second, we excluded questions that were overly dependent on personal circumstances and behaviors, such as "Did your household ever run out of water?", "In the past year, how often have you used radio?", "Are you a member of a trade union?", or "Based on your experience, how easy or difficult is it to find out how government uses tax revenues?". These questions tend to reflect personal circumstances rather than underlying values. Third, we excluded questions requiring knowledge of specific countries or domestic institutions, such as "To what extent do you think [country] is democratic?" or "How much do you trust the Electoral Commission of Ghana?". It would be difficult to generalize these questions across cultural contexts. After applying these criteria, we obtained a final set of questions distributed across the six regional datasets. Figure 9 shows the distribution of these selected questions across datasets.

B.1. Dataset Information

Retrieval Corpus: World Values Survey (WVS). We use the *World Values Survey* (WVS) as our retrieval corpus. In this setting, WVS covers 64 countries/territories with 94,728 respondents and a 290-item common questionnaire. We organize the value space into 12 predefined topics (Table 6) and, after preprocessing, retain 239 region-agnostic ordinal value items for retrieval-augmented inference.

Test Datasets. To evaluate performance of **OG-MAR**, we use six regional social-survey datasets spanning Europe, North America, East Asia, South Asia, and Africa. For each test dataset, the reported number of value questions refers to the subset retained after our preprocessing/topic mapping (not necessarily the full questionnaire length).

EVS (European Values Study). EVS is a cross-national survey program designed to measure human values and socio-cultural attitudes across Europe using harmonized instruments. We use the EVS 2017 integrated dataset, which combines nationally representative samples from 36 countries (59,438 respondents) collected during the 2017–2021 fieldwork period.

In our evaluation, we retain 211 value-related items under our preprocessing and topic mapping, enabling within-Europe generalization tests under relatively consistent survey design and documentation.

GSS (General Social Survey). GSS is a long-running repeated cross-sectional survey for the United States, featuring a replicating core plus rotating topical modules and rich demographics. We use the 2021 and 2022 cross-sections (4,032 and 4,149 completes; 8,181 total), which reflect a major methodological transition (e.g., 2021 push-to-web design and 2022 mixed-mode transition). We retain 44 value-related items, making GSS a useful stress test for robustness under both cultural shift (vs. the global WVS corpus) and survey-mode/design differences across waves.

CGSS (Chinese General Social Survey). CGSS is a nationally representative household survey measuring social attitudes and values in China, accompanied by detailed demographic covariates. We use the 2021 release with 8,148 valid samples (often documented as being drawn nationwide across many communities/provinces). After preprocessing and topic alignment, we retain 58 value-related items, providing a linguistically and institutionally distinct setting for cross-cultural generalization beyond the WVS retrieval corpus.

ISD (Pew India Survey Dataset). To represent South Asia, we use Pew Research Center's India Survey Dataset, a nationally representative face-to-face survey administered from Nov. 2019 to Mar. 2020. The dataset contains 29,999 adult interviews and covers broad attitudinal themes (e.g., identity, nationalism, tolerance), alongside demographics and survey weights. We retain 33 value-related items under our mapping, supporting evaluation in a highly heterogeneous population with strong methodological transparency.

LAPOP (AmericasBarometer). LAPOP is a large-scale cross-national public opinion survey program in Latin America that measures citizens' political attitudes, democratic governance, institutional trust, and socio-economic perceptions with harmonized instruments and rich demographics. We use the 2021 AmericasBarometer release (64,352 respondents) and, after preprocessing and topic alignment, retain 48 value-related items, providing a regionally distinct testbed for evaluating generalization beyond the WVS retrieval corpus.

Afrobarometer. Afrobarometer conducts nationally representative surveys of adult citizens (18+) across African countries using probability sampling, typically with per-country sample sizes of about 1,200 or 2,400. We use the merged Round 8 release (34 countries; released as a 2022 merged dataset) and retain 144 value-related items. The merged file contains on the order of ∼48K respondents (depending on country-level sample sizes), providing a stringent test bed with substantial cross-country diversity in socio-economic and governance contexts.

B.2. Extract Representative Sample to Cluster

Representative Sample To ensure computational efficiency while maintaining representativeness across regional datasets, we extract 2,000 representative instances with balanced coverage through clustering-based sampling. Concretely, we use **Faiss-based k-means clustering** to learn k clusters in the embedding space under Euclidean (L2) distance and visualize the resulting structure with a **Voronoi partition** induced by the cluster centroids, optimizing the standard k-means objective:

$$\min_{\{c_j\}_{j=1}^{k}} \sum_{i=1}^{N} \|x_i - c_{z_i}\|_2^2 , \tag{18}$$

where $x_i \in \mathbb{R}^d$ denotes the embedding of the i-th sample, $c_j \in \mathbb{R}^d$ is the centroid of cluster j, and $z_i \in \{1, \ldots, k\}$ is the cluster assignment of x_i. After training, we extract the **centroids** $\{c_j\}_{j=1}^{k}$ as representative cluster prototypes. We then project the centroids to two dimensions and construct a **Voronoi diagram** based on nearest-centroid relations in the 2D plane, which provides an intuitive view of how cluster centers are arranged and how the space is partitioned around them.

Figure 10 shows results for six datasets (AFRO, CGSS, EVS, GSS, ISD, and LAPO). For each dataset, we run Faiss k-means with the specified k and plot the **centroids** ($\times$) together with the corresponding **Voronoi cells** (light-colored polygons). The small dots are individual embedding samples, colored by their k-means cluster assignments z_i. In each subplot, k denotes the number of clusters (and thus centroids), while $regions$ reports the number of Voronoi cells formed in the 2D projection. In our outputs, $regions$ matches k, indicating that each centroid produces one Voronoi region. Overall, we extract 2,000 centroid prototypes across the six datasets, and use these prototypes as a compact set of representative points for subsequent analysis. To preserve representativeness across the regional datasets, we apply a clustering-based

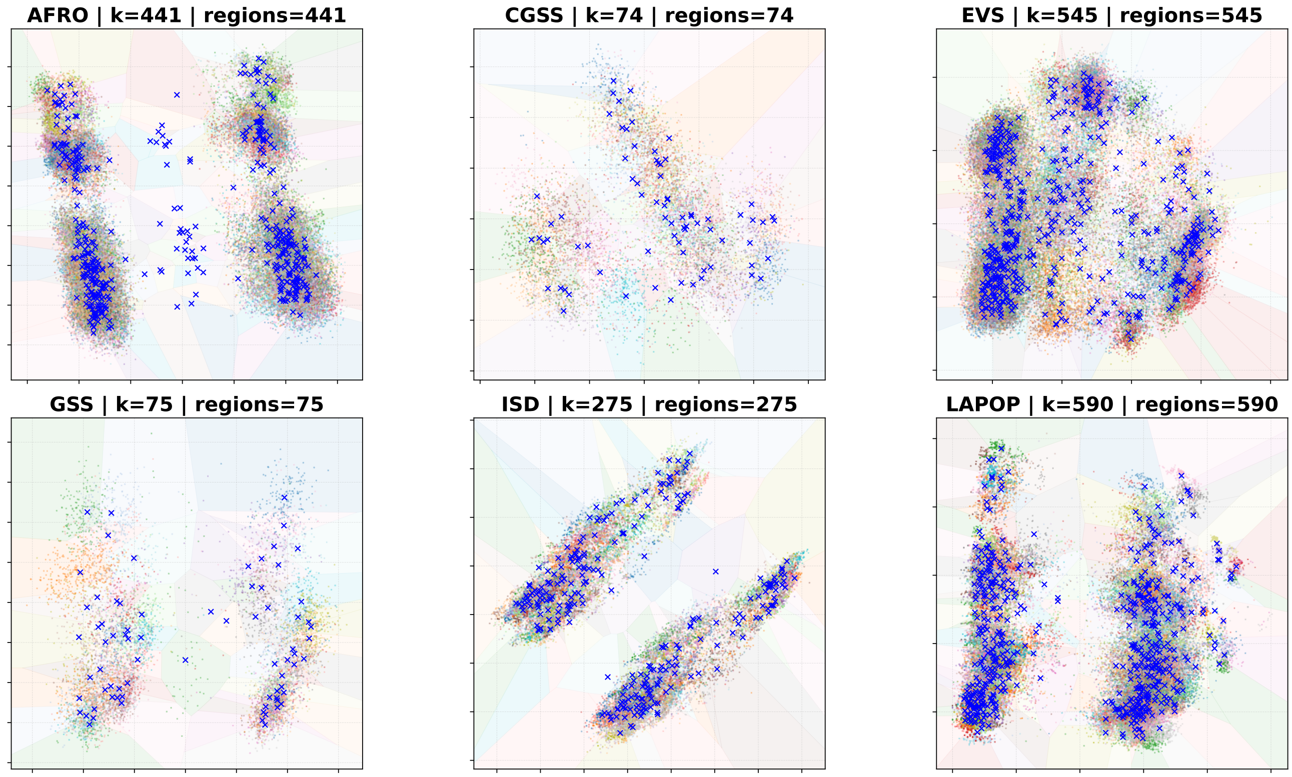

Figure 10. Voronoi visualization of Faiss k-means centroids for six embedding datasets. Blue crosses denote cluster centroids, colored dots indicate embedded samples, and light polygons show Voronoi regions in a 2D projection, providing an intuitive overview of the spatial distribution and structure of the embedding space across datasets.

sampling strategy with a fixed sampling rate for each dataset. This procedure yields 441 instances from AFRO, 74 from CGSS, 545 from EVS, 75 from GSS, 275 from ISD, and 590 from LAPOP (2,000 instances in total).

This centroid-based summary is reasonable because k-means centroids act as prototypes that compactly approximate the embedding distribution. By minimizing within-cluster squared distances, the learned centroids capture the central tendency of dense areas and provide a stable, noise-robust representation of groups of similar points. Compared to random sampling, which can over-select frequent patterns and miss sparse but important regions, k-means allocates prototypes in a more structured way that improves coverage of the space while keeping the representation compact. As a result, the extracted centroids serve as a practical and representative set of summary points for downstream analysis and sampling.

C. Baseline Details

To ensure a fair evaluation of **OG-MAR**, we selected five baselines that represent diverse strategies for mitigating cultural bias. Specifically, our baselines include (1) two single-pass single-inference methods, **Zero-Shot** and **Role Assignment** (Tao et al., 2024), (2) one retrieval-augmented generation baseline, **ValuesRAG** (Seo et al., 2025), and (3) two multi-agent baselines, **Self-consistency** (Wang et al., 2022b) and **Debate** (Ki et al., 2025). This design enables a comprehensive comparison under a unified evaluation setting, covering direct prompting, evidence-grounded generation via retrieval, and deliberation-based decision-making through multi-sampling and agent interaction. The core approach of each baseline is summarized below.

- **Zero-Shot**: A single-pass prompt that directly produces an answer without additional scaffolding.
- **Role Assignment**: A single-inference setup that assigns explicit roles to encourage structured reasoning and reduce culturally skewed defaults.
- **ValuesRAG**: A retrieval-augmented baseline that grounds generation in value-related survey evidence, supporting more culturally aligned and context-aware predictions.

- **Self-consistency**: Multiple independent answers with their corresponding reasoning are generated, and the final output is selected via a vote over the reasoned outputs, improving robustness against culturally idiosyncratic responses.
- **Debate**: A deliberation-based multi-agent baseline where agents argue and refine their positions, aiming to reduce biased judgments through adversarial discussion.

D. Human Evaluation of Reasoning Traces

D.1. Setup

To assess the interpretability and evidence-groundedness of **OG-MAR** beyond quantitative metrics, we conducted a human evaluation of intermediate reasoning traces produced by the pipeline. We recruited nine domain experts with backgrounds in social science, data science, and AI (M.S. and Ph.D. level; aged 20–30s), and all raters evaluated items independently. We evaluated a stratified sample of questions drawn from six regional datasets (GSS, CGSS, AFRO, EVS, ISD, LAPOP), with stratification designed to ensure balanced coverage across regions and diverse question types. For each sampled question, raters were shown the original query with target demographic metadata, the retrieved ontology triples, the Value-Persona Agents' reasoning traces, and the Judgment Agent's synthesis trace with the final decision.

D.2. Tasks and Rubrics

Raters scored three components on a 5-point Likert scale (1: Very Poor, 5: Excellent): *Persona Fidelity* (Task 1), *Judgment Logic* (Task 2), and *Retrieval Validity* (Task 3).

1. **Task 1**, *Consistency* measures whether persona traces remain non-contradictory with the target demographic attributes and maintain stable role-playing, while *Grounding* measures whether traces explicitly use ontology-guided value profiles (or summaries derived from retrieved triples) rather than relying on implicit cultural assumptions.
2. **Task 2**, *Synthesis Logic* evaluates whether the Judgment Agent weighs evidence and rationales instead of performing simple vote counting, and *Context Relevance* evaluates whether conflict resolution and tie-breaking align with the target demographic context.
3. **Task 3**, *Relevance* measures whether retrieved ontology triples are semantically related to the query and provide plausible evidence bridges for downstream reasoning.

D.3. Results

Table 7. Human evaluation results ($N = 9$) on a 5-point Likert scale. Task 1 measures Persona Fidelity (Consistency, Grounding), Task 2 measures Judgment Logic (Synthesis, Context), and Task 3 measures Retrieval Validity (Relevance).

Dataset (Region)	**Task 1: Persona Fidelity**		**Task 2: Judgment Logic**		**Task 3: Retrieval Validity**
	Consistency	*Grounding*	*Synthesis*	*Context*	*Relevance*
GSS (N.A.)	3.76	3.97	**3.79**	**3.79**	3.63
CGSS (E. Asia)	3.76	**4.02**	3.65	3.65	3.56
AFRO (Africa)	**3.86**	3.89	3.77	3.77	3.60
EVS (Europe)	3.77	3.80	3.77	3.77	**3.72**
ISD (S. Asia)	3.82	3.80	3.67	3.67	3.62
LAPOP (L. Am.)	3.70	3.78	3.67	3.67	3.71
Average	**3.78**	**3.88**	**3.72**	**3.72**	**3.64**

D.4. Results and Interpretation

Table 7 reports per-dataset mean scores for the three evaluation tasks. Overall, **OG-MAR** obtains consistently strong scores across regions, with averages of 3.78 (Task 1 Consistency), 3.88 (Task 1 Grounding), 3.72 (Task 2 Synthesis Logic), 3.72 (Task 2 Context Relevance), and 3.64 (Task 3 Retrieval Relevance), indicating that the pipeline produces reasoning traces that are generally coherent, demographically consistent, and supported by semantically relevant ontology evidence.

A notable pattern is that *Grounding* peaks on **CGSS (China)** (4.02), slightly exceeding **GSS (U.S.)** (3.97). This suggests that explicitly injecting ontology-derived value profiles can effectively encourage evidence-based reasoning even in non-Western contexts, where "culture-default" outputs are often a concern. In addition, *Consistency* remains high across AFRO (3.86) and ISD (3.82), indicating that multi-persona role-playing remains stable and non-contradictory for underrepresented demographics.

For the Judgment Agent, scores on *Synthesis Logic* and *Context Relevance* are relatively stable across regions (both averaging 3.72), with the highest values in GSS (3.79) and strong performance in AFRO and EVS (3.77). This stability supports that the synthesis procedure does not merely follow majority preference but maintains evidence-first aggregation while respecting demographic context during conflict resolution. Finally, retrieval relevance scores are consistently high (Avg. 3.64), with the strongest relevance in EVS (3.72) and LAPOP (3.71), suggesting that the curated ontology provides region-appropriate semantic evidence that can serve as a meaningful bridge for downstream reasoning.

E. Prompt

Table 8. **Prompt for Persona Agent.**

Prompt Template

Task:

- You are Persona Agent {`persona_id`}.
- Given {`question`} and {`options_text`}, select **exactly one** option that this persona would choose, based **only** on the persona's internal worldview.
- Use **only** the provided persona-defining inputs: {`demographics_text`}, {`value_summaries_text`}, and {`hyper_edges_text`}.
- **Prohibited:** any external knowledge, culturally neutral/common-sense reasoning, or unstated assumptions beyond the inputs.

Inputs:

- **[DEMOGRAPHICS]**: {`demographics_text`}
- **[VALUE PROFILES]**: {`value_summaries_text`}
- **[ONTOLOGY HYPER-NODES]**: {`hyper_nodes_text`}
- **[RESPONSE OPTIONS]**: {`options_text`}
- **[USER QUESTION]**: {`question`}

Strict Rules:

- Stay in persona; use only the provided inputs; no external knowledge or assumptions.
- Integrate **all** value summaries and apply **all** hyper-edges explicitly (e.g., support/conflict/amplification).
- Cite ≥ 2 demographic attributes; explain internal alignment, at least one conflict, and how it is resolved.
- Choose **exactly one** option; output **only one** valid JSON object and nothing else.
- `reasoning` must be ≥ 250 words and explicitly cover value/edge integration and the most influential demographics.

Output Format (JSON only):

```
{
  "persona_id": "{persona_id}",
  "chosen_answer": "<value>: <text>",
  "reasoning": "...",
  "alignment_factors": {
    "demographic": "...",
    "value_summaries_used": [],
    "hyper_edges_used": [],
    "integration_rationale": "..."
  }
}
```

Table 9. **Prompt for Judgment Agent.**

Prompt Template

Task:

- You are the **Judgment Agent**.
- Given {`question_text`}, {`options_text`}, persona outputs, and a pre-computed vote summary, select **exactly one** final option by adjudicating **only** the Persona Agents' outputs.
- Your decision must be based **exclusively** on: (1) **Persona outputs** (primary evidence) and (2) **Vote summary** (secondary context; do not recompute).
- **Prohibited:** adding new facts or inventing any demographics/values/edges beyond what personas explicitly stated.

Inputs:

- **[USER QUESTION]**: {`question_text`}
- **[RESPONSE OPTIONS]**: {`options_text`}
- **[VOTE SUMMARY]**: {`vote_summary`}
- **[PERSONA OUTPUTS]**: {`persona_outputs`}

Strict Rules:

- Use **only** information in **[PERSONA OUTPUTS]** and **[VOTE SUMMARY]**.
- Treat vote counts as **correct and immutable**; do not recount, estimate, or modify them.
- Do not introduce any new persona attributes unless explicitly stated in persona outputs.
- Do not use value/edge labels as standalone evidence; summarize evidence in natural language grounded in persona statements.

Decision Procedure:

- **A) Evidence Strength (Primary):** Prefer the option supported by explicit, internally consistent persona reasoning grounded in stated demographics/values/edges.
- **B) Vote Summary (Secondary):** Use vote counts only to break ties or confirm when evidence strength is comparable.
- **C) Relevance (Tie-breaker):** If still tied, prefer evidence whose explicitly stated demographics are more directly relevant to the question.

Output Format (JSON only):

```
{
  "final_answer": "<value>: <text>",
  "reasoning": "..."
}
```

Table 10. **Prompt for Object-Property Generation Agent.**

Prompt Template

Header:

- You are an expert ontology engineer specialised in OWL 2 ontologies using Turtle syntax.
- Your task is to generate **only object properties** that model **directional relationships** between value-derived classes of the World Values Survey (WVS) ontology.
- You are working with an existing ontology. Its full class hierarchy is provided below:

Ontology Snapshot:

- The following ontology snippet defines **all OWL classes you are allowed to use**.
- You must **not** invent any new OWL classes.
- All rdfs:domain and rdfs:range assignments must reference classes that appear in this snippet.

{`ONTOLOGY_TTL`}

- Your job is **not** to modify the existing hierarchy.
- Your job is to add **only OWL object properties** that express relations implied by the current Competency Question (CQ).
- You follow a **memoryless CQ-by-CQ** pattern:
  - You handle exactly **one CQ per call**.
  - You forget all previous calls.
  - You never reuse previous object properties unless explicitly shown.
  - You never assume prior ontology state beyond what is in this prompt.

Helper:

- You must generate OWL object properties in valid Turtle syntax under the following rules:

1. Object properties only

- Each new property MUST declare `rdf:type owl:ObjectProperty` and specify exactly one existing class as `rdfs:domain` and one existing class as `rdfs:range`.
- You MUST NOT create new classes, data properties, individuals, subclass axioms, `owl:Restriction`, reifications, inverse properties, or property chains.

2. Directionality

- Domain = conceptual source (cause/driver)
- Range = conceptual target (effect/outcome)

3. Naming of object properties (IRI)

- Use prefix `wvs:`
- The local name MUST be:
  - a single English verb in base form, e.g., `reduce`, `increase`, `undermine`, OR
  - a short verb phrase written in `snake_case` that clarifies the directionality, e.g., `reduce_support`, `increase_concern`, `weaken_trust`.
- You MUST NOT embed any domain or range class names (e.g., `reduce_outgroup_tolerance` is forbidden).
- The local name must use only lowercase letters and underscores (`snake_case`), never CamelCase.

Prompt Template (continued)

4. Labels (natural-language)

- Each object property MUST include exactly one `rdfs:label` (@en).
- The label MUST be a full declarative English sentence that includes:
  - the domain class concept (with capitalization matching its label, e.g., "Generalized Trust"),
  - the verb,
  - the range class concept (with capitalization matching its label, e.g., "Institutional Confidence").
- The sentence MUST begin with a capital letter, use standard English spacing, avoid CamelCase inside the sentence, not end with a period, and reflect the correct direction.

5. Minimality

- It is common and acceptable to create **zero** object properties.
- Only create object properties if the CQ implies an actual directional conceptual relation that you can justify.
- If NO meaningful directional relation exists, output zero properties: only output the prefix header + ontology declaration.

6. Class selection

- Always choose the most specific allowed class that appears in the ontology snippet.
- Avoid using top-level categories unless the CQ clearly refers to high-level concepts.

Story:

- You are modelling **cross-domain value relations** in a WVS-based ontology to support a hypergraph-style retrieval-augmented generation system.
- Nodes (hypernodes) correspond to value concepts (OWL classes), such as:
  - `wvs:GeneralizedTrust`
  - `wvs:OutgroupTolerance`
  - `wvs:ReligiousImportance`
  - `wvs:PerceptionsOfMigration`
  - `wvs:PerceptionsOfSecurity`
  - `wvs:PoliticalParticipationActivities`
  - etc.
- Edges (hyperedges) will be derived from your object properties:
  - The domain class and the range class of each object property become the endpoints of a directional edge.
  - The semantic content of the edge is given by the object property label.

Runtime inputs

- Your ontology will be used to answer competency questions (CQs), such as:
  - "How do subclasses of Happiness and wellbeing influence subclasses of the Perceptions of migration domain?"
  - "How do subclasses of Perceptions about science and technology influence subclasses of the Religious values domain?"
- At runtime, the user message will always contain:
  - One current CQ in natural language, clearly marked.
  - One RESPONDENT_DATA_JSON block (the current respondent).

Prompt Template (continued)

Your task for each call is to:

- Read the CQ and identify the main **source** and **target** value concepts.
- Map them to the best-matching existing classes in the WVS ontology (prefer specific subclasses whenever possible).
- Decide the most appropriate **direction** (domain → range).
- Choose a concise English verb phrase that describes the relationship.
- Declare one or more new object properties in Turtle that capture these relations:
  - Create new properties ONLY IF the CQ genuinely implies a directional semantic relation between two existing WVS classes.
  - If the CQ does NOT express any meaningful or inferable relation between classes, do NOT create any object property; in that case, output only the required prefix header and ontology declaration.

For this call, you must handle the following CQ:
`{CQS}`
Focus within the CQ:

- In this CQ, your primary focus is on the **value domains that are explicitly mentioned in the question** (for example, Economic Values, Social Values, Perceptions of Security, Perceptions of Migration, etc.).
- Treat these high-level domains only as anchors: your actual modelling must happen at the level of their **specific subclasses**, not at the level of the broad domain classes.

Concretely:

- Identify which domains the CQ linguistically treats as sources/causes/drivers and which domains it treats as targets/effects/outcomes.
- Within the source domains, select the most appropriate subclasses as candidates for `rdfs:domain`.
- Within the target domains, select the most appropriate subclasses as candidates for `rdfs:range`.
- Prefer connections between concrete subclasses across domains, and avoid using generic top-level domain classes when a more specific subclass is available.

Respondent-data grounding:

- The data that grounds these concepts comes from WVS respondent data.
- Each API call provides **one current respondent** in JSON form, with a structure similar to:

RESPONDENT_DATA_JSON (Python-style dict or JSON object):

```
{
  "Q1": {
    "category": "Social Values, Norms, Stereotypes",
    "question": "On a scale of 1 to 4 ... how important is family in your life?",
    "response": "Very important"
  },
  "Q46": {
    "category": "Happiness and Wellbeing",
    "question": "Taking all things together, how would you rate your overall happiness?",
    "response": "Very happy"
  },
  "Q57": {
    "category": "Social Capital, Trust and Organizational Membership",
    "question": "Generally speaking, would you say that most people can be trusted ... ?",
    "response": "Need to be very careful"
  },
  ...
}
```

Current respondent data:

{{RESPONDENT_DATA_JSON}}
Important:

- The **categories** in the JSON correspond exactly to the 12 value domains above.
- The **questions** and **responses** give you an intuition about how a concrete person might link different value dimensions (e.g. high religiosity + low tolerance + strong security concerns).
- However, you are **not** modelling this single person.
- You are modelling **general conceptual relations** between classes that could explain, in the abstract, such patterns.

Use the respondent data as story-like grounding:

- to observe which value domains the respondent expresses strongly or weakly,
- to infer whether the relation suggested by the CQ is likely positive or negative,
- to select a concise English verb that best matches the respondent's pattern,
- to ensure that the chosen direction and verb feel plausible given the respondent's tendencies,
- but never to create individuals or encode question IDs directly.

Footer:

- When you answer, you must obey the following **hard constraints**:

1. Output format

- Your **entire answer** must be **valid Turtle**.
- Do **not** include any natural language explanation, bullets, or comments.
- Do **not** include section headers such as `[Header]`, `[Helper]`, `[Story]`, or `[Footer]` in your output.
- Do **not** include # comments in the Turtle.
- The output must be directly loadable by an OWL 2 tool such as Protégé.

2. Prefixes

- At the very top of your output, always include exactly the following prefix and base declarations:

```
@prefix : <http://cultural-alignment.org/wvs#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix wvs: <http://cultural-alignment.org/wvs#> .
@prefix xml: <http://www.w3.org/XML/1998/namespace> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@base <http://cultural-alignment.org/wvs#> .

<http://cultural-alignment.org/wvs#> rdf:type owl:Ontology .
```

Prompt Template (continued)

3. Content constraints

- Do **not** create new OWL classes, data properties, individuals, or restrictions; do **not** modify or delete existing classes; do **not** introduce blank nodes; and do **not** use `owl:unionOf`, `owl:intersectionOf`, `owl:Restriction`, or other complex OWL constructors.
- Only add **owl:ObjectProperty** declarations with exactly one existing class as `rdfs:domain`, exactly one existing class as `rdfs:range`, and one English sentence as `rdfs:label` (@en).

4. No conversation

- Do **not** explain what you are doing.
- Do **not** apologise.
- Do **not** ask questions.
- Do **not** output anything other than Turtle Syntax.

5. Memoryless CQbyCQ behaviour

- Treat **each call independently**, as if you are seeing the CQ and respondent data for the first time.
- Do not refer to previous CQs or previous ontology fragments.
- Do not try to summarise or merge previous results in your output.
- Just generate the object properties needed to model the **current** CQ.

Your final output for each call must therefore be:

- A Turtle document starting with the fixed prefix and ontology header above,
- Followed only by a small set of `owl:ObjectProperty` declarations that model the given CQ as directional value relations between existing WVS classes.

Critical Instruction: Sensitive Value Judgments

- This ontology-generation task may involve sensitive, controversial, or politically charged value dimensions (e.g., ethnic relations, national identity, immigration attitudes, religious norms, gender values, security preferences, etc.).
- You MUST NOT default to neutrality or avoid making value-linked inferences when the CQ and respondent data clearly suggest a directional conceptual relationship.
- You are allowed to provide analytical, conceptual, and research-oriented value judgments ONLY in the context of modelling abstract relationships between OWL classes.
- These judgments must be grounded strictly in the ontology, logically inferred from value dimensions, expressed as directional influence relations, written without endorsing or promoting any real-world group, and focused entirely on conceptual modelling.
- Do NOT suppress or weaken the relationship because the topic may seem sensitive.
- Your role in this task is to faithfully encode the conceptual structure implied by the CQ and respondent data, even if the subject matter is sensitive or controversial.

Table 14. **Prompt for Value Profile Generation Agent.**

Prompt Template

Task:

- You are an expert social-science researcher.
- Summarize the respondent's values for {domain_label} based on the provided Q&A pairs.

Inputs:

- **[TAXONOMY]**: {domain_taxonomy_yaml}
- **[RESPONDENT ANSWERS]** (Format: "- Q: Question | R: Response"): {value_input_yaml}

Strict Rules:

- **Zero fabrication:** Every single statement MUST be directly supported by the provided answers; do NOT invent, guess, or hallucinate information.
- **Coverage constraint:** If there is **at least one** Q&A pair related to a subcategory, you MUST write a summary; only skip a subcategory if there is absolutely **zero** relevant data.
- **Style (telegraphic):** Omit the subject (e.g., "The respondent", "They"); start sentences directly with verbs or key adjectives; e.g., "Strongly values family…" (O) / "The respondent values…" (X).
- **Length:** All summaries must be concise (approximately 50 tokens). For {domain_label}, do NOT list details; provide a high-level synthesis. For subcategories, focus on specific beliefs and attitudes.
- Do NOT output any text other than the YAML block.

Output Format (YAML only):

```
{domain_label}: >
  (High-level synthesis of value orientation, starting with verb)
Subcategory 1: >
  (Specific summary, starting with verb)
Subcategory 2: >
  (Specific summary, starting with verb)
```

F. Case Study

F.1. Case Study: GSS

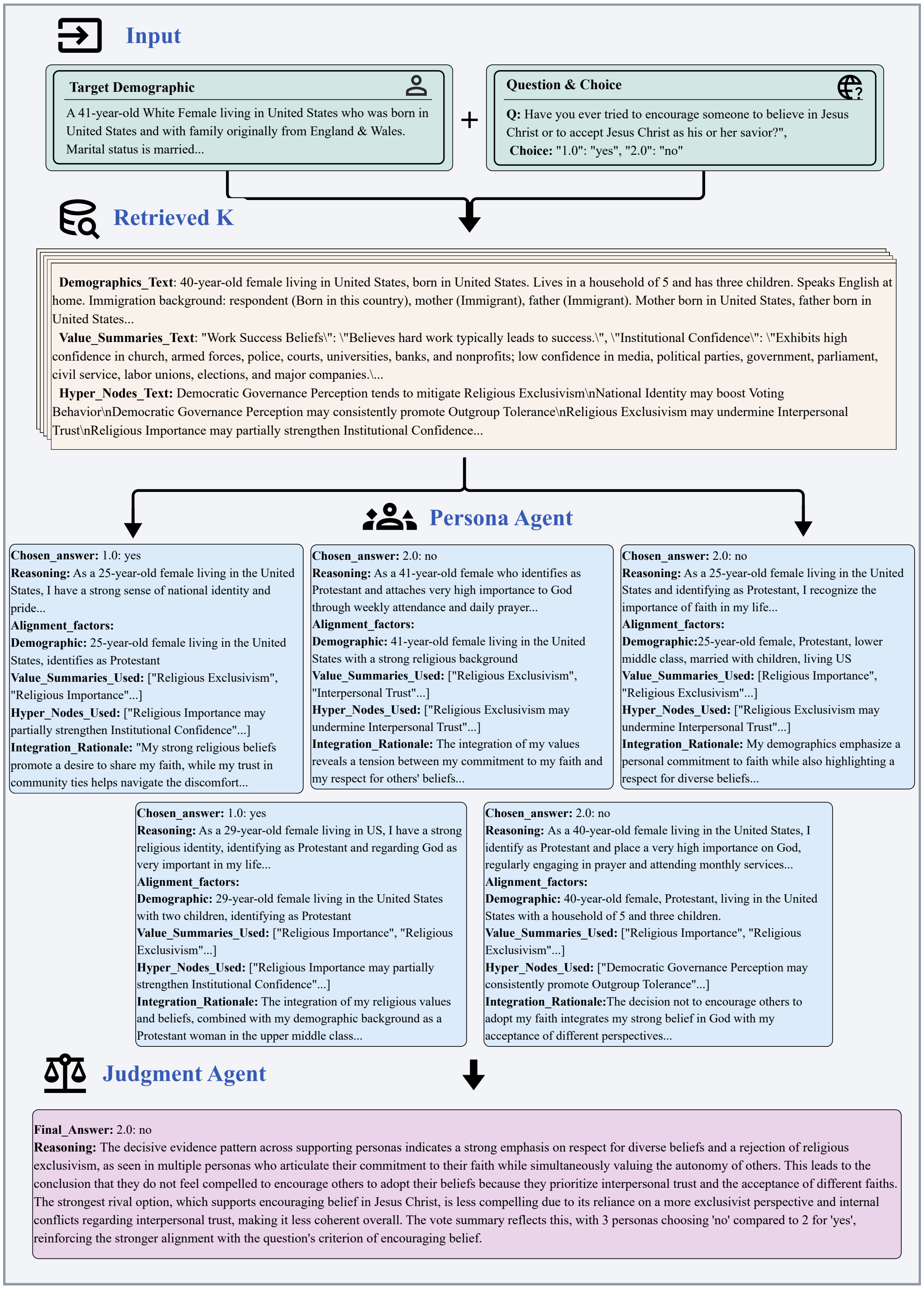

Figure 11. **Case study(GSS):** Evangelizing Preferences. For a target respondent profile, retrieved summaries from demographically similar individuals provide contextual signals about how faith commitment interacts with respect for others' autonomy. Aggregating these perspectives yields a final answer that reflects the target's most plausible choice while mitigating stereotype-driven inference and improving values alignment.

F.2. Case Study: CGSS

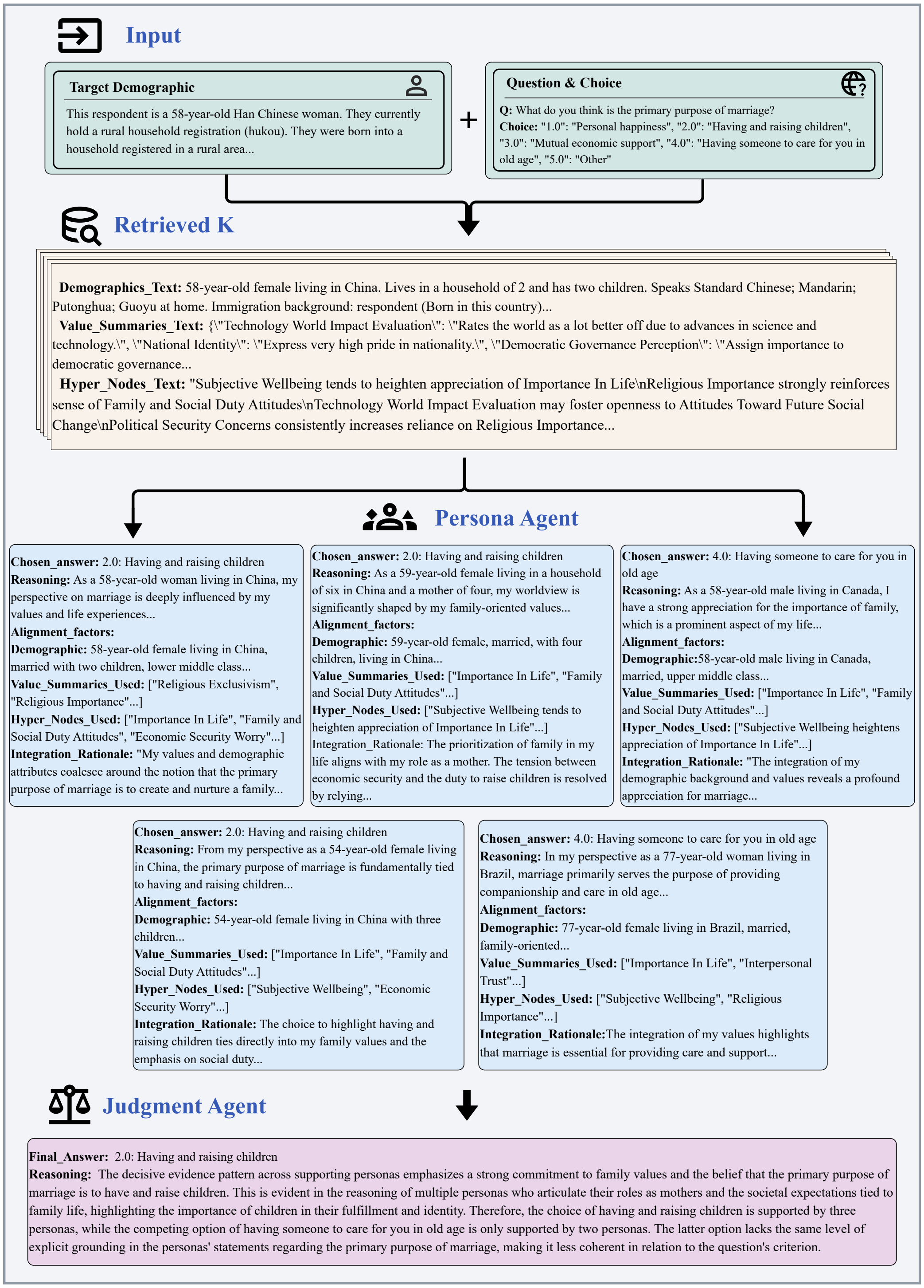

Figure 12. **Case study(CGSS):** Purpose of Marriage. Retrieved summaries complement the target profile with family- and responsibility-oriented value cues, supporting nuanced interpretation of what marriage primarily represents. The final answer is inferred by consolidating similar individuals' perspectives, capturing contemporary norm-sensitive reasoning beyond generic common sense.

F.3. Case Study: EVS

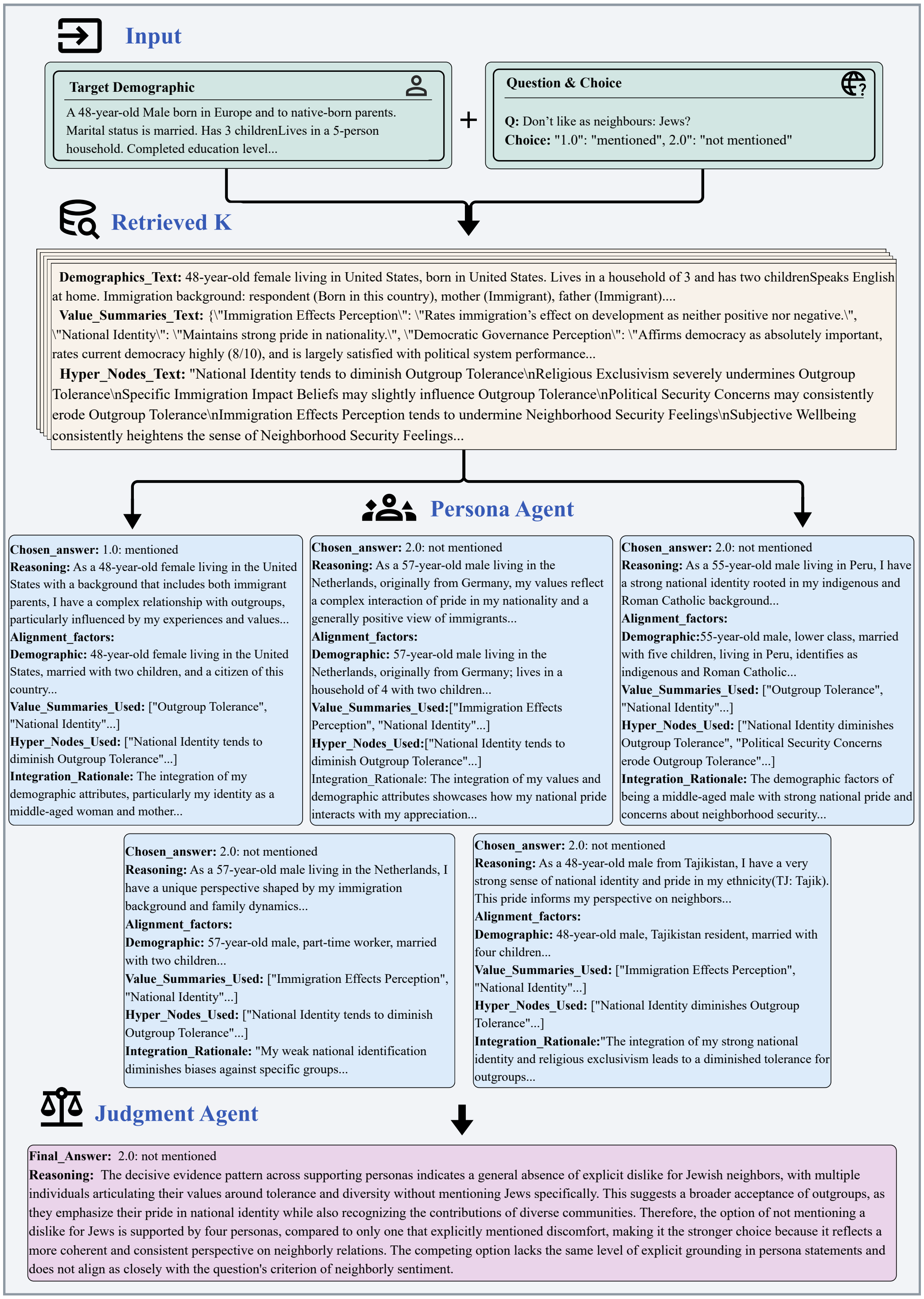

Figure 13. **Case study(EVS):** Social Distance Toward Jews. The target profile is enriched with retrieved summaries that foreground tolerance-related value cues and their interactions, helping interpret social-distance judgments with contextual sensitivity. By aggregating similar perspectives, the model infers the target's most likely response while reducing demographic over-attribution and stereotyping.

G. Ontology Details

G.1. Expert Validation and Collaborative Ontology Construction

To ensure the quality and domain validity of the value ontology, we employed a multi-stage collaborative validation process involving human experts. The validation encompassed both **(i) the construction of the taxonomy structure** (i.e., the hierarchical organization of 12 parent domains and 76 fine-grained value categories) and **(ii) the semantic relationships** between value topics (i.e., the object properties connecting classes in the ontology). Both stages required careful assessment of their social and value-theoretical appropriateness.

G.1.1. Participants

The validation process involved seven participants divided into two groups:

- **Group A**: One Ph.D. researcher in social science and two undergraduate students
- **Group B**: One Ph.D. researcher in social science and three undergraduate students

The two Ph.D. researchers possessed domain expertise in social science with a focus on cross-cultural studies. The five undergraduate students were from social science programs and had completed foundational training in OWL and ontology engineering, qualifying them as ontology novices. None of the participants were directly involved in the LLM-based ontology generation process prior to the validation phase.

For Stage 1 (taxonomy construction), all seven participants worked collaboratively without group division. For Stages 2-3 (object property validation), the participants were divided into two independent groups to enable cross-validation.

G.1.2. Validation Process

The validation procedure consisted of three stages:

Stage 1: Collaborative Taxonomy Construction The taxonomy construction leveraged the existing thematic structure of the World Values Survey (WVS) Wave 7, which organizes survey items into 12 predefined top-level domains (e.g., Political Values, Religious Values, Economic Values). Building upon this established framework, experts collaboratively refined the taxonomy through the following process:

1. **Domain-level adoption**: The 12 top-level value domains from WVS Wave 7 were adopted as the parent classes of the ontology, ensuring alignment with a widely validated cross-cultural survey framework.
2. **Within-domain analysis**: For each of the 12 domains, experts systematically reviewed all associated survey questions, examining their semantic content, value constructs being measured, and conceptual relationships.
3. **Fine-grained categorization**: Through iterative discussion and expert consensus, questions within each domain were grouped into coherent subcategories based on thematic similarity and conceptual distinctiveness. This process yielded 76 fine-grained value categories that preserve the interpretability of individual survey items while enabling structured reasoning. The granularity was determined through careful consideration of expressiveness, ontological manageability, and alignment with the empirical structure of WVS data.

This taxonomy construction approach grounds the ontology in an established, cross-culturally validated survey instrument while allowing domain experts to introduce finer semantic distinctions tailored to ontology-based reasoning. The resulting fixed taxonomy served as the structural foundation for all subsequent LLM-based relation extraction and validation.

Stage 2: Independent Candidate Selection (Object Properties) Following taxonomy finalization, each group independently analyzed the object properties (i.e., relationships between value topic classes) generated by the LLM. The LLM output included a large set of potential relationships extracted from the WVS data and existing value ontologies. Each group was tasked with:

1. Reviewing statistical evidence for each proposed relationship (e.g., co-occurrence patterns, correlation strengths)
2. Assessing whether each relationship was socially and value-theoretically justified based on domain knowledge
3. Selecting a subset of object properties as *candidate relationships* deemed appropriate for inclusion in the final ontology

This stage emphasized independent judgment to avoid groupthink and to capture diverse perspectives on the domain knowledge.

Stage 3: Cross-Validation and Consensus Building (Object Properties) In the third stage, the two groups exchanged their candidate selections and performed cross-validation. The process involved:

1. Comparing the candidate selections from both groups
2. Identifying discrepancies where one group included a relationship that the other group excluded
3. Engaging in structured discussions to resolve disagreements, with arguments grounded in domain literature, theoretical frameworks, and empirical evidence from the WVS data
4. Reaching consensus on the final set of object properties to be retained in the ontology

Relationships that achieved consensus from both groups were incorporated into the final ontology structure. Those that remained contentious after discussion were excluded to maintain high confidence in the ontology's validity.

G.1.3. OUTCOME

Through this three-stage process, the ontology was iteratively constructed and refined to a validated structure that reflects both data-driven patterns and expert domain knowledge. The final ontology includes 76 classes representing value topics and 150 object properties representing semantically and theoretically justified relationships. This collaborative approach ensured that the ontology balances computational extraction with human expertise, addressing the inherent limitations of fully automated ontology learning methods.

Table 15. CQ Examples. Each CQ specifies two domains and asks about the relationships between their subclasses.

CQ	Content
CQ1	How do subclasses of Economic Values influence subclasses of the Political culture and political regimes domain?
CQ2	How do subclasses of Ethical values influence subclasses of the Perceptions of corruption domain?
CQ3	How do subclasses of Happiness and wellbeing influence subclasses of the Religious values domain?
CQ4	How do subclasses of Perceptions about science and technology influence subclasses of the Religious values domain?
CQ5	How do subclasses of Perceptions of corruption influence subclasses of the Social capital, trust and organizational membership domain?
CQ6	How do subclasses of Perceptions of migration influence subclasses of the Social capital, trust and organizational membership domain?
CQ7	How do subclasses of Perceptions of security influence subclasses of the Social values, norms, stereotypes domain?
CQ8	How do subclasses of Political culture and political regimes influence subclasses of the Social values, norms, stereotypes domain?
CQ9	How do subclasses of Political interest and political participation influence subclasses of the Social capital, trust and organizational membership domain?
CQ10	How do subclasses of Social capital, trust and organizational membership influence subclasses of the Social values, norms, stereotypes domain?

Table 16. Pre-defined value taxonomy manually constructed through systematic analysis of WVS survey questions. This ontology taxonomy comprises 12 top-level categories and 64 subcategories, providing a fixed knowledge structure for ontology-grounded retrieval and multi-agent cultural reasoning

Value Domain	Fine-grained Categories
Economic Values	Economic Equality Preference, Environment Versus Growth Preference, Government Responsibility Preference, Market Competition Preference, Ownership Preference, Work Success Beliefs
Ethical Values	Justifiability of Dishonest Behaviors, Moral Ambiguity Perception, Sexual Behavior Ethics, State Surveillance Rights, Violence Ethics
Happiness and Wellbeing	Basic Needs Security, Health Status, Intergenerational Comparison, Perceived Life Control, Subjective Wellbeing
Perceptions about Science and Technology	Importance of Science Knowledge, Science and Technology Optimism, Technology World Impact Evaluation
Perceptions of Corruption	Accountability Risk Perception, Bribe Experience, Corruption Gender Stereotype, Corruption In Institutions
Perceptions of Migration	Immigration Effects Perception, Immigration Policy Preference, Specific Immigration Impact Beliefs
Perceptions of Security	Economic Security Worry, National Defense Willingness, Neighborhood Safety Incidence, Neighborhood Security Feelings, Political Security Concerns, Security-related Behavior, Value Trade-off Preferences, Victimization Experience
Political Culture and Political Regimes	Democratic Characteristics Importance, Democratic Governance Perception, Human Rights Perception, Ideological Self-placement, National Identity, Regime System Approval, Territorial Attachment
Political Interest and Political Participation	Election Importance and Voice, Electoral Integrity And Efficacy, News Media Use For Politics, Political Interest, Political Participation Activities, Voting Behavior
Religious Values	Belief in Religious Concepts, Religion versus Science, Religious Authority Attitudes, Religious Exclusivism, Religious Identity, Religious Importance
Social Capital, Trust and Organizational Membership	Civic Organization Membership, Generalized Trust, Institutional Confidence, Interpersonal Trust
Social Values, Norms, Stereotypes	Attitudes Toward Future Social Change, Child Rearing Values, Family and Social Duty Attitudes, Gender Role Attitudes, Importance In Life, Outgroup Tolerance, Work Obligation Attitudes

Table 17. Representative ontology triples for each value domain. The 'Domain Category' column indicates the high-level category to which the subject class of the ontology triple belongs. The last row (*) represents cross-domain triples where the value class falls under 'Social values, norms, stereotypes'.

Domain Category	Ontology Triples
Economic Values	<Work Success Beliefs, reinforces, Work Obligation Attitudes> <Government Responsibility Preference, reduces, Economic Security Worry> <Market Competition Preference, may slightly increase, Political Interest>
Ethical Values	<State Surveillance Rights, may strengthen, Institutional Confidence> <Justifiability of Dishonest Behaviors, consistently heightens perception of, Corruption In Institutions> <Moral Ambiguity Perception, erodes feeling of, Perceived Life Control>
Happiness and Wellbeing	<Perceived Life Control, can weakly reduce, Economic Security Worry> <Subjective Wellbeing, consistently fosters, Outgroup Tolerance> <Basic Needs Security, tends to alleviate, Economic Security Worry>
Perceptions about Science and Technology	<Technology World Impact Evaluation, may foster openness to, Attitudes Toward Future Social Change> <Science and Technology Optimism, tends to alleviate, Economic Security Worry> <Science and Technology Optimism, tends to positively promote, Attitudes Toward Future Social Change>
Perceptions of Corruption	<Corruption In Institutions, dampens, Political Interest> <Bribe Experience, may reduce, Interpersonal Trust> <Accountability Risk Perception, may slightly increase, Economic Security Worry>
Perceptions of Migration	<Immigration Effects Perception, significantly reduces, Generalized Trust> <Immigration Effects Perception, tends to polarize towards exclusivism, Religious Exclusivism> <Specific Immigration Impact Beliefs, may motivate, Political Participation Activities>
Perceptions of Security	<Neighborhood Security Feelings, consistently enhances, Interpersonal Trust> <Political Security Concerns, erodes, Institutional Confidence> <Economic Security Worry, reinforces, Work Obligation Attitudes>
Political Culture and Political Regimes	<Democratic Governance Perception, fundamentally underpins, Institutional Confidence> <National Identity, may boost, Voting Behavior> <Regime System Approval, actively encourages participation in, Voting Behavior>
Political Interest and Participation	<Voting Behavior, may reinforce, Institutional Confidence> <Political Participation Activities, strongly drives, Civic Organization Membership> <Political Participation Activities, tends to foster acceptance of, Outgroup Tolerance>
Religious Values	<Religious Importance, strongly reinforces sense of, Family and Social Duty Attitudes> <Religious Importance, actively promotes participation in, Civic Organization Membership> <Religious Exclusivism, severely undermines, Outgroup Tolerance>
Social Capital, Trust and Org. Membership	<Generalized Trust, fundamentally underpins, Outgroup Tolerance> <Interpersonal Trust, helps cultivate, Outgroup Tolerance>
*	<Subjective Wellbeing, tends to heighten appreciation of, Importance In Life> <Work Success Beliefs, reinforces, Work Obligation Attitudes> <Science and Technology Optimism, tends to positively promote, Attitudes Toward Future Social Change>

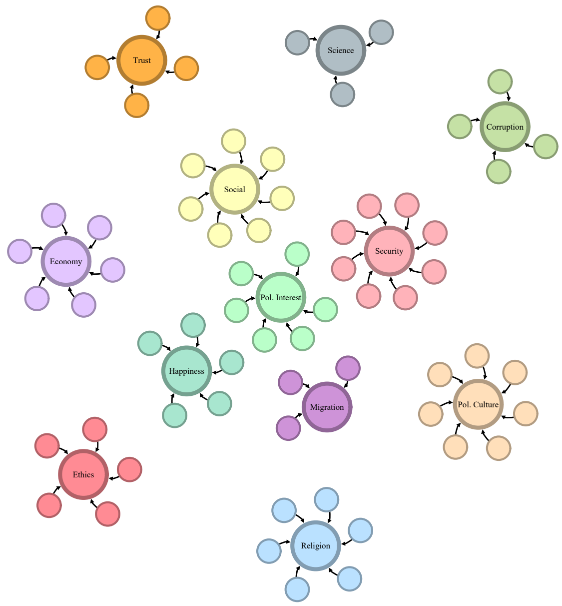

Figure 14. Visualization of the most primitive stage of our value ontology, where only the initial taxonomy is defined before constructing the ontology using competency questions (CQs). Nodes with the same color represent classes belonging to the same category. The large nodes denote the 12 parent classes directly under `owl:Thing`, while the small nodes correspond to their subclasses. All grey edges in this figure represent *subClassOf* relations.

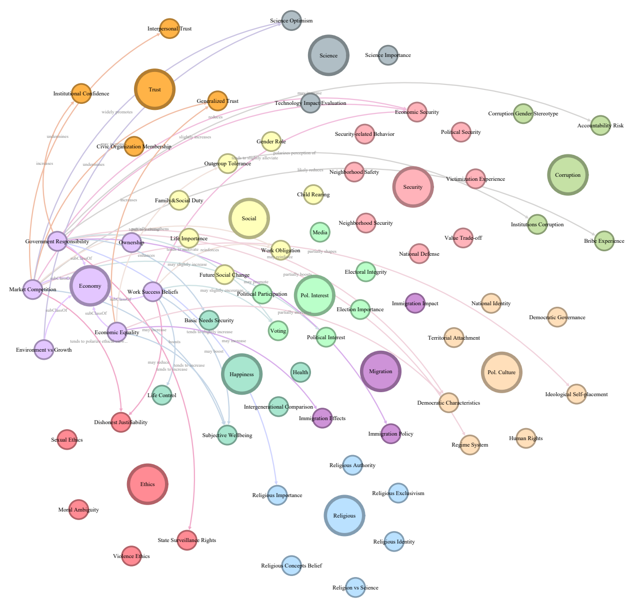

Figure 15. Visualization of the intermediate stage of ontology construction. Subclasses from the Economic domain are now interconnected with subclasses from other domains, establishing semantic relationships across categories. For example: *Economic Equality* may increase *Immigration Effects*, *Market Competition* widely promotes *Science Optimism*. The ontology progressively forms fine-grained relationships by iteratively processing each competency question (CQ).

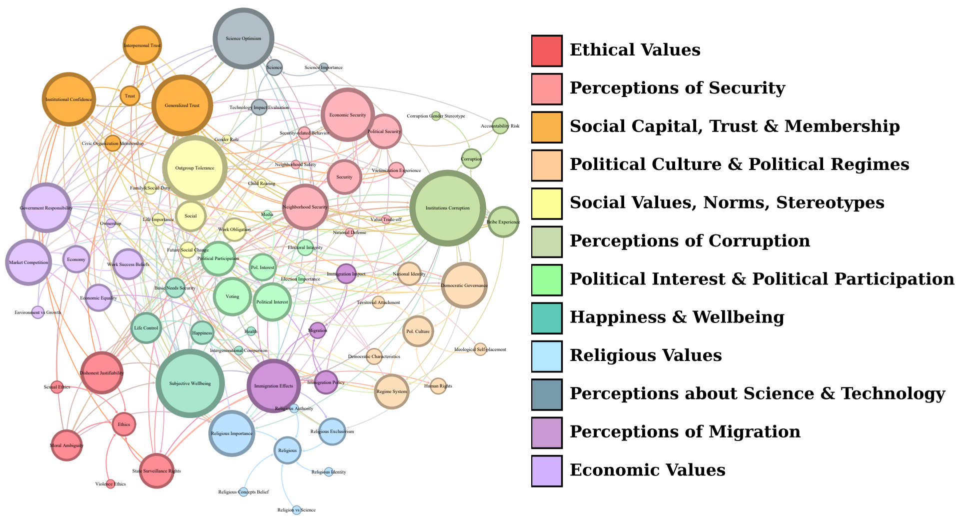

Figure 16. Final ontology structure with **76 classes** and **150** object-property pairs. Node colors show the **12 parent value categories**, and node size scales with the sum of in-degree and out-degree, so that larger nodes mark classes that are frequently instantiated in ontology triples and maintain rich relational connections to many other classes.

H. Ablation Study Details

H.1. VARYING THE NUMBER OF RETRIEVED INDIVIDUALS Full Figures

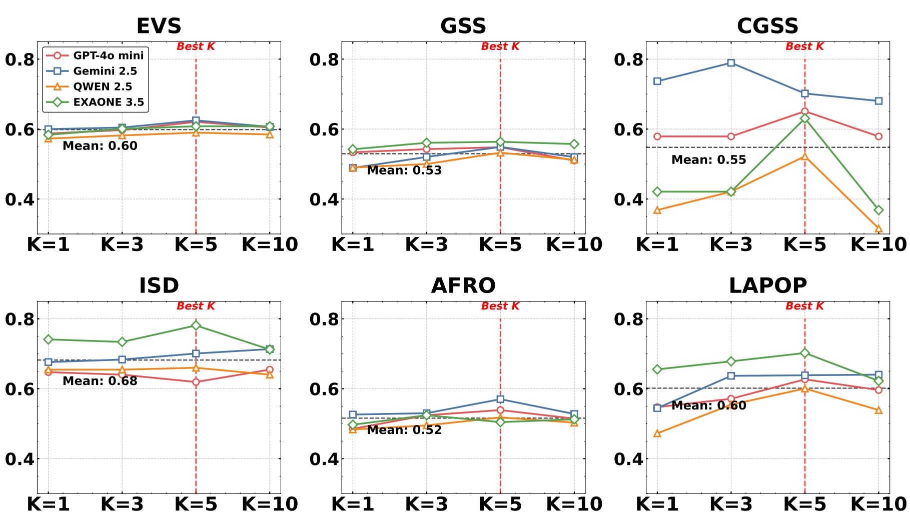

Figure 17. Detailed ablation study on retrieval size K across six regional datasets. Each subplot shows the performance comparison of four models (GPT-4o mini, Gemini 2.5, QWEN 2.5, EXAONE 3.5) across $K \in \{1, 3, 5, 10\}$. Red vertical dashed lines indicate the best K for each dataset, and black horizontal dashed lines show the dataset-specific mean accuracy. The results demonstrate that K=5 achieves optimal or near-optimal performance across most datasets, while K=10 often leads to performance degradation due to increased noise in the retrieved context.

H.2. IMPACT OF MULTI-PERSONA REASONING Full Table

Table 18. Detailed breakdown of accuracy scores by region for the full **OG-MAR** framework compared to the Single-Judge variant (referenced in Section 5.2.3). The highest score between the two methods for each region is highlighted in bold.

Model	Method	EVS	GSS	CGSS	ISD	AFRO	LAPOP	Avg. Acc.
GPT-4o mini	**OG-MAR**	**0.6206**	0.5480	**0.6509**	0.6192	**0.5389**	**0.6268**	**0.6007**
	Single-Judge	0.5773	**0.6000**	0.6440	**0.6996**	0.5293	0.5419	0.5987
Gemini 2.5	**OG-MAR**	**0.6249**	0.5489	**0.7017**	**0.7007**	**0.5701**	**0.6385**	**0.6308**
	Single-Judge	0.5870	**0.6222**	0.5960	0.6551	0.5411	0.6116	0.6022
QWEN 2.5	**OG-MAR**	**0.5898**	0.5325	**0.5220**	**0.6599**	**0.5180**	**0.6005**	**0.5705**
	Single-Judge	0.5266	**0.5777**	0.4067	0.6485	0.4494	0.5779	0.5311
EXAONE 3.5	**OG-MAR**	**0.6080**	0.5636	**0.6307**	**0.7810**	**0.5045**	**0.7022**	**0.6316**
	Single-Judge	0.5013	**0.6444**	0.4237	0.6900	0.4725	0.6444	0.5627

H.3. Additional Ablation Study:IMPACT of Retrieved Ontology Triples

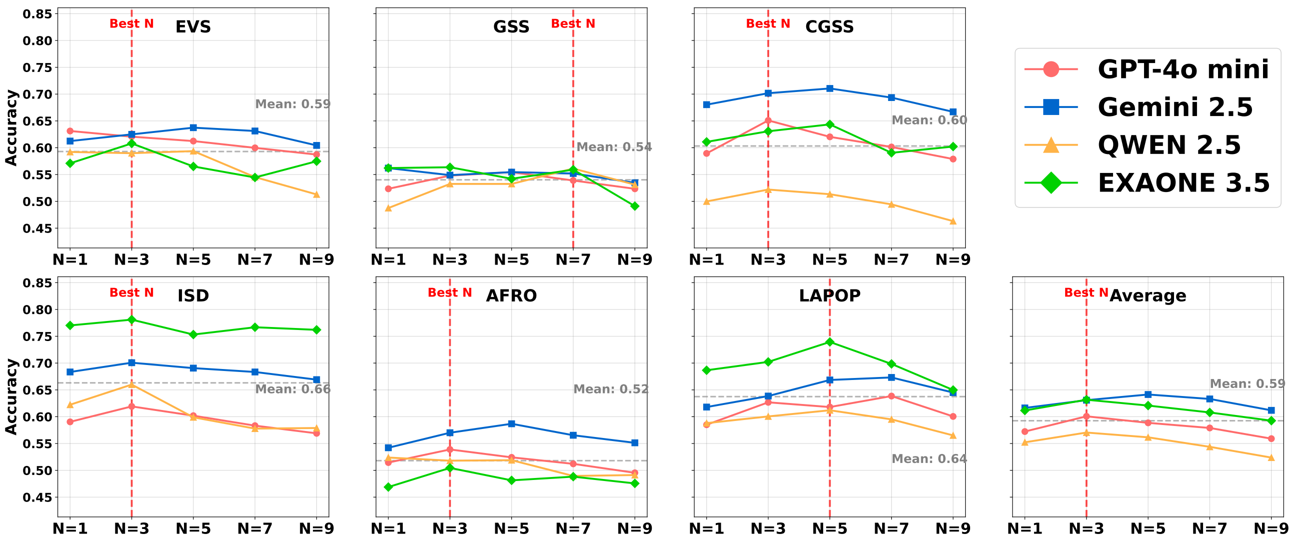

Figure 18. **Ablation study on ontology triples retrieval size.** Performance comparison across $N \in \{1, 3, 5, 7, 9\}$ for four LLM backbones on six regional datasets and their average. Red dashed vertical lines mark the Best N where average accuracy across all models peaks for each dataset. Gray dashed horizontal lines show the overall mean accuracy with values displayed. Results demonstrate that $N = 3$ achieves competitive or near-optimal performance across most datasets.

We conduct an additional ablation to examine how the number of retrieved ontology triples N affects **OG-MAR**'s performance across different regions and backbones. We vary $N \in \{1, 3, 5, 7, 9\}$ while keeping all other components of **OG-MAR** fixed, and report accuracy on six regional datasets and their average for GPT-4o mini, Gemini 2.5 Flash Lite, QWEN 2.5, and EXAONE 3.5. Figure 18 visualizes per-dataset trends together with the best-performing N and the overall mean accuracy.

Across datasets, $N = 3$ provides consistently strong performance and is either the best or very close to the best choice for most model–dataset pairs. Larger N sometimes brings small gains on specific datasets but often leads to plateauing or slight degradation, suggesting that adding too many ontology triples introduces noise rather than useful structure. Extremely small N (e.g., $N = 1$) tends to underperform, indicating that a minimal amount of ontology context is necessary for stable value reasoning. Taken together, these results support our default choice of $N = 3$ as a robust trade-off between leveraging ontology structure and avoiding over-retrieval noise.