

LLM Project Write-Up Part 2

Yitong Liu

1 Methods

This section examines how demographic personas shape the internal reasoning patterns of a large language model. For each General Social Survey (GSS) question from 1972 to 2022, the model is prompted to “think step by step” what the response of *Average American* or as one of six race–gender personas (Black Male, White Male, Other Male, Black Female, White Female, Other Female) will be and the reasoning behind it. This produces 120,022 effective persona-based chain-of-thought reasoning responses from 17,146 sets of survey questions.

Note: I filter out LLM responses that try to return one overall reasoning for all persona or those less than 7 reasoning from downstream analysis.

The aim of this report is to investigate how persona cues shift the model’s reasoning at several levels. My analysis proceeds in three steps:

1. **Embedding:** how far persona-conditioned reasoning embeddings move away from the baseline in latent space.
2. **Keyword detection:** which keywords become more distinct to reasoning under each persona, using TF-IDF shift.
3. **Moral and affective framing:** how personas differ in moral framings and sentiment, using dictionary methods.

All reasoning texts are embedded using `intfloat/e5-large-v2`, yielding 1024-dimensional vectors normalized to unit length. These embeddings allow two complementary comparisons. First, for every (year, question) item, the Euclidean distance between a persona’s reasoning and the corresponding baseline response is computed, capturing within-item semantic deviations. Second, persona centroids are constructed by averaging embeddings across all items, enabling measurement of systematic differences between race and gender groups independently of year and question content. These two representations together trace how demographic modifiers alter the model’s internal configuration of reasoning.

Lexical distinctiveness is measured through a global TF-IDF space built over the entire corpus. For each persona and question, TF-IDF values are compared to the baseline for that same question, producing a shift vector that records how strongly each word becomes more or less central to the

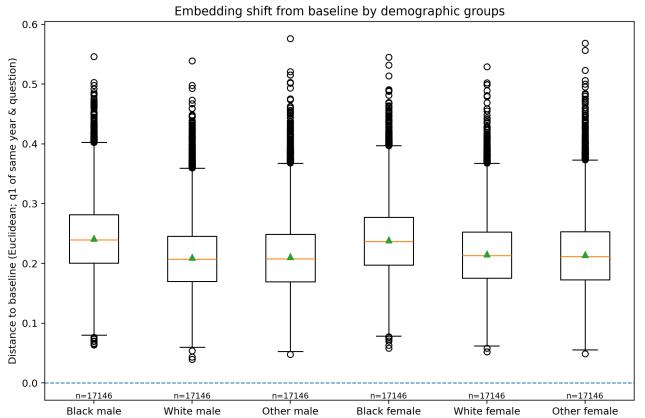


Figure 1: Embedding shifts from baseline by demographic group. Persona reasoning consistently diverges from baseline, indicating that demographic prompting induces systematic semantic changes.

model’s reasoning under a given demographic identity. Averaging these shifts across all questions yields a stable persona-level lexical profile. From these profiles, I extract the top distinct keywords for each persona and examine representative reasoning passages to understand how these lexical differences reflect broader thematic and interpretive tendencies.

To connect these lexical differences to normative content, all reasoning texts are analyzed using the Moral Foundations Dictionary (MFD 2.0). Each persona’s proportion of tokens belonging to the categories of Care, Fairness, Loyalty, Authority, and Sanctity is computed, separating virtue and vice terms. These proportions reveal which moral dimensions are emphasized by each persona and how moral framing contributes to the broader patterns observed in embedding space. Sentiment is evaluated using VADER, producing positive, negative, neutral, and compound scores that summarize the affective tone of persona-conditioned reasoning.

2 Empirical Findings

I first ask whether demographic prompting materially shifts the model’s internal representations. Figure 1 shows the distribution of Euclidean distances between persona embeddings and the baseline embeddings for the same questions.

All personas exhibit substantial separation from the baseline, with median distances clustered between roughly 0.20

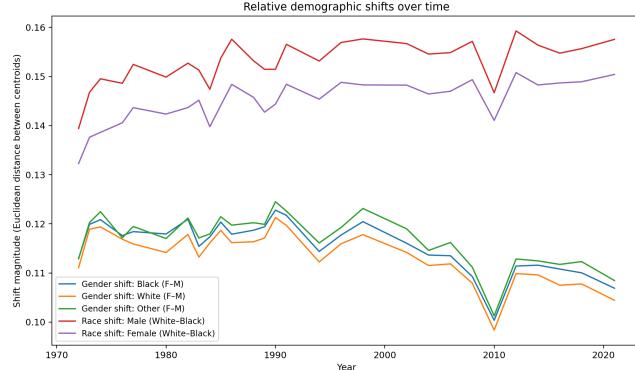


Figure 2: Relative race and gender shifts over time (1972–2022). Race produces larger and more persistent semantic displacements than gender.

and 0.27. This confirms that demographic information consistently alters the semantics of the model’s reasoning, rather than simply modifying surface-level phrasing. The median distances are remarkably similar across groups, suggesting that the magnitude of persona influence is broadly comparable across race and gender.

I also compare how persona centroids have shifted along the lines of gender and race by comparing persona centroids sharing the same race or gender categories. Figure 2 traces the Euclidean distances between race-specific and gender-specific centroids over time.

Across all fifty years, gender differences within race form a tight band centered around 0.11–0.12, whereas race differences within gender maintain a larger and more stable separation around 0.14–0.15. These patterns indicate that the semantic influence of race is consistently stronger than that of gender, producing more sustained shifts in the latent space of reasoning. The fact that these curves remain stable across decades suggests that the underlying semantic dimensions tied to race and gender are encoded in ways that generalize across survey content and historical variation.

I next examine persona-specific lexical patterns. By comparing TF-IDF values for each persona to the baseline across all questions, I identify the words whose salience increases most strongly under each demographic identity. Table 1 summarizes the top 30 entries for each persona.

These keywords reveal clear thematic contrasts. Black personas tend to foreground concepts of justice, vulnerability, protection, collective experience, and structural inequity, while White personas emphasize individual responsibility, institutional authority, and personal obligation. Other-race personas often fall between these two poles but display elevated markers of interpersonal cohesion.

These lexical patterns align with themes observed in representative excerpts, where LLM reason with particular emphasis on the cultural variation and historical discrimination of non-white groups and on maternal and family-related com-

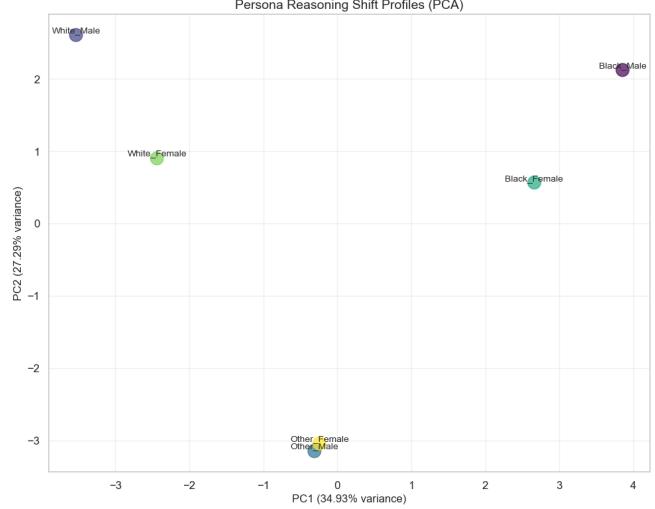


Figure 3: Principal components of persona-level reasoning signatures. PCA1 reflects racial differences, while PCA2 captures gender-modulated stylistic variation.

mitments among female groups. As shown in Table 2, although the LLM consistently estimates around one to two earners for all personas when asked about the average number of earners per household in 1972, its reasoning processes differ across personas. While the model applies the traditional breadwinner framework to all groups, it highlights historical labor-market constraints specifically for the Black male persona, yet emphasizes a slightly higher expected number of earners for the Black female persona. This suggests that for the Black female persona, the LLM is attempting to reconcile two historical patterns: structural discrimination affecting Black households and the comparatively higher labor-force participation of Black women—patterns that are not simultaneously activated in the Black male reasoning. Similarly, for other race personas, the model introduces additional considerations related to cultural variation. This example illustrates that persona conditioning not only complicates the LLM’s reasoning about social questions, but that specific persona combinations can also lead to distinct shifts in how the model justifies its answers.

To understand how these lexical differences relate to broader structural patterns, I apply principal component analysis to the persona-by-word TF-IDF shift matrix. Figure 3 plots the first two principal components.

The first component, which explains just over 32% of total variance, cleanly separates Black personas on one side and White personas on the other, with Other personas positioned between them. This confirms that the dominant axis of variation in reasoning is racialized. The second component, accounting for about 27% of the variance, captures subtler stylistic differences between male and female personas within each racial group. The geometric arrangement shows that race defines the primary direction along which persona-

Table 1: Top 30 distinctive reasoning keywords per persona (TF-IDF shift).

Persona	Top 30 Distinctive Keywords
Black Male	blk, oversamp, darker, intermixing, tracking, commissions, tips, dismantling, apparent, half-black, low-wage, hurdles, firsthand, unhealthy, dark, stalled, underreported, tones, non-identification, workgroup-based, practicality, arrests, verbal, dismantle, non-confrontational, reallocation, grp, hazardous, formation, reverence
Black Female	blk, oversamp, blocs, reliably, caters, husbands, darker, insufficiently, verbally, dark, commissions, tips, herbs, self-definition, newcomers, doubly, drive-related, workgroup-based, bloc, hazardous, health-specific, indies, monitored, grp, shopper, reinforced, compound, non-disclosing, anemia, sickle
White Male	compensations, romney, white-majority, law-abiding, legislative, whiteness, trigonometry, edges, gave, nationality-based, hosting, resp, romanian, reward-based, satisfactory, instruction, tones, impede, individualized, wwi, reinforced, str, entrenched, ancestor, camping, post-obama, science-faith, layoff, self-satisfaction, boarder
White Female	marketing, compensations, german-american, host, garden, mirroring, negligible, comprehensiveness, juggle, hazardous, str, well-served, romney, altruistic, combine, health-specific, layers, moderated, infrequency, multitask, insurmountable, gender-inclusive, groomed, congregational, health-consciousness, shopper, hard-working, small-town, privacy-conscious, non-disclosure
Other Male	parallels, tracking, romanian, hosting, invested, logic, clustering, non-affiliation, mirrored, fables, non-denominationalism, behave, speculated, sevrl, ticketing, dishonest, fitting, downplaying, instruments, subgroup, advertise, government-supportive, pragmatically, signals, decreases, disconnection, accommodating, liberal-conservative, third-generation, entrenched
Other Female	non-irish, tracking, offerings, non-denominationalism, fables, non-identification, sevrl, symptom, empathizes, male-centric, invested, clustering, speculated, distinctiveness, grp, non-affiliation, positioning, dishonest, adult-only, caveat, integrative, non-polarizing, clinics, subgroup, layers, instruments, breadth, combine, decreases, advertise

Table 2: Answers and COT reasonings on the question of the average number of earners in 1972, per persona .

Persona	Answer and Reasoning
Typical American	In 1972, most families still relied on a single male breadwinner, though women were increasingly entering the workforce. The average response would likely fall between 1 and 2 earners, reflecting the shift toward dual-income households.
Black Male	Similar to the general population, a typical Black male respondent would likely report 1 to 2 earners, shaped by both the traditional breadwinner model and employment constraints affecting Black workers at the time.
White Male	A typical White male respondent would similarly report around 1 to 2 earners, as white families were still predominantly single-breadwinner households but dual-income arrangements were becoming more common.
Other Male	A typical male respondent identifying as neither Black nor White would likely indicate 1 to 2 earners, with variation depending on specific cultural or economic circumstances.
Black Female	A typical Black female respondent might lean slightly higher within the 1 to 2 earner range, reflecting the comparatively higher labor force participation of Black women in 1972.
White Female	A typical White female respondent would likely report around 1 to 2 earners, consistent with predominantly single-income households during this period but with increasing female workforce participation.
Other Female	A typical female respondent identifying as neither Black nor White would likely report 1 to 2 earners, with potential cultural variation influencing household earning patterns.

conditioned reasoning varies, while gender modulates this variation along a secondary, largely orthogonal dimension.

To understand LLM’s framing for persona-based reasoning more systematically, I apply dictionaries-based methods to detect normative framing and emotion intensiveness. Figure 4 reports the distribution of moral framings across personas.

Results with Moral Foundations Dictionary show that reasoning under Black persona has elevated rates of fairness. This reflects LLM’s attention to racial exclusion, as well as group-based vulnerability and marginality in its reasoning. White personas display the highest rates of authority , reflecting LLM reasoning’s greater emphasis on legitimacy, and institutional trust. Interestingly, despite I define other-race as groups with races other than black or white in the prompt, LLM reasoning for other-race persona exhibit the strongest presence of loyalty-related terms, signaling a heightened focus of group cohesion.

Patterns for gender are consistent with well-documented gender stereotypes. All three female personas—White Fe-

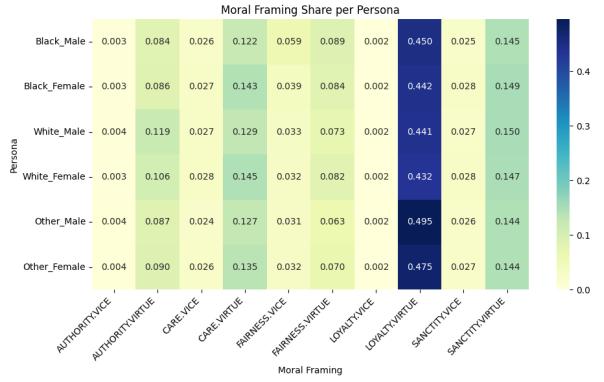


Figure 4: Moral framing shares by persona. Black personas emphasize fairness and vulnerability; White personas emphasize authority; Other-race personas emphasize loyalty.

male, Black Female, and Other Female—score the highest

on care-related moral terms. In contrast, I do not observe a corresponding inverse pattern for male personas. This asymmetry suggests that the model’s reasoning incorporates gendered expectations more strongly for women than for men, amplifying culturally familiar associations between femininity, caregiving, and emotional concern.

These differences align closely with the thematic distinctions documented in table 1, demonstrating that persona-conditioned reasoning diverges not only in content but also in the moral frameworks within which explanations are constructed.

Finally, I evaluate tones and emotions. Figure 5 presents VADER sentiment metrics by persona.

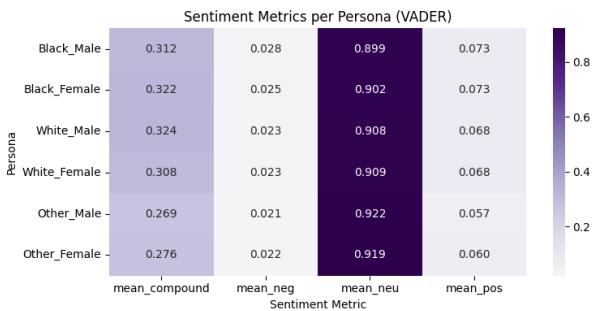


Figure 5: Sentiment metrics per persona (VADER). Black personas show slightly more positive tone; Other personas are more neutral.

Black personas tend to produce slightly more positive and emotionally engaged reasoning, while other-race personas exhibit more neutral affect. Although these differences are modest, they contribute to the overall picture of personas guiding not only the thematic framing but also the emotional tone of the model’s reasoning.

3 Conclusion

Across embedding geometry, lexical distinctiveness, structural dimensionality, and normative framing, a consistent pattern emerges: persona cues reliably shift the model’s reasoning in systematic and interpretable ways. Race exerts the strongest influence, defining the primary direction of variation in both semantic space and moral framing, while gender introduces smaller but coherent adjustments within racial groups. Reasoning for Black personas consistently adopts a framework of social justice, structural inequity, and group-based vulnerability. Reasoning for White personas emphasizes authority, legitimacy, and institutional trust, whereas Other-race personas lie between these poles but lean toward themes of group cohesion.

Gender patterns reveal an additional asymmetry. Female personas systematically express more care-related and gender-specific considerations than male personas, yet male personas do not show a corresponding consolidation around

an opposing moral dimension. This indicates a gendered bias in how the model constructs identity-linked reasoning, with femininity more strongly associated with caregiving, emotion, and social context than masculinity.

Taken together, these results demonstrate that LLMs do not treat persona instructions as superficial stylistic cues. Instead, they internalize demographic identities and the societal biases and stereotypes that accompany them in ways that reproduce recognizable sociocultural patterns of reasoning.