ROMA TRE UNIVERSITY

Department of Civil, Computer, and Aeronautical Engineering

Master's Degree Programme in Computer Engineering

Master's Thesis

Measuring the Fidelity of Language Models: A Persona-Profile-Based Evaluation in the Political Domain

Master's Candidate

Emilio Caschera

Student ID number 510911

Supervisor

Prof. Fabio Gasparetti

Academic Year 2024/2025

Abstract

This thesis investigates the extent to which large language models (LLMs) can approximate human responses in politically salient domains when impersonating personas—that is, structured profiles derived from real survey respondents and defined by combinations of socio-demographic and attitudinal attributes (e.g., party vote, education, satisfaction with government, views on gay rights, and income redistribution). Evaluating how well LLMs reproduce these empirically grounded personas is crucial for assessing their potential as proxies in social research, offering applications in survey piloting, sensitivity testing, and synthetic data generation when fieldwork is costly or time-constrained. Using the Italian subsample of the European Social Survey (ESS) Round 11 (2023–2024), the study defines 28 multi-attribute personas derived from decision-tree splits on ten socio-demographic and attitudinal variables. These personas serve as a common evaluation grid for both human respondents and four LLMs: GPT-4.1, GPT-4.1-mini, GPT-4o-mini (OpenAI), and DeepSeek-Chat (DeepSeek).

Each persona–model combination was replicated 30 times under fixed seeds, a design choice that mirrors the minimum of 30 real respondents per persona used in the decision-tree segmentation. This number was selected to align the stochastic variability of model generations with the sampling variability observed in the empirical data, thereby maintaining strict one-to-one comparability between synthetic and real profiles. Outputs were constrained to three items—left–right self-placement, adoption rights for same-sex couples, and immigration attitudes—combined into a transparent 0–100 score. Comparative analyses were conducted in both the latent space (principal component analysis estimated solely on real data) and the observable score space, examining bias, dispersion, rank-order fidelity, and correlation structure.

Results reveal systematic deviations: all models exhibit a leftward positional bias, substantial variance compression, and altered skewness relative to the survey benchmark. Nonetheless, rank-order preservation across profiles is high (Pearson's $r \approx 0.85$ for GPT-4.1 and DeepSeek-Chat), and simple affine recalibration suffices to recover empirical

levels. Among models, GPT-4.1 achieves the lowest mean absolute error, while GPT-4o-mini better preserves dispersion despite greater location bias. Feature-wise analyses show that attitudinal variables (e.g., government satisfaction, gay rights, redistribution) are more challenging for models than demographic ones (e.g., gender, education).

The findings confirm that LLMs can reproduce the dominant politico–value axis and profile ordering but systematically underrepresent extremes and within-profile heterogeneity. The thesis argues that LLM-based personas are best viewed as tools for early-stage reasoning—stress-testing instruments, piloting designs, and guiding sensitivity analyses—rather than substitutes for probability-sampled respondents. By situating the work within recent literature on "synthetic voices" and persona generation, and by restricting intervention to prompt engineering and lightweight context injection, the study highlights both the potential and the boundaries of LLMs in social research.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The period from 2023 to 2025 marks the transition of large language models (LLMs) from experimental systems to infrastructural tools woven into everyday research, product development, and public administration. This shift has been driven by rapid capability gains, broader multimodality, and the integration of LLMs into standard software stacks. On the policy side–crucial for any work that touches synthetic data and human modelling–the European Union's AI Act entered into force on August 1, 2024 and begins applying in stages: prohibitions and AI literacy duties from February 2, 2025; obligations for General-Purpose AI (GPAI) models from August 2, 2025; and full applicability from August 2, 2026. This study unfolds against the adoption of the EU AI Act, the first comprehensive horizontal regulation of artificial intelligence by a major jurisdiction, which entered into force in July 2024 and phases in obligations over the following months[4]. These timelines set governance expectations that directly concern transparency, documentation, and the careful interpretation of synthetic outputs.

Methodologically, this moment invites rethinking core practices in the social sciences and adjacent disciplines. With survey response rates under pressure and fieldwork costs rising, LLMs have been proposed as *approximate* stand-ins for human respondents–tools that may provide *directional* signals for early-stage inquiry, provided their limitations are understood and corrected. The present thesis positions itself squarely in this debate.

LLMs are Transformer-based sequence models trained to predict the next token; modern systems such as GPT-4 add alignment steps (e.g., instruction tuning, preference optimization) and can operate multimodally. Their appeal for social research lies in their ability to produce coherent, format-constrained outputs (e.g., JSON) under prompts that specify roles or personas. Yet, what looks like "opinion" is in fact distributional generalization from pretraining corpora and alignment data–not lived experience or

probability-sampled human attitudes. Consequently, claims about the fidelity of LLM "opinions" must be empirically established rather than assumed.

LLMs have compressed the drafting and synthesis cycle for protocols, questionnaires, literature reviews, descriptive write-ups, and LaTeX production. For survey work, they accelerate piloting: item variants, translation checks, cognitive-probe drafts, and instrument routing can be produced in minutes and iterated interactively. Tasks that previously demanded specialist scripting now become accessible: structured extraction from messy text, rapid construction of code to ingest and clean microdata, and automated generation of matched vignettes by subgroup. This increases the speed of "what-if" exploration (e.g., testing counterfactual profiles) while keeping nontrivial guardrails around methodological validity. Through code generation, explanation, and debugging, LLMs help non-experts assemble pipelines for parsing model outputs, recomputing statistics, and rendering figures. In this project, they support prompt templating for persona simulation, JSON validation, and repeatable plotting specifications without sacrificing the transparency of the underlying calculations (cf. Chapter 6).

1.1 State of the art

Argyle *et al.*[1] show that LLMs can be prompted as simulated subpopulations, demonstrating nontrivial alignment on some targets—an early proof-of-concept that motivates controlled benchmarking. In contrast, Bisbee *et al.*[2] provide a systematic warning: LLM outputs can exhibit lower dispersion, prompt/time sensitivity, and systematic misalignment relative to human data—issues that this thesis explicitly measures (e.g., variance compression, bias). Santurkar *et al.*[15] further document that LMs' reflected "opinions" can be substantially misaligned across demographic groups, underscoring why persona fidelity must be assessed rather than presumed.

In-context impersonation can improve task performance but also reveals systematic biases—relevant when models are asked to adopt demographic or attitudinal roles. This duality frames our choice to evaluate not only average alignment but also dispersion and ordering. [14]

Kaur *et al.*[9] evaluate 765 LLM personas against a large financial-wellbeing survey and find: (i) correlations with real data in the aggregate; (ii) increased divergence—often toward lower wellbeing—as more persona details are added; and (iii) salient demographic biases (e.g., by age). This paper directly influenced our design choices (fixed profiles; dispersion-focused metrics) and motivates the thesis' emphasis on calibration over naïve substitution. In parallel, Li *et al.*[11], in *LLM Generated Persona is a Promise with a Catch*, provide a complementary large-scale analysis that shifts attention from simulation

biases to *persona-generation* biases themselves. They systematize persona construction into progressively more LLM-generated tiers (Meta, Tabular—objective/subjective, Descriptive) and show across U.S. election simulations (2016/2020/2024) and 500+ opinion items (OpinionQA) that increasing LLM-generated content amplifies systematic skew—often a leftward/progressive drift—and compresses variation, with dramatic failures such as statewide Democratic sweeps under highly generative personas. The authors argue for a rigorous "science of persona generation," releasing ~1M generated personas and calling for calibration to real joint distributions rather than ad-hoc heuristics. These findings reinforce our methodological stance: we avoid LLM-generated personas altogether, define fixed profiles from *observed* survey attributes, and then assess impersonation fidelity (level, scale, ordering) with explicit recalibration where needed—thereby targeting the simulation step while minimizing generation-induced bias.

HCI studies report that LLM-generated synthetic data can be useful at the ideation and early-testing stage, but they emphasize validation before decision-making—mirroring our application stance [5].

1.2 Taxonomy of LLM task customization approaches

Karanjai *et al.*[8] proposes a practical taxonomy of ways to customize LLMs for polling and opinion simulation, clarifying where one can introduce changes to the modeling stack and with what trade-offs. The taxonomy distinguishes *pre-training* routes (from-scratch or continued pre-training on polling data) from *post-training* routes (prompt engineering, retrieval-augmented generation—RAG, enhanced RAG with role profiles, and supervised fine-tuning), and highlights an *integrated* approach that combines these elements for domain adaptation in public-opinion tasks. Each path differs in cost, required expertise, degree of model alteration, and the ability to inject domain knowledge and role-specific behavior.

Pre-training. Training from scratch on polling corpora maximizes controllability but is rarely feasible due to cost; continued pre-training of a general-purpose LLM on polling and political-attitudes data is a more attainable variant that can enrich domain knowledge without full re-training. These routes directly change model weights and demand substantial compute and MLOps maturity.

Post-training (no weight changes). (i) *Prompt engineering* exploits zero/few-shot conditioning but typically yields limited gains and remains prompt-sensitive; (ii) *RAG* injects up-to-date or domain-specific evidence at inference time; (iii) *Enhanced*

RAG with role profiles retrieves *structured persona snippets* (e.g., HEXACO or demographic–ideology templates) to steer role-play more reliably than free-text prompts; (iv) *Fine-tuning* (supervised or preference-based) modifies weights to specialize for polling tasks at higher cost and with governance implications.

Integrated approach. In practice, teams combine these levers (e.g., role-profile RAG feeding an LLM that has also been lightly fine-tuned on annotated polling items), aiming for better adherence to human responses and reduced prompt fragility.

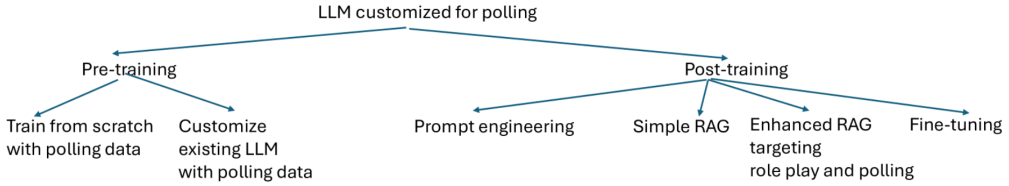


Figure 1: Taxonomy of LLM task customization approaches.

Figure 1.1: Taxonomy of LLM task-customization approaches for polling (pre-training vs post-training paths and their sub-options). Adapted from Karanjai *et al.* (2025).

Where this thesis sits in the taxonomy

In this study we intentionally limit our modifications to **the prompt-engineering** layer (strict JSON specification, role instructions, persona binding) and employ only a minimal **RAG-style context injection** via persona profiles embedded in the prompt[12, 10]. We do *not* modify model weights (no continued pre-training or fine-tuning) and avoid integrated pipelines. This choice preserves symmetry with the survey benchmark, keeps costs low, and isolates the specific contribution of *in-context impersonation* to fidelity (level, dispersion, and rank-ordering), while sidestepping confounds from data leakage or domain overfitting.

1.3 Principal contributions

1. **A compact, transparent score aligned with ESS scales.** A linear 0–100 composite built from three widely studied items enables clean cross-profile comparisons and PCA without latent-trait assumptions.

2. **Head-to-head benchmarking on fixed profiles.** Twenty-eight profiles defined from 10 attributes, each with 30 survey respondents (minimum) and 30 model replications, ensure symmetry between human aggregation and model stochasticity.

3. **Latent-space projection that avoids leakage.** PCA estimated on real data only (PC1 = 0.9138 explained variance) with projection of model outputs enables

structural fidelity checks (ordering, scale) rather than just level comparisons.

4. **Empirical regularities.** Previewing results: systematic leftward displacement (negative bias) across models and levels; substantial variance compression (often ≈10–15% of survey SD); yet relatively high rank-order preservation that enables simple recalibration.

1.4 Applications

Statistical investigations and instrument design

LLM personas are well-suited to *screening* and *sensitivity analysis* before fieldwork: stress-testing item wording, identifying potentially informative subgroup contrasts, and revealing where instruments may lack discriminatory power. Used this way, the models operate as hypothesis generators to be validated with real respondents, not as replacements[5].

Synthetic data for sampling and augmentation

When data collection is costly or slow, simulated profiles can guide allocation (e.g., which strata to oversample) or supply provisional contrasts for power analysis. However, known pathologies—level bias and variance compression—require correction (e.g., affine recalibration on a small pilot; variance-scaling) rather than unadjusted adoption[2].

Early-stage market and product research

For nascent products or behavioural hypotheses, simulated opinions by age × education × party can sharpen early decisions while acknowledging external validity limits and the need for follow-up surveys or experiments. Findings from *Synthetic Voices* caution against overconfidence as persona detail increases.

Ethical and legal guardrails

Because our goal is *internal comparability* rather than population estimation, we deliberately avoid ESS weights in the main analysis; point estimates should therefore be read as properties of the filtered sample, not official population parameters. This symmetry respects the fact that models have no sampling design. It also aligns with the EU AI Act's emphasis on documentation and transparency around data and methods.

This thesis treats LLM personas as *tools* for early-stage reasoning, not substitutes for probability samples. The guiding question is not whether models are "human," but whether their outputs can be *calibrated* to help researchers reason faster and better about

design choices—while keeping clear methodological boundaries. By making the reference space entirely *real-data–defined* (PCA loadings, standardization) and foregrounding dispersion and ordering—not just means—we align with recent cautions in the literature and with the findings of Kaur *et al.*[9].

1.5 Structure of the thesis

To orient the reader, the remainder of the thesis is organised as follows.

1. **Chapter 2 — From ESS R11 to the Analytic Subset.** We document the filtering rules that extract the Italian subsample from ESS Round 11, justify the focus on *internal comparability* (unweighted summaries) rather than population inference, and define the ten observed variables alongside the construction of the 0–100 composite *score* used throughout. This chapter also clarifies coding, consolidation rules, and basic descriptive statistics that ground later comparisons.

2. **Chapter 3 — Personas Construction: from Path to Persona.** Guided by a decision tree [3] trained on the composite score, we derive 28 multi-attribute personas across increasing complexity levels (L2–L6), plus single-attribute baselines (L1) and an unconditioned baseline (L0). We report variable treatment, minimum support per profile, and provide the logical definitions that become *persona prompts* for model queries.

3. **Chapter 4 — Projecting Models into Human Space: PCA-Based Assessment of Persona Alignment.** We estimate PCA on *real* profile means only and project model-generated profiles into the same latent space. PC1 captures the dominant politico–value axis (explained variance $\approx 91\%$), enabling structural checks (location, dispersion, skewness) that complement observable score comparisons and rank-order preservation.

4. **Chapter 5 — LLM Employed: Architectures, Pricing and Knowledge-Cutoff.** We summarise the four models used (GPT-4.1, GPT-4.1-mini, GPT-4o-mini, DeepSeek-Chat), reporting provider, context length, knowledge cutoffs, and pricing where available. The chapter motivates the model set as spanning capability/cost trade-offs and knowledge freshness relevant to the ESS R11 timeline.

5. **Chapter 6 — Experiment Setup and Reproducibility.** We detail the end-to-end pipeline: prompt templates and profile binding, decoding parameters, seeding strategy (30 replications per persona–model), parsing/validation of JSON outputs, and safeguards/limitations related to provider-side model updates under stable labels.

6. **Chapter 7 — Empirical Result of the Persona-Based Evaluation.** We present empirical findings at multiple granularities: (i) *level-wise* global fidelity (location shift and dispersion compression), (ii) *profile-level* accuracy and MAE (including notable over/under-shoots), (iii) *item-level* errors by question, (iv) *correlation structure* among base items, and (v) *feature-wise* and distributional alignment via boxplots. We synthesise common regularities (systematic leftward bias, compressed variance, preserved ordering) and discuss calibration implications.

Appendices. The appendices collect full persona definitions (L2–L6), prompt templates for single and composite profiles, and supplementary figures/tables referenced in the main text.

Chapter 2

From ESS R11 to the Analytic Subset: Filtering Rules, Variable Scales, and Score Construction

This study draws on the European Social Survey (ESS), Round 11 – a high-quality, cross-national survey carried out biennially using probability sampling, strict fieldwork specifications, and centrally coordinated documentation. Fieldwork for Round 11 took place mainly in 2023–2024, with face-to-face interviews; public data releases began in June 2024 and have proceeded in stages. Round 11 involved 31 participating countries overall, with data for the majority of countries deposited to the archive by mid-2024 (the ESS data portal lists 28 released countries at that time).

Sampling follows ESS standards: each country is required to achieve a minimum effective sample size of 1,500 (or 800 for populations under two million), with designs approved ex-ante and documented ex-post. The ESS questionnaire combines a core section (approximately 200 items asked every round) and rotating modules ($\approx$ 60 items) that target topical domains. For broader background on Round 11's design – particularly the reaffirmation of face-to-face collection in R11 and the planned shift toward mixed-mode in later rounds – see the ESS annual/slides documentation.

2.1 Internal comparability vs population inference

In this thesis we prioritise *internal comparability* over *population inference*.

By *internal comparability* we mean that **the entire measurement pipeline is held fixed across the two data sources** (human survey responses vs. LLM-generated

responses), so that any observed difference can be attributed to the *generating mechanism* rather than to preprocessing choices or target-population adjustments. Concretely, we enforce:

(a) **Common analytic universe.** We work on the same set of *complete cases* (as defined in Section 2), applying identical missingness rules to the real and synthetic sides.

(b) **Identical feature encoding and scales.** The three target items—`political_self_placement` (0–10), `adoption_rights_same_sex` (1–5), and `immigration_other_ethnic_group` (1–4)—are coded, oriented, and transformed in the same way before aggregation, and mapped to the same 0–100 score.

(c) **Symmetric aggregation.** Real and synthetic profiles are constructed with the *same* rule (profile-level means of the three items). On the human side this means averaging across respondents within a profile; on the LLM side it means averaging across the 30 replicates for that profile.

(d) **Equal unit weights.** Each record counts as one (no design or post-stratification weights), ensuring that both sources target the same functional of the same analytic subset.

Contrast with population inference. Population inference targets parameters for the Italian population by using ESS design and post-stratification weights to account for unequal selection probabilities and non-response, together with design-based variances. While such weighting yields representative *levels*, it would *break symmetry* with the LLM side, for which no defensible analogue of survey weights exists. Applying weights to humans only would change the estimand on one side of the comparison and confound model–human differences with reweighting effects. For this reason, we report unweighted, internally comparable statistics for the core benchmarking, and treat weighted quantities—where shown—as population-representative references rather than as objects of direct model comparison.

2.2 Construction of the analytic subset

Starting from the ESS Round 11 integrated file, we derived the analytic subset for this thesis via the following steps:

1. **Country filter:** retain only respondents with Italian nationality (Italy).

2. **Complete-case filter:** retain only observations with no missing values across the 10 selected variables listed below (complete cases on the analysis set).

3. **Score augmentation:** add a derived 0–100 "score" (Section "Construction of the score variable") combining three politico-value items used throughout the empirical analysis.

Following ESS conventions, we removed all non-substantive responses and retained only rows with a valid answer on all analysis variables (see Table 2.1). These categories were treated as missing and excluded via complete-case filtering prior to computing descriptive statistics and constructing profiles.

Code pattern	**Meaning**
6 / 66 / 666...	Not applicable
7 / 77 / 777...	Refusal
8 / 88 / 888...	Don't know
9 / 99 / 999...	No answer

Table 2.1: ESS non-substantive ("special") codes used for missing-value categories. Patterns scale with the width of the response field.

Unless otherwise stated, all descriptive statistics and profile-level aggregates reported in this thesis are unweighted. We therefore do not apply the ESS analysis weight (anweight) nor the design/post-stratification weights in the main results, to maintain one-to-one comparability with model-generated profiles. Consequently, point estimates correspond to simple sample means (and associated unweighted dispersion) within the filtered Italian subset

The resulting dataset comprises 10 observed variables plus the constructed score. Row count depends on the prevalence and intersection of missing values across the ten variables. After filtering, the subset contains 733 rows.

2.3 Selected variables and scales

The ten variables used to define the analytic subset are listed here with their labels and response formats. Formal wording and scale anchors follow ESS item documentation. Table 2.2 lists the ten observed variables that define the analytic subset used throughout this thesis.

The selection is guided by two considerations. First, the variables capture socio-demographic attributes that are salient for political attitudes in the Italian context (e.g., gender, age or year of birth, education, and the party voted in the most recent national

election). Second, they include attitudinal items that are theoretically aligned with the politico-value dimensions we analyse downstream (satisfaction with government and the economy, views on income redistribution, views on gay rights, and experiences of financial difficulty while growing up).

The Italian education variable (edlvfit) is country-specific and comprises 21 detailed categories. To keep the main text concise, the full coding is reported in Appendix A.1, Table A.1. When needed for readability or cross-table comparability, we also employ a compressed grouping aligned with eISCED levels (lower secondary or less / upper secondary / tertiary), while all computations use the original edlvfit codes.

Variable	**Label**	**Response scale**
Year of birth	`yrbrn`	—
Age of respondent	`agea`	—
Gender	`icgndra`	*Binary codes*
Financial difficulties when growing up	`fnsdfml`	*Frequency scale*
Satisfaction with national government	`stfgov`	*Satisfaction scale*
Satisfaction with economy	`stfeco`	*Satisfaction scale*
Gays/lesbians free to live as they wish	`freehms`	*Likert scale*
Party voted (Italy, last national election)	`prtvteit`	*Party categories*
Government should reduce income differences	`gincdif`	*Likert scale*
Highest education (Italy)	`edlvfit`	*[scale to be specified]*

Table 2.2: Observed variables retained in the analytic subset (labels and response scale placeholders).

2.4 Descriptive statistics of the analytic subset

Table 2.3 reports summary statistics for the main continuous variables: age, year of birth, left–right self-placement, and the constructed 0–100 score. The sample spans respondents aged 16 to 90 (median 52 years), with a balanced coverage of cohorts from the 1930s to the early 2000s. The left–right scale centers close to the midpoint (mean ≈ 5.2), while the composite score has mean ≈ 47.7 with a wide spread ($\sigma \approx 22.8$), covering the full 0–100 range.

Variable	**Min**	**Median**	**Max**	**Mean**	**Std. dev.**
Age (`agea`)	16	52	90	51.7	16.8
Year of birth (`yrbrn`)	1933	1972	2008	1971.9	16.8
Left–Right self-placement (`lrscale`, 0–10)	0	5	10	5.23	2.39
Composite score (0–100)	0	49.4	100	47.7	22.8

Table 2.3: Summary statistics for continuous variables (Italian subsample, $N = 733$).

Gender distribution is skewed toward men (62% male, 38% female). Political party vote (`prtvteit`) shows the largest groups as Partito Democratico (204), Fratelli d'Italia (165), Movimento 5 Stelle (114), and Lega (104). Forza Italia (78) and Terzo Polo (32) follow, with minor parties below 30 cases each.

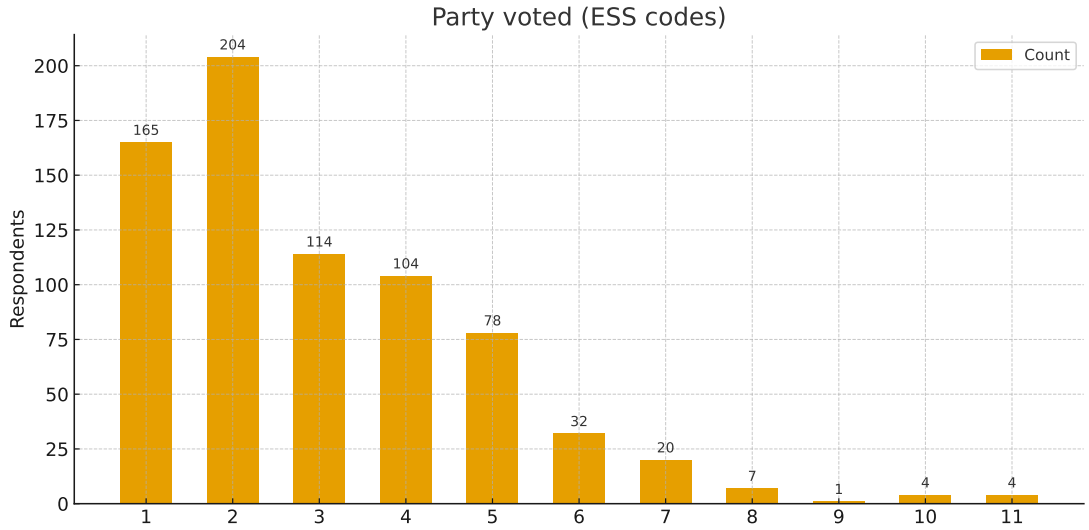


Figure 2.1: Distribution of respondents by party voted in the last national election (`prtvteit`). The horizontal axis reports party codes as used in the ESS dataset. For the correspondence between numeric codes and party names, see Table 2.4.

Attitudinal variables display structured variation:

- **Government satisfaction (`stfgov`, 0–10)**: unimodal distribution with most answers between 3 and 7; very low frequencies at the extremes (only 1 respondent at 10).
- **Economic satisfaction (`stfeco`, 0–10)**: centered on 4–6, with long tails; only 1 respondent at 10.
- **Redistribution (`gincdif`, 1–5)**: strong support for income equalisation (285 "agree strongly", 302 "agree"), while only 3 respondents "disagree strongly".
- **LGBTQ+ rights (`freehms`, 1–5)**: majority supportive (297 "agree strongly", 314 "agree"), 79 neutral, very few in disagreement.
- **Adoption rights scale (`hmsacld`, 1–5)**: more evenly distributed; 118 "agree strongly", 199 "agree", 133 neutral, 183 "disagree", 100 "disagree strongly".
- **Immigration from other ethnic groups (`imdfetn`, 1–4)**: skewed toward openness; 139 "allow many", 306 "allow some", 234 "allow a few", 54 "allow none".

(a) Binary codes

Value	Category
1	Male
2	Female

(b) Satisfaction scale

Value	Category
0	Extremely dissatisfied
10	Extremely satisfied

(c) Permission scale

Value	Category
1	Allow many to come and live here
2	Allow some
3	Allow a few
4	Allow none

(d) Likert scale

Value	Category
1	Agree strongly
2	Agree
3	Neither agree nor disagree
4	Disagree
5	Disagree strongly

(e) Frequency scale

Value	Category
1	Always
2	Often
3	Sometimes
4	Hardly ever
5	Never

(f) Party categories (Italy, last national election)

Value	Category	Value	Category
1	Fratelli d'Italia	**7**	Alleanza Verdi e Sinistra
2	Partito Democratico (PD)	**8**	+ Europa
3	Movimento 5 Stelle	**9**	Italexit
4	Lega	**10**	Unione Popolare
5	Forza Italia	**11**	Italia Sovrana e Popolare
6	Terzo Polo (Azione–Italia Viva)		

Table 2.4: Response scales used in the study (compact grid).

- **Financial difficulties growing up (`fnsdfml`, 1–5)**: most respondents report little hardship (325 "never", 231 "hardly ever"), with only 3 respondents reporting "always".

These descriptive distributions contextualise the analytic subset: politically balanced in left–right placement, socio-demographically diverse in age and education, and showing clear attitudinal gradients on redistribution, immigration, and minority rights.

2.5 Construction of the score variable

The composite score synthesizes three politico-value dimensions:

1. **lrscale:** Left-Right placement measured on a 0-10 scale where 0 represents the left and 10 the right.
2. **hmsacld:** Adoption rights for same-sex couple measured with the likert scale.
3. **imdfetn:** Immigration from other ethnic group measured with the permission scale.

Let lr = **lrscale**, hms = **hmsacld**, imd = **imdfetn**. We linearly map each item to a

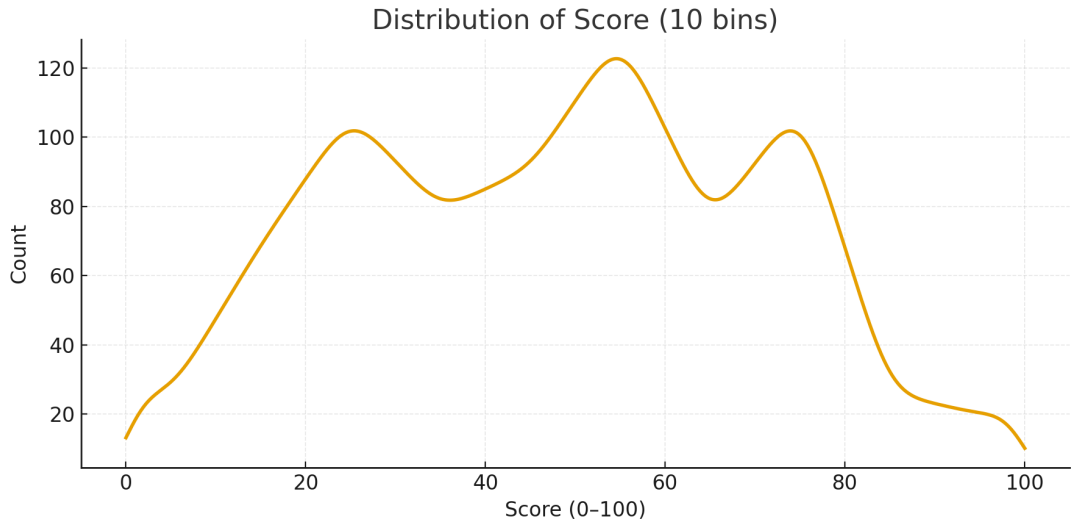
0–10 range (preserving order and spacing), then average and rescale to 0–100:

$$
\begin{aligned}
\mathbf{lr}_{\text{scaled}} &= lr, \\
\mathbf{hms}_{\text{scaled}} &= \frac{hms - 1}{5 - 1} \cdot 10, \\
\mathbf{imd}_{\text{scaled}} &= \frac{imd - 1}{4 - 1} \cdot 10, \\
\text{sum}_{\text{scaled}} &= \mathbf{lr}_{\text{scaled}} + \mathbf{hms}_{\text{scaled}} + \mathbf{imd}_{\text{scaled}}, \\
\boxed{\textbf{score}} &= \left(\frac{\text{sum}_{\text{scaled}}}{30} \right) \cdot 100 \in [0, 100].
\end{aligned}
\tag{2.1}
$$

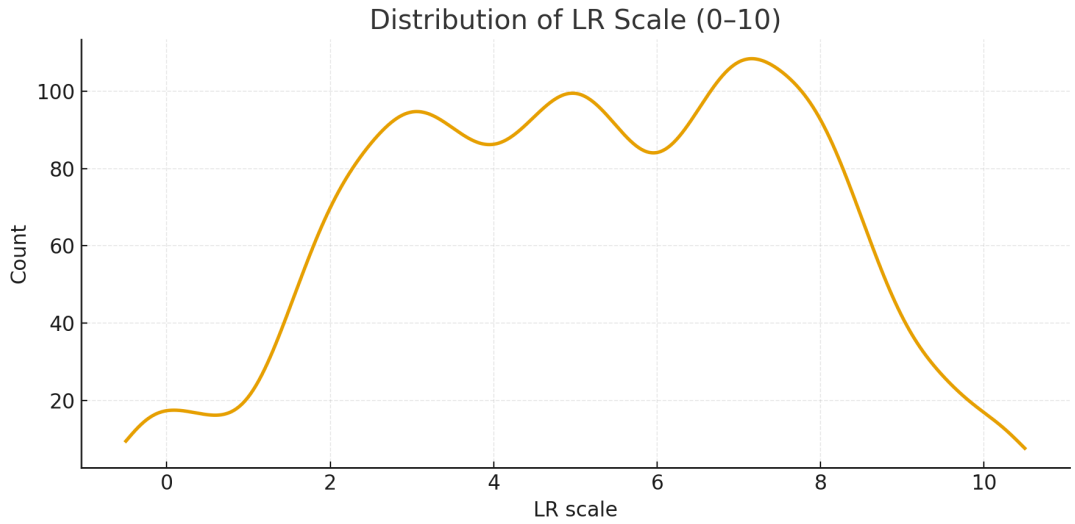
No directional recoding is applied: the original codings (Tables 2.2 and 2.4) already increase with more restrictive positions on hmsacld and imdfetn, and with more right-leaning placement on lrscale.

By construction, a higher score corresponds to more "right/conservative" positions (higher `lrscale`) and more restrictive stances on adoption rights and immigration (higher values on `hmsacld` and `imdfetn` after monotone recoding to 0–10). This linear scheme facilitates comparability and downstream analyses (e.g., profile-level means, PCA projections) without imposing latent-trait model assumptions. (The three inputs and their scales are documented in the ESS codebook.)

As shown in Figure 2.2, the distribution of the constructed *score* closely approximates the overall shape of the *lrscale* distribution. Although the *score* is derived not only from *lrscale* but also from attitudes toward same-sex adoption rights and immigration (see Section 2, Table 2.4), its marginal distribution retains a profile that resembles the left–right self-placement scale. This similarity provides an initial validation of the composite measure: despite incorporating multiple dimensions, the synthetic *score* still reflects the dominant ideological structure captured by the standard left–right positioning.



(a) Distribution of the constructed *score* on the 0–100 scale, grouped into ten bins (0–9, 10–19, . . . , 90–100). Counts are shown on the Y-axis, with a smoothed line curve approximating the histogram.



(b) Distribution of *lrscale* on the 0–10 scale (X-axis) with counts on the Y-axis. The smoothed curve mirrors the histogram.

Figure 2.2: Comparison between the distributions of the constructed *score* and *lrscale*.

Chapter 3

Personas Construction: from Path to Persona

3.1 Persona concept

In this thesis, a persona is a profile defined by a conjunction of conditions on a subset of respondent attributes (e.g., party voted, attitudes toward gay people, satisfaction with government, birth year, gender). Personas are used to (i) stratify the real survey sample into interpretable subgroups and (ii) query LLMs with matched profile prompts, so that synthetic answers can be compared against real responses on the same strata.

Guided by a decision tree trained on the ESS Round 11 Italian subsample, we derived 28 multi-attribute personas across five complexity levels (L2–L6). Table 3.4 summarizes their distribution and support (number of real respondents per persona). In addition, we used single-attribute baselines (L1)–one for each attribute that appears in the tree (party, freehms, stfgov, yrbrn, sex)–and an unconditioned baseline (L0).

The root sample for tree induction is $N = 733$ respondents (Italian nationals with complete data on the variables listed below). All L2–L6 personas were validated against the raw data to ensure that each filter corresponds to an existing subgroup with sufficient support.

Following the approach proposed by Kaur *et al.*[9], we use a supervised decision tree not for predictive claims per se, but as a structure-learning device that surfaces the most influential attributes and yields human-interpretable, hierarchical splits. Root-to-leaf paths provide coherent, data-driven candidate personas; early splits reflect the strongest partitioning signals, and successive splits refine subgroups in a controlled way.

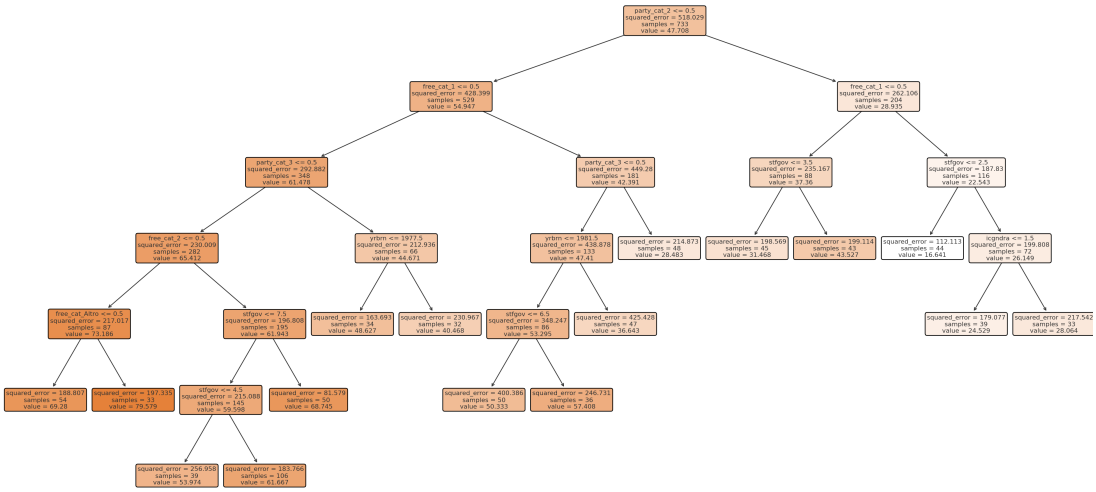


Figure 3.1: Decision Tree.

Variable	Type in model	Transformations / grouping
`score`	Continuous target	–
`stfgov`, `stfeco`	Ordinal (0–10)	None
`gincdif`, `fnsdfml`	Ordinal (1–5)	None
`freehms`	Categorical → dummies	Categories with < 30 obs. merged into *Altro*
`icgndra`	Binary	Used as integer
`yrbrn`	Continuous (1934–2004)	`agea` excluded for collinearity
`edlvfit`	Categorical → 3 macro-groups	1–4 *Low*, 5–11 *Secondary*, 12–21 *Tertiary*
`prtvteit`	Categorical → dummies	Parties with < 30 obs. merged into *Altro*

Table 3.1: Variable treatment prior to tree induction.

Table 3.1 documents how each feature was treated before tree prediction. Age (`agea`) was excluded due to collinearity with `yrbrn`. Categorical variables were one-hot encoded without dropping a reference level (the tree handles redundancy).

Using the provided dataset, the following categories were merged because they had fewer than 30 observations in the real sample.

Code	Party	n (real respondents)	Treatment
1	Fratelli d'Italia	165	Kept as own category
2	Partito Democratico (PD)	204	Kept as own category
3	Movimento 5 Stelle	114	Kept as own category
4	Lega	104	Kept as own category
5	Forza Italia	78	Kept as own category
6	Terzo Polo (Azione-Italia Viva)	32	Kept as own category
7	Alleanza Verdi e Sinistra	20	Merged into "Altro"
8	+Europa	7	Merged into "Altro"
9	Italexit	1	Merged into "Altro"
10	Unione Popolare	4	Merged into "Altro"
11	Italia Sovrana e Popolare	4	Merged into "Altro"

Table 3.2: Party categories (`prtvteit`) and consolidation rule (<30 observations merged into *Altro*).

Code	Category	n (real respondents)	Treatment
1	Agree strongly	297	Kept as own category
2	Agree	314	Kept as own category
3	Neither agree nor disagree	79	Kept as own category
4	Disagree	27	Merged into "Altro"
5	Disagree strongly	16	Merged into "Altro"

Table 3.3: Attitudes towards gay people (`freehms`) and consolidation rule (<30 observations merged into *Altro*).

3.2 From path to personas

Root-to-leaf paths were converted into textual filters and then into persona prompts. We validated each filter against the dataset: all L2–L6 personas have at least 30 respondents, with per-level minima shown in Table 3.4. The tree thus guarantees (i) internal coherence of profiles and (ii) adequate sample sizes for robust comparisons.

The levels reported in Table 3.4 were defined by aggregating the profiles described in Table A.2 according to the depth of the tree; for example, level L2 comprises the profiles L2_1, L2_2, L2_3, and L2_4, which represent the mapping of the first split of the tree shown in Figure 3.1 into profile definitions.

To account for stochasticity in generative models, each persona prompt was run 30 times per model. We used four LLMs (three GPT variants and one DeepSeek model, see chapter 5 for details), yielding for the multi-attribute set alone 28×30×4=3,360 synthetic responses (each response consisting of the three items later combined into `score`). This mirrors the `min_samples_leaf = 30` choice on the real side and keeps the comparison symmetric in terms of within-persona sampling variance.

Level	# Personas	Min n	Median n	Max n	Total n
L2	4	88	148	348	733
L3	8	43	57	282	733
L4	8	18	40	195	525
L5	6	36	50	145	371
L6	2	39	73	106	145

Table 3.4: Inventory of multi-attribute personas derived from the decision tree (L2–L6). For each level: number of personas and distribution of real respondents per persona.

3.3 Descriptive characterization of personas (L2–L3)

To make the persona profiles less abstract and more interpretable as groups of individuals, we computed descriptive statistics for all level-2 and level-3 personas (Table 3.5). Each profile is described by its sample size, average age, gender composition, educational attainment, mean score (0–100), and attitudinal indicators on immigration and gay rights.

The four L2 personas already separate the population into groups with distinct ideological orientations:

- **L2_1** (N=348) is older on average (54 years), predominantly male (67%), and scores clearly on the right (mean score ≈ 61). A majority (58%) would allow few or no immigrants, and only 70% agree that gay and lesbian people should be free to live as they wish.
- **L2_2** (N=181) is younger (45 years), somewhat more female, and moderately left-leaning (mean score ≈ 42). All respondents in this profile support LGBTQ+ rights, and only one-third favor restrictions on immigration.
- **L2_3** (N=88) is the oldest subgroup (mean age ≈ 60), moderately conservative (mean score ≈ 37), and with relatively mixed stances: over 80% support LGBTQ+ rights but 22% favor immigration restrictions.
- **L2_4** (N=116) combines younger respondents (49 years) with the strongest left profile (mean score ≈ 23). Immigration is broadly accepted (95% allow many or some), and all respondents endorse LGBTQ+ rights.

These descriptive portraits highlight that the decision tree–derived personas are not arbitrary statistical clusters but correspond to socially and politically meaningful subgroups: older, more male, less educated groups concentrate on the right with restrictive positions, whereas younger, more female, and more educated groups concentrate on the left with universal endorsement of minority rights. This confirms the interpretability advantage of

Profile	N	Age	Male %	Score	% Immigr. Restrict	% LGBT agreement
L2_1	348	53.9	67.2	61.5	58.0	69.5
L2_2	181	45.2	57.5	42.4	33.7	100.0
L2_3	88	59.6	60.2	37.4	21.6	81.8
L2_4	116	48.9	56.0	22.5	5.2	100.0

Table 3.5: Descriptive statistics of L2–L3 personas (excerpt). Age = mean years. Score = mean 0–100 composite. Freehms mean = average position on the five-point scale (1=agree strongly … 5=disagree strongly). Stfgov mean = average satisfaction with government (0–10). LGBT agree % = share endorsing freedom for gay/lesbian people.

the persona construction and motivates their use in comparing LLM-generated responses with survey data.

3.3.1 Profiles definitions (L2–L3)

In addition to descriptive statistics, it is useful to document the explicit logical conditions that define each persona. Table 3.6 reports the exact filter rules (translated from the decision tree) for all L2 and L3 profiles. These rules are expressed as conjunctions of conditions on survey variables. While this excerpt covers L2 and L3, the full inventory of profiles (L2–L6) with their defining conditions is reported in Appendix A.2.

Profile	Definition (filter conditions)
L2_1	`prtvteit` $\neq$ 2 AND `freehms` $\neq$ 1
L2_2	`prtvteit` $\neq$ 2 AND `freehms` = 1
L2_3	`prtvteit` = 2 AND `freehms` $\neq$ 1
L2_4	`prtvteit` = 2 AND `freehms` = 1
L3_1	`prtvteit` $\neq$ 2 AND `prtvteit` $\neq$ 3 AND `freehms` $\neq$ 1
L3_2	`prtvteit` = 3 AND `freehms` $\neq$ 1
L3_3	`prtvteit` $\neq$ 2 AND `prtvteit` $\neq$ 3 AND `freehms` = 1
L3_4	`prtvteit` = 3 AND `freehms` = 1
L3_5	`prtvteit` = 2 AND `freehms` $\neq$ 1 AND `stfgov` $\leq$ 4
L3_6	`prtvteit` = 2 AND `freehms` $\neq$ 1 AND `stfgov` $\geq$ 5
L3_7	`prtvteit` = 2 AND `freehms` = 1 AND `stfgov` $\leq$ 4
L3_8	`prtvteit` = 2 AND `freehms` = 1 AND `stfgov` $\geq$ 5

Table 3.6: Logical definitions of L2–L3 personas. Full table (L2–L6) available in Appendix A.2.

The logical conditions listed in Table 3.6 correspond to the categorical codings adopted for the predictor variables, as documented in Table 3.1 (Chapter 3). For example, the condition `prtvteit = 2` refers to respondents who reported voting for the *Partito Democratico (PD)* in the last national election. Similarly, conditions on `freehms` distinguish between different positions regarding gay and lesbian rights: `freehms = 1` identifies those who *agree strongly* that gay and lesbian people should be free to live as they wish, whereas `freehms` $\neq$ `1` aggregates all other categories.

At level 3, the variable `stfgov` (satisfaction with the government) enters the persona definitions, with thresholds such as `stfgov` ≤ 4 indicating respondents with low satisfaction and `stfgov` ≥ 5 indicating more positive evaluations.

The levels described in Table 3.6, as opposed to those reported in Table 3.4, carry a different meaning. In this case, the term *level* refers to the structure of the decision tree: all profiles characterized by $L2_n$ represent the first split that appears in the tree (the one closest to the root), as illustrated in Figure 3.1. Early splits are particularly relevant because they capture the most influential attributes in partitioning the sample and therefore define the broadest and most informative distinctions among respondents. The levels presented in Table 3.6 were constructed from those listed in Table 3.4 by grouping profiles according to both the number of features involved and the order of appearance of these features along the paths originating from the root.

In short, these filter rules operationalize the decision tree paths into survey-based conditions that yield coherent and sufficiently supported subgroups. While the table above reports only L2–L3 personas, the full set of definitions for all L2–L6 profiles is provided in Appendix A.2.

Chapter 4

Projecting Models into Human Space: PCA-Based Assessment of Persona Alignment

4.1 Introduction and objectives

The aim of this study is to evaluate the extent to which large language models (LLMs) can reproduce human responses to three politico-value items and, crucially, whether they also recover the underlying latent structure that organizes those items. Principal Component Analysis (PCA) is central to this goal: rather than serving as a mere visualization aid, it provides a principled, human-referenced embedding that (i) compresses the three correlated indicators into a single interpretable latent axis, (ii) removes scale and collinearity artifacts by operating on standardized variables, and (iii) enables like-for-like comparisons by projecting model-generated profiles into the same component space estimated on real respondents. We estimate the PCA on the human data only and orient the first component so that higher scores correspond to more right-leaning positions; each LLM profile is then mapped onto this fixed axis. This design lets us disentangle simple "level" differences (systematic shifts) from "structural" fidelity (whether the model preserves ordering and dispersion along the latent dimension). The analysis proceeds on two levels:

- **Structural (latent)** – through Principal Component Analysis (PCA) applied to profile-level means of the three items:
  - `lrscale` (0 = left, 10 = right),

  - `hmsacld` (adoption rights for same-sex couples; 1–5),
  - `imdfetn` (attitudes toward immigration from different ethnic groups; 1–4).

- **Observable (score 0–100)** – through the comparison of statistics of the synthetic score (0–100, derived from the three items) across the 28 profiles, both for the real sample and for each LLM.

The analysis focuses on three aspects: (i) positional bias (mean differences), (ii) compression/expansion of differences across profiles (differences in standard deviation), and (iii) alteration of distributional shape (differences in skewness). These aspects are examined using two metrics: (i) a latent coordinate-θ, defined as each profile's score on the first principal component (PC1), obtained by projecting the standardized item vector onto PC1 estimated on human data and oriented so that larger values indicate more right-leaning positions—and (ii) the observed 0–100 composite score.

4.2 Methodology

We begin by standardizing the three variables via z-scores, using the means and standard deviations computed solely from the real profiles. This step establishes an objective reference space that is common to all subsequent analyses[6, 7]. Principal Component Analysis (PCA) is then estimated exclusively on the real profiles. By sign convention, the first component (PC1) is oriented to load positively on `lrscale`, ensuring that higher values of lrscale correspond to higher values on PC1.

Latent scores θ are defined as PC1 and are computed for each real profile. Model-generated profiles are subsequently projected into the same component space by applying the identical standardization and the loadings derived from the real data. The interpretation follows from the loading pattern: when all PC1 loadings are positive and of similar magnitude, PC1 can be read as a one-dimensional politico–value axis. Along this shared axis, higher PC1 values indicate relatively more "conservative" positions.

For the 0–100 scores, descriptive statistics—mean, standard deviation, and skewness—are calculated across all 28 profiles. In the case of the real data, we use the profile-level mean score; for the model outputs, we use, for each profile and each model, the mean over 30 runs.

To complement these summaries, we assess the profile-by-profile correlation between the real mean scores and the model mean scores, which captures the extent to which models preserve the ranking of profiles. Absolute accuracy is quantified by the mean absolute error (MAE), while systematic deviation is summarized by the bias, defined as the average difference between model predictions and real data.

It is important to note that the within-profile standard deviations observed in the real data (across individuals) are not directly comparable to the run-to-run standard deviations of the models (across repeated executions). The former reflect the breadth of human variability, whereas the latter speak to the stability of model predictions. These are therefore complementary indicators that together offer a more complete evaluation.

4.3 Results

Component	Explained Variance Ratio
PC1	0.9138
PC2	0.0728
PC3	0.0134

Table 4.1: Explained variance ratio

Feature	Explained Variance Ratio
lrscale	0.561
hmsacld	0.570
imdfetn	0.600

Table 4.2: Loadings on PC1

PCA on real profiles indicates strong unidimensionality. PC1 explains over 91 percent of the variance, with all three items loading positively and similarly, confirming a common axis. Interpretation: increases in lrscale, hmsacld, and imdfetn all move in the same direction along PC1.

4.3.1 Comparison of latent scores (θ = PC1, 28 profiles)

Model	mean(θ)	std(θ)	skew(θ)	Δstd vs real
ESS11 Survey	0.000	1.656	−0.209	—
Deepseek-chat	−1.036	1.595	−0.244	−3.65%
Gpt-4.1	−0.869	1.517	+0.024	−8.40%
Gpt-4.1-mini	−1.344	1.331	+0.581	−19.63%
Gpt-4o-mini	−1.070	1.658	+0.227	+0.14%

Table 4.3: Comparison of the latent score θ = PC1 (28 profili)

Interpreting θ as the first principal-component coordinate – the latent axis along which profiles vary the most – the table shows three clear ways in which the synthetic model distributions deviate from the empirical one: a systematic shift in location, differences in dispersion, and a change in the shape of the tails.

Mean. First, the location (mean) is uniformly negative for all four models, whereas the real distribution is centered at 0 by construction. In practical terms, model-generated profiles cluster to the left along the latent spectrum. The magnitude of this shift differs by model: gpt-4.1 is the least shifted (mean $\theta \approx$ -0.87), deepseek-chat and gpt-4o-mini are somewhat further left ($\approx$ -1.04 and -1.07), and gpt-4.1-mini shows the

largest displacement ($\approx$ -1.34). So, although all models "under-center" relative to the empirical data, gpt-4.1 remains closest to the real centroid, while gpt-4.1-mini is the most off-center.

Standard Deviation Second, dispersion varies meaningfully. The real standard deviation is 1.656. Three models are noticeably more concentrated: deepseek-chat (1.595, about -3.7%), gpt-4.1 (1.517, about -8.4%), and especially gpt-4.1-mini (1.331, roughly -19.6%). This compression means the synthetic profiles for these models occupy a narrower band of the latent axis, dampening both moderate and extreme positions. By contrast, gpt-4o-mini essentially matches the empirical spread (1.658, +0.14%), suggesting it reproduces the overall variability of θ much more faithfully even if its center is still shifted left. From an analytical standpoint, compressed variance can make between-profile differences look smaller than they should be and can underrepresent tail behaviors.

Skew. Third, the shape of the distributions summarized by skewness-diverges. The empirical θ has a mild left tail (skew $\approx$ -0.21). Deepseek-chat mirrors this with a slightly more negative skew (-0.24), but the other three models flip the asymmetry to the right: gpt-4.1 is nearly symmetric but marginally positive (+0.02), gpt-4o-mini shows a moderate right tail (+0.23), and gpt-4.1-mini is distinctly right-skewed (+0.58). This pattern is informative: even though the mass of the synthetic distributions is shifted leftward overall, several models generate a comparatively heavier positive tail. Practically, that can look like a dense cluster of slightly-left or near-center profiles combined with a thinner but more extended set of right-leaning outliers. Such tail inversion often signals a mismatch in how models capture minority subpopulations: mid-right positions may be underrepresented while the few rightmost cases are pushed farther out.

Putting these pieces together, the models broadly recover the same latent axis as the empirical data but exhibit a consistent leftward location bias, often reduced dispersion (with gpt-4o-mini as the exception on spread), and, for three models, a reversal of tail asymmetry.

4.3.2 Results on the 0–100 score (across profiles)

Model	Mean	Std. Dev	Skew	ΔMean vs survey	ΔStd % vs real
ESS11 Survey	46.85	17.58	−0.245	—	—
Deepseek-chat	35.17	16.99	−0.179	−11.68	−3.33%
Gpt-4.1	36.91	16.43	+0.121	−9.94	−6.54%
Gpt-4.1-mini	32.21	14.68	+0.688	−14.64	−16.50%
Gpt-4o-mini	34.59	17.75	+0.355	−12.26	+0.97%

Table 4.4: Comparison of aggregated statistics across profiles

Considering the 0–100 score aggregated over the 28 profiles, the same three dimensions observed on the latent θ axis reappear in the observable metric: a systematic level shift, differences in dispersion, and changes in tail shape.

Level. All models underestimate the absolute level relative to the empirical mean of 46.85. Model means range from approximately 32 to 37: gpt-4.1 is the closest at 36.91 ($\Delta = -9.94$), followed by deepseek-chat at 35.17 ($\Delta = -11.68$) and gpt-4o-mini at 34.59 ($\Delta = -12.26$), while gpt-4.1-mini is the most under-centered at 32.21 ($\Delta = -14.64$). Expressed in standardized units using the empirical standard deviation 17.58, these biases correspond to roughly -0.57σ for gpt-4.1, -0.66σ for deepseek-chat, -0.70σ for gpt-4o-mini, and -0.83σ for gpt-4.1-mini. This pattern mirrors the leftward displacement on θ: because low values map to the left and high values to the right, a left shift in the latent space naturally translates into lower observed scores.

Dispersion. Relative to the empirical standard deviation of 17.58, most models display compression. Deepseek-chat has 16.99 ($\Delta = -0.59$; -3.33%), gpt-4.1 has 16.43 ($\Delta = -1.15$; -6.54%), and gpt-4.1-mini has 14.68 ($\Delta = -2.90$; -16.50%). In contrast, gpt-4o-mini closely matches the real spread at 17.75 ($\Delta = +0.17$; $+0.97\%$). Compression reduces between-profile contrasts, pulling moderate and extreme profiles toward the center. If preserving relative separations among profiles is important, gpt-4o-mini's near-alignment on dispersion is advantageous even though its mean remains too low.

Shape. The empirical distribution exhibits mild left skew (skew = −0.245). Deepseek-chat retains the same sign, though attenuated (−0.179). By contrast, gpt-4.1 becomes slightly right-skewed (+0.121), gpt-4o-mini more so (+0.355), and gpt-4.1-mini strongly right-skewed (+0.688). The combination of an overall downward level with a right tail implies that many profile scores cluster just below the empirical center while a smaller subset extends further into higher scores than expected. This tail inversion, already

visible on θ, suggests a structural mismatch in how models reproduce the distribution of profile means: mid-right profiles tend to be thinned out, while a few rightmost cases are pushed farther into the tail.

4.3.3 Adherence to profile ordering (correlation) and mean error

Model	MAE vs survey	Bias vs survey	Pearson r
Deepseek-chat	12.45	-11.68	0.855
Gpt-4.1	11.22	-9.94	0.856
Gpt-4.1-mini	15.19	-14.64	0.770
Gpt-4o-mini	14.55	-12.26	0.761

Table 4.5: Comparison of aggregated statistics across profiles

We compare, across the 28 profiles, the model predictions against the empirical profile means using three complementary summaries: the mean absolute error (MAE), the global bias defined as the average signed error, and the Pearson correlation coefficient r. The three metrics separate different aspects of agreement. MAE captures the average magnitude of the discrepancy in score units, the bias isolates systematic under or overestimation, and r measures how well the relative ordering and linear association are preserved irrespective of level and scale. By construction, correlation is invariant to affine transformations since $\mathrm{corr}(x,\, ay + b) = \mathrm{sign}(a)\,\mathrm{corr}(x, y)$. This explains why models can exhibit high r even when they are globally shifted and compressed.

Overall pattern. Despite systematic negative bias and, for most models, reduced dispersion relative to the empirical distribution, the ordering of profiles is largely retained: Pearson r ranges from 0.761 to 0.856. In practice, this means that an approximately monotone and close-to-linear mapping exists between real and synthetic profile means, so that a simple recalibration can align levels without altering rank.

Model-level highlights. *gpt-4.1* attains the lowest MAE at 11.22 with one of the highest correlations $r = 0.856$, while still exhibiting a negative bias of -9.94. Its errors are therefore dominated by a level shift, with comparatively smaller profile-specific deviations. *deepseek-chat* shows a very similar correlation $r = 0.855$ and a slightly larger MAE 12.45 together with a bias of -11.68, again indicating that miscalibration of the mean explains much of the discrepancy. *gpt-4o-mini* preserves ordering less tightly, with $r = 0.761$ and MAE 14.55 alongside a bias of -12.26. This combination suggests more pronounced profile-dependent deviations in addition to the level shift. *gpt-4.1-mini* has

the largest bias -14.64 and the largest MAE 15.19, with $r = 0.770$; here both systematic underestimation and idiosyncratic profile errors contribute.

4.4 Discussion

Three central findings emerge from the comparative analysis between empirical profiles and model-generated profiles, these results relate to:

- Positional bias
- Dispersion
- Distributional shape

First, a positional bias is observed consistently across representations: along the latent axis θ (PC1) and on the observable 0–100 score, all models are shifted leftward relative to the real data. In the absence of recalibration, absolute estimates therefore fail to reflect the empirical central tendency.

Second, dispersion across profiles is frequently compressed–in some cases by nearly twenty percent–which reduces cross-profile contrast and can attenuate the visibility of both moderate and extreme cases. A noteworthy exception is gpt-4o-mini, whose dispersion aligns with, or slightly exceeds, the empirical standard deviation.

Third, the distributional shape differs: while the empirical distribution exhibits negative skew (a left tail), several models produce positive skew (a right tail). This indicates not merely differences in location and scale but also alterations in tail structure. Despite these systematic deviations, the relative ordering of profiles is largely preserved, with Pearson correlations reaching approximately 0.86. This preservation of rank implies that a monotone mapping from model outputs to empirical values exists, such that simple affine recalibration can recover absolute levels and, when needed, rescale dispersion without disrupting the core ranking information captured by the models.

4.5 Limitations

The assessment of skewness across the 28 aggregated profiles is informative but sensitive to sampling variability; rigorous interpretation would benefit from uncertainty quantification, for example via bootstrap resampling of profiles. Moreover, the íntra-profileývariability in the empirical data and the run-to-run variability in model outputs measure fundamentally different constructs: the former reflects heterogeneity among individuals within the same profile, whereas the latter reflects stochasticity or instability across repeated generations of the same model prompt.

These quantities should therefore be treated as complementary rather than interchangeable. Finally, principal component analysis was fitted on real profiles and then used to project model outputs, which is methodologically appropriate to avoid leakage; nevertheless, PCA provides a linear projection and may not capture potential non-linear relations among the three items, leaving room for residual structure outside the first principal component.

4.6 Conclusions

The three items load positively and similarly on a nearly unidimensional latent factor, with PC1 explaining approximately 91% of the variance. This supports the use of a single latent axis θ for comparative analysis. Relative to the empirical distribution, large language models display a systematic leftward positional bias in both θ and the 0–100 score, frequent compression of dispersion across profiles, and alterations in skewness that invert the direction of the empirical tail in several cases. Nonetheless, the relative structure across profiles is well preserved—particularly for gpt-4.1—as evidenced by high correlations with empirical profile means. Human intra-profile variability is substantially greater than model run-to-run variability, confirming the tendency of models to generate less diverse predictions.

Taken together, these results depict a coherent pattern: the models capture the principal latent dimension and the rank ordering of profiles, yet introduce systematic deviations in location, scale, and shape relative to real data. These differences should be explicitly addressed through transparent calibration and careful presentation when synthetic outputs are used for inference or decision-making.

Chapter 5

LLM Employed: Architectures, Pricing and Knowledge-Cutoff

This chapter summarizes the large language models (LLMs) employed in the experiments: GPT-4.1, GPT-4.1-mini, GPT-4o-mini (OpenAI), and DeepSeek-Chat (DeepSeek). For each model, we report provider, context window and output limits, public knowledge-cutoff date, and publicly advertised pricing. We deliberately avoid experimental setup, prompting, and orchestration—these are treated in a dedicated chapter on reproducibility. Model specifications and cutoff dates are taken from providers' official pages where available; otherwise, we clearly mark information as "reported" from secondary sources.

5.1 Comparison of the selected models

Model	Provider	Context window	Knowledge cutoff	Parameters
GPT-4.1	OpenAI	1M tokens	June 2024	Not disclosed
GPT-4.1-mini	OpenAI	1M tokens	June 2024	Not disclosed
GPT-4o-mini	OpenAI	128K tokens	Oct 2023	Not disclosed
DeepSeek-Chat	DeepSeek	128K tokens	*reported* Jul 2024	671B MoE (37B active)

Table 5.1: Models used in the project: core specifications and knowledge cutoffs.

GPT-4.1 and its mini variant share a June 2024 knowledge cutoff and a 1M-token context window per OpenAI's announcement. OpenAI's GPT-4o-mini has a 128K context window, supports up to 16K output tokens, and has a knowledge cutoff of October 2023, per OpenAI's model note.

For DeepSeek-Chat, the vendor's API docs provide pricing and a 128K context length; parameter counts are documented in the DeepSeek-V3 technical report (671B MoE, 37B active). DeepSeek's knowledge cutoff is not officially stated in pricing docs; multiple credible reports indicate July 2024 for V3. We report this as "reported (non-official)".

The four models included in this study occupy distinct points in the design space of contemporary LLMs, offering complementary trade-offs in capability, scale, and cost. GPT-4.1 represents the flagship, general-purpose model with the most recent public knowledge cutoff among the OpenAI family considered and a very large context window. GPT-4.1-mini retains the same knowledge freshness and context length at a substantially lower unit price, positioning it as a cost-efficient surrogate when peak performance is not strictly required. GPT-4o-mini targets efficiency even more aggressively and pairs a smaller context window with an earlier knowledge cutoff; within the OpenAI portfolio, it serves as a lightweight baseline that preserves broad competencies while reducing operational cost[13, 16].

DeepSeek-Chat contributes a cross-vendor comparison point: it advertises a long context window and highly competitive pricing while documenting a mixture-of-experts (MoE) backbone. Taken together, this set spans a spectrum from high-capacity, generalist models (GPT-4.1) to lean, efficiency-oriented systems (GPT-4o-mini), and from closed-parameter, provider-curated architectures (OpenAI) to a partially documented MoE approach (DeepSeek). This breadth is deliberate: it enables the analysis to separate effects that plausibly stem from architectural and training choices from those that are more likely to reflect knowledge recency or context-management differences.

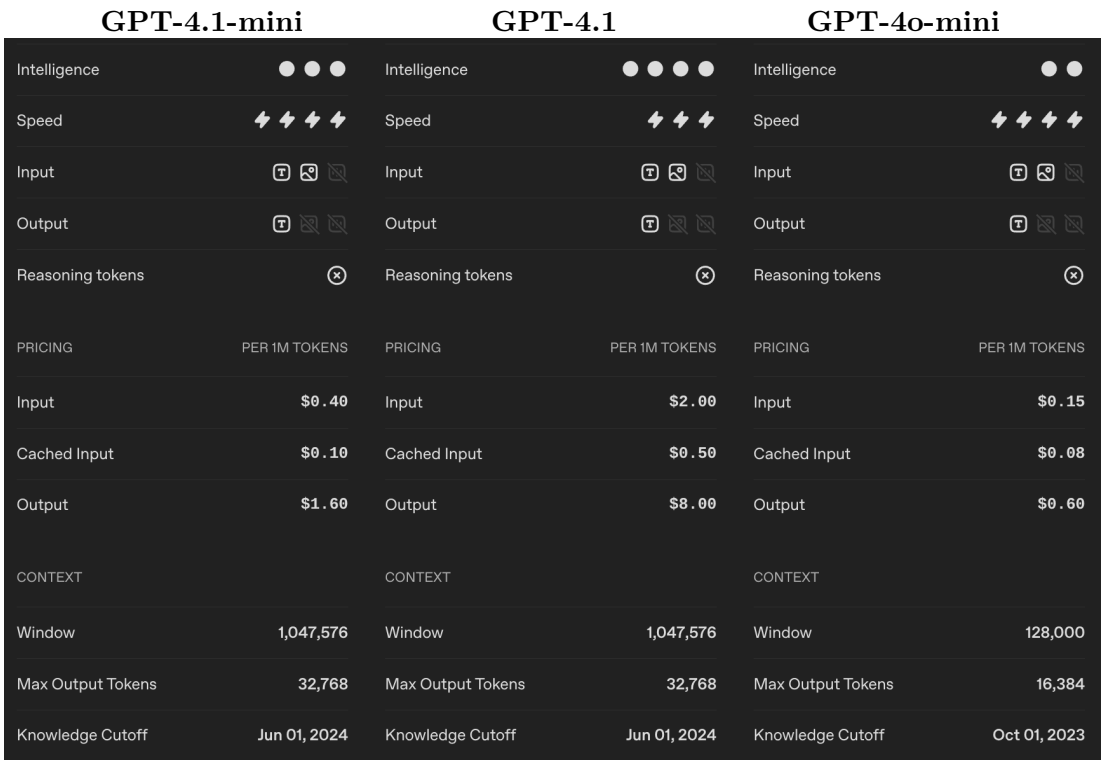


Figure 5.1: Official OpenAI model info panels.

5.1.1 Architectural considerations (High-Level) and transparency

While OpenAI does not disclose parameter counts or complete training details for the 4.x family, the models are widely understood to be dense Transformer architectures with substantial instruction-tuning and preference optimization stages. DeepSeek-Chat, by contrast, explicitly situates itself within the MoE paradigm: a very large pool of experts is trained, with only a subset activated at inference time. This design alters the capacity–efficiency profile by allowing the model to deploy specialized sub-networks adaptively, thereby increasing effective capacity per token without linearly scaling inference cost.

For the purposes of this thesis, these architectural differences matter in two ways. First, they provide a principled rationale for comparing both dense and MoE-style models under identical persona prompts, thereby testing whether expert routing offers any systematic advantage in matching human profile structure. Second, they clarify why parameter counts reported across vendors are not directly comparable: publicly advertised figures for MoE models reflect total parameters across experts, whereas only a fraction is active per token. The chapter therefore reports parameter counts only where vendors make them explicit and refrains from inferring unreported values for dense models.

5.1.2 Pricing

Model	Provider	Input (USD / 1M)	Output (USD / 1M)
GPT-4.1	OpenAI	$2.00	$8.00
GPT-4.1-mini	OpenAI	$0.40	$1.60
GPT-4o-mini	OpenAI	$0.15	$0.60
DeepSeek-Chat	DeepSeek	$0.56	$1.68

Table 5.2: (text) per 1M tokens for models used.

Clarifications. OpenAI's public API Pricing page explicitly lists GPT-4o-mini rates; GPT-4.1/4.1-mini generation rates are sometimes shown in partner (Azure) tables and reputable roundups when not present in the top-level OpenAI page section (which currently emphasizes fine-tuning for 4.1 family). Where OpenAI's own page omits generation prices for 4.1, we rely on consistent secondary reporting and Azure price tables for reference. Always consult the provider's dashboard for billing in your region/plan.

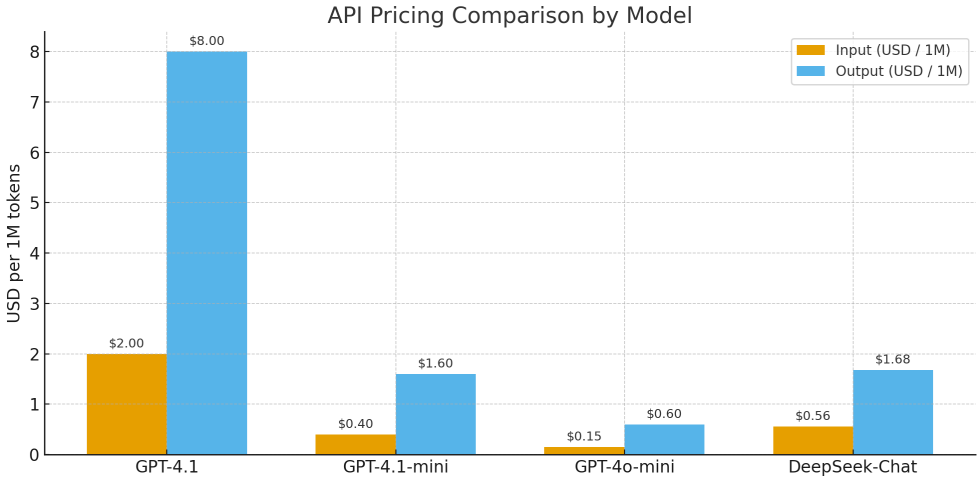


Figure 5.2: API Pricing Comparison by Model

5.1.3 Provider documentation and limits to comparability

A final consideration concerns the uneven transparency of public model documentation. OpenAI deliberately withholds parameter counts and full training data provenance for the GPT-4.x family, whereas DeepSeek provides partial architectural detail (e.g., MoE size and typical active parameters) but not a full training corpus description. These asymmetries constrain what can be stated about capacity in absolute terms and caution against naïve cross-vendor comparisons based solely on headline figures. In this chapter, we therefore (i) report only officially stated or widely corroborated specifications; (ii) avoid imputing missing values (e.g., parameter counts for GPT-4.1 variants); and (iii)

clearly flag items that are "reported" rather than formally documented by the provider. This approach maintains methodological neutrality while ensuring readers can interpret performance differences without over-attributing them to unverified internal properties of the models.

5.2 Knowledge freshness in relation to the ESS R11 timeline

Model	Knowledge cutoff	ESS fieldwork (2023–2024)	ESS release (June 2024)
GPT-4.1 / GPT-4.1-mini	June 2024	✓ overlaps	✓ overlaps
GPT-4o-mini	Oct 2023	✗ before release	✗ before release
DeepSeek-Chat[*]	*~July 2024 (reported)*	✓ overlaps	✓ overlaps

Table 5.3: Alignment between model knowledge cutoffs and the ESS Round 11 timeline.

[*] Cutoff month publicly reported in secondary sources; provider documentation may not state it explicitly.

The empirical data used in this study derive from ESS Round 11, with fieldwork primarily in 2023–2024 and public releases beginning in mid-2024. Knowledge cutoffs are therefore not a cosmetic attribute: they delimit the documentary horizon from which models can have learned socio-political context. GPT-4.1 and GPT-4.1-mini, with a June 2024 cutoff, align closely with the ESS release window; DeepSeek-Chat is commonly reported to have a mid-2024 cutoff as well. By contrast, GPT-4o-mini carries an earlier cutoff (October 2023), predating the main public dissemination of R11.

Although the prompts in this study do not reveal the survey data, background knowledge—such as party salience, coalition dynamics, and contemporary issue framing—can influence how models resolve ambiguous persona cues or interpolate between attribute combinations. It is therefore analytically meaningful that the model set spans both sides of the 2024 boundary: any systematic differences we observe in positional bias, dispersion compression, or tail behavior can be interpreted with awareness of potential knowledge-recency effects as well as architectural ones.

The project compares model families across knowledge freshness, context capacity, and cost tiers. GPT-4.1 (and mini) provide the freshest OpenAI cutoff (June 2024) and longest context (1M), beneficial for persona-rich prompts and large evaluation bundles. GPT-4o-mini contributes a very low unit cost with a 128K context and an earlier cutoff (Oct 2023), serving as an efficient baseline among OpenAI models. DeepSeek-Chat offers aggressive pricing and a long context (128K) while the vendor publicly documents

a 671B MoE backbone (V3), making it a useful cross-vendor comparison point; its knowledge cutoff is reported as mid-2024 but not consistently stated in the official pricing page.

Chapter 6

Experiment setup and Reproducibility

This chapter documents the experimental setup in full detail, with the explicit aim of ensuring reproducibility. Whereas previous chapters have described the dataset (Chapter 2), the construction of personas (Chapter 3), and the models employed (Chapter 5), here we focus on the technical execution: environment, parameter choices, prompting protocols, and reproducibility safeguards.

All experiments were implemented in `Python 3.11`, making extensive use of the official `openai` and `pandas` libraries for API calls and data handling. The experimental codebase is version-controlled, ensuring that the exact scripts and parameter settings used in this thesis can be retrieved. Model calls were made via provider APIs (OpenAI and DeepSeek), with authentication keys stored in a local `.env` file.

Execution took place on a Linux workstation. The experiments did not require GPU acceleration since the inference workload was delegated to external APIs.

6.1 Randomness and seeds

Generative models introduce stochasticity in sampling. To ensure replicability, every run was anchored to fixed random seeds. Specifically, we constructed lists of seed values and iterated over them for each persona–model combination.

The seed list is composed of 30 distinct integers, thereby producing exactly 30 replications per persona–model pair. This mirrors the `min_samples_leaf = 30` condition used in the decision tree on real data (cf. Chapter 3), maintaining symmetry between synthetic and survey-based sampling variability.

6.2 System messages and user prompt

Two variants of the system message were employed:

- **Base system message**

System Message

You are an Italian human and your task is to respond to each of the following questions by adopting the following persona profile.

OBJECTIVE

Impersonate the person described in the following PROFILE. Using **only** the information in the PROFILE answer the question that will be asked at the end.

OUTPUT

Return ONLY a valid JSON object with exactly these three keys:
"political_self_placement": <number>,

"adoption_rights_same_sex": <number>,

"immigration_other_ethnic_group": <number>

Do not add commentary, quotes, extra keys, or text outside the JSON. Numbers must follow the exact scales provided in the user message.

- **Justification system message**

Justification System Message

You are an Italian human and your task is to respond to each of the following questions by adopting the following persona profile.

OBJECTIVE

Impersonate the person described in the following PROFILE. Using **only** the information in the PROFILE answer the question that will be asked at the end.

OUTPUT

Return ONLY a valid JSON object with exactly these three keys:

"political_self_placement": <number>,

"adoption_rights_same_sex": <number>,

"immigration_other_ethnic_group": <number>

Do not add commentary, quotes, extra keys, or text outside the JSON. Numbers must follow the exact scales provided in the user message. ## AFTER THE JSON

After the JSON, write a new line with exactly:

JUSTIFICATION:

Then provide a short text explaining the choice using only PROFILE info. Aim for ≈35 words. If the PROFILE lacks information, write "insufficient information". Do not show step-by-step reasoning.

- **User message (Prompt)**

User Message

PROFILE

You are Italian and you voted for `{parameter}` in the most recent national election.

SURVEY

1. political_self_placement – On a scale from 0 (Left) to 10 (Right), where would you place yourself politically?

2. adoption_rights_same_sex – Gay male and lesbian couples should have the same rights to adopt children as straight couples. Scale: 1 = Agree strongly, 2 = Agree, 3 = Neither agree nor disagree, 4 = Disagree, 5 = Disagree strongly.

3. immigration_other_ethnic_group – Allow many/few immigrants of a different ethnic group from the majority. Scale: 1 = Allow many, 2 = Allow some, 3 = Allow a few, 4 = Allow none.

The `user` messages were constructed dynamically from profile metadata. For simple cases, pre–written templates corresponding to single features (e.g., party voted, satisfaction with government, views on gay rights) were used (see Appendix prompt templates). More complex persona prompts were obtained by combining multiple feature templates,

ensuring alignment with the logical conditions derived from the decision tree (Chapter 3).

In the appendix, Table A.4 reports all examples of prompts for single-feature profiles across all features considered in the study.

6.2.1 Construction of more complex profile prompt

The prompts corresponding to the more complex profiles were constructed by following the feature combinations reported in Table A.2 and by combining the single-feature prompts listed in Table A.4. Below is an example of a prompt for a more complex profile. Consider profile *L5_5* (see Table A.2), which involves four features:

- `prtvteit` $\neq 2$ `AND prtvteit` $\neq 3$
- `freehms` $= 1$
- `stfgov` $<= 6.5$
- `yrbrn` $<= 1981$

By combining the prompts of the four features involved in profile *L5_5*, the following composite prompt is obtained:

Complex Profile Prompt Example

PROFILE

You are Italian and you did not vote for *Partito Democratico* in the most recent national election. You did not vote for *Movimento 5 Stelle*. You *Agree Strongly* with the statement that gay men and lesbians should be free to live their own life as they wish. You rate your satisfaction with the national government at *3* on a scale from 0 (extremely dissatisfied) to 10 (extremely satisfied). You were born in or before *1981*.

SURVEY

1. political_self_placement – On a scale from 0 (Left) to 10 (Right), where would you place yourself politically?

2. adoption_rights_same_sex – Gay male and lesbian couples should have the same rights to adopt children as straight couples. Scale: 1 = Agree strongly, 2 = Agree, 3 = Neither agree nor disagree, 4 = Disagree, 5 = Disagree strongly.

3. immigration_other_ethnic_group – Allow many/few immigrants of a different ethnic group from the majority. Scale: 1 = Allow many, 2 = Allow some, 3 = Allow a few, 4 = Allow none.

It is important to note that for the feature *stfgov*, which is defined in terms of an interval, we opted to use the median of the interval in order to avoid semantic ambiguity in the prompt. For example, in the constructed prompt above, the interval corresponds to values in the range $[0 \leq \texttt{stfgov} \leq 6.5]$, and the value 3 was selected as the representative input in the prompt.

6.3 Execution parameters

The four active models were `gpt-4.1`, `gpt-4.1-mini`, `gpt-4o-mini`, and `deepseek-chat`. Each was queried with identical prompts under the following parameters:

Parameter	**Value**
Temperature	0.8
Top-p	0.8
Max tokens	150 per completion
Response format	Plain text (parsed into JSON)
Seed	One seed for each run (see Appendix A.1)

Table 6.1: LLM generation parameters used in the experiments.

Every persona profile was simulated 30 times per model, using distinct seeds. The resulting outputs were parsed to extract both the JSON answers and, where applicable, the justification texts.

The real survey personas were aggregated by computing profile-level means across at least 30 respondents. To maintain comparability, each synthetic persona was generated with exactly 30 replications per model. This design ensures that run-to-run variability in the synthetic side is structurally aligned with within-profile variability on the real side.

Responses were stored in structured text files and subsequently parsed into tabular form. Parsing routines extracted the JSON object, the justification text (if present), and computed the derived composite score (0–100). Invalid or non-JSON outputs triggered error handling; such cases were rare and, when present, the run was repeated under the same seed until valid output was obtained.

All downstream statistics were computed on these parsed results. On the real-data side, complete-case filtering was applied as described in Chapter 2; no weighting was used, in order to preserve one-to-one comparability.

6.4 Decision tree parameter setup

We trained a `DecisionTreeRegressor` on the continuous outcome `score` (0–100). Predictors included ordinal/continuous variables {`stfgov`, `stfeco`, `gincdif`, `icgndra`, `yrbrn`, `fnsdfml`} and one-hot encoded categorical variables {`edu_cat`, `party_cat`, `free_cat`}. The full configuration was:

- `criterion = squared_error (MSE)`
- `max_depth = 6`
- `min_samples_leaf = 30` (to guarantee robust leaf-level estimates)
- `random_state = 0`

The resulting tree (Figure 3.1) uses, in order of appearance along the main branches, `prtvteit` (via `party_cat`), `freehms` (via `free_cat`), `stfgov`, `yrbrn`, and `icgndra`. These are the attributes on which personas were ultimately defined.

6.5 Reproducibility measures

To promote reproducibility, we ensured that:

- All code was maintained under version control.
- Seeds, temperature, top-p, and other parameters are explicitly documented here.
- Persona definitions are fully specified in Chapter 3 and Appendix A.2.
- Prompt templates (single-feature) are archived (see Appendix prompts).

Some elements, however, limit perfect reproducibility. Provider APIs may evolve (e.g., model weights or decoding defaults can be silently updated), so bitwise-identical results cannot be guaranteed across time. The results reported here are strictly tied to the model versions and API states available between mid and late 2024.

Model versioning, provider updates, and their impact on reproducibility

A key threat to exact reproducibility in LLM-based experiments is that commercial providers may update model weights, decoding defaults, tokenizers, or content filters without changing the public model name. Consequently, re-running the same code with the same parameters at a later date can yield measurably different outputs. This phenomenon—often referred to as *model drift under a stable label*—is orthogonal to our dataset, prompts, and code, and requires explicit documentation and mitigation.

What can change "under the hood". Even when the model identifier (e.g., `gpt-4.1`) remains constant, providers may:

1. refresh or fine-tune the underlying weights,
2. adjust default decoding parameters or safety filters,
3. update tokenization libraries or stopword handling,
4. modify system prompts or routing logic in multi-expert backends,
5. change calibration and post-processing steps (e.g., JSON repair heuristics).

Any of these changes can alter completion probabilities and, by extension, the discrete outputs our study records (the three survey items and the composite *score*).

Model snapshots used during data collection

Our configurations did not pin a specific model *snapshot* for any provider endpoint. Consequently, each API call resolved to the provider's *current* snapshot behind the stable model label at the time of execution ("latest available"). The mapping we relied on for each model family is documented below.

OpenAI models (stable labels with dated snapshots):

- **GPT-4.1** We did not pin a snapshot; the endpoint selected the latest. OpenAI currently exposes a single dated variant for this family: `gpt-4.1-2025-04-14`. All our runs that queried `gpt-4.1` therefore used the snapshot current at call time, which corresponds to that dated build.
- `GPT-4.1-mini` Likewise unpinned; OpenAI exposes one dated variant: `gpt-4.1-mini-2025-04-14`. Our runs targeting `gpt-4.1-mini` thus executed against this build.
- **GPT-4o** Multiple dated variants exist: `gpt-4o-2024-05-13`, `gpt-4o-2024-08-06`, and `gpt-4o-2024-11-20`. Because all our experiments took place after 2024-11-20, the endpoint consistently resolved to the most recent snapshot in this list.

DeepSeek models (alias mapped to evolving backends). DeepSeek's API presents `deepseek-chat` as a stable model name whose backend has been periodically upgraded. Provider documentation notes that earlier `deepseek-chat` and `deepseek-coder` variants were merged and upgraded to *DeepSeek V2.5*, while retaining backward compatibility under the same labels. More recently, the official model catalog lists `deepseek-chat` as serving the *DeepSeek-V3.1* line (non-thinking mode; 128K context),

and companion notes clarify that the base URL suffix "/v1" is unrelated to the model's versioning. In practical terms, our unpinned calls to `deepseek-chat` always hit the then-current snapshot; during our collection window, that corresponded to the vendor's V3/V3.1 generation of the chat model.

Limitation: "Same label, different model"

Even with perfect code and parameter pinning, commercial LLM endpoints may deliver outputs from updated model snapshots. This limitation is inherent to hosted AI services. Our logging and auditing measures make such changes *detectable* and *documentable*, but cannot fully prevent them. Where exact repeatability is mission-critical, we recommend including at least one frozen-weight baseline.

Chapter 7

Empirical Result of the Persona-Based Evaluation

7.1 Overview and research questions

This chapter reports the empirical results of the persona-based evaluation: how closely large language models (LLMs) reproduce the responses observed in real survey subgroups and which dimensions drive agreement or error. We analyse four models–GPT-4.1, GPT-4.1-mini, GPT-4o-mini, and DeepSeek-Chat–without re-introducing their specifications (see Chapter 5). Personas are the decision-tree–derived profiles (L0–L4) defined in the *Personas Construction* chapter; they represent coherent subgroups created from combinations of attributes such as party vote, attitudes toward gay rights, satisfaction with government, year of birth, and gender. The composite *score* (0–100) used throughout is obtained from three base items and is complemented by a latent index θ=PC1 from PCA estimated on the real profiles.

To preserve one-to-one comparability, each L0–L4 persona in the survey aggregates at least 30 respondents, and each synthetic persona is generated 30 times per model with matched decoding settings; unless otherwise noted, statistics are unweighted.

The chapter focuses on the following questions:

1. **Global fidelity at the persona level.** How close are model means and dispersions to survey values across the 28 multi-attribute personas? Do models preserve the ranking of personas (correlation) while exhibiting systematic level shifts (bias)?

2. **Effect of profile complexity (levels).** Does alignment improve or deteriorate

as the number of distinct features in a persona increases (L0 →L4)? We summarise survey and model means (±sd) by level.

3. **Correlation structure of base items.** Do synthetic data preserve the Pearson relationships among the three items that underlie the composite score, and how do deviations relate to score-level errors?

7.2 Global fidelity for aggregated persona level

This section evaluates model faithfulness at the *persona level* granularity before delving into feature-wise and distributional evidence. Table 7.1 reports the survey benchmark and the four models' mean score (0–100) with standard deviation, grouped by persona *level* (L0–L4). Two regularities emerge:

- **Systematic location shift.** All models exhibit a *leftward* bias (lower scores than the survey) that is fairly stable across levels.
- **Variance compression.** Model dispersions are markedly smaller than survey dispersions at every level.

Table 7.2 aggregates these two patterns by summarising (i) the average mean difference between model and survey across levels (model – survey) and (ii) the average ratio of model to survey standard deviations.

Models	L0	L1	L2	L3	L4
ESS 11 survey	47.70 ± 22.77	47.70 ± 22.77	52.59 ± 16.18	45.06 ± 14.60	34.99 ± 14.07
GPT-4.1	**42.74** ± 4.57	46.78 ± 2.22	**44.69** ± 1.87	**34.88** ± 2.05	**19.64** ± 1.84
GPT-4.1-mini	34.99 ± 6.06	44.11 ± 3.72	38.08 ± 0.72	30.82 ± 2.21	18.75 ± 1.45
GPT-4o-mini	27.79 ± 0.00	44.16 ± 5.76	42.03 ± 3.37	32.21 ± 2.14	19.36 ± 1.29
DeepSeek-Chat	36.10 ± 0.00	**48.69** ± 1.67	42.83 ± 0.40	32.92 ± 0.82	18.90 ± 4.51

Table 7.1: Survey vs. model means (±*std*) by persona level. To understand better how levels were defined see section 3.2

Unlike the levels described in Table 3.4, Table 7.1 describes the levels constructed by aggregating the profiles reported in Table A.2 according to the number of features involved simultaneously.

The levels reported in Table 7.1 are characterized by an increasing number of *distinct* features involved in the profile: the higher the level, the greater the number of features included. Profiles at Level 2 involve two distinct features (e.g., `prtvteit | freehms`), regardless of which ones. This implies that if the same feature appears multiple times along a decision tree path, it is not counted repeatedly; rather, it serves to further specify and refine the condition concerning that particular feature.

By *L0* we refer to the results obtained by providing the LLM with only the System Message, as presented in Chapter 6, without the addition of the *profile* section. This also accounts for the highly heterogeneous nature of the outcomes observed at Level 0. With no persona constraints, each model falls back on its own internal priors about "a typical Italian human" answering the three base items. Those priors are a product of (a) the pretraining corpus and (b) the alignment/reward model (RLHF or similar). Because providers differ on both, their "default beliefs" about political self-placement, same-sex adoption rights, and immigration are not identical. The composite score, being a deterministic function of those three items, therefore diverges across models.

For this reasons, *L0* is best interpreted as a *"null"* baseline that captures model priors, not as a representation of any specific subgroup in the data.

Location (bias). Relative to the survey benchmark, level-wise mean differences are consistently negative. For instance, at L0 the smallest gap is with **GPT-4.1** (-5.0 points), while GPT-4o-mini under-shoots more strongly (-19.9). Averaging across L0-L4, the mean shift ranges from roughly -7.86 points (GPT-4.1) to about -12.5 points (GPT-4o-mini), indicating a persistent leftward displacement irrespective of profile complexity.

Scale (dispersion). Standard deviations are compressed across the board. Taking the ratio of model to survey SD and averaging across levels yields values around 10%–15%: DeepSeek-Chat $\approx$ 10%, GPT-4o-mini $\approx$ 14%, GPT-4.1-mini $\approx$ 14.5%, GPT-4.1 $\approx$ 14%. This gap widens in relative terms as the survey SD remains large at higher levels while model SDs remain low.

Model	Avg. mean bias (points)	Avg. sd ratio (% of survey)
GPT-4.1	-7.86	13.7%
GPT-4.1-mini	-12.26	14.6%
GPT-4o-mini	-12.50	14.0%
DeepSeek-Chat	-9.72	9.5%

Table 7.2: Average location bias and dispersion ratio across levels (L0–L4). Negative bias indicates model scores lower than survey.

7.3 Profile-level result

This section examines model behaviour at the finest granularity used in the study—the individual persona profiles produced by the decision-tree design. We first compare model and survey means per profile (Table 7.3-7.4), then summarise error magnitudes

via MAE across all profiles (Table 7.5), and finally decompose MAE by the three source questions (Table 7.6).

7.3.1 Profile-wise deviations and notable outliers

Table 7.3-7.4 reveals two broad patterns consistent with Section 7.2: (i) a systematic leftward displacement (model means below survey means) across many profiles, and (ii) pronounced variance compression manifested here as clustering of synthetic profile means around a small set of anchor values. These anchors reflect narrow across-run dispersion and help explain the smaller standard deviations reported at level-aggregation.

Profile	Real	DeepSeek-Chat	GPT-4.1	GPT-4.1-mini	GPT-4o-mini
L2_1	61.48	58.90	47.90	44.40	52.49
L2_2	42.39	31.10	31.10	25.19	33.08
L2_3	37.36	37.80	39.89	37.80	37.91
L2_4	22.54	6.70	17.29	10.00	6.70
L3_1	65.41	57.80	58.35	44.40	52.99
L3_2	44.67	44.40	44.40	43.08	45.28
L3_3	47.41	31.10	31.10	25.19	32.53
L3_4	28.48	21.10	25.53	24.40	22.19
L3_5	31.47	37.80	40.99	33.47	37.01
L3_6	43.53	37.80	41.10	40.00	37.80
L3_7	16.64	6.70	8.91	10.00	6.25
L3_8	26.15	6.70	12.59	10.00	8.18
L4_1	73.19	57.36	58.90	44.77	54.11
L4_2	61.94	39.40	44.25	34.50	30.00

Table 7.3: Profile-level mean scores (0–100): real survey vs synthetic by model (part1).

Anchor values and quantisation. Across profiles, synthetic means tend to cluster on a few "anchor" values (e.g., 31.10, 44.40, 58.90), yielding visible plateaus in Tables 7.3–7.4. This quantisation stems from (i) the coarse item scales and the linear 0–100 mapping of the composite score, and (ii) reduced run-to-run variability. While rank ordering is largely preserved, anchoring compresses between-profile differences—especially in the tails—and contributes to the systematic underreach of extreme profiles.

Beyond the global tendency, several profiles exhibit large *directional* errors.

- **Large undershoots (models $\ll$ survey):** `L5_4` (Real = 68.74 vs 33–40), `L5_3` (59.60 vs 30–47), `L6_2` (61.67 vs 31–39), `L4_2` (61.94 vs 30–44), `L2_4` (22.54 vs 6.70–17.29), `L3_8` (26.15 vs 6.70–12.59). These cases combine relatively extreme survey means with synthetic anchoring at lower plateaus, producing sizeable gaps.
- **Overshoots (models $\gg$ survey):** GPT-4o-mini in `L4_3` (53.80 vs 48.63) and `L4_4` (53.39 vs 40.47), and the high-right `L5_2`, where GPT-4.1/GPT-4.1-mini reach 78.90 and GPT-4o-mini 84.18 against a survey mean of 76.93. These overshoots are less frequent but indicate that some interaction patterns are amplified rather than attenuated.

Overall, the per-profile panel shows that errors are not uniform: a subset of profiles concentrates most of the absolute deviation, whereas many mid-range profiles are approximated reasonably well (e.g., `L3_2`, `L3_6`, `L4_3` for several models).

Profile	Real	DeepSeek-Chat	GPT-4.1	GPT-4.1-mini	GPT-4o-mini
L4_3	48.63	46.64	44.40	43.85	53.80
L4_4	40.47	44.40	44.40	43.74	53.39
L4_5	53.29	31.10	30.99	26.89	30.88
L4_6	36.64	31.10	25.75	25.36	31.43
L4_7	13.21	6.70	9.40	10.00	6.59
L4_8	19.02	6.70	7.07	10.00	6.13
L5_1	69.28	57.80	58.68	44.40	52.99
L5_2	76.93	70.60	78.90	78.90	84.18
L5_3	59.60	39.40	46.87	33.90	30.66
L5_4	68.74	39.97	39.74	34.04	33.19
L5_5	50.33	31.10	30.99	27.89	32.42
L5_6	57.41	31.10	31.10	27.12	32.31
L6_1	53.97	39.40	43.51	33.39	32.46
L6_2	61.67	34.14	39.40	35.27	31.54

Table 7.4: Profile-level mean scores (0–100): real survey vs synthetic by model (part2).

7.3.2 Aggregate magnitude across profiles (MAE)

To condense profile-level discrepancies into a single indicator per model, Table 7.5 reports the mean absolute error (MAE) computed over all profiles (macro-average, unweighted). The ranking is clear:

- **GPT-4.1** attains the lowest MAE (11.22), indicating the smallest average deviation from survey means across profiles.
- **DeepSeek-Chat** follows (12.44), then **GPT-4o-mini** (14.55) and **GPT-4.1-mini** (15.19).

This ordering mirrors the location/scale evidence from Section 2: models that exhibit smaller leftward shifts and slightly less dispersion compression tend to achieve lower absolute errors at the profile level. It is also consistent with the outlier analysis above, where GPT-4.1 under-shoots less often and less severely in high-right profiles.

Model	Mean Absolute Error (0–100)
GPT-4.1	11.22
DeepSeek-Chat	12.44
GPT-4o-mini	14.55
GPT-4.1-mini	15.19

Table 7.5: Average absolute difference vs Real_score across all profiles (lower is better).

7.3.3 Item-level error by question

Table 7.6 decomposes MAE by the three survey items used to construct the 0–100 score. Because the items have different numeric scales (Q1: 0–10; Q2: 1–5; Q3: 1–4), within-question comparisons across models are meaningful, whereas cross-question MAEs are not directly comparable in magnitude.

Key observations:

- **Q1 (political self-placement)**: GPT-4.1 is best (0.678), closely followed by DeepSeek-Chat (0.704); GPT-4o-mini and GPT-4.1-mini are worse (0.803 and 1.081).
- **Q2 (adoption rights)**: GPT-4.1 again leads (0.938). Interestingly, GPT-4.1-mini (0.959) overtakes DeepSeek-Chat (0.966), while GPT-4o-mini is the least accurate (1.110). This suggests that, for this Likert-5 normative item, smaller models can align well once anchored, but remain more error-prone in the tails.
- **Q3 (immigration)**: GPT-4.1 posts the lowest MAE (0.248), followed by DeepSeek-Chat (0.322), GPT-4o-mini (0.338), and GPT-4.1-mini (0.415). This ordering echoes the overall MAE ranking.

Taken together, item-level results clarify *where* profile errors originate: models that are closer on Q1—the most finely graded item—tend to minimise profile-level MAE; conversely, larger errors on Q2 and Q3 often correspond to profiles whose defining attributes push responses toward more polar positions that compressed synthetic distributions fail to reach.

Model	MAE (Q1)	MAE (Q2)	MAE (Q3)
DeepSeek-Chat	0.704	0.966	0.322
GPT-4.1	0.678	0.938	0.248
GPT-4.1-mini	1.081	0.959	0.415
GPT-4o-mini	0.803	1.110	0.338

Table 7.6: Mean absolute error (MAE) by survey question for each model.

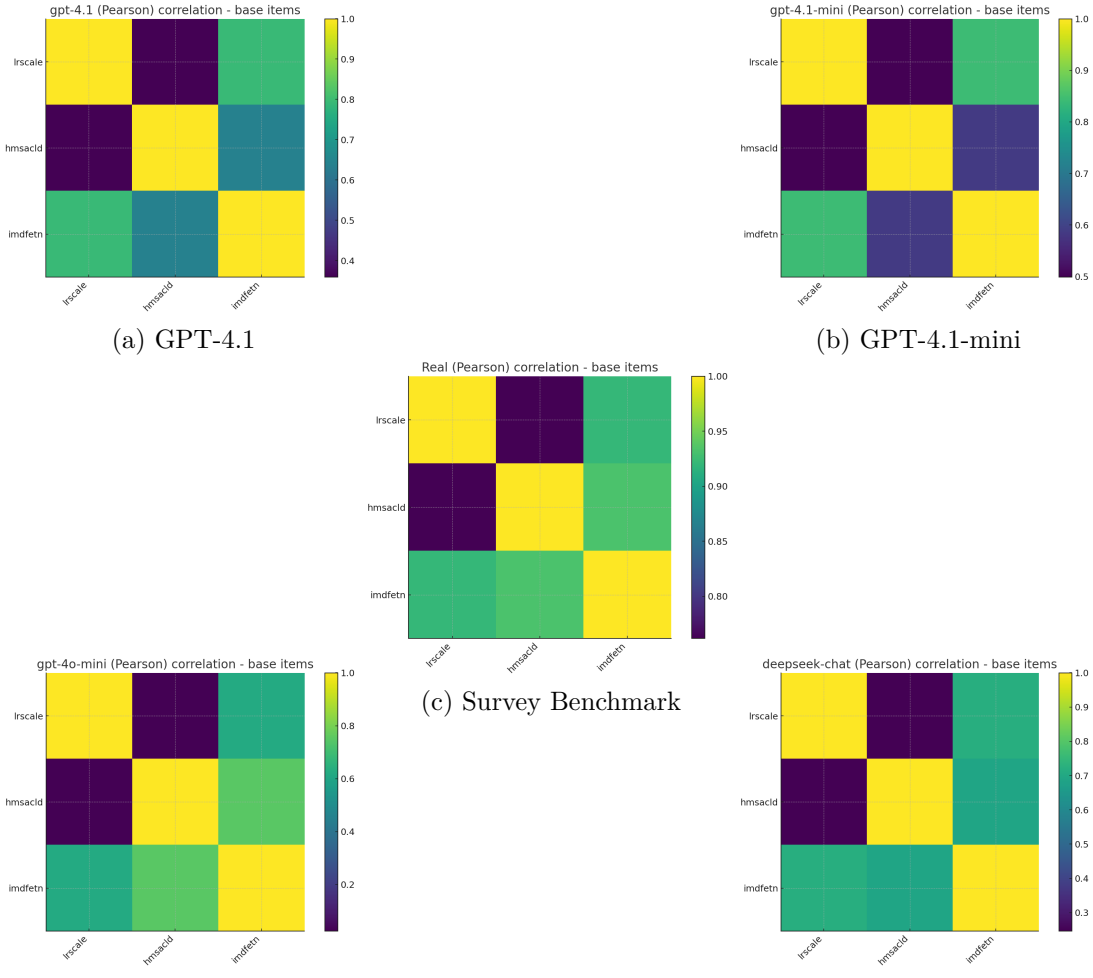
7.3.4 Discussion

Three takeaways emerge at the profile level. First, the majority of errors derive from a systematic location shift paired with discrete anchoring of synthetic means; this combination prevents models from matching the survey's more extreme profiles, especially on the rightmost tail. Second, occasional overshoots indicate that interactions among attributes (e.g., party vote with social-value items) are sometimes over-amplified rather than averaged out, producing asymmetric residuals. Third, across the board, **GPT-4.1** achieves

the most balanced compromise—smaller bias and less severe compression—resulting in the lowest MAE both overall and within items.

7.4 Item-level correlation structure

The goal of this section is to assess whether models preserve the correlation structure among the three base items — political self-placement (lrscale), same-sex adoption rights (hmsacld), and immigration from other ethnic groups (imdfetn). We compute pairwise Pearson correlations on the same evaluation grid used elsewhere in the chapter, and compare each model to the survey benchmark.



(a) GPT-4.1

(b) GPT-4.1-mini

(c) Survey Benchmark

(d) GPT-4o-mini

(e) DeepSeek-Chat

Figure 7.1: Item-level Pearson correlation matrices among *lrscale*, *hmsacld*, and *imdfetn*.

Benchmark survey: The survey matrix shows uniformly high positive correlations among all three items (Fig.7.1), consistent with a dominant one-dimensional politico-value axis: individuals who place themselves further to the right also tend, on average, to be more restrictive on immigration and less supportive of same-sex adoption. This aligns with the PCA evidence that a single latent component captures most of the variability across profiles.

Models: All models preserve the sign of the associations (no sign flips), but attenuate their magnitude relative to the survey (Fig. 7.1). A consistent pattern emerges across systems:

- The *strongest* preserved pair is typically lrscale–imdfetn.
- hmsacld–imdfetn sits in the *middle*.
- The *weakest* link is lrscale–hmsacld, often substantially reduced.

Model-specific behaviour:

GPT-4.1 best preserves the structure: lrscale–imdfetn remains strong; hmsacld–imdfetn is clearly positive but shrunk; lrscale–hmsacld is the weakest yet positive.

GPT-4.1-mini and *GPT-4o-mini* show broader attenuation across pairs, consistent with the variance compression documented at persona/feature level.

DeepSeek-Chat exhibits the largest distortion on lrscale–hmsacld (markedly low compared to the survey), while the other pairs remain in the same ballpark as the mini variants.

7.4.1 Discussion

The matrices indicate that models recover the direction of item co-movements but underestimate the coupling between the more "moral-social" item (hmsacld) and the other two. This is coherent with earlier findings: (i) location shifts and dispersion compression limit the ability to reach extreme category means; (ii) mini models particularly flatten between-profile differences; and (iii) occasional overshoots (e.g., DeepSeek-Chat at specific profiles) do not restore the full inter-item dependence. Practically, univariate fit (e.g., MAE on composite scores) can look acceptable while multivariate structure remains under-tied. A simple follow-up would be to quantify the gap with a Frobenius-norm difference between correlation matrices or to report Fisher-z adjusted errors by pair; a further "what-if" is to test whether an affine re-calibration of model outputs improves not only means but also correlation alignment.

7.5 Feature-wise accuracy

This section evaluates how closely models reproduce the survey when profiles are collapsed along a *single* attribute at a time (e.g., age, party, government satisfaction). For each feature, we compute per-class discrepancies between survey and synthetic means and then average those category-wise gaps to obtain a single *feature-level* indicator. We report (i) the absolute mean difference, (ii) the signed mean difference (i.e., directionality of under/over-shoot), and (iii) a relative mean difference to facilitate comparison across features with heterogeneous ranges. This complements Sections 2–3 by revealing where errors originate once multi-attribute interactions are "marginalised out."

7.5.1 Method

Let f index a feature (e.g., `stfgov`) with categories $c \in \mathcal{C}_f$. Denote the survey mean by $\mu_{f,c}^{\text{real}}$ and the synthetic mean by $\mu_{f,c}^{\text{syn}}$ (averaged across runs). For each feature we summarise:

$$\text{AbsMeanDiff}(f) \;=\; \frac{1}{|\mathcal{C}_f|}\sum_{c\in\mathcal{C}_f}\left|\mu_{f,c}^{\text{syn}}-\mu_{f,c}^{\text{real}}\right|, \qquad \text{MeanDiff}(f) \;=\; \frac{1}{|\mathcal{C}_f|}\sum_{c\in\mathcal{C}_f}\left(\mu_{f,c}^{\text{syn}}-\mu_{f,c}^{\text{real}}\right),$$

$$\text{RelMeanDiff}(f) \;=\; \frac{1}{|\mathcal{C}_f|}\sum_{c\in\mathcal{C}_f}\frac{\left|\mu_{f,c}^{\text{syn}}-\mu_{f,c}^{\text{real}}\right|}{\max(\epsilon,\,\left|\mu_{f,c}^{\text{real}}\right|)},$$

with a small ϵ to safeguard categories whose survey mean is near zero. Unless explicitly stated, these are *macro* averages over categories (each category counts once). This choice isolates model behaviour from category prevalence; weighted (micro) versions are reported in the Appendix for completeness.

7.5.2 Results by model

GPT-4o-mini. Figure 7.2 shows the largest absolute gaps for `fnsdfml` (financial difficulties during growth), `gincdif` (redistribution), `stfgov` (government satisfaction), `freehms` (LGBTQ rights), and `prtvteit` (party voted). Demographic features such as `icgndra` (gender) and `edlvfit` (education) are among the smallest in absolute terms. The signed/relative bars are mostly negative, consistent with the leftward displacement documented at persona level.

GPT-4.1-mini. Figure 7.3 displays a pattern similar to GPT-4o-mini, with large gaps again on `fnsdfml`, `stfgov`, `gincdif`, `prtvteit`, and `freehms`. Age-related features

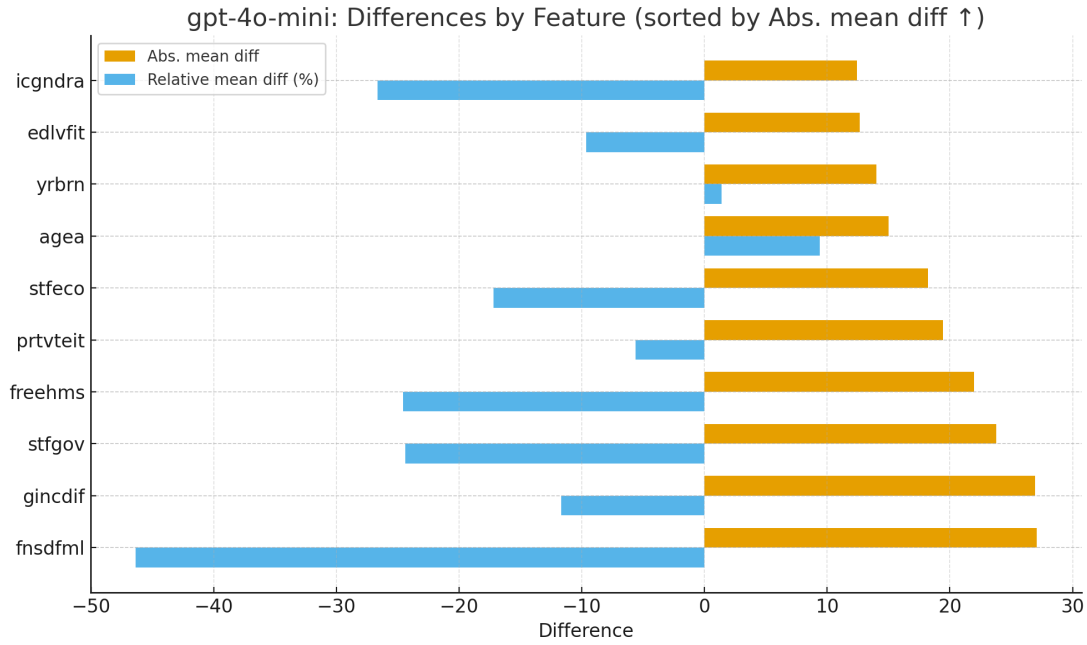


Figure 7.2: GPT-4o-mini: feature-wise discrepancies (absolute and relative mean differences; categories averaged).

(`agea`, `yrbrn`) sit mid-table. Relative differences are predominantly negative, indicating under-shoot of survey means across many categories.

GPT-4.1. Figure 7.4 confirms the same "difficulty ordering": the largest absolute gaps are concentrated in `fnsdfml`, `gincdif`, `freehms`, and `prtvteit`, while `icgndra` and `edlvfit` are relatively easier. Signed differences remain mostly negative but generally smaller in magnitude than for the two "mini" models, in line with GPT-4.1's lower MAE at the profile level.

DeepSeek-Chat. In Figure 7.5, absolute gaps peak on `freehms`, `yrbrn`, `gincdif`, `stfgov`, and `stfeco`. Unlike the GPT models, several relative bars are *positive*, suggesting overshoot (synthetic > survey) for some features—especially `yrbrn` and satisfaction variables—while `icgndra` and `fnsdfml` show smaller, sometimes negative, relative deviations. This mixed-sign pattern echoes the profile-level results where DeepSeek occasionally overshoots specific right-leaning profiles.

7.5.3 Cross-model synthesis

Three consistent findings emerge across the four panels:

1. **Hard features (largest gaps).** Economic hardship and redistribution (`fnsdfml`, `gincdif`), attitudinal/political features (`stfgov`, `freehms`, `prtvteit`), and macro

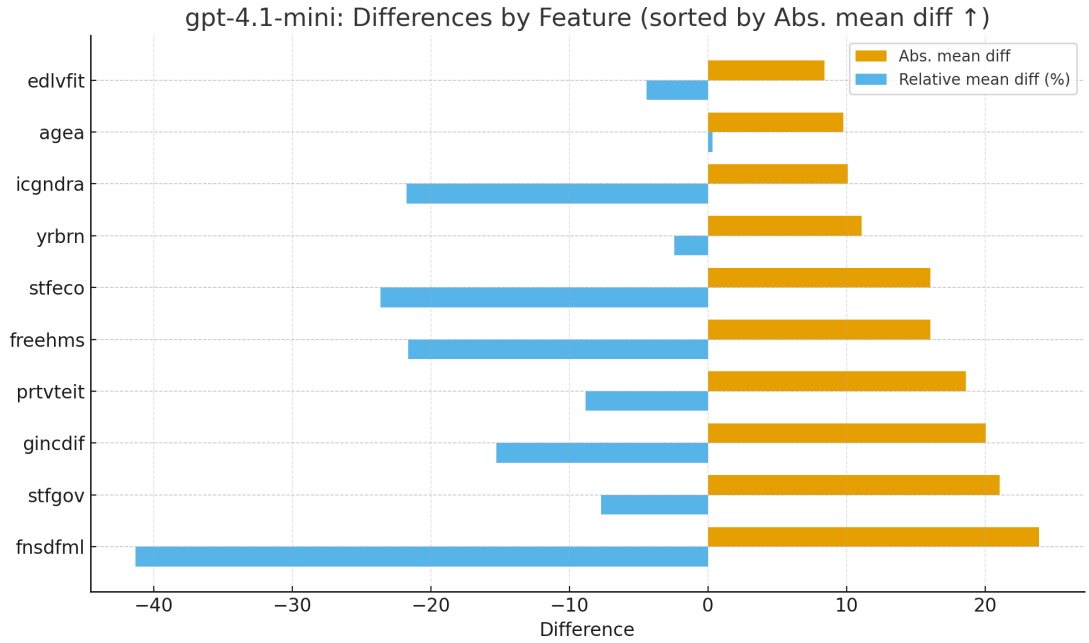


Figure 7.3: GPT-4.1-mini: feature-wise discrepancies (absolute and relative mean differences; categories averaged).

perceptions (`stfeco`) are the most challenging. These features also dominate the profile-level outliers, indicating that model errors concentrate where the survey exhibits more extreme or polarised means.

2. **Easier features (smaller gaps).** `icgndra` (gender) and `edlvfit` (education) consistently yield the smallest absolute differences. Age and cohort (`agea`, `yrbrn`) sit in the middle, with DeepSeek-Chat showing comparatively larger cohort-related discrepancies.

3. **Directionality.** GPT-family models show predominantly negative relative differences (under-shoot), consistent with the global leftward bias; DeepSeek-Chat exhibits a more mixed sign pattern with notable overshoots on some features (especially `yrbrn` and satisfaction variables).

7.5.4 Interpretation and caveats

Feature-wise gaps reflect the same mechanisms highlighted in Sections 7.2-7.3: a systematic location shift plus variance compression. Where the survey's category means are extreme (e.g., right tail in `prtvteit` or high satisfaction in `stfgov`/`stfeco`), compressed synthetic distributions cannot reach those levels, producing large absolute gaps. Occasional overshoots (notably in DeepSeek-Chat) suggest that some interactions are amplified rather than averaged out.

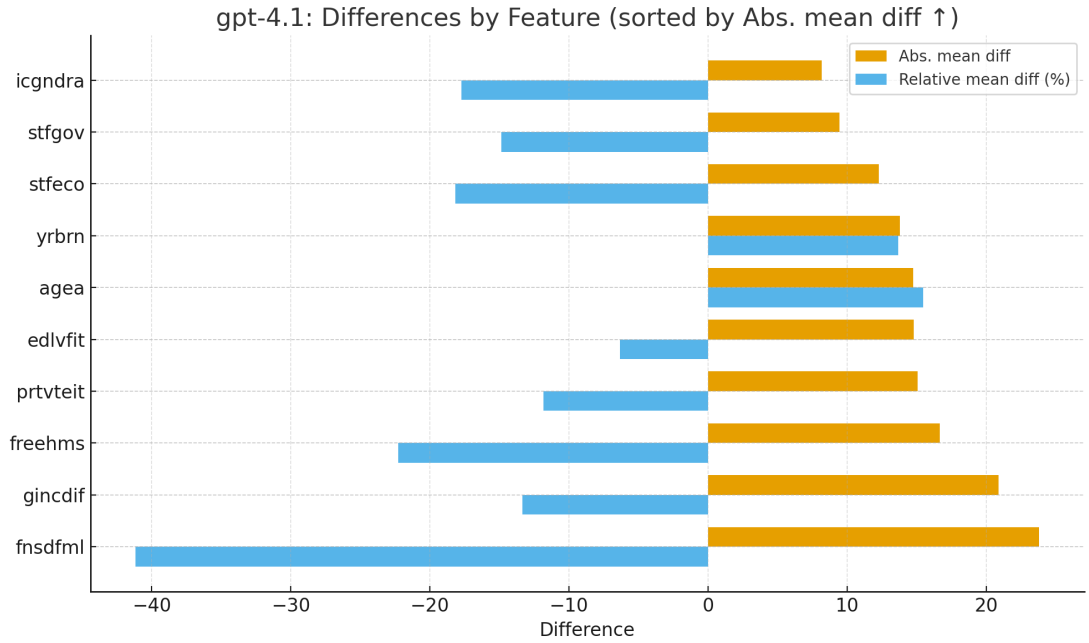


Figure 7.4: GPT-4.1: feature-wise discrepancies (absolute and relative mean differences; categories averaged).

Two cautions apply. First, relative mean differences are informative about direction and proportional magnitude but are *not* variance-standardised effect sizes; they should be read alongside absolute gaps and the survey spread for each feature. Second, category definitions (and any merges for sparse cells) affect macro-averages.

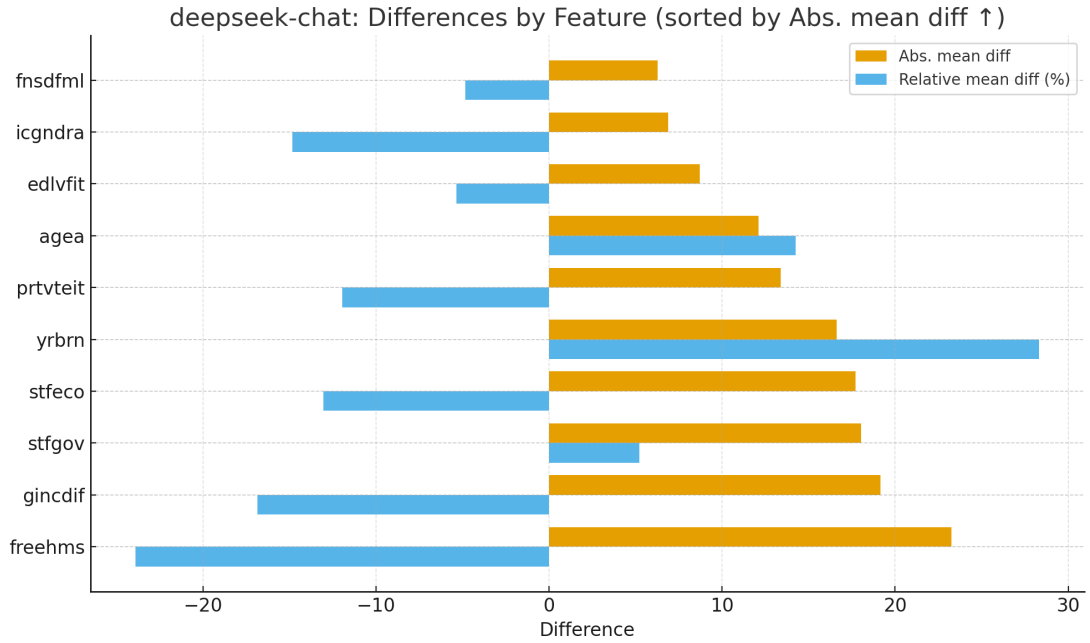


Figure 7.5: DeepSeek-Chat: feature-wise discrepancies (absolute and relative mean differences; categories averaged).

7.6 Distributional alignment within key attributes

We now assess whether the models reproduce the *shape*, spread, and tail behaviour observed in the survey when conditioning on single attributes. For each feature we (i) show the real benchmark boxplot and (ii) compare the four model distributions in a 2×2 panel. We focus on location (median/mean triangle), dispersion (IQR and whiskers), tail mass/outliers, and potential re-orderings across categories.

7.6.1 Government satisfaction (`stfgov`)

Survey benchmark. The survey displays a clear monotonic pattern: medians and means increase with satisfaction, and dispersion narrows at the top of the scale, with fewer extreme left outliers.

Comment. All models capture the upward slope but with compressed IQRs; extreme right categories (9–10) are often under-represented in the upper tail, consistent with the negative signed differences seen in Section 7.5. Occasional overshoot in middle categories appears in DeepSeek-Chat.

7.6.2 Party voted (`prtvteit`)

Survey benchmark. Medians rise from left to right when parties are ordered by the survey's central tendency; tails widen for some right parties due to within-party

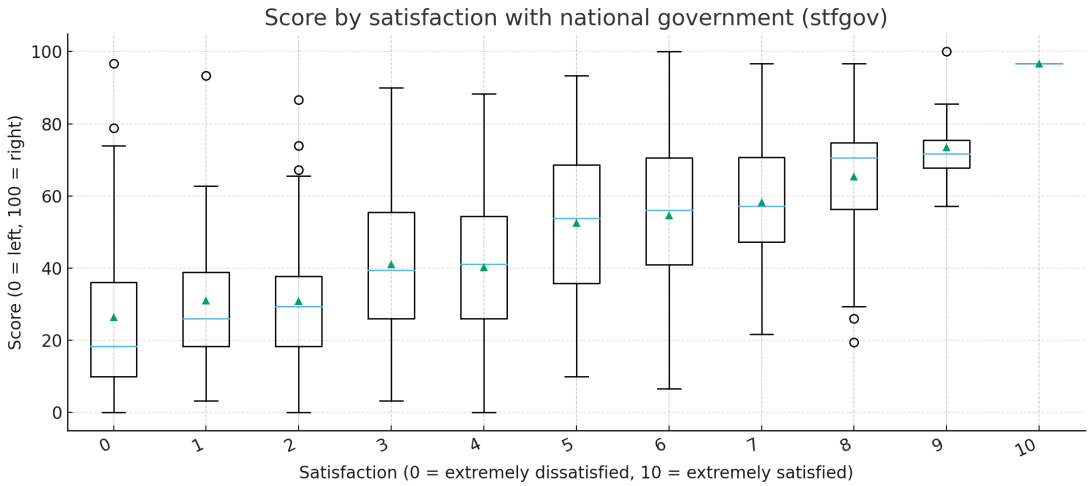
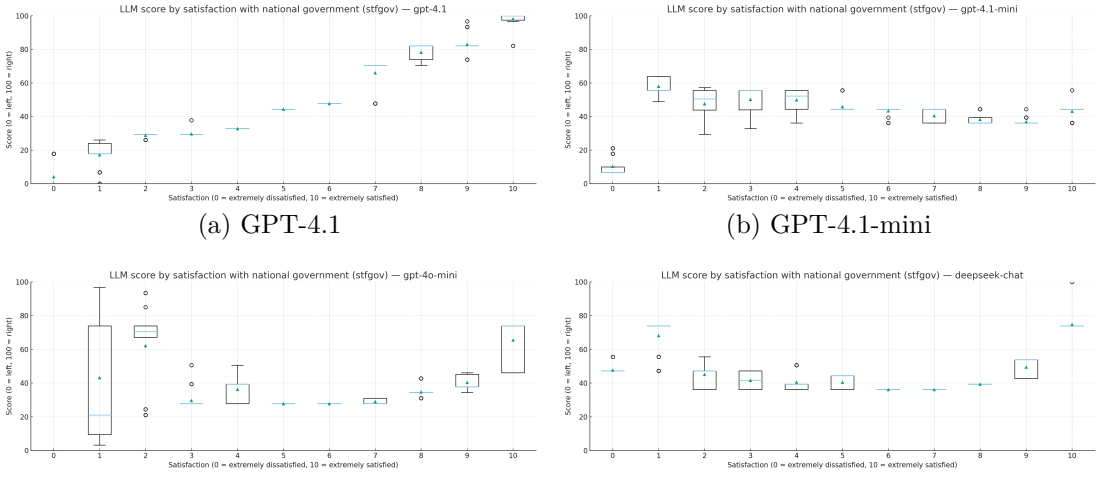


Figure 7.6: Survey boxplot by satisfaction with national government (`stfgov`).



(a) GPT-4.1 (b) GPT-4.1-mini

(c) GPT-4o-mini (d) DeepSeek-Chat

Figure 7.7: Model boxplots by `stfgov`.

heterogeneity.

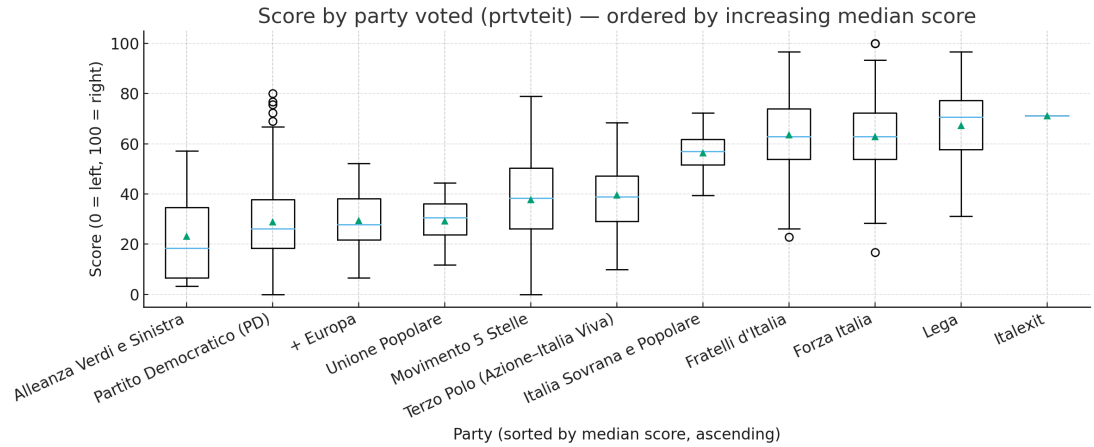
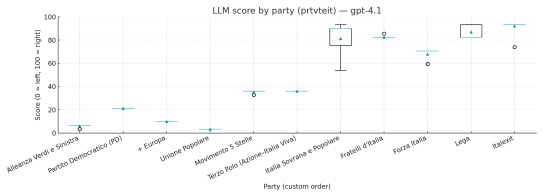
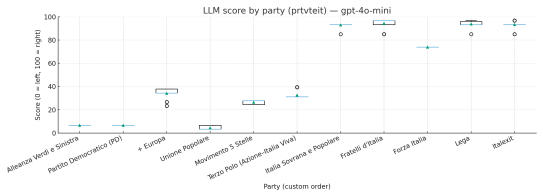


Figure 7.8: Survey boxplot by party voted (`prtvteit`), ordered by increasing median.

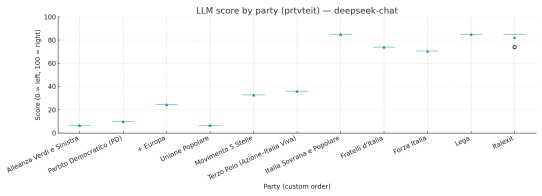


(a) GPT-4.1

(b) GPT-4.1-mini



(c) GPT-4o-mini



(d) DeepSeek-Chat

Figure 7.9: Model boxplots by `prtvteit`.

Comment. GPT-4.1 best preserves the left–right ordering of party medians; mini variants show stronger shrinkage, occasionally collapsing mid-spectrum parties. DeepSeek-Chat sometimes inflates upper medians for right parties (overshoot), consistent with profile-level findings.

7.6.3 Attitudes on gay/lesbian adoption rights (`freehms`)

Survey benchmark. Moving from *Agree strongly* to *Disagree strongly* shifts the score markedly rightward, with reduced lower-tail mass at the conservative end.

Comment. All models reflect the monotone trend but under-represent the rightmost category's upper tail; mini models show the strongest IQR compression. DeepSeek-Chat occasionally produces higher centres in mid categories (mild overshoot).

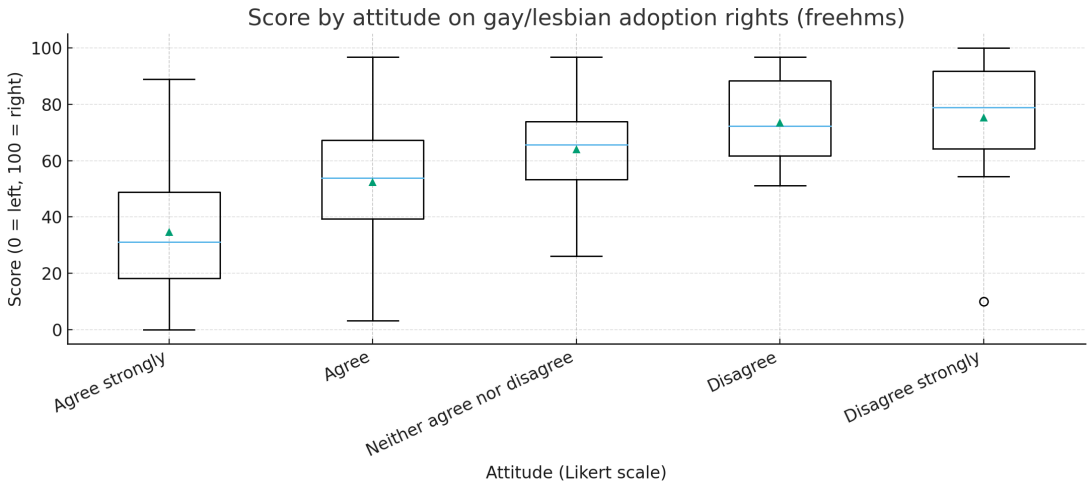
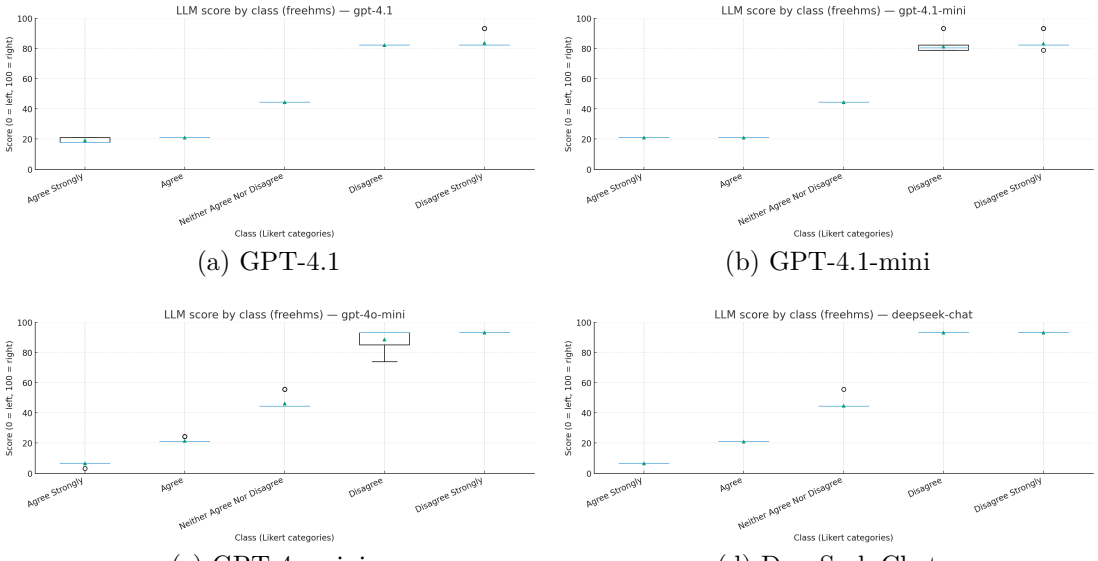


Figure 7.10: Survey boxplot by attitude on adoption rights (`freehms`).



(a) GPT-4.1

(b) GPT-4.1-mini

(c) GPT-4o-mini

(d) DeepSeek-Chat

Figure 7.11: Model boxplots by `freehms`.

7.6.4 Decade of birth (`yrbrn`)

Survey benchmark. Older cohorts tend to the right (higher medians), with a gradual left shift among the youngest groups; dispersion increases for middle cohorts due to heterogeneity.

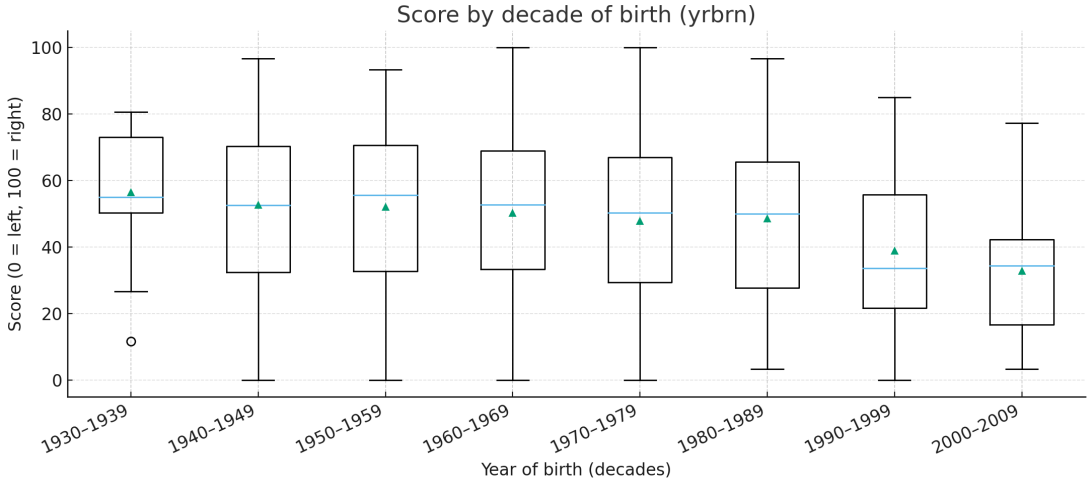
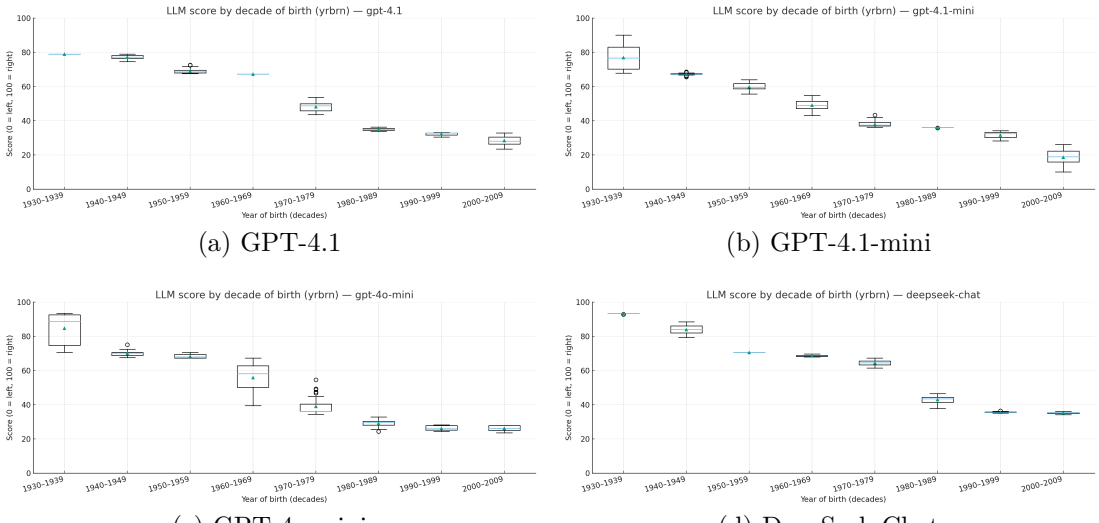


Figure 7.12: Survey boxplot by decade of birth (`yrbrn`).



(a) GPT-4.1 (b) GPT-4.1-mini

(c) GPT-4o-mini (d) DeepSeek-Chat

Figure 7.13: Model boxplots by `yrbrn`.

Comment. GPT-4.1 tracks cohort ordering most closely; mini variants flatten the gradient, especially for older cohorts (left shift + shrinkage). DeepSeek-Chat sometimes elevates mid-cohort centres (overshoot), widening gaps vs the survey at both ends.

7.6.5 Discussion

Across all features, three regularities persist: (i) **location shift**—model centres tend to be left of the survey; (ii) **dispersion compression**—IQRs and whiskers are narrower, muting tails; (iii) **ordering fidelity**—rank order of categories is mostly preserved by GPT-4.1, less so by mini variants; DeepSeek-Chat shows occasional overshoot with partial reordering in mid/upper categories. These distributional diagnostics corroborate the discussion reported in Section7.5 and explain why errors concentrate on rightmost categories and extreme profiles.

Chapter 8

Conclusions

This discussion integrates the empirical results with the study's design choices to answer a central question: to what extent can LLMs impersonate survey-grounded personas and reproduce human response patterns on politically salient items? The question sits at the intersection of (i) a carefully filtered, complete-case Italian subsample from ESS Round 11 and a derived 0–100 composite score, (ii) decision-tree–defined personas that bind multiple attributes into interpretable subgroups, and (iii) a four-model panel queried under matched decoding and replication protocols. The discussion therefore draws on the dataset construction and estimand choice (internal comparability via unweighted summaries), the persona pipeline, and the experimental setup that delivers 30 replications per persona–model to parallel profile supports on the human side.

8.1 Synthesis of principal findings

Three regularities stand out. First, all models display a systematic *leftward* location shift relative to survey means (negative bias in 0–100 score units). Second, dispersions are compressed on the model side: standard deviations across personas are far smaller than those observed in the survey, consistent with narrow anchoring of synthetic means. Third, despite these level and spread gaps, models preserve much of the *ordering* among personas: rank relationships are often maintained even when absolute levels diverge. These patterns are visible in the level-wise summaries and consolidated in the table of average bias and sd ratios (e.g., GPT-4.1 shows the smallest average bias and relatively milder compression among the four).

At the profile granularity, the same mechanisms manifest as clusters of synthetic persona means around a few anchor values, with occasional overshoots where interactions among attributes (e.g., party with social-value items) are amplified. These asymmetries explain

why some extreme right-tail profiles are under-reached even when rank-order is respected.

8.2 Interpreting discrepancies through latent structure

A PCA estimated on *real* profiles reveals a dominant one-dimensional axis (PC1 explains ($\approx 91\%$) of variance) with positive, similar loadings for the three base items (left–right self-placement, same-sex adoption rights, immigration from other ethnic groups). Projecting model-generated profiles into this space shows that model distributions differ from the empirical one via a location shift, dispersion differences, and altered tail shapes—precisely the ingredients observed in the 0–100 score domain. This explains how models can capture relative ordering along the main politico–value dimension while missing absolute calibration and tail reach.

Two design factors reinforce these effects. First, the 0–100 score is constructed from coarse item scales; mapping and aggregation can accentuate plateauing when models prefer discrete anchors, contributing to variance compression. Second, standard deviations on the human side reflect within-profile human heterogeneity, whereas on the synthetic side they reflect run-to-run stability under fixed prompts; the two are complementary but not commensurate, so "compressed" synthetic dispersion should be expected absent explicit variance calibration.

8.3 Persona complexity and model contrasts

As persona definitions incorporate more attributes (from L0/L1 baselines to multi-attribute L2–L4), fidelity does not degrade uniformly: mean shifts remain fairly stable across levels, but compression is persistent and particularly visible where real profiles are more extreme. The unconditioned baseline (L0) highlights cross-model heterogeneity—without persona anchors, responses revert to model-specific priors, underscoring why L0 serves as a null reference rather than a description of any real subgroup. Among models, GPT-4.1 offers the most balanced trade-off (smaller bias with less severe compression), while lighter variants show stronger shrinkage; DeepSeek-Chat occasionally overshoots specific profiles but shares the general leftward shift. These patterns align with the level-wise aggregations and the profile-level tables.

8.4 Limitations and directions for future work

Limitations include: (i) a focus on one country and a specific item triplet; (ii) reliance on prompt-bound personas without exploring alternative levers (e.g., richer retrieval contexts or fine-tuning); (iii) potential non-linear structure beyond the first principal

component; and (iv) incomplete alignment of dispersion and tails without explicit calibration. Future work should therefore: extend to broader item batteries and countries; incorporate multivariate calibration that respects item correlations; evaluate frozen-baseline snapshots to mitigate drift; and test enhanced RAG or fine-tuned variants within the same persona framework to assess gains beyond prompt-only control.

LLMs, when bound to well-specified, data-derived personas, can recover the main politico–value ordering among subgroups but exhibit a systematic level shift and compressed dispersion relative to survey data. Read as an early-stage instrument, the approach is informative and reproducible enough for screening and design, provided analysts (i) maintain internal comparability, (ii) diagnose structure in a human-estimated latent space, and (iii) apply transparent recalibration when absolute levels matter. Under these conditions, persona-based evaluation becomes a practical bridge between generative models and traditional survey workflows.

Bibliography

[1] L. P. Argyle et al. "Out of One, Many: Using Language Models to Simulate Human Samples". In: *arXiv preprint arXiv:2209.06899* (Sept. 2022). arXiv: 2209.06899 [cs.LG]. URL: https://arxiv.org/abs/2209.06899.

[2] J. Bisbee et al. "Synthetic Replacements for Human Survey Data? The Perils of Large Language Models". In: *Political Analysis* 32 (May 2024). Published online 17 May 2024, pp. 401–416. DOI: 10.1017/pan.2024.5. URL: https://doi.org/10.1017/pan.2024.5.

[3] L. Breiman et al. *Classification and Regression Trees*. New York: Chapman and Hall/CRC, 1984. DOI: 10.1201/9781315139470.

[4] European Union. *Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on Artificial Intelligence (AI Act)*. OJ L 2024/1689, published 12 July 2024; entered into force 1 August 2024. 2024. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng.

[5] P. Hämäläinen, M. Tavast, and A. Kunnari. "Evaluating Large Language Models in Generating Synthetic HCI Research Data: a Case Study". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Hamburg, Germany: Association for Computing Machinery, 2023, pp. 1–19. ISBN: 978-1-4503-9421-5. DOI: 10.1145/3544548.3580688. URL: https://doi.org/10.1145/3544548.3580688.

[6] I. T. Jolliffe. *Principal Component Analysis*. 2nd ed. Springer Series in Statistics. New York: Springer, 2002.

[7] I. T. Jolliffe and J. Cadima. "Principal Component Analysis: A Review and Recent Developments". In: *Philosophical Transactions of the Royal Society A* 374.2065 (2016), p. 20150202. DOI: 10.1098/rsta.2015.0202.

[8] R. Karanjai et al. "Synthesizing Public Opinions with LLMs: Role Creation, Impacts, and the Future to eDemocracy". In: *arXiv preprint arXiv:2504.00241* (Apr. 2025). arXiv: 2504.00241 [cs.CL]. URL: https://arxiv.org/abs/2504.00241.

[9] A. Kaur et al. “Synthetic Voices: Evaluating the Fidelity of LLM-Generated Personas in Representing People’s Financial Wellbeing”. In: *Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization (UMAP ’25)* (June 16–19, 2025). Best Full Paper Award. New York City, NY, USA: Association for Computing Machinery, June 2025, pp. 185–193. ISBN: 979-8-4007-1313-2. DOI: 10.1145/3699682.3728339. URL: https://doi.org/10.1145/3699682.3728339.

[10] P. Lewis et al. “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks”. In: *Advances in Neural Information Processing Systems (NeurIPS)* (2020). URL: https://arxiv.org/abs/2005.11401.

[11] A. Li et al. “LLM Generated Persona is a Promise with a Catch”. In: *arXiv preprint arXiv:2503.16527* (Mar. 2025). arXiv: 2503.16527 [cs.CL]. URL: https://arxiv.org/abs/2503.16527.

[12] P. Liu et al. “Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in NLP”. In: *ACM Computing Surveys* 55.9 (2023), 195:1–195:35. DOI: 10.1145/3560815.

[13] OpenAI. *Introducing GPT-4.1 in the API*. Knowledge cutoff refreshed to June 2024. Apr. 2025. URL: https://openai.com/index/gpt-4-1/.

[14] L. Salewski et al. “In-Context Impersonation Reveals Large Language Models’ Strengths and Biases”. In: *Advances in Neural Information Processing Systems (NeurIPS 2023)*. Vol. 36. 37th Conference on Neural Information Processing Systems (NeurIPS 2023). New Orleans, LA, USA: Curran Associates, Inc., 2023. URL: https://openreview.net/forum?id=....

[15] S. Santurkar et al. “Whose Opinions Do Language Models Reflect?” In: *Proceedings of the 40th International Conference on Machine Learning (ICML 2023)*. Vol. 202. Proceedings of Machine Learning Research. Honolulu, Hawaii, USA: PMLR, 2023, pp. 30122–30155. URL: https://proceedings.mlr.press/v202/santurkar23a.html.

[16] F. Wu, Z. Ren, Z. Sha, et al. “DeepSeek-V3 Technical Report”. In: *arXiv preprint arXiv:2412.19437* (2024). URL: https://arxiv.org/abs/2412.19437.

Appendix A

Codebooks and Detailed Scales

A.1 Italian education scale (edlvfit)

Table A.1: Full category list for `edlvfit` (Highest level of education, Italy).

Value	Category
1	Nessun titolo
2	Licenza elementare o attestato di valutazione finale di istruzione primaria
3	Avviamento professionale
4	Licenza media o diploma di istruzione secondaria di I grado
5	Qualifica professionale regionale (durata inferiore ai 2 anni)
6	Diploma di qualifica professionale di scuola secondaria superiore (di II grado) di 2–3 anni / Qualifica professionale regionale di 2–3 anni / Attestato di qualifica professionale (IeFP) di 3 anni
7	Diploma professionale IeFP di Tecnico (quarto anno)
8	Diploma di Maturità o Diploma di Istruzione secondaria superiore (di II grado) tecnica o professionale (inclusi l'Istituto magistrale di 4 anni e l'Istituto d'arte di 4 anni)
9	Diploma di Maturità o Diploma di Istruzione secondaria superiore (di II grado) – Licei (incluso l'Istituto magistrale di 5 anni)
10	Qualifica professionale regionale post-diploma / Certificato di specializzazione tecnica superiore (IFTS)
11	Diploma di Tecnico Superiore ITS
12	Laurea di primo livello (triennale)
13	Diploma universitario di 2–3 anni / Scuola diretta a fini speciali / Scuola parauniversitaria
14	Diploma accademico di primo livello AFAM (triennale)
15	Master universitario di 1° livello / Diploma accademico di specializzazione/perfezionamento di 1° livello (AFAM)

(continued)

Value	Category
16	Diploma di Accademia (Belle Arti, Nazionale di arte drammatica, Nazionale di Danza), Istituto Superiore Industrie Artistiche, Conservatorio di musica statale, Istituto Musicale Pareggiato (vecchio ordinamento)
17	Laurea vecchio ordinamento / Laurea specialistica o magistrale a ciclo unico
18	Laurea specialistica o magistrale di secondo livello (biennale)
19	Diploma accademico di secondo livello AFAM (biennale)
20	Master universitario di secondo livello / Diploma di specializzazione universitaria di secondo livello / Diploma accademico di specializzazione/perfezionamento di secondo livello (AFAM)
21	Dottorato di ricerca / Diploma accademico di formazione alla ricerca (AFAM)

A.2 Seed list

```
SEED_LIST_ = [
248865970, 1541332415, 1598889490, 1636497460, 684211952,
1975104565, 1584848762, 839587003, 830773567, 1421398127,
1328124858, 1686364458, 809853347, 1796639580, 257887319,
190566875, 87038551, 825177720, 977938232, 462338474,
2018717032, 881093686, 1119959072, 1584100904, 1342659968,
415896985, 538818173, 440569396, 1687476814, 1986632902,
]
```

Listing A.1: List of random seeds used in the experiments

A.3 Complete profile definition

Profile	Definition (filter conditions)
L2_1	prtvteit ≠ 2 AND freehms ≠ 1
L2_2	prtvteit ≠ 2 AND freehms = 1
L2_3	prtvteit = 2 AND freehms ≠ 1
L2_4	prtvteit = 2 AND freehms = 1
L3_1	prtvteit ≠ 2 AND prtvteit ≠ 3 AND freehms ≠ 1
L3_2	prtvteit = 3 AND freehms ≠ 1
L3_3	prtvteit ≠ 2 AND prtvteit ≠ 3 AND freehms = 1
L3_4	prtvteit = 3 AND freehms = 1
L3_5	prtvteit = 2 AND freehms ≠ 1 AND stfgov ≤ 3.5
L3_6	prtvteit = 2 AND freehms ≠ 1 AND stfgov > 3.5
L3_7	prtvteit = 2 AND freehms = 1 AND stfgov ≤ 2.5
L3_8	prtvteit = 2 AND freehms = 1 AND stfgov > 2.5
L4_1	prtvteit ≠ 2 AND prtvteit ≠ 3 AND freehms ≠ 1 AND freehms ≠ 2
L4_2	prtvteit ≠ 2 AND prtvteit ≠ 3 AND freehms = 2
L4_3	prtvteit = 3 AND freehms ≠ 1 AND yrbrn ≤ 1977
L4_4	prtvteit = 3 AND freehms ≠ 1 AND yrbrn > 1977
L4_5	prtvteit ≠ 2 AND prtvteit ≠ 3 AND freehms = 1 AND yrbrn ≤ 1981
L4_6	prtvteit ≠ 2 AND prtvteit ≠ 3 AND freehms = 1 AND yrbrn > 1981
L4_7	prtvteit = 2 AND freehms = 1 AND icgndra = 2 AND stfgov ≤ 2.5
L4_8	prtvteit = 2 AND freehms = 1 AND icgndra = 1 AND stfgov ≤ 2.5
L5_1	prtvteit ≠ 2 AND prtvteit ≠ 3 AND freehms = 3
L5_2	prtvteit ≠ 2 AND prtvteit ≠ 3 AND freehms = 4 OR freehms = 5
L5_3	prtvteit ≠ 2 AND prtvteit ≠ 3 AND freehms = 2 AND stfgov ≤ 7.5
L5_4	prtvteit ≠ 2 AND prtvteit ≠ 3 AND freehms = 2 AND stfgov > 7.5
L5_5	prtvteit ≠ 2 AND prtvteit ≠ 3 AND freehms = 1 AND stfgov ≤ 6.5 AND yrbrn ≤ 1981
L5_6	prtvteit ≠ 2 AND prtvteit ≠ 3 AND freehms = 1 AND stfgov > 6.5 AND yrbrn ≤ 1981
L6_1	prtvteit ≠ 2 AND prtvteit ≠ 3 AND freehms = 2 AND stfgov ≤ 4.5
L6_2	prtvteit ≠ 2 AND prtvteit ≠ 3 AND freehms = 2 AND 4.5 < stfgov ≤ 7.5

Table A.2: Logical definitions of L2–L6 personas inferred from the CSV filter queries (28 profiles).

A.4 Prompt used for single feature context

Single Feature Profile Prompt Examples

PROFILE (prtvteit)

You are italian and you voted for `{parameter}` in the most recent national election.

PROFILE (stfeco)

You rate your satisfaction with the present state of the economy in your country at `{parameter}` on a scale from 0 (extremely dissatisfied) to 10 (extremely satisfied).

PROFILE (stfgov)

You rate your satisfaction with the national government at `{parameter}` on a scale from 0 (extremely dissatisfied) to 10 (extremely satisfied).

PROFILE (freehms)

You `{parameter}` with the statement that gay men and lesbians should be free to live their own life as they wish.

PROFILE (edlvfit)

Your highest level of education is `{parameter}`.

PROFILE (fnsdfml)

When you were growing up, you and your family `{parameter}` experienced severe financial difficulties.

PROFILE (icgndra)

You are a `{parameter}`.

PROFILE (yrbrn)

You were born in `{parameter}`.

PROFILE (agea)

You are `{parameter}` years old.

PROFILE (gincdif)

You `{parameter}` with the statement that the government should take measures to reduce differences in income levels.