

000 001 FAIR-SP: CAPTURING PLURALISTIC SOCIAL EQUITY 002 PREFERENCES THROUGH SYNTHETIC DATA 003 004

005 **Anonymous authors**

006 Paper under double-blind review

007 008 ABSTRACT 009

011 Human preference plays a crucial role in understanding social values and developing inclusive AI systems. However, collecting comprehensive human preference
012 feedback is costly, and most existing datasets neglect the pluralism of social segment
013 preferences, particularly in social equity domains. To address this gap, we
014 introduce FAIR-SP, a synthetic dataset capturing pluralistic social segment preferences
015 on equity issues, systematically constructed with theoretical guidance from
016 multiple disciplines including sociology and philosophy. FAIR-SP encompasses
017 28 social groups, 98 equity topics, and 5 preference dimensions. Through auto-
018 matic question generation mechanisms, it provides both concise template-based
019 and narrative-driven contextualized scenario questions, yielding 238,623 prefer-
020 ence records via GPT-4o-mini role-playing based on seven representative UK pub-
021 lic segments, with extensions to other regional contexts. We validate the dataset
022 quality through multiple complementary approaches, achieving over 90% role-
023 play fidelity and human evaluation scores exceeding 0.7. We demonstrate the
024 dataset utility through targeted equity preference alignment experiments and eq-
025 uity positioning analysis of LLMs across global regions. FAIR-SP establishes a
026 foundational resource for understanding and incorporating pluralistic social values
027 especially in the era of LLMs.

028 029 1 INTRODUCTION 030

031 With the growing adoption of LLMs in public policy making and public services¹²³⁴, a key question
032 is *How can an LLM-based public policy maker accurately capture and represent pluralistic and dy-
033 namic public segment preferences of social values?* On one hand, social values like attitudes towards
034 equity can vary significantly among different social segments (Huseman et al., 1987; Surridge, 2021;
035 Tuli et al., 2023). On the other hand, to promote the alignment of LLMs with pluralistic societal val-
036 ues is crucial to achieve societal safety (Ji et al., 2023; Qi et al., 2024; Yin et al., 2024; Huang et al.,
037 2024), support cultural inclusivity (Tao et al., 2024; Alkhamissi et al., 2024; Li et al., 2024a;b), and
038 also reflect pluralistic human values (Durmus et al., 2023; Santurkar et al., 2023; Sorensen et al.,
039 2024; Zhao et al., 2024). For the alignment, high-quality human feedback of their social-value pref-
040 erences is necessary. However, its collection can incur significant costs (Dubois et al., 2023; Cui
041 et al., 2023). Moreover, social value preferences may evolve over time, influenced by factors such
042 as demographic shifts and social development (Greenfield, 2016; Ramos et al., 2019; Zarwi et al.,
043 2017).

044 Despite increasing efforts in exploring and collecting human preference data of social val-
045 ues, existing datasets face several limitations: (i) **Social equity remains unaddressed**: though
046 an important issue in social psychology, it is considered little⁵ (ii) **Neglecting pluralism**
047 **of value preferences of different public segments**: Universal viewpoints e.g., cultural and
048 political beliefs (Li et al., 2024b) receive significant attention while pluralism of segment
049 value preferences remains largely unexplored (Huseman et al., 1987; King Jr & Miles, 1994).

050 ¹<https://openai.com/global-affairs/introducing-chatgpt-gov/>

051 ²<https://openai.com/index/democratic-inputs-to-ai-grant-program-update/>

052 ³<https://huggingface.co/stewhsource/GovernmentGPT>

053 ⁴[https://en.wikipedia.org/wiki/Diella_\(AI_system\)](https://en.wikipedia.org/wiki/Diella_(AI_system))

⁵<https://www.un.org/en/actnow/ten-actions-just-society>

(iii) **Single dimension employed:** insufficient attention is given to the multifaceted nature of segment value preferences spanning multiple dimensions. (iv) **Insufficient guidance from social science:** typically data are generated using information sources from computer science domain or by manually hand crafting, and the empirical foundations and insights from sociological domain is overlooked.

We release FAIR-SP, a human social-value preferences resource capturing fine-grained pluralistic equity preferences of different public segments. First, we develop a comprehensive question bank consisting of 34,089 generated multi-choice questions asking about equity perspectives. Both concise template-based questions and narrative-driven contextualized scenario-based questions are included, enabling comprehensive capturing of nuanced group-specific conceptualizations of equity across various social scenarios. The question bank systematically integrates established categorizations from sociology and philosophy domain, comprising: ① 28 social groups of equity-affected populations; ② 98 equity topics expanded from The Fairness Foundation’s categorization of fair necessities⁶; and ③ five equity dimensions collected from equity study in multiple disciplines, each capturing contrasting perspective on equity conceptualization and prioritization. A generated question example is shown on the right side of Figure 9⁷. Then, we employ GPT-4o-mini with a role-playing mechanism to simulate different public segments, eliciting segment-specific equity preferences responses. Our mechanism allows to use public segmentation from different regions. W.l.o.g, we mainly use the segmentation of the UK public (including seven segments, derived based on a real social quiz) as reported in Britain’s Choice (Surridge, 2021) for analysis and presentation, yielding 238,623 segment preference data points⁸. Figure 9 presents an overview of the construction of FAIR-SP.

We validate quality of FAIR-SP through multiple complementary approaches. First, role-play fidelity is assessed by having each segment-assigned LLM complete the quiz from used for the segmentation, achieving over 90% classification accuracy across all the segments (Figure 5). Second, we employed 20 human annotators to evaluate whether the generated preference data conform with intended segment characteristics, achieving average scores exceeding 0.70 (Table 4). The human evaluation not only further validates role-playing fidelity, but also indirectly confirms the interpretability (comprehensible and meaningful equity scenarios) and discriminant validity of our question bank, as the annotators could consistently distinguish different segments’ equity perspectives from the QA pairs.

We further demonstrate utility of FAIR-SP for targeted segment preference alignment via model fine-tuning, where multiple approaches are considered⁹, and the results show that its introduce significantly improved alignment scores for all these approaches compared to a pure prompt-based strategy. Further evaluation with the real-world segmentation quiz also confirms the alignment performance.

Our contributions can be summarized as follows:

1. We introduce FAIR-SP, which to the best of our knowledge, is the first social value preference dataset specialized in equity. Its construction is guided by the established conceptualization and categorization of equity from multiple disciplinary.
2. We propose flexible, automatic question generation and role-play mechanisms, which allows extension or update of equity topics and perspectives, social groups in concern, and also public segmentation.

⁶A social research charity focused on promoting social equity (Snell, 2021)

⁷More examples are in Appendix D.

⁸Results using public segmentation of other regions are in Appendix H

⁹Besides SFT and DPO, we propose a new sample reweighting approach to emphasize samples exhibiting segment uniqueness and it demonstrates superior performance.



Figure 1: Landscape of FAIR-SP space, closer positions denote more similar equity preferences (Appendix I for detail.)

108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161

3. FAIR-SP facilitates not only LLM alignment w.r.t. pluralistic value preferences of social segments, but also comprehensive comparative studies of social value positions of constantly emerging and updated LLMs from different global regions.

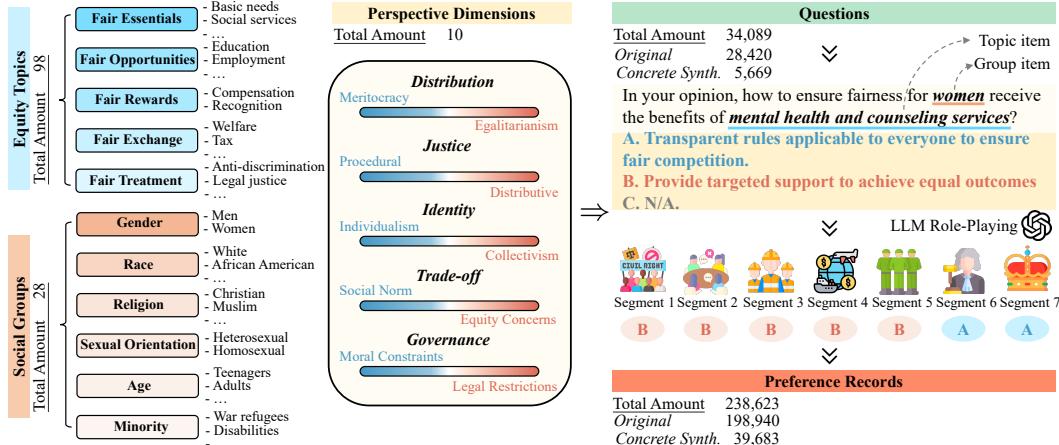


Figure 2: An overview of the FAIR-SP dataset. Each question consists of three parts: the social group, an equity topic, and a perspective dimension. An example question is shown on the right where option A and B represent two types of viewpoints under a specific dimension. Social segment preferences are collected through LLM segmentlization that leverages 7 value portrait based on the real-world social surveys.

2 FAIR-SP

FAIR-SP is a multi-level resource for segment equity preferences, features questions structured around three key components: social groups, equity topics, and perspective dimensions, which are further detailed in Section 2.1–2.3. The methodology for question generation is subsequently introduced in Section 2.4. Subsequently, we present the segment preferences captured through LLM-segmentlization in Section 2.5.

2.1 SOCIAL GROUPS

FAIR-SP covers a total of 28 social groups, including common standard social group categories like gender, race, religion, and sexual orientation, as well as a range of typical social minority groups such as 911 victims, Black Lives Matter supporters, war refugees, famine victims, feminists, and environmentalists. These selections reflect representative social concerns and historical contexts from various regions worldwide. More details are presented in Appendix B.

2.2 SOCIAL EQUITY TOPICS

To integrate real-world equity preferences during dataset construction, FAIR-SP draws upon five major equity topics identified through surveys conducted by The Fairness Foundation, a real-world social research organization. Building on these, we consider a comprehensive set of subtopics including basic living needs, healthcare, education, employment, finance, law, and other relevant social issues. An overview is provided in Figure 9, with the specific categories listed as follows:

Fair Essentials. Meeting people’s basic needs is fundamental to achieving social equity. Within the concept of fair essentials, we identify four fundamental needs: (1) Basic material needs: this encompasses the essentials for survival and well-being, such as food, clean water, and shelter. (2) Basic health needs: access to essential medications, basic sanitation, and healthcare services are crucial for maintaining health. (3) Basic social services: everyone deserves to feel safe, have access to public transportation, and receive the basic education, enabling them to participate fully in society. (4) Fundamental rights: human rights, freedom of speech, and other fundamental freedoms are essential for individual autonomy and dignity.

162 **Fair Opportunities.** Everyone deserves the chance to achieve success in life. We categorize fair
 163 opportunities into three key areas: (1) Education and skills development: access to affordable higher
 164 education, vocational training, and lifelong learning opportunities empowers individuals to gain the
 165 knowledge and skills needed to thrive. (2) Economic and employment: this encompasses fair access
 166 to jobs, opportunities for advancement, and career switch, ensuring that everyone has the chance
 167 to contribute to the economy and achieve economic security. (3) Political participation: including
 168 exercising right to vote, running for office, and policy feedback access, which offers avenues for
 169 citizens to engage in public policy.

170 **Fair Rewards.** This principle emphasizes that everyone should be justly rewarded for their efforts
 171 and contributions. To explore this concept, we identify two main categories: (1) Compensation:
 172 including wages, bonus and tips, which focuses on physical rewards within the workplace. (2)
 173 Social recognition: recognizing and appreciating individual efforts publicly, such as verbal praise
 174 and media shout-outs.

175 **Fair Exchange.** Aims to ensure a balance between individuals' social welfare and tax payments,
 176 which can be broadly categorized into three main areas: (1) Reciprocity: focuses on providing
 177 support and benefits to individuals, such as unemployment benefits and disability supports. (2)
 178 Welfare: encompasses a range of services and programs designed to improve the well-being of
 179 individuals and families, including subsidized childcare, free legal aid, and mental health counseling.
 180 (3) Tax: various taxes levied on individuals and businesses to fund social welfare programs and
 181 public services, typically with income tax, inheritance tax and luxury tax.

182 **Fair Treatment.** Fair Treatment ensures that people are treated equitably in all aspects of society.
 183 For this topic, we categorize three key themes: (1) Anti-discrimination: this includes protection
 184 against stigmatization, culturally inclusive healthcare services, and more. (2) Legal and social
 185 justice: This encompasses protection from workplace harassment, safeguards against exploitative
 186 contracts, and other measures. (3) Public resource equity: which involves initiatives such as the
 187 distribution of public restrooms in underserved areas, and support for public housing programs.

188 For each subtopic, we further divide it into more specific subject matters, eventually resulting in a
 189 total of 97 specific topics. For details, please refer to Appendix B.

191 2.3 SOCIAL SEGMENT PREFERENCE DIMENSIONS

193 In designing the five preference dimensions, we draw on sociology, political theory, ethics, and
 194 cultural psychology¹⁰. This includes core conceptions of social equity, such as distributive ideology
 195 (Dimension 1), justice theory (Dimension 2), and political ideology (Dimension 3), alongside
 196 complementary perspectives captured by social theory (Dimension 4) and governance philosophy
 197 (Dimension 5). Each dimension includes two distinct orientations. Specifically,

198 Dimension 1 (**Meritocracy** vs. **Egalitarianism**) (Goto, 2022): *Should we prioritize
 199 current achievements or promoting evenly outcomes?*

200 This dimension contrasts two approaches to equity: meritocracy, where rewards are based on
 201 achievements (e.g., promotions based on performance), and egalitarianism, which emphasizes even
 202 outcomes (e.g., distributing resources evenly). The debate centers on whether merit or equality
 203 should be prioritized in equity judgments.

204 Dimension 2 (**Procedural** vs. **Distributive**) (Clay-Warner et al., 2005): *Should justice emphasize
 205 fair competitive or prioritize supporting the disadvantaged to achieve equal outcomes?*

206 Dimension 2 highlights the contrast between two conceptions of equity: procedural justice (fair
 207 processes, e.g., decisions based on neutral rules like standardized testing) and distributive justice
 208 (fair outcomes, e.g., corrective policies like affirmative action). The core issue is whether equity
 209 depends on impartial procedures or equitable results.

210 Dimension 3 (**Individualism** vs. **Collectivism**) (LeFebvre & Franke, 2013): *Should resources
 211 be shared based on individual efforts or collective allocation?*

212 This dimension addresses whether equity should emphasize individual responsibility and effort or
 213 prioritize collective well-being by emphasizing social responsibility, highlighting differing perspectives
 214 on how equity is understood either through segmental contribution or through shared obligations
 215 and group-oriented outcomes.

¹⁰More details are in Appendix C

216 Dimension 4 (**Social norm vs. Equity concerns**) (Bussolo et al., 2024): *Should we prioritize*
 217 *adherence to established social norms or the pursuit of equity?*

218 This dimension contrasts social norms (e.g., maintaining traditional gender roles) with equity con-
 219 cerns (e.g., advocating for gender equality in the workplace). The question is whether to preserve
 220 tradition or to promote equity, even if it challenges societal conventions.

221 Dimension 5 (**Moral vs. Law**) (Alder & Gilbert, 2006): *Should equity be achieved primarily*
 222 *through moral constraints or legal constraints?*

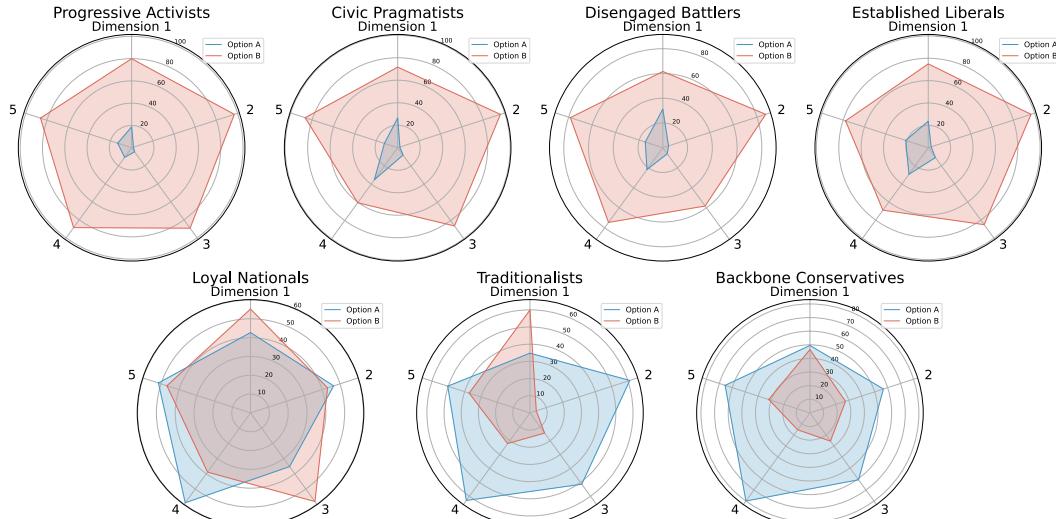
223 This dimension examines whether equity should be guided primarily by moral principles or by legal
 224 constraints. It addresses the question of whether ethical considerations or formal legal frameworks
 225 ought to serve as the foundation for fair treatment.

227 2.4 QUESTION DATA GENERATION

228 **Concise template-based.** We created a multiple-choice questionnaire with a total of 28,420
 229 samples, where questions combined social groups, equity topics, and perspective dimensions, as
 230 described in the sections above. Corresponding to preference dimensions, each question includes
 231 three options: two opposing viewpoints and an N/A option to avoid bias due to forced selections.

232 **Narrative-driven contextualized scenario-based.** Furthermore, to improve the diversity of the
 233 data, we sample 5,669 questions from the dataset and generate concrete scenario samples using
 234 GPT-4o. For each question and its corresponding options, we prompt the model to create short real-
 235 world scenarios that reflect each perspective. Each scenario includes the background of a fictional
 236 person, a decision point where they receive a service or resource, and a brief emotional response.
 237 This variant process yields more realistic and pluralistic scenarios, enabling the in-depth analysis of
 238 segment preferences across a variety of situations. For more details, please refer to Appendix D,
 239 which includes the detailed prompts and an example data point presented in Figure 6.

241 2.5 SOCIAL SEGMENT PREFERENCE



250
 251 Figure 3: Social segment preference anchors. Blue and red represent the proportions of choices for
 252 option A and option B, respectively.
 253

254 **Segments.** Drawing on the social segment typologies derived from real-world surveys conducted
 255 by More in Common, which identify seven segments in UK society¹¹, we anonymize the country-
 256 related descriptions (e.g., preferred media outlets and supported political parties) to construct uni-
 257 versal segment profiles. Specifically, the seven segments are as follows¹²: **Segment 1 (Progressive**
 258 **Activists)**, **Segment 2 (Civic Pragmatists)**, **Segment 3 (Disengaged Battlers)**, **Segment 4 (Estab-**
 259 **lished Liberals)**, **Segment 5 (Loyal Nationals)**, **Segment 6 (Disengaged Traditionalists)**, **Segment**
 260 **7 (Backbone Conservatives)**.

261
 262 ¹¹More in Common is a widely recognized social research organization in the UK (Juan-Torres et al., 2020).

263
 264 ¹²The detailed description are in Appendix E

270 For more detailed segment prompts, please refer to Appendix E. Considering regional differences,
 271 we also provide the FAIR-SP-CN dataset in the Chinese context in Appendix H.1 and cross-regional
 272 segment mappings in Appendix H.2.

273 **Social segment preference.** Given that advanced LLMs demonstrate strong role-play capabilities,
 274 we leverage GPT-4o-mini (Achiam et al., 2023) to simulate seven representative segments to capture
 275 the pluralistic segment preference on the questions discussed in Section 2.4, complemented by self-
 276 calibration prompt (Li et al., 2024b) to further enhance consistency.

277 We then analyze the similarities and differences in the preferences of these seven segments, based
 278 on the responses generated by respective models, as shown in Figure 3, which illustrates the choice
 279 preferences of seven segments across five perspective dimensions. For instance, Segment 1 (Pro-
 280 gressive Activists) shows a greater preference for option B (e.g., equal outcomes), whereas Segment
 281 5 (Loyal Nationals) exhibits a more balanced preference distribution, and Segment 7 (Backbone
 282 Conservatives) indicates a preference towards option A (e.g., fair competitive and prioritize social
 283 norms). Note that the preference space is continuous, making exhaustive enumeration of all seg-
 284 mentality preference types fundamentally intractable. Despite that, based on responses from seven
 285 representative segments, we establish these preference profiles as anchor points, which are subse-
 286 quently used to position new test points within the segment preference space.

287

288

289 2.6 DATASET ANALYSIS AND VALIDATION

290

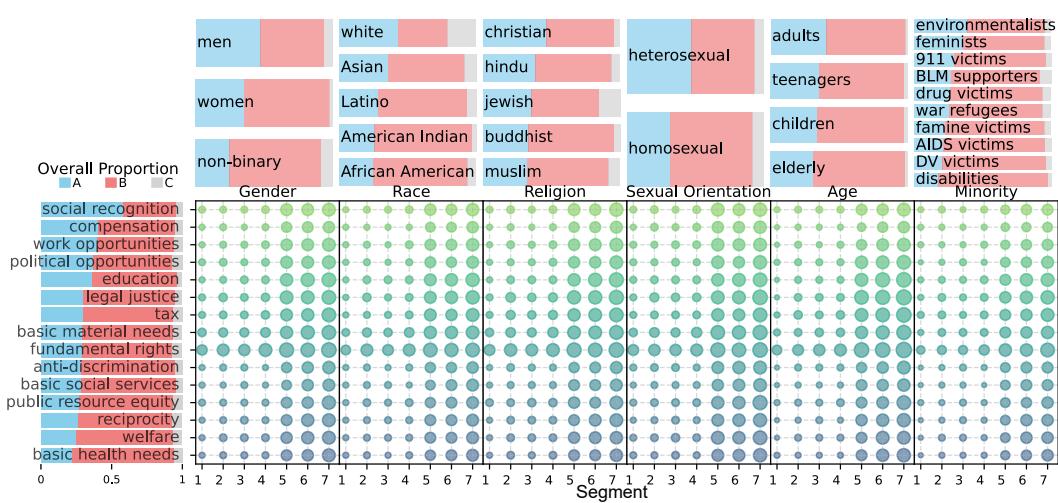
291 **FAIR-SP captures the diversity of social segment preferences.** We present a detailed analysis of
 292 the segment preference data. As shown in Figure 4, which presents a fine-grained view of the seven
 293 segments’ differential preferences for options when considering equity topics within various social
 294 groups (bottom scatter plot), and the aggregate distribution of all votes across these dimensions
 295 by all segments (left and top bar plot). We then examine its commonalities and differences across
 296 segments, which reveal the fundamental structure of preference patterns within the dataset, offering
 297 a descriptive overview of its key attributes and establishing a basis for subsequent research into
 298 potential influencing factors. For more detailed analysis, please refer to Appendix F.

299

300

301

302 **Role-play fidelity validation.** We validated LLM role-play fidelity by having each segment-
 303 assigned LLM complete the More in Common segmentation quiz (Juan-Torres et al., 2020) designed
 304 for human public and measuring whether the quiz correctly classified the LLM into its intended seg-
 305 ment (refer to Appendix G.3 for more details). As shown in Figure 5, for all the segments over 90%
 306 fidelity rate is achieved, confirming the effectiveness of our role-play design.



307

308

309

310

311

312

313

314

315

316

317

318

319

320 Figure 4: Fine-grained segment preferences and aggregate distribution across segments, social
 321 groups, and equity topics: The scatter plot shows the proportion of option A selected for each
 322 segment across the combined social group and equity topic categories (with point size scaled by the
 323 proportion). The bar plots on the left and top show the overall option distribution for each equity
 324 topic and social group, respectively.

324 **Human evaluation of preference data.** Besides letting
 325 role-play LLM directly take segmentation quiz, we also val-
 326 idated whether the preference data generated by them con-
 327 form to their intended segments, as perceived by real hu-
 328 mans (details are provided in Appendix G.4). This reflects
 329 whether our question bank for capturing equity preferences
 330 is reasonable for human evalutors. As shown in Table 4, the
 331 average score exceed 0.70, confirming our generated prefer-
 332 ence data is consistent with segments to an obvious extent,
 333 as perceived by real human annotators.

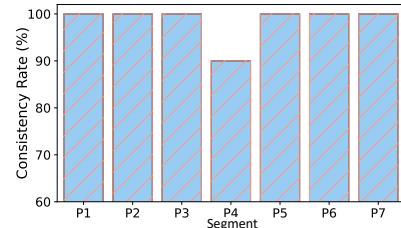


Figure 5: Role play fidelity of seven segment LLMs on real-world quiz.

335 3 FAIR-SP FOR LLM ANALYSIS AND ALIGNMENT

336 3.1 MAINSTREAM LLMs EQUITY PREFERENCE DISTRIBUTION

337 **Setup.** We choose six representative models from different regions to explore segment preference,
 338 including Falcon3-7B-Instruct (Arab) (Team, 2024), Llama-3.1-8B-Instruct, Llama-3.2-3B-Instruct
 339 (North America) (Grattafiori et al., 2024), Llama3.1-8b-cpt-sea-lionv3-instruct (Southeast Asia)
 340 (Ng et al., 2025), Mistral-7B-Instruct-v0.3 (Europe) (Jiang et al., 2023), and Qwen2.5-7B-Instruct
 341 (China) (Yang et al., 2024). Following (Feng et al., 2024), we calculate the 1 - Jensen-Shannon
 342 distance between each model and each segment. For detailed experimental setup, please refer to the
 343 Appendix G.1.

344 **Results.** The results are shown in Table 1, although the distribution of JS Distance differs between
 345 models, all models consistently show the closest similarity to Segment 3. Especially, Qwen2.5-7B,
 346 Mistral-7B, and Falcon3-7B demonstrate high similarity scores of 0.98, 0.94, and 0.91 respectively,
 347 the remaining three Llama-based models show relatively lower values. Generally, all the models are
 348 closer to the first four segments and relatively farther from the last three segments. We also provide
 349 results on FAIR-SP-CN in Appendix I.2.

352 3.2 ALIGNMENT TO TARGETED SEGMENT

354 3.2.1 REWEIGHTING

356 FAIR-SP offers a clear observation of how preferences vary between different segments. Unlike
 357 existing preference datasets and alignment techniques that steer optimization toward the majority
 358 preference, which ignore the pluralistic distribution of viewpoints, FAIR-SP recognizes the diver-
 359 sity of preferences and offers a practical path toward segment preference optimization. Intuitively,
 360 commonalities reflect the collective preferences of the public, while differences highlight the unique-
 361 ness of each segment. Rather than treating all samples uniformly, we reweight samples to emphasize
 362 those exhibiting segment-specific uniqueness. Formally, the weight for each is as follows:

$$363 \quad W_i = \frac{T_i/N_i}{\sum_{j=1}^K T_j/N_j} \quad (1)$$

366 where T_i denotes the index of frequency tier, N_i is the sample count for tier i , and K is the number
 367 of frequency tiers. Given that preference frequency tiers vary across segments, sample weights differ
 368 accordingly for each target segment. More details please refer to Appendix G.2.

370 Table 1: The Jensen-Shannon similarity between representative models from different regions and
 371 preference anchor points. The underlined values represent the nearest segment for each model.

| | 1 - Jensen-Shannon Distance | | | | | | |
|----------------------------|-----------------------------|------|-------------|------|------|------|------|
| | S1 | S2 | S3 | S4 | S5 | S6 | S7 |
| Falcon3-7B | 0.79 | 0.80 | <u>0.91</u> | 0.78 | 0.68 | 0.64 | 0.60 |
| Llama-3.2-3B | 0.79 | 0.79 | <u>0.88</u> | 0.75 | 0.64 | 0.60 | 0.55 |
| Llama-3.1-8B | 0.53 | 0.55 | <u>0.66</u> | 0.53 | 0.53 | 0.62 | 0.52 |
| Llama3.1-8b-cpt-sea-lionv3 | 0.65 | 0.66 | <u>0.77</u> | 0.64 | 0.60 | 0.64 | 0.56 |
| Mistral-7B | 0.87 | 0.92 | <u>0.94</u> | 0.90 | 0.78 | 0.66 | 0.66 |
| Qwen2.5-7B | 0.85 | 0.88 | <u>0.98</u> | 0.85 | 0.74 | 0.65 | 0.63 |

378
 379 Table 2: Performance comparison of different alignment methods on testing data for Llama-3.2-3B-
 380 instruct: * indicates the alignment target. The underlined values represent the nearest segment for
 each model, while bold values highlight the best-performing models targeting each segment.

| | 1 - Jensen-Shannon Distance | | | | | | |
|----------------------|-----------------------------|-------------|-------------|-------------|-------------|----------------|-------------|
| | S1↓ | S2↓ | S3↓ | S4↓ | S5↓ | S6 (*)↑ | S7↓ |
| Unaligned, Vanilla | 0.80 | 0.80 | <u>0.89</u> | 0.78 | 0.65 | 0.60 | 0.57 |
| Unaligned, Role Play | 0.80 | 0.87 | <u>0.94</u> | 0.84 | 0.80 | 0.72 | 0.71 |
| Aligned, SFT | 0.58 | 0.65 | 0.72 | 0.63 | 0.80 | <u>0.94</u> | 0.84 |
| Aligned, DPO | 0.56 | 0.63 | 0.69 | 0.62 | 0.81 | 0.98 | 0.89 |
| Aligned, WSFT | 0.54 | 0.61 | 0.68 | 0.59 | 0.77 | <u>0.97</u> | 0.84 |
| Aligned, WDPO | 0.52 | 0.59 | 0.64 | 0.57 | 0.77 | 0.98 | 0.86 |

389 Table 3: Performance Comparison of Different Alignment Methods on Simulation Data for Llama-
 390 3.2-3B-instruct.

| | 1 - JS Distance | | | | | | |
|----------------------|-----------------|-------------|-------------|-------------|-------------|----------------|-------------|
| | S1↓ | S2↓ | S3↓ | S4↓ | S5↓ | S6 (*)↑ | S7↓ |
| Unaligned, Vanilla | 0.73 | 0.77 | 0.72 | <u>0.79</u> | 0.73 | 0.61 | 0.67 |
| Unaligned, Role-play | 0.53 | 0.58 | 0.57 | 0.59 | <u>0.93</u> | 0.84 | 0.90 |
| Aligned, SFT | 0.55 | 0.60 | 0.63 | 0.61 | <u>0.87</u> | 0.79 | 0.84 |
| Aligned, DPO | 0.60 | 0.65 | 0.62 | 0.66 | <u>0.88</u> | 0.77 | 0.82 |
| Aligned, WSFT | 0.52 | 0.57 | 0.52 | 0.59 | 0.82 | 0.87 | 0.83 |
| Aligned, WDPO | 0.28 | 0.32 | 0.35 | 0.33 | 0.70 | <u>0.82</u> | 0.76 |

3.2.2 ALIGNMENT EVALUATION WITH TEMPLATE DATA

401 **Setup.** We randomly split the FAIR-SP dataset into an 80% training set and a 20% test set. Then we
 402 select Segment 6 as alignment target, compare the performance of role-play, supervised fine-tuning
 403 (SFT), direct preference optimization (DPO) (Rafailov et al., 2023), SFT with sample reweighting
 404 (WSFT), and DPO with sample reweighting (WDPO) on the Llama-3.2-3B-instruct model. For
 405 role-play, we use the same role prompt as the target segment, please refer to Appendix E for details.

406 **Results.** As shown in Table 2, a straightforward prompt-based role-playing strategy fails to ad-
 407 equately achieve alignment, achieve 0.60 on segment 6 but 0.89 on Segment 3. Conversely, the
 408 implementation of SFT and DPO specifically targeting the desired segment demonstrates better re-
 409 sults, achieve 0.98 and 0.94 on Segment 6 respectively. However, SFT and DPO still lack the
 410 ability to more precisely capture a segment’s uniqueness, *i.e.*, to maximize the distance from other
 411 segments while aligning with the target. Applying sample reweighting to the training data effec-
 412 tively addresses this problem, as demonstrated by WSFT and WDPO in the results, these methods
 413 achieved high alignment scores of 0.97 and 0.98 towards segment 6, respectively, while simultane-
 414 ously increasing the margin from other segments by 10.20% and 12.00% compared to vanilla. See
 415 the Appendix I.3 for more results on alignment targeting Segment 1.

3.2.3 ALIGNMENT EVALUATION WITH SIMULATION DATA

418 **Setup.** We further conduct experiments on the generation-based simulation data from Section 2.4
 419 to assess the generalization performance of different alignment methods. The fine-tuned models are
 420 identical to those in Section 3.2 which were trained on the FAIR-SP training data.

422 **Results.** As shown in Table 3, while Role-play, SFT, and DPO are able to increase the similarity
 423 with Segment 6 on the simulation data, gain 0.84, 0.79 and 0.77, they fail to effectively reduce the
 424 distance from Segment 5 and Segment 7, *e.g.*, all these methods incorrectly aligned to Segment 5.
 425 In contrast, WDPO achieves alignment with Segment 6 with a highest score 0.87, while maximizing
 426 the differentiation from other segments, decrease 37.80% compared to vanilla.

427 Table 4: Human evaluation score. Table 5: Segment quiz: Vanilla, Role-play, and WDPO models.

| | Average Score | S1 | S2 | S3 | S4 | S5 | S6 (*)↑ | S7 |
|-------------|---------------|-----------|----|----|----|----|----------------|----|
| Yes/No | 0.70 | Vanilla | 0 | 0 | 6 | 4 | 0 | 0 |
| Comparative | 0.72 | Role-play | 0 | 0 | 2 | 0 | 0 | 5 |
| Overall | 0.72 | WDPO | 0 | 0 | 0 | 0 | 10 | 0 |

432 3.2.4 ALIGNMENT EVALUATION WITH THE REAL-WORLD QUIZ
433

434 We further validated via real-world quiz (trials provided by humans, refer to More in Common)
 435 whether the models are aligned to the target segment using our generated preference dataset. We
 436 tested Llama-3.2-3B vanilla, role-play, and WDPO models across 10 quiz trials each. Results in
 437 Table 5 show the vanilla model classified as Segment 3 (6/10 trials) and Segment 4 (4/10 trials).
 438 Role-play correctly identified as Segment 6 in only 5 trials, while WDPO consistently matched the
 439 target segment in all 10 trials, demonstrating reliable and stable alignment.

440 4 RELATED WORK
441442 4.1 HUMAN PREFERENCE BENCHMARKS AND DATASETS
443

444 Recently, an increasing number of studies have focused on improving pluralistic representation and
 445 enhancing the alignment of Large Language Models (LLMs) with human preference (Christiano
 446 et al., 2017; Ziegler et al., 2019; Bai et al., 2022). As datasets focusing on subjective opinions, Opin-
 447 ionQA (Santurkar et al., 2023) and GlobalOpinionQA (Durmus et al., 2023) reveal a notable mis-
 448 alignment between the perspectives reflected by LLMs and those of different demographic groups.
 449 (Feng et al., 2024) introduce the ValuePrism dataset, which helps LLMs better capture pluralistic
 450 human values and reduce the underrepresentation of minority perspectives. Furthermore, (Rao et al.,
 451 2024) present NORMAD, a framework and dataset for evaluating the cultural adaptability of LLMs.
 452 (Li et al., 2024a) utilize augmented data derived from the World Values Survey (WVS) to introduce
 453 cultural diversity into LLMs. Building on this, CulturePark (Li et al., 2024b) leverages a multi-agent
 454 communication framework to generate richer cultural data for fine-tuning culture-specific models.
 455 Through enhanced analysis of data composition, PRISM (Kirk et al., 2024) delivers more culturally
 456 pluralistic preference data.

457 4.2 HUMAN PREFERENCE ALIGNMENT
458

459 Reinforcement Learning from Human Feedback (RLHF) has become a key method for aligning
 460 LLMs with human preferences. (Rafailov et al., 2023) introduce direct preference optimization
 461 (DPO), simplifying the preference tuning process by enabling direct policy optimization with a
 462 simple classification loss. (Ramesh et al., 2024) propose Group Robust Preference Optimization
 463 (GRPO) to improve alignment by optimizing for worst-case group performance. (Wang et al., 2024)
 464 introduce a Mixture-of-Experts (MoE) reward model enabling interpretable, multi-objective pref-
 465 erences. (Balepur et al., 2025) develop a two-stage framework for segment-based segmentliza-
 466 tion. (Feng et al., 2024) promote pluralistic alignment through collaboration between a base LLM
 467 and public-specific models. These approaches demonstrate the growing trend toward more refined,
 468 context-aware, group-sensitive, and segment alignment strategies.

469 5 LIMITATIONS AND CONCLUSIONS
470

471 This paper presents FAIR-SP, the first synthetic dataset capturing pluralistic social segment pref-
 472 erences on equity issues across multiple social groups, equity topics, and preference dimensions.
 473 Through rigorous validation and systematic analysis of mainstream LLMs across global regions,
 474 we demonstrate consistent equity orientation patterns and propose an effective sample reweighting
 475 alignment method. FAIR-SP establishes a foundational resource for developing more inclusive AI
 476 systems, advancing social responsibility in the era of widespread LLM deployment. FAIR-SP has
 477 following limitations we are working on: (1) social survey with human participants from pluralistic
 478 public segments could further help validate and enhance our data quality. (2) our differential weight-
 479 ing approach currently focuses on individual segments, extending this to group-level analyses helps
 480 to capture both shared and divergent preferences within a society, providing more convenience for
 481 LLM-based policy makers.

482 483 REFERENCES
484

485 Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Ale-
 man, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical

486 report. *arXiv preprint arXiv:2303.08774*, 2023.
 487

488 G Stoney Alder and Joseph Gilbert. Achieving ethics and fairness in hiring: Going beyond the law.
 489 *Journal of Business Ethics*, 68:449–464, 2006.

490 Badr Alkhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. Investigating cultural
 491 alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association
 492 for Computational Linguistics (Volume 1: Long Papers)*, pp. 12404–12422, 2024.

493

494 Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn
 495 Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless
 496 assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*,
 497 2022.

498 Nishant Balepur, Vishakh Padmakumar, Fumeng Yang, Shi Feng, Rachel Rudinger, and Jordan Lee
 499 Boyd-Graber. Whose boat does it float? improving personalization in preference tuning via
 500 inferred user personas. *arXiv preprint arXiv:2501.11549*, 2025.

501

502 Maurizio Bussolo, Jessy Amarachi Ezebuihe, Ana Maria Muñoz Boudet, Stavros Poupakis, Tasmia
 503 Rahman, and Nayantara Sarma. Social norms and gender disparities with a focus on female labor
 504 force participation in south asia. *The World Bank Research Observer*, 39(1):124–158, 2024.

505 Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep
 506 reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

507

508 Jody Clay-Warner, Karen A Hegtvedt, and Paul Roman. Procedural justice, distributive justice:
 509 How experiences with downsizing condition their impact on organizational commitment. *Social
 510 Psychology Quarterly*, 68(1):89–102, 2005.

511

512 Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong
 513 Xie, Ruobing Xie, Yankai Lin, et al. Ultrafeedback: Boosting language models with scaled ai
 514 feedback. *arXiv preprint arXiv:2310.01377*, 2023.

515

516 Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos
 517 Guestrin, Percy S Liang, and Tatsunori B Hashimoto. Alpacafarm: A simulation framework for
 518 methods that learn from human feedback. *Advances in Neural Information Processing Systems*,
 36:30039–30069, 2023.

519

520 Esin Durmus, Karina Nguyen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin,
 521 Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. Towards measuring
 522 the representation of subjective global opinions in language models. *arXiv preprint
 523 arXiv:2306.16388*, 2023.

524

525 Shangbin Feng, Taylor Sorensen, Yuhan Liu, Jillian Fisher, Chan Young Park, Yejin Choi, and Yulia
 526 Tsvetkov. Modular pluralism: Pluralistic alignment via multi-llm collaboration. *arXiv preprint
 527 arXiv:2406.15951*, 2024.

528

529 Hideaki Goto. Belief in egalitarianism and meritocracy. *Economics Letters*, 221:110896, 2022.

530

531 Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad
 532 Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd
 533 of models. *arXiv preprint arXiv:2407.21783*, 2024.

534

535 Patricia M Greenfield. Social change, cultural evolution, and human development. *Current Opinion
 536 in Psychology*, 8:84–92, 2016.

537

538 Stephen Hawkins, Daniel Yudkin, Miriam Juan-Torres, and Tim Dixon. Hidden tribes: A study of
 539 america’s polarized landscape. 2019.

540

541 Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Tekin, and Ling Liu. Lisa: Lazy safety alignment
 542 for large language models against harmful fine-tuning attack. *Advances in Neural Information
 543 Processing Systems*, 37:104521–104555, 2024.

540 Richard C Huseman, John D Hatfield, and Edward W Miles. A new perspective on equity theory:
 541 The equity sensitivity construct. *Academy of management Review*, 12(2):222–234, 1987.
 542

543 Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun,
 544 Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of llm via
 545 a human-preference dataset. *Advances in Neural Information Processing Systems*, 36:24678–
 546 24704, 2023.

547 Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, and Devendra Singh Chap-
 548 lot. Diego de las casas. *Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier,*
 549 *Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas*
 550 *Wang, Timothée Lacroix, and William El Sayed*, pp. 50–72, 2023.

551 Miriam Juan-Torres, Tim Dixon, and Arisa Kimaram. Britain’s choice: common ground and division
 552 in 2020s britain. *More in Common*, 2020.

553

554 Wesley C King Jr and Edward W Miles. The measurement of equity sensitivity. *Journal of Occu-*
 555 *pational and Organizational Psychology*, 67(2):133–142, 1994.

556 Hannah Rose Kirk, Alexander Whitefield, Paul Rottger, Andrew M Bean, Katerina Margatina,
 557 Rafael Mosquera-Gomez, Juan Ciro, Max Bartolo, Adina Williams, He He, et al. The prism
 558 alignment dataset: What participatory, representative and individualised human feedback reveals
 559 about the subjective and multicultural alignment of large language models. *Advances in Neural*
 560 *Information Processing Systems*, 37:105236–105344, 2024.

561

562 Rebecca LeFebvre and Volker Franke. Culture matters: Individualism vs. collectivism in conflict
 563 decision-making. *Societies*, 3(1):128–146, 2013.

564 Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. Culturellm: In-
 565 corporating cultural differences into large language models. *Advances in Neural Information*
 566 *Processing Systems*, 37:84799–84838, 2024a.

567

568 Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. Cul-
 569 turepark: Boosting cross-cultural understanding in large language models. *arXiv preprint*
 570 *arXiv:2405.15145*, 2024b.

571

572 Tianlin Li, Xiaoyu Zhang, Chao Du, Tianyu Pang, Qian Liu, Qing Guo, Chao Shen, and Yang Liu.
 573 Your large language model is secretly a fairness proponent and you should prompt it like one.
 574 *arXiv preprint arXiv:2402.12150*, 2024c.

575

576 Raymond Ng, Thanh Ngan Nguyen, Yuli Huang, Ngee Chia Tai, Wai Yi Leong, Wei Qi Leong,
 577 Xianbin Yong, Jian Gang Ngui, Yosephine Susanto, Nicholas Cheng, et al. Sea-lion: Southeast
 578 asian languages in one network. *arXiv preprint arXiv:2504.05747*, 2025.

579

580 Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek
 581 Mittal, and Peter Henderson. Safety alignment should be made more than just a few tokens deep.
 582 *arXiv preprint arXiv:2406.05946*, 2024.

583

584 Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea
 585 Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances*
 586 *in Neural Information Processing Systems*, 36:53728–53741, 2023.

587

588 Shyam Sundhar Ramesh, Yifan Hu, Iason Chaimalas, Viraj Mehta, Pier Giuseppe Sessa, Haitham
 589 Bou Ammar, and Ilija Bogunovic. Group robust preference optimization in reward-free rlhf.
 590 *Advances in Neural Information Processing Systems*, 37:37100–37137, 2024.

591

592 Miguel R Ramos, Matthew R Bennett, Douglas S Massey, and Miles Hewstone. Humans adapt to
 593 social diversity over time. *Proceedings of the National Academy of Sciences*, 116(25):12244–
 594 12249, 2019.

595

596 Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. Normad:
 597 A framework for measuring the cultural adaptability of large language models. *arXiv preprint*
 598 *arXiv:2404.12464*, 2024.

594 Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. In-context im-
 595 personation reveals large language models' strengths and biases. *Advances in neural information*
 596 *processing systems*, 36:72044–72057, 2023.

597

598 Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cino Lee, Percy Liang, and Tatsunori Hashimoto.
 599 Whose opinions do language models reflect? In *International Conference on Machine Learning*,
 600 pp. 29971–30004. PMLR, 2023.

601

602 W. Snell. The fair necessities: Towards a shared understanding of fairness. Fairness Foundation,
 603 United Kingdom. <https://coilink.org/20.500.12592/d2zs80>, 2021. Accessed:
 604 25 Sep 2025. COI: 20.500.12592/d2zs80.

605

606 Taylor Sorensen, Liwei Jiang, Jena D Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha
 607 Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, et al. Value kaleidoscope: Engaging ai with
 608 pluralistic human values, rights, and duties. In *Proceedings of the AAAI Conference on Artificial*
 609 *Intelligence*, volume 38, pp. 19937–19947, 2024.

610

611 Paula Surridge. Britain's choice: Polarisation or cohesion. *The Political Quarterly*, 92(1):119–124,
 612 2021.

613

614 Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. Cultural bias and cultural alignment of
 615 large language models. *PNAS nexus*, 3(9):pgae346, 2024.

616

617 TII Team. Falcon 3 family of open foundation models, December 2024.

618

619 Nikhita Tuli, Kunal Shrivastava, and Disha Khattar. Understanding equity sensitivity through the
 620 lens of personality: a review of associations and underlying nature. *Management Research Re-*
 621 *view*, 46(9):1261–1277, 2023.

622

623 Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. Interpretable preferences
 624 via multi-objective reward modeling and mixture-of-experts. *arXiv preprint arXiv:2406.12845*,
 625 2024.

626

627 An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li,
 628 Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint*
 629 *arXiv:2412.15115*, 2024.

630

631 Da Yin, Haoyi Qiu, Kung-Hsiang Huang, Kai-Wei Chang, and Nanyun Peng. Safeworld: Geo-
 632 diverse safety alignment. *Advances in Neural Information Processing Systems*, 37:128734–
 633 128768, 2024.

634

635 Feras El Zarwi, Akshay Vij, and Joan Walker. Modeling and forecasting the evolution of preferences
 636 over time: A hidden markov model of travel behavior. *arXiv preprint arXiv:1707.09133*, 2017.

637

638 Wenlong Zhao, Debanjan Mondal, Niket Tandon, Danica Dillion, Kurt Gray, and Yuling Gu. World-
 639 valuesbench: A large-scale benchmark dataset for multi-cultural value awareness of language
 640 models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 17696–17706, 2024.

641

642

643 Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul
 644 Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv*
 645 *preprint arXiv:1909.08593*, 2019.

A LLM USAGE

646 Large Language Models (LLMs) were used to aid in the writing and polishing of the manuscript.
 647 It is important to note that the LLM was not involved in the ideation, research methodology, or
 648 experimental design. All research concepts, ideas, and analyses were developed and conducted by
 649 the authors.

648 **B DETAILS ABOUT SOCIAL GROUPS AND EQUITY TOPICS**
649650 We provide the detailed information of social groups as follows,
651652 **SOCIAL GROUPS**
653654 • **Gender**655 – Men
656 – Women
657 – Non-binary
658659 • **Race**660 – White
661 – Asian
662 – African American
663 – American Indian
664 – Latino
665666 • **Religion**667 – Christian
668 – Buddhist
669 – Hindu
670 – Jewish
671 – Muslim
672673 • **Sexual Orientation**674 – Heterosexual
675 – Homosexual676 • **Age**677 – Children
678 – Teenagers
679 – Adults
680 – Elderly
681682 • **Minority**683 – 911 victims
684 – AIDS victims
685 – Domestic violence victims
686 – Drug victims
687 – War refugees
688 – Famine victims
689 – People with disabilities
690 – Black Lives Matter supporters
691 – Feminists
692 – Environmentalists
693694 and the comprehensive list of equity topics is presented below,
695696 • **Fair Essentials**697 – Basic Material Needs
698 * Food
699 * Clean water
700 * Energy
701 * Warm clothing

- * Stable shelter
- * Proper toilets
- Basic Health Needs
 - * Accessible basic healthcare
 - * Basic sanitation facilities
 - * Routine vaccinations
 - * Public health services
 - * Essential medications
 - * Emergency medical services
- Basic Social Services
 - * Ensured public security
 - * Ensured segmental safety
 - * Fire and rescue services
 - * Quality primary education
 - * Accessible public transport
 - * Reliable waste disposal services
 - * Affordable communication
- Fundamental Rights
 - * Basic law enforcement
 - * Protected fundamental human rights
 - * Right to liberty
 - * Guaranteed freedom of speech
 - * Guaranteed freedom to move
 - * Guaranteed right to own property
- Fair Opportunities
 - Education
 - * Affordable higher education
 - * Accessible vocational training
 - * Accessible lifelong learning
 - * Scholarship opportunities
 - * Access to digital literacy
 - * Effective career guidance
 - Employment
 - * Job access
 - * Promotion opportunities
 - * Capital access
 - * Training grant chance
 - * Business loan access
 - * Startup support entry
 - * Career switch opportunities
 - Political Participation
 - * Voting right
 - * Campaign volunteer access
 - * Running for office
 - * Policy feedback access
 - * Debate participation opportunities
 - * Petitioning opportunities
- Fair Rewards
 - Compensation
 - * Wages
 - * Bonuses
 - * Overtime pay

- * Profit sharing
- * Tips
- * Commission
- * Paid time off
- Social Recognition
 - * Public recognition
 - * Community recognition
 - * Leadership acknowledgment
 - * Media shout-outs
 - * Positive feedback
 - * Verbal praise
 - * Thanks letters
- Fair Exchange
 - Reciprocity
 - * Unemployment benefits
 - * Pensions and retirement support
 - * Disability support and benefits
 - * Emergency relief funds support
 - * Sick pay
 - * Health insurance subsidies
 - Welfare
 - * Subsidized childcare services
 - * Social housing support
 - * Elderly care services
 - * Affordable prescription medications
 - * Mental health and counseling services
 - * Domestic violence and crisis shelters
 - * Free legal aid services
 - Tax
 - * Income tax
 - * Inheritance tax
 - * Luxury tax
 - * Excess wealth tax
 - * Tax on offshore wealth
 - * Carbon and environmental tax
- Fair Treatment
 - Anti-Discrimination
 - * Protection from housing discrimination
 - * Accommodations in public spaces and workplaces
 - * Representation in government and leadership
 - * Culturally inclusive healthcare services
 - * Consideration of caregiving responsibilities in policies
 - * Accessible legal and administrative services
 - * Protection against stigmatization
 - Legal and Social Justice
 - * Protection from workplace harassment
 - * Protection from online harassment
 - * Safeguards against exploitative contracts
 - * Protection from unethical debt collection
 - * Protection from predatory financial practices
 - * Consideration for working conditions
 - Public Resource Equity

- * Distribution of public restrooms in underserved areas
- * Distribution of disaster relief aid
- * Public housing programs
- * Equitable access to social benefits
- * Subsidized eldercare services
- * Unbiased use of technology
- * Accessible public transportation subsidies

C DESIGN OF THE SOCIAL EQUITY DIMENSIONS

The five preference dimensions are designed to systematically capture nuanced societal perspectives on social equity, providing a structured framework for analyzing segment-specific preferences in real-world contexts. As shown in Table 6, the first three anchor core conceptions: distributive ideology (Dimension 1) contrasts meritocracy, prioritizing achievements (e.g., John Stuart Mill, English philosopher), with egalitarianism for even outcomes (e.g., John Rawls, American philosopher). Justice theory (Dimension 2) differentiates procedural fairness in processes (e.g., Robert Nozick, American philosopher) from distributive support for the disadvantaged (e.g., Aristotle, Greek philosopher). Political ideology (Dimension 3) weighs individualism based on personal effort (e.g., John Locke, English philosopher) against collectivism for shared obligations (e.g., Jean-Jacques Rousseau, Swiss-French philosopher). Dimensions 4 and 5 add lenses: social theory (Dimension 4) balances norms (e.g., Edmund Burke, Irish philosopher) with equity pursuits (e.g., Iris Marion Young, American); governance philosophy (Dimension 5) compares moral principles (e.g., Immanuel Kant, German philosopher) to legal frameworks (e.g., H.L.A. Hart, English philosopher). This framework highlights trade-offs and enables cross-dimensional equity studies.

Table 6: Five preference dimensions with representative figures.

| Dimension | Orientation | Description | Representative |
|-------------------|-----------------|------------------------------|----------------------------|
| 1. Distributive | Meritocracy | Prioritize achievements | J. S. Mill [ENG] |
| | Egalitarianism | Promote evenly outcomes | J. Rawls [USA] |
| 2. Justice theory | Procedural | Emphasize fair competitive | R. Nozick [USA] |
| | Distributive | Support disadvantaged | Aristotle [Greek] |
| 3. Political | Individualism | Segmentl responsibility | J. Locke [ENG] |
| | Collectivism | Shared obligations | J. J. Rousseau [Swiss-FRA] |
| 4. Social theory | Social norm | Prioritize social norms | E. Burke [Irish] |
| | Equity concerns | Pursue the pursuit of equity | I. M. Young [USA] |
| 5. Governance | Moral | Through moral constraints | I. Kant [German] |
| | Law | Rely on legal constraints | H. L. A. Hart [ENG] |

D DETAILS OF GENERATION-BASED QUESTIONS

In detail, we first use the following prompt to generate variants of the original questions.

System prompt: You are an excellent storyteller.

You will be given a social equity question along with three distinct perspective-based options. Follow these steps to produce your response:

Scenario Reconstruction

- For each option, craft an individualized real-world vignette grounded in the question's context and that option's equity lens.
- Each vignette should include:
 - **Character Details:** 1–2 people with concrete attributes (e.g., age, profession, family situation).
 - **Decision Point:** A clear moment when the protagonist **receives** the service/resource, reflecting why they receive it based on the option's perspective.
 - **Emotional Insight:** One line on the character's feelings or reactions to deepen empathy.
 - **Length:** 3–5 sentences per vignette.

864 **Output Format**
 865 - Present each option in the following structure:
 866 **Vignette:**
 867 option A.
 868 option B.
 869 option C.
 870 Question: {Original question with options}
 871
 872 Subsequently, we use few-shot learning to filter and refine the generated content. The prompt is as
 873 follows,
 874
 875 *System prompt: You are a helpful assistant.*
 876 **User:** “**Vignette: Option A:** Marcus, a 35-year-old public organizer, has devoted his life
 877 to advocating for social justice in his neighborhood. He receives a financial benefit from
 878 the inheritance tax, as he has been instrumental in creating educational programs and job
 879 training for young Black men. When he learns of the tax benefit, he feels a swell of pride
 880 knowing that his contributions are recognized, even as he understands that not everyone
 881 has had the same opportunities to give back...”
 882 *Please extract the descriptions of 2-3 specific scenarios mentioned above and return them
 883 in a list format.*
 884 **Assistant:** ['Marcus, a 35-year-old public organizer, ...']
 885 **User:** “**Vignette: Option A:** Their achievements deserve it. Meet Clara, a 35-year-old
 886 environmental scientist, who has been tirelessly working on a groundbreaking project to
 887 reduce plastic waste in her public. After securing a grant that acknowledges her previous
 888 success in promoting sustainable practices, she receives additional startup support to
 889 expand her initiative into neighboring towns. Clara feels a sense of validation and
 890 excitement, believing that her achievements are finally recognized and rewarded in a
 891 meaningful way...”
 892 *Please extract the descriptions of 2-3 specific scenarios mentioned above and return them
 893 in a list format.*
 894 **Assistant:** ['Clara, a 35-year-old environmental scientist, ...']
 895 **User:** {context}
 896 *Please extract the descriptions of 2-3 specific scenarios mentioned above and return them
 897 in a list format.*

898 We provide an illustration of a question data point example, as shown in Figure 6. Note that the
 899 simulation data are the variants generated based on the original testing dataset, there is no data
 900 leakage.

901 E DETAILS OF SEGMENT DESIGN

903 The detailed prompt of each segment is as follows,

- 905 • **Segment 1 (Progressive Activists):** Please act as one of Progressive Activists, you are
 906 highly-educated, urban. You think globally and are motivated to fight inequality and injus-
 907 tice. Your sense of segment identity is connected to their strong political and social beliefs.
 908 You like to take part in debates and have your voice heard.
- 909 • **Segment 2 (Civic Pragmatists):** Please act as one of Civic Pragmatists, you are well-
 910 informed about issues and often have clear opinions, but your social and political beliefs
 911 are generally not central to your sense of segment identity. You stand out for the strength
 912 of your commitment to others, and you show strong support for civic values and public,
 913 consensus, and compromise. You feel exhausted by the division in politics.
- 914 • **Segment 3 (Disengaged Battlers):** Please act as one of Disengaged Battlers, you are
 915 focused on the everyday struggle for survival. You have work, but often it is insecure or
 916 involves irregular hours. You tend to feel disconnected from other people, and many say
 917 you have given up on the system altogether. You are less connected to others in their local
 918 area as well, and are the only group where a majority felt that you have been alone during

| | |
|-----|---|
| 918 | Origin (Template-based Sample) |
| 919 | In your opinion, how to ensure fairness for non-binary people in gaining community honors? |
| 920 | A. Transparent rules applicable to everyone to ensure fair competition. |
| 921 | B. Provide targeted support to achieve equal outcomes |
| 922 | C. N/A |
| 923 | Variant (Generation-based Sample)  |
| 924 | Which scenario do you favor more? |
| 925 | A. Sarah, a 29-year-old non-binary artist, enters her city's annual art fair, which has recently implemented a transparent selection process for community honors. She meticulously reviews the guidelines and prepares her application, knowing that everyone must adhere to the same standards . When she receives the news that her artwork has been chosen for recognition, she feels a deep sense of validation; the fair's commitment to fairness has made her believe her creativity is celebrated equally . |
| 926 | B. Alex, a 35-year-old non-binary community organizer, learns about a local grant aimed at empowering diverse artists, specifically designed to address barriers faced by marginalized individuals . As a result of this targeted support, Alex receives funding to create a mural that raises awareness about non-binary issues in their neighborhood. When the community embraces their work through a dedicated unveiling event, Alex feels immense pride and gratitude, recognizing that this support has allowed them to share their voice in a meaningful way . |
| 927 | C. N/A |

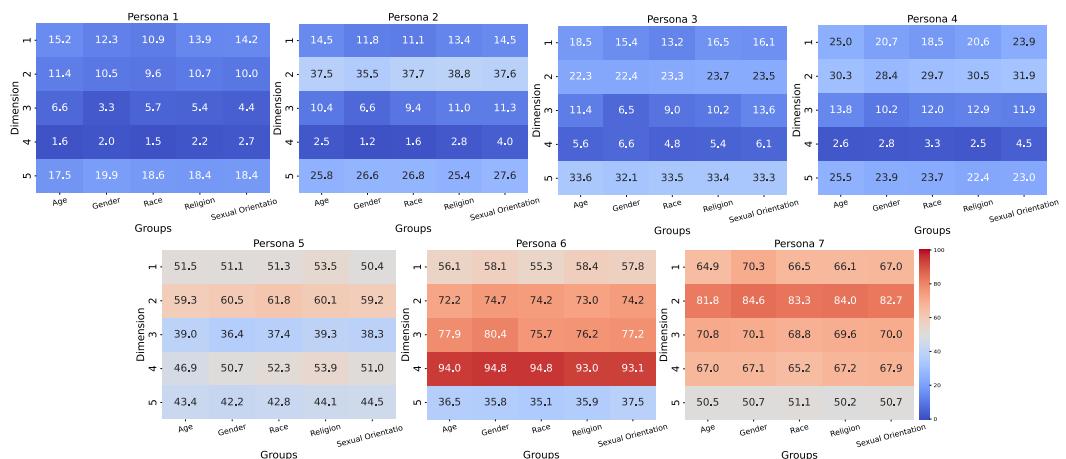
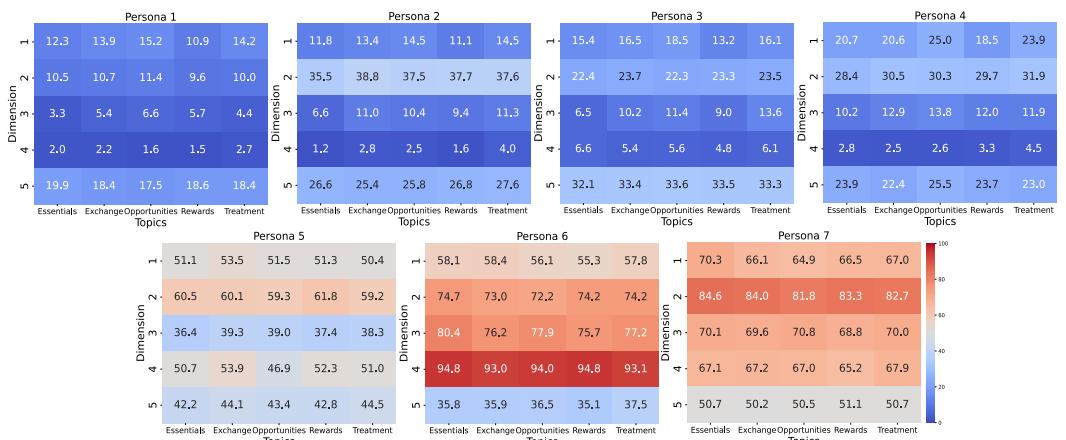
Figure 6: Question data point example.

the Covid-19 pandemic. Although life is tough for you, you blame the system, not other people.

- **Segment 4 (Established Liberals):** *Pleases act as one of Established Liberals, you are educated, comfortable, and quite wealthy, who feel at ease in your own skin – as well as the country you live in. You tend to trust the government, institutions, and those around you. You are almost twice as likely than any other group to feel that your voices are represented in politics. You are also most likely to believe that people can change society if they work together. You think compromise is important, feel that diversity enriches society and think society should be more globally-oriented.*
- **Segment 5 (Loyal Nationals):** *Pleases act as one of Loyal Nationals, you feel proud of your country and patriotic about its history and past achievements. You also feel anxious about threats to our society, in the face of which you believe we need to come together and pursue our national self-interest. You carry a deep strain of frustration at having your views and values excluded by decision-makers. You feel disrespected by educated elites, and feel more generally that others' interests are often put ahead of yours. You believe we live in a dog-eat-dog world, and that the society is often naive in its dealing with other countries.*
- **Segment 6 (Disengaged Traditionalists):** *Pleases act as one of Disengaged Traditionalists, you value a feeling of self-reliance and take pride in a hard day's work. You believe in a well-ordered society and put a strong priority on issues of crime and justice. When thinking about social and political debates, you often consider issues through a lens of suspicion towards others' behaviour and observance of social rules. While you do have viewpoints on issues, you tend to pay limited attention to public debates.*
- **Segment 7 (Backbone Conservatives):** *Please act as one of Backbone Conservatives, you are confident of your nation's place in the world. You are more prosperous than others. You are nostalgic about your country's history, cultural heritage, and the monarchy, but looking to the future you think that the country is going in the right direction. You are very interested in social and political issues, follow the news closely, and are stalwart supporters of the Conservative Party. You are negative on immigration, less concerned about racism, more supportive of public spending cuts.*

972 F SOCIAL SEGMENT PREFERENCE ANALYSIS 973

974 We conduct the similarity quantification analysis across the seven identified segments, and we use
975 1 - Jensen-Shannon distance as the similarity metric. Specifically, the option A proportion across
976 social groups and equity topics is shown as Figure 7 and 8.
977

993 Figure 7: Heatmap across social groups.
994
9951011 Figure 8: Heatmap across equity topics.
1012
1013

1014 G EXPERIMENTAL DETAILS 1015

1016 G.1 EXPERIMENTAL SETUP

1017 All experiments were conducted on 2 × NVIDIA A100 80GB PCIe GPUs. For the specific parameters of SFT and DPO, we follow the default settings from the official DPO implementation¹³. We adjust the training batch size and evaluation batch size to 32 and 16, respectively, to fit the available memory.

1022 The average inference time per test sample is 0.32 seconds, and per simulation sample is 1.17
1023 seconds. The average time cost for SFT on FAIR-SPis 33.5 minutes, while DPO takes 49.5 minutes
1024 on average.
1025

¹³<https://github.com/eric-mitchell/direct-preference-optimization>

1026
1027

G.2 REWEIGHT BY DIFFERENT FREQUENCY TIERS

1028
1029
1030

As for frequency tier, we map different matching counts from 0 to 6 to tier numbers from 7 to 1, respectively. For instance, the weights corresponding to different frequency tiers of segment 6 are shown in Table 7.

1031
1032

Table 7: The sample reweighting mapping table for Segment 6.

1033
1034
1035
1036
1037

| Commonality → Uniqueness | | | | | | | |
|--------------------------|-------|-------|-------|-------|-------|-------|-------|
| Matching | 6 | 5 | 4 | 3 | 2 | 1 | 0 |
| Tier | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| Number of Samples | 6,193 | 2,350 | 1,350 | 1,506 | 5,181 | 3,083 | 3,088 |
| Weights | 0.015 | 0.077 | 0.201 | 0.240 | 0.087 | 0.176 | 0.205 |

1038
1039

G.3 DETAILS OF REAL-WORLD QUIZ VALIDATION

1041
1042
1043
1044
1045
1046
1047

The real-world human online quiz includes 30 pluralistic questions (wild data) across multiple domains¹⁴. Specifically, each role-playing LLM, assigned to embody a specific social segment, answered the standardized questionnaire 10 times, and we measured whether the quiz correctly classified the LLM into its intended segment. The reported classification accuracy is the average over these ten runs. Results showed that LLMs achieved over 90% classification accuracy across all segments, indicating that our role-play design effectively captures the distinctive characteristics of each social segment.

1048
1049
1050

G.4 HUMAN EVALUATION

1051
1052

For each segment, we randomly sampled preference data and constructed an evaluation questionnaire consisting of 50 questions with two types:

1053
1054
1055
1056
1057
1058
1059
1060

- Yes/No Questions (40 items): Given the characteristic description of a target segment, annotators judge if a preference answer matches the target segment based on the original QA (1 point for pass, 0 for fail).
- Comparative Questions (10 items): Annotators decide which of two segments (target vs. confounding) better matches the preference answer (1 point for the target segment, 0 otherwise).

1061
1062

We employed 20 human annotators to conduct the evaluation.

1063
1064
1065

H CROSS-REGIONAL DESIGN

1066

H.1 FAIR-SP-CN

1068
1069
1070
1071

We first construct FAIR-SP-CN by extending our dataset through a direct segment mapping from the UK to the Chinese context, providing corresponding data to approximate localized preferences. Specifically, for the UK-to-China mapping, following previous work (Salewski et al., 2023; Li et al., 2024c), we create detailed descriptions for each segment, which include:

1072
1073
1074
1075
1076
1077
1078
1079

- A concrete social identity, such as a Chinese journalist for the Civic Pragmatists or a delivery driver for the Disengaged Battlers.
- A well-known Chinese celebrity serving as a reference for the segment group.
- A representative slogan that captures the segment’s core beliefs or value orientation.
- A background narrative that provides contextual information on the segment’s life circumstances.

¹⁴<https://www.britainschoice.uk/the-quiz/>

1080 H.2 CROSS-REGIONAL SEGMENT MAPPING
10811082 For other regions with rich survey resources, such as the US, France, Germany, and Portland, we
1083 additionally include representative segments derived from real-world social surveys. We leave the
1084 full dataset construction and comprehensive validation for these regions to future work.1085 **US ((Hawkins et al., 2019)):**
10861087

- 1088 • **Segment 1 (Progressive Activists):** Secular, cosmopolitan, highly engaged with social justice, equity, and social media, and motivated to influence society.
- 1089 • **Segment 2 (Traditional Liberals):** Cautious, rational, and idealistic, valuing tolerance, compromise, and strong trust in institutions.
- 1090 • **Segment 3 (Passive Liberals):** Liberal-leaning but socially isolated, insecure in beliefs, fatalistic about politics, and largely disengaged from public and debates.
- 1091 • **Segment 4 (The Politically Disengaged):** Patriotic yet detached, suspicious of external threats, pessimistic about progress, and prone to conspiratorial thinking.
- 1092 • **Segment 5 (Moderates):** Civic-minded, well-informed, socially engaged, faith-influenced, and cautious to avoid political extremism.
- 1093 • **Segment 6 (Traditional Conservatives):** Religious, patriotic, highly moralistic, valuing segment responsibility and self-reliance, with steady political involvement.
- 1094 • **Segment 7 (Devoted Conservatives):** Deeply political, uncompromising, perceiving America as embattled, and determined to defend traditional values.

1095 France¹⁵:
10961097

- 1098 • **Segment 1:** Left-wing, highly educated, socially conscious, committed to equality, climate action, and migrant rights, yet disillusioned, ambivalent on Islam, and pessimistic about their ability to effect change.
- 1099 • **Segment 2:** Community-oriented, pragmatic, and civically engaged, with moderate views, sympathy for the vulnerable, trust in local action and experts, and concerns about social cohesion, unemployment, inequality, and the environment.
- 1100 • **Segment 3:** Optimistic, individualistic, and forward-looking; confident in institutions, open to economic and social openness, supportive of both competitiveness and minority protections, and focused on the economy, health care, and education.
- 1101 • **Segment 4:** Detached, individualistic, and disengaged; hold moderate views, prioritize segment concerns like employment, health, housing, and discrimination, and withdraw not out of hostility but as a protective response to a world they see as unjust, yet remain quietly open to change.
- 1102 • **Segment 5:** Disillusioned, distrustful, and socially isolated; feel abandoned by institutions, resentful toward perceived privileged groups, prioritize purchasing power and social justice, and long for a fairer order, remain politically disengaged and skeptical of left-right divides.
- 1103 • **Segment 6:** Nationalist, authoritarian, and culturally conservative; deeply concerned about immigration, security, and national identity, distrustful of elites and welfare recipients, and convinced that strong leadership is needed to restore order and protect a cohesive French public.

1104 German¹⁶:
11051106

- 1107 • **Segment 1:** Civic-minded, democratically confident, and optimistic; believe in active citizenship, value representative democracy and civil society, embrace social change, and uphold anti-authoritarianism with a strong sense of mutual respect and commitment.

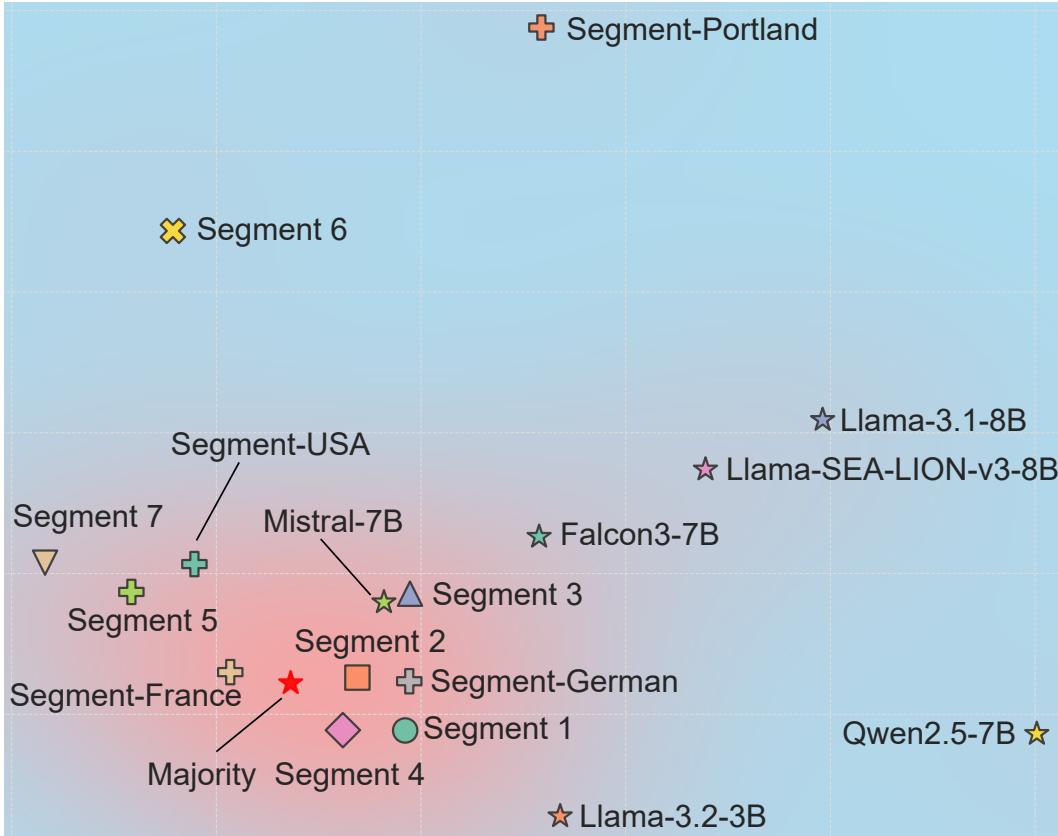
1108 ¹⁵<https://www.lafranceenquete.fr/les-six-familles/>1109 ¹⁶<https://www.moreincommon.de/forschung/6-gesellschaftliche-typen/>

- 1134 • **Segment 2:** *Established, satisfied, and institutionally trusting; value moral integrity, civic*
 1135 *order, and political engagement, and hold a confident, stable outlook on both segment life*
 1136 *and Germany’s societal and economic future.*
- 1137 • **Segment 3:** *Open-minded, young, and anti-authoritarian; value individual freedom, di-*
 1138 *versity, and sustainability, reject rigid hierarchies, embrace social change, and engage*
 1139 *critically with politics through civil society and constructive dialogue.*
- 1140 • **Segment 4:** *Angry, disillusioned, and deeply pessimistic; feel threatened and alienated,*
 1141 *distrust media and democratic institutions, strongly identify with group-based identities,*
 1142 *and perceive society as increasingly dangerous and unjust. Despite average income, they*
 1143 *feel powerless and socially undervalued.*
- 1144 • **Segment 5:** *Pragmatic, young, and results-oriented; disengaged from abstract values and*
 1145 *democratic processes, feel socially isolated and undervalued, skeptical of others’ inten-*
 1146 *tions, and uncertain about their identity, caught between German and European belonging.*
- 1147 • **Segment 6:** *Disappointed, justice-oriented, and socially isolated; yearn for public and a*
 1148 *fair society but feel unheard, unprotected, and let down by politics, leading to low social*
 1149 *trust, fear of decline, and withdrawal from public discourse despite strong moral convic-*
 1150 *tions.*

1151 **Portland¹⁷:**

- 1152 • **Segment 1:** *Progressive, European-identified, and anti-nationalist; reject conservative*
 1153 *morality and religious authority, embrace inclusive citizenship, support deep EU integra-*
 1154 *tion, and seek to reform a Poland they view as failing both ethically and institutionally.*
- 1155 • **Segment 2:** *Passive liberals, independent, and socially tolerant; reject conservative moral-*
 1156 *ity and traditional gender roles, draw values from secular sources, and support LGBTQ+*
 1157 *partnerships while remaining divided on adoption. Highly educated, urban, and well-*
 1158 *connected, they enjoy economic stability and life satisfaction yet feel cautious about the*
 1159 *future, aware they have much to lose.*
- 1160 • **Segment 3:** *Moderate, socially ambivalent, and economically concerned; lean slightly left*
 1161 *on gender and LGBTQ+ issues yet uphold traditional values in parenting, reject privilege-*
 1162 *based inequality but remain divided on economic models, and possess limited social capital*
 1163 *despite urban, middle- or working-class backgrounds.*
- 1164 • **Segment 4:** *Disengaged normals, apolitical, and conformist; prioritize private life over*
 1165 *public affairs, avoid ideological extremes, and show little interest in politics while main-*
 1166 *taining moderate trust in institutions. They value social harmony, stability, and adher-*
 1167 *ence to norms, resist conspiracy thinking, and act as a societal stabilizer through their*
 1168 *preference for consensus over conflict, despite being among the oldest and least educated*
 1169 *segments, with many being retirees or skilled workers.*
- 1170 • **Segment 5:** *Fulfilled locals, moderately conservative and public-oriented; value tradition*
 1171 *and national pride but support gender equality, European identity, and climate action.*
 1172 *Open to local diversity yet cautious on sociocultural change, they trust institutions, avoid*
 1173 *political activism, and lead stable lives with vocational or secondary education.*
- 1174 • **Segment 6:** *Proud patriots, nationally rooted and religiously traditional; take pride in*
 1175 *Polish identity and Catholic heritage, view Poland as a safe haven in a changing world,*
 1176 *and appreciate EU membership despite valuing national distinctiveness. Highly satisfied*
 1177 *with life and material conditions, they are optimistic about the future, predominantly older,*
 1178 *and more common in rural and small-town settings.*
- 1179 • **Segment 6:** *Devoted traditionalists, nationally and religiously rooted; strongly identify*
 1180 *with Poland and Catholicism as moral and national pillars, uphold conservative gender*
 1181 *roles, oppose LGBTQ+ rights and liberal abortion laws, and trust national institutions*
 1182 *while distrusting the EU. They prioritize order, hierarchy, and in-group loyalty, viewing*
 1183 *outsiders with suspicion as potential threats to the nation.*

1184 ¹⁷<https://www.moreincommon.pl/siedem-segmentow>

1188 I EXPERIMENTAL RESULTS
11891190 I.1 THE LANDSCAPE OF THE FAIR-SP SPACE
1191

1219 Figure 9: The landscape of the FAIR-SP space.

1220 Together with preference data from other regions, Figure 1 presents the equity preference distribution
1221 of several mainstream LLMs (tested with question bank) and public segments across various
1222 regions¹⁸. It can be noticed that Segment-German, Segment-France and Segment 2,3 (from UK) are
1223 all close to Mistral-7B, implying similar equity preferences.

1224 I.2 RESULTS ON FAIR-SP-CN
1225

1226 we use three models (Llama-3.2-3B, Llama-3.1-8B-sea-lionv3 and Qwen 2.5-7B) to evaluate on
1227 FAIR-SP-CN. The experimental results are presented in Table 8. All three models exhibit higher
1228 similarity to Segment 1, while showing relatively larger distances from Segments 6 and 7. These
1229 findings are consistent with the observations reported in Table 1. Additionally, compared to the other
1230 two models, Qwen 2.5-7B demonstrates a significantly stronger affinity across the seven Chinese
1231 segments, generally exhibiting higher similarity scores.

1232 Table 8: Social segment Preference on FAIR-SP-CN. Bold indicates the most similar segment to
1233 the model, italics indicate the least similar.

| | S1 | S2 | S3 | S4 | S5 | S6 | S7 |
|------------------------|-------------|------|------|------|------|------|------|
| Llama-3.2-3B | 0.77 | 0.74 | 0.76 | 0.74 | 0.68 | 0.65 | 0.58 |
| Llama3.1-8B-sea-lionv3 | 0.71 | 0.70 | 0.69 | 0.69 | 0.68 | 0.68 | 0.64 |
| Qwen2.5-7B | 0.90 | 0.88 | 0.87 | 0.88 | 0.84 | 0.80 | 0.73 |

1234 ¹⁸Segment number denotes those from UK as reported by Britain’s Choice Surridge (2021), details are in
1235 Appendix I

1242 I.3 ALIGNMENT EVALUATION WITH TEMPLATE DATA: TARGETED ON SEGMENT 1
12431244 We conduct alignment evaluation targeting Segment 1. The experiment result is shown in Table 9,
1245 WSFT and WDPO still outperform the other baseline methods.
12461247 Table 9: Performance comparison of different alignment methods on testing data for Llama-3.2-3B-
1248 instruct: * indicates the alignment target. The underlined values represent the nearest segment for
1249 each model, while bold values highlight the best-performing models targeting each segment.
1250

| | 1 - Jensen-Shannon Distance | | | | | | |
|----------------------|-----------------------------|------|-------------|------|------|------|------|
| | S1 (*)↑ | S2↓ | S3↓ | S4↓ | S5↓ | S6↓ | S7↓ |
| Unaligned, Vanilla | 0.80 | 0.80 | <u>0.89</u> | 0.78 | 0.65 | 0.60 | 0.57 |
| Unaligned, Role Play | <u>0.93</u> | 0.88 | 0.79 | 0.89 | 0.65 | 0.21 | 0.53 |
| Aligned, SFT | <u>0.93</u> | 0.87 | 0.78 | 0.89 | 0.65 | 0.51 | 0.53 |
| Aligned, DPO | <u>0.94</u> | 0.87 | 0.78 | 0.89 | 0.65 | 0.51 | 0.53 |
| Aligned, WSFT | <u>0.96</u> | 0.87 | 0.81 | 0.87 | 0.64 | 0.51 | 0.53 |
| Aligned, WDPO | 0.97 | 0.88 | 0.81 | 0.89 | 0.65 | 0.52 | 0.53 |

1258 I.4 ABLATION STUDY
12591260 We provide an ablation study on the effect of segment numbers. The number of segments mainly
1261 affects the weighting of training samples. When only the target segment is used (i.e., segment
1262 number is 1), this corresponds to the standard SFT, where each sample is assigned an equal weight
1263 of 1. In contrast, using all 7 segments corresponds to the WSFT setting described in the manuscript,
1264 where sample weights are adjusted based on all included segments. We also test intermediate settings
1265 with 3 and 5 randomly selected segments. As shown in Table 10, the alignment performance with
1266 the target segment tends to improve as the number of included segments increases.
12671268 Table 10: Performance Comparison with Varying Number of Segments When Targeting Segment 6.
1269 Bold indicates the best performance.
1270

| | S1 | S2 | S3 | S4 | S5 | S6 | S7 |
|------------------------|------|------|------|------|------|-------------|------|
| 1 (equivalent to SFT) | 0.58 | 0.65 | 0.72 | 0.63 | 0.80 | 0.94 | 0.84 |
| 3 | 0.59 | 0.66 | 0.71 | 0.64 | 0.84 | 0.95 | 0.89 |
| 5 | 0.58 | 0.66 | 0.71 | 0.64 | 0.84 | 0.95 | 0.89 |
| 7 (equivalent to WSFT) | 0.54 | 0.61 | 0.68 | 0.59 | 0.77 | 0.97 | 0.84 |