In Your Own Words: Free-Text Descriptions of Identity Reveal Information Beyond Census Categories

Anonymous authors
Paper under double-blind review

Extended Abstract

Categories for race, gender, and sexual orientation are essential for demographic measurement but often fail to reflect the full complexity of individuals' identities. We explore whether free-text responses, paired with modern language modeling techniques, can offer more granular and contextually relevant representations of identity. We show that free-text descriptions contain information that categorical variables do not: themes extracted using sparse autoencoders are not well-explained by conventional identity categories, and free-text-based features—such as *mentions feelings of not fully belonging or being out of place related to race/ethnicity* and *explicitly states a lack of cultural or ethnic identity or traditions*—significantly improve the prediction of key outcomes, including life satisfaction and mental health. Our results demonstrate the potential of language modeling to capture nuanced dimensions of identity that are often flattened by fixed-choice categories. These findings have implications for demographic and public opinion research, where identity categories are often used as explanatory variables. By incorporating identity themes learned from free-text responses, researchers can better capture the social meaning and lived experiences of identity, enabling richer measurement and potentially more accurate modeling of attitudes, well-being, and behavior.

Collecting free-text data. We first build the *In Your Own Words* dataset, comprising responses from 1,004 English-speaking participants from the United States who provided both categorical and open-ended descriptions of their race, gender, and sexual orientation identities. Our categorical identity questions follow Census and Pew Research Center best practices and existing literature Hughes et al. (2022), while our free-text survey question asked participants: *In at least 2-3 sentences, how would you describe your {race and/or ethnicity, gender identity, sexual orientation}?* We also collect additional information on life outcomes including self-perceived physical health, mental health, and life satisfaction, as well as measures of discrimination. We will make this dataset available to other researchers and plan to develop a web interface to facilitate continued free-text identity data collection over time from a broader sample.

Computationally extracting themes from text. While identity free-text responses may contain rich information, analyzing them at scale is challenging. Identity researchers have historically relied on time-intensive manual coding Fraser et al. (2020), or have excluded open-ended responses from quantitative analysis altogether Morgan et al. (2020); Bates et al. (2019). Our goal is to extract meaningful themes that accurately capture how participants describe their identities in free-text responses. We do this by (1) learning an interpretable representation of the free-text responses using a sparse autoencoder (SAE) and then (2) applying a large language model (LLM) to interpret each dimension of the representation by identifying common themes among the free-text examples which score highly along that dimension. Figure 1 describes our high-level workflow.

A major advantage of our approach is that it can capture both category-specific and cross-cutting identity themes without requiring a predefined classification step or extensive manual work; the only manual step involves reading the themes to ensure that they do not pick up on writing style or other text-related artifacts that are not useful to the researcher. Unlike methods that sort individuals into a single cluster or group, our model allows each response to reflect multiple identity themes at once, acknowledging that identities are often multifaceted and overlapping. For example,

> "I am racially fully Korean, where I can speak/read/write in Korean fluently. I moved to the states at 5 years old, so despite Korean being my first language, I identify as American and I think in English. My preferred language is English."

The participant's response activates five themes: *"mentions speaking English or American English as a primary language"*, *"mentions being first, second, or third generation American or immigrant"*, *"mentions specific regions or countries of ancestral origin"*, *"mentions speaking or understanding multiple languages or specific non-English languages"*, and *"mentions the languages spoken by themselves or their family"*, capturing aspects of their racial identity that range from multilingualism to immigration and ethnic background.

This example of multi-theme interpretation would be difficult to recover using traditional clustering or topic modeling techniques, which often assign a response to a single group or rely on broad word patterns that require manual interpretation Blei et al. (2003); Grootendorst (2022). Additionally, compared to fully LLM-driven approaches Pham et al. (2024)—which can be costly and require careful prompt design—our approach offers a middle ground. See Figure 2 for the full set of themes extracted from race free-text responses.

Themes reveal dimensions of identity not captured by survey categories. To the right of each row in Figure 2, we display the R^2 value, which indicates how well the theme is explained by the race categories. A higher R^2 means that the theme closely aligns with a self-reported race category, while a lower R^2 suggests that the theme is expressed across multiple groups or inconsistently within a single group.

We focus our analysis on themes expressed across multiple identity groups and not well explained by race categories alone. Several relate to nationality and migration history, like *"mentions being first, second, or third generation American or immigrant"* ($R^2 = 0.06$), and language, *"mentions the languages spoken by themselves or their family"* ($R^2 = 0.09$). These patterns suggest that respondents draw on cultural, multilingual, and generational reference points when describing their racial identity, dimensions that standard demographic categories likely miss. We also observe more affective themes, such as *"mentions feelings of not fully belonging or being out of place related to race/ethnicity"* ($R^2 = 0.05$), which articulate ambiguity or marginalization within one's assigned race category.

We notice similar cross-cutting patterns in gender and sexual orientation themes. Some themes align with specific groups (e.g., *"mentions being nonbinary"*), but many highlight shared identity experiences. For instance, themes such as *"mentions how gender identity influences decisions, safety, or interactions in life"* ($R^2 = 0.06$) appear across a range of gender categories—including both cisgender and transgender respondents. Similarly, in the sexual orientation analysis, themes like *"mentions discomfort or rejection of labels for their sexual orientation"* ($R^2 = 0.34$) are activated across queer and heterosexual respondents.

Themes improve prediction of well-being and social outcomes. We assess whether using the SAE themes improves prediction of five outcomes (life satisfaction, physical health, mental health, everyday discrimination, and identity importance) relative to only using traditional Census categories, using adjusted R^2 as a prediction metric and using a nested F-test to assess the statistical significance of the increase in predictive power. We find that using SAE themes does improve predictive power relative to using conventional categories alone. For instance, Table 1 shows that adding race-based SAE themes improves prediction of mental health by over 90%, and gender-based themes increase explained variance in identity salience by 1.5x.

This work demonstrates that free-text responses to identity questions can be systematically analyzed using modern language modeling techniques to uncover themes that are interpretable, predictive, and not well-captured by standard survey categories. Our themes improve prediction of well-being and social outcomes and offer researchers a new way to surface shared narratives across identity categories. Our findings point to the value of integrating free-text into survey design, especially in public opinion and social science research where identity is central (and multifaceted) but often distilled to a set of categories.

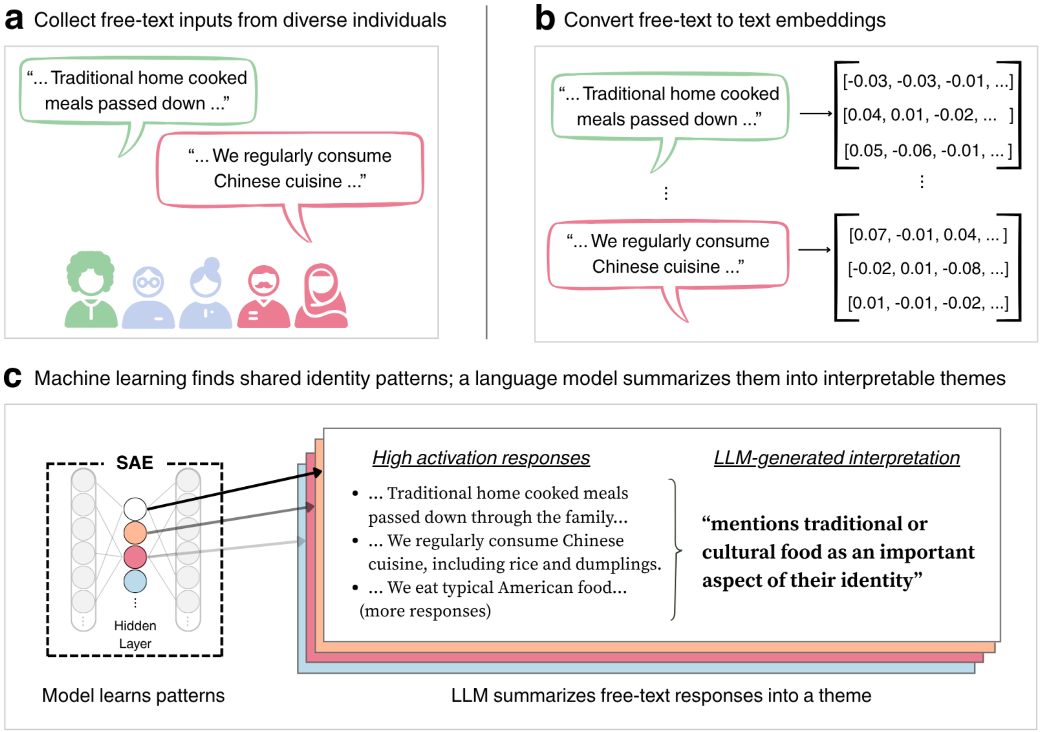
a Collect free-text inputs from diverse individuals
"... Traditional home cooked meals passed down ..."
"... We regularly consume Chinese cuisine ..."
b Convert free-text to text embeddings
"... Traditional home cooked meals passed down ..."
[-0.03, -0.03, -0.01, ...]
[0.04, 0.01, -0.02, ...]
[0.05, -0.06, -0.01, ...]
"... We regularly consume Chinese cuisine ..."
[0.07, -0.01, 0.04, ...]
[-0.02, 0.01, -0.08, ...]
[0.01, -0.01, -0.02, ...]
c Machine learning finds shared identity patterns; a language model summarizes them into interpretable themes
SAE
Hidden Layer
High activation responses
• ... Traditional home cooked meals passed down through the family...
• ... We regularly consume Chinese cuisine, including rice and dumplings.
• ... We eat typical American food...
(more responses)
LLM-generated interpretation
"mentions traditional or cultural food as an important aspect of their identity"
Model learns patterns
LLM summarizes free-text responses into a theme

Figure 1: **Overview of the *In Your Own Words* computational pipeline.** **a**, Participants respond to open-ended questions about their race, gender, and sexual orientation. **b**, we convert free-text identity responses into embeddings using OpenAI's *text-embedding-3-large* model. **c**, we apply an SAE to compress each embedding into a low-dimensional vector in which only a small number of dimensions are active per response; prior work has shown that this tends to result in embeddings with interpretable dimensions Movva et al. (2025). Each dimension, often called a "neuron", captures a recurring pattern in how identity is expressed, such as references to cultural heritage, language, or childhood experiences. To interpret each neuron, we retrieve high-activation texts and prompt a large language model (LLM) to generate short natural language summaries. We also perform a manual step to exclude neurons that are uninterpretable or reflect prompt artifacts.

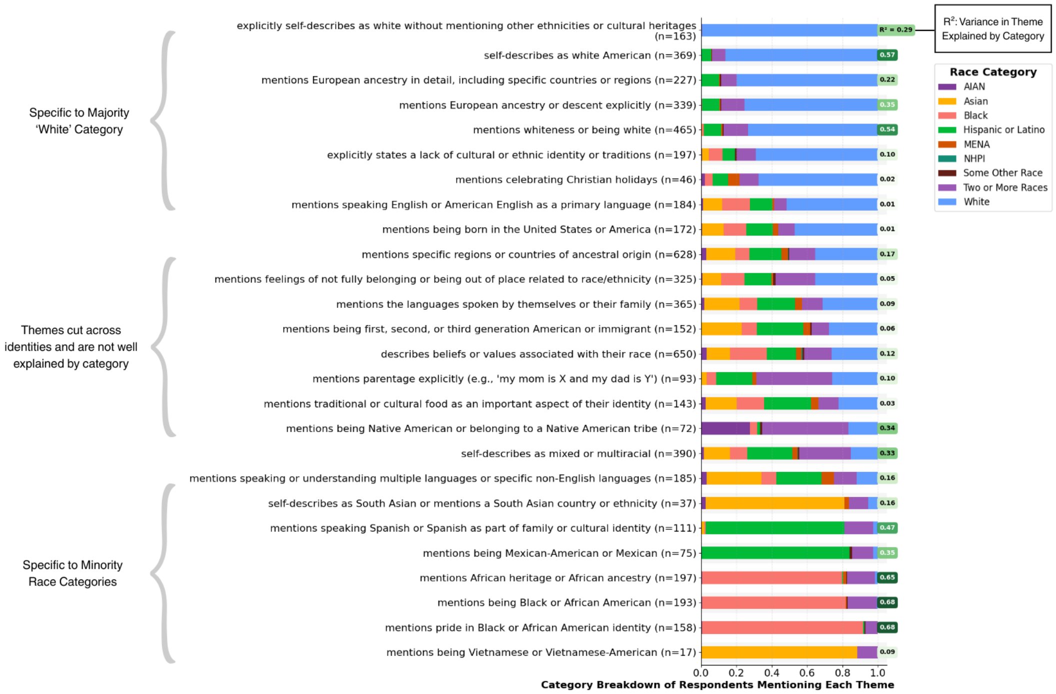

Figure 2: **Some identity themes align with race categories, but many cut across them**. Each row represents an interpretable theme learned from a sparse autoencoder trained on free-text race responses. Bars indicate the proportion of participants in each race category whose responses aligned with that theme. Themes are sorted by the proportion of majority identity (e.g. 'White') respondents in each theme. Themes are generated from a k=4-sparse autoencoder with 32 latent dimensions (bottleneck layer). We manually remove themes that are specific to the LLM's prompt and those with low interpretation fidelity, a metric borrowed from Movva et al. (2025).

		Well-Being Outcomes			Social Outcomes	
Identity		**Life Satisfaction**	**Physical Health**	**Mental Health**	**Everyday Discrimination**	**Identity Importance**
Race	Model Improv.	0.035 (***)	0.039 (***)	0.043 (***)	0.007 (–)	0.046 (***)
	% Increase	328.8%	123.6%	90.3%	30.4%	15.6%
Gender	Model Improv.	0.029 (***)	0.024 (**)	0.032 (**)	0.015 (*)	0.082 (***)
	% Increase	45.7%	43.8%	43.3%	45.7%	156.2%
Sexual Orientation	Model Improv.	0.014 (*)	0.022 (*)	0.030 (***)	–0.005 (–)	0.077 (***)
	% Increase	16.9%	31.3%	32.0%	–11.9%	222.6%

Table 1: **Identity dimensions help explain significantly more variation in well-being outcomes and identity salience.** Improvement in explained variance (Model Improv.) and percent increase (% Increase) are calculated by comparing the variance explained by models with interpretable SAE identity dimensions to baseline category-only models. % increase is computed as: (Model Improv. / Baseline Explained Variance) × 100%. Significance levels are Benjamini-Hochberg corrected: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

References

Nancy Bates, Yazmín A García Trejo, and Monica Vines. Are sexual minorities hard-to-survey? insights from the 2020 census barriers, attitudes, and motivators study (cbams) survey. *Journal of Official Statistics*, 35(4):709–729, 2019.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

Gloria Fraser, Joseph Bulbulia, Lara M Greaves, Marc S Wilson, and Chris G Sibley. Coding responses to an open-ended gender measure in a new zealand national sample. *The Journal of Sex Research*, 57(8):979–986, 2020.

Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

Jennifer L Hughes, Abigail A Camden, Tenzin Yangchen, Gabrielle PA Smith, Melanie M Domenech Rodríguez, Steven V Rouse, C Peeper McDonald, and Stella Lopez. Guidance for researchers when using inclusive demographic questions for surveys: Improved and updated questions. *Psi Chi Journal of Psychological Research*, 27(4):232–255, 2022.

Rachel E Morgan, Christina Dragon, Gemirald Daus, Jessica Holzberg, Robin Kaplan, Heather Menne, Amy Symens Smith, Maura Spiegelman, et al. Updates on terminology of sexual orientation and gender identity survey measures. Technical report, United States. Federal Committee on Statistical Methodology, 2020.

Rajiv Movva, Kenny Peng, Nikhil Garg, Jon Kleinberg, and Emma Pierson. Sparse autoencoders for hypothesis generation, 2025. URL https://arxiv.org/abs/2502.04382.

Chau Minh Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. TopicGPT: A prompt-based topic modeling framework. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2956–2984, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.164. URL https://aclanthology.org/2024.naacl-long.164/.

A Appendix: SAE Themes

A.1 Race Themes (32-dimensional)

Following Movva and Peng et al. (2025), we evaluate each interpretation's fidelity as the F1 score when predicting the presence of a theme, using an LLM (GPT-4o-mini) to annotate 100 positive and 100 negative samples for the presence of the concept. Table 2 shows the 32 themes initially generated. We exclude themes with F1 fidelity scores below 0.50, as well as those that reflect prompt artifacts. Excluded themes are indicated with strikethrough in Table 2, thus resulting in 26 interpretable themes.

#	Interpretation	F1 Fidelity
1	mentions celebrating Christian holidays	0.78
2	~~uses the phrase ‘I would describe my race’ or a close variation of it~~	0.91
3	mentions speaking English or American English as a primary language	0.90
4	mentions being first, second, or third generation American or immigrant	0.75
5	mentions pride in Black or African American identity	0.88
6	~~uses the phrase ‘I am’ followed by a race or ethnicity descriptor without elaboration~~	0.75
7	self-describes as mixed or multiracial	0.79
8	explicitly self-describes as white without mentioning other ethnicities or cultural heritages	0.86
9	mentions specific regions or countries of ancestral origin	0.69
10	mentions European ancestry or descent explicitly	0.64
11	mentions being Black or African American	0.98
12	mentions speaking or understanding multiple languages or specific non-English languages	0.71
13	mentions being Native American or belonging to a Native American tribe	0.97
14	self-describes as South Asian or mentions a South Asian country or ethnicity	0.75
15	~~uses single-word or very short descriptions of race/ethnicity (1-3 words)~~	0.69
16	mentions feelings of not fully belonging or being out of place related to race/ethnicity	0.67
17	mentions being Mexican-American or Mexican	0.99
18	explicitly states a lack of cultural or ethnic identity or traditions	0.97
19	mentions European ancestry in detail, including specific countries or regions	0.68
20	mentions parentage explicitly (e.g., ’my mom is X and my dad is Y’)	0.86
21	mentions the languages spoken by themselves or their family	0.78
22	~~explicitly uses the phrase ‘identify as’ or ‘identify with’ to describe their race or ethnicity~~	0.74
23	~~mentions speaking Arabic language or dialects~~	0.46
24	describes beliefs or values associated with their race	0.70
25	mentions whiteness or being white	0.61
26	~~mentions celebrating Lunar New Year or similar cultural holidays~~	0.39
27	mentions being Vietnamese or Vietnamese-American	0.51
28	mentions being born in the United States or America	0.82
29	mentions speaking Spanish or Spanish as part of family or cultural identity	0.80
30	mentions traditional or cultural food as an important aspect of their identity	0.62
31	mentions African heritage or African ancestry	0.72
32	self-describes as white American	0.76

Table 2: Natural language interpretations of SAE neurons trained on race-related free text response embeddings and their associated fidelity scores (F1).

A.2 Gender Themes (32-dimensional)

Similar to the 32-dimensional race themes, we manually exclude features that are uninterpretable or reflect prompt artifacts. For gender, this process results in a set of 27 interpretable dimensions (see Table 3). The themes learned from gender-related free text responses reveal both category-specific patterns and cross-cutting identity narratives. At one end of the spectrum (top half of themes in Figure 3), we observe themes that align closely with conventional gender categories. For example, “self-describes as male or man and mentions alignment with gender assigned at birth” ($n = 287$, $R^2 = 0.78$) and “mentions being born female and identifying as female” ($n = 412$, $R^2 = 0.59$) both show strong alignment with cisgender men and women, respectively, and yield high explained variance. Higher explained variance suggests that these themes are more consistently expressed among individuals within a given demographic group.

We also observe themes that capture gendered social roles and expectations, particularly among the cisgender individuals. For example, “mentions giving birth or having children as a significant aspect of their identity” is almost exclusively expressed by cisgender women ($n = 60$, $R^2 = 0.03$), while “mentions traditionally masculine activities or traits such as sports, fixing things, or being a provider/protector” ($n = 41$, $R^2 = 0.04$) aligns closely with cisgender men. Themes such as “mentions responsibility to provide for family” ($n = 34$, $R^2 = 0.02$) and “mentions roles or responsibilities within a family or household” ($n = 86$, $R^2 = 0.01$) appear across both cisgender men and women (illustrated in Figure 3 with high proportions of red and blue bars), reflecting traditional gender roles primarily shaping the experiences of cisgender individuals. Notably, the low explained variance reflects that these

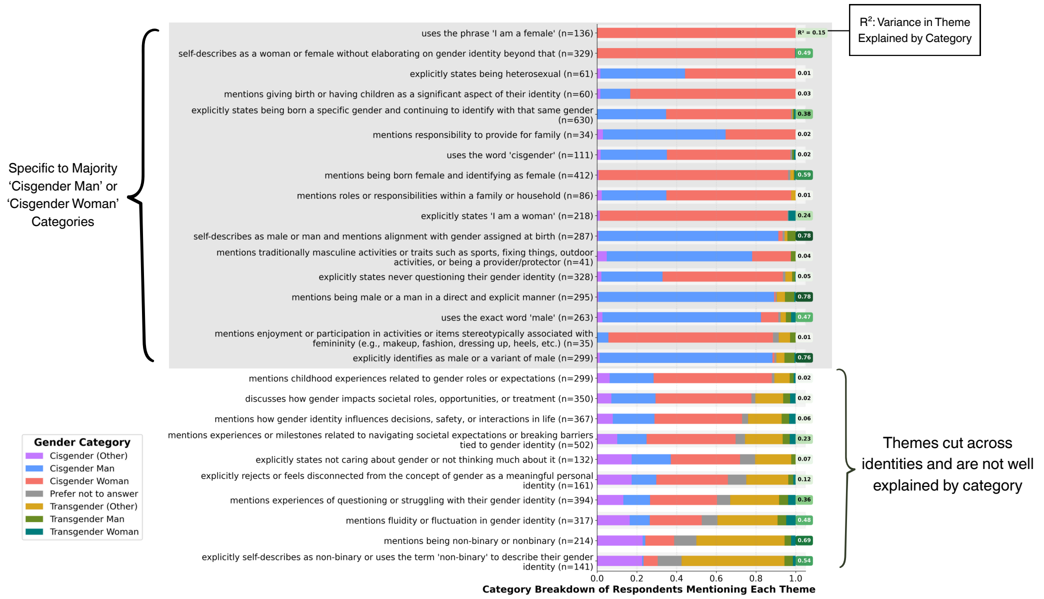

Figure 3: **Gender-related themes differ across cisgender men and women, and other groups**. Each row represents an interpretable theme learned from a sparse autoencoder trained on free text gender responses. Bars indicate the proportion of participants in each gender category whose responses aligned with that theme. Themes are sorted by the joint proportion of cisgender man and cisgender woman identity respondents (decreasing size of red and blue bars).

themes, while associated with cisgender identities, are not uniformly expressed across all individuals in those groups.

Many other themes cut across gender categories. Respondents from multiple groups describe their identity using themes like "mentions childhood experiences related to gender roles or expectations," "mentions experiences of questioning or struggling with their gender identity," and "mentions fluidity or fluctuation in gender identity." These themes are activated by all gender categories, but respondents identifying as transgender, nonbinary, or selecting "prefer not to answer" largely activated themes related to social expectations, safety, and navigation of gendered spaces.

A.3 Sexual Orientation Themes (32-dimensional)

Again, we manually exclude features that are uninterpretable or reflect prompt artifacts. For sexual orientation, this process results in a set of 28 interpretable dimensions (see Table 4). Themes drawn from sexual orientation responses exhibit a similar pattern: some themes align strongly with conventional categories, while many others reveal more complex or overlapping forms of identification. For example, "explicitly self-describes as straight or heterosexual without additional context" is expressed almost exclusively by participants identifying as straight, and shows high explained variance ($R^2 = 0.69$). Other high-R^2 themes include "explicitly uses the term 'heterosexual'" and "explicitly states never questioning their sexual orientation."

In contrast, many minority-aligned or nuanced themes span across sexual orientation categories and show low R^2 values. Themes such as "mentions discomfort or rejection of labels for their sexual orientation," "mentions uncertainty or questioning," and "mentions romantic orientation as distinct from sexual orientation" appear among a wide range of identity groups, from bisexual and queer respondents to those identifying as asexual or pansexual. Additionally, themes like "mentions age or life stage when they realized or came out" or "refuses to provide a clear or specific answer" reflect lived experiences that

#	Interpretation	F1 Fidelity
1	uses the exact word ’male’	0.91
2	mentions childhood experiences related to gender roles or expectations	0.78
3	mentions responsibility to provide for family	0.55
4	mentions use of multiple pronouns or experimenting with pronouns	0.46
5	mentions physical appearance or specific physical traits	0.38
6	mentions giving birth or having children as a significant aspect of their identity	0.86
7	single-word self-description of gender	0.56
8	explicitly states never questioning their gender identity	0.82
9	explicitly states being born a specific gender and continuing to identify with that same gender	0.70
10	mentions how gender identity influences decisions, safety, or interactions in life	0.72
11	self-describes as male or man and mentions alignment with gender assigned at birth	0.90
12	uses the phrase ’I would describe my gender identity as ...’	0.77
13	mentions enjoyment or participation in activities or items stereotypically associated with femininity (e.g., makeup, fashion, dressing up, heels, etc.)	0.57
14	mentions traditionally masculine activities or traits such as sports, fixing things, outdoor activities, or being a provider/protector	0.50
15	uses the word ’cisgender’	0.98
16	explicitly states being heterosexual	0.90
17	mentions being non-binary or nonbinary	0.81
18	mentions fluidity or fluctuation in gender identity	0.78
19	uses the phrase ’I am a female’	0.81
20	mentions being assigned a gender at birth	0.38
21	mentions experiences of questioning or struggling with their gender identity	0.75
22	mentions experiences or milestones related to navigating societal expectations or breaking barriers tied to gender identity	0.65
23	mentions roles or responsibilities within a family or household	0.56
24	self-describes as a woman or female without elaborating on gender identity beyond that	0.76
25	mentions being male or a man in a direct and explicit manner	0.86
26	explicitly rejects or feels disconnected from the concept of gender as a meaningful personal identity	0.71
27	explicitly states not caring about gender or not thinking much about it	0.91
28	explicitly identifies as male or a variant of male	0.77
29	discusses how gender impacts societal roles, opportunities, or treatment	0.85
30	explicitly states ’I am a woman’	0.93
31	mentions being born female and identifying as female	0.80
32	explicitly self-describes as non-binary or uses the term ’non-binary’ to describe their gender identity	0.98

Table 3: Natural language interpretations of SAE neurons trained on gender-related free text response embeddings and their associated fidelity scores (F1).

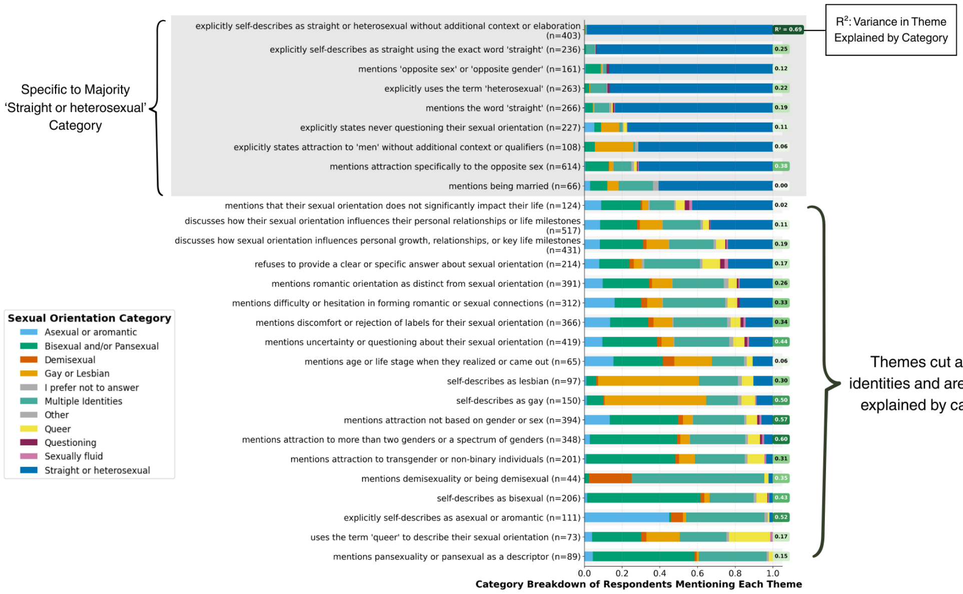

Figure 4: **Sexual orientation-related themes differ between heterosexual and non-heterosexual groups**. Each row represents an interpretable theme learned from a sparse autoencoder trained on free text sexual orientation responses. Bars indicate the proportion of participants in each sexual orientation category whose responses aligned with that theme. Themes are sorted by the proportion of straight or heterosexual identity respondents.

are rarely captured in standard sexual orientation labels, but emerge naturally in free-text descriptions. Compared to the gender themes in Figure 3, Figure 4 shows fewer themes dominated by the majority group, 'Straight or heterosexual'.

#	Interpretation	F1 Fidelity
1	refuses to provide a clear or specific answer about sexual orientation	0.64
2	explicitly states attraction to ’men’ without additional context or qualifiers	0.91
3	mentions demisexuality or being demisexual	0.89
4	mentions being married	0.91
5	explicitly self-describes as asexual or aromantic	0.99
6	explicitly self-describes as straight or heterosexual without additional context or elaboration	0.80
7	explicitly self-describes as straight using the exact word ’straight’	0.99
8	discusses how sexual orientation influences personal growth, relationships, or key life milestones	0.68
9	uses the term ’queer’ to describe their sexual orientation	0.71
10	explicitly states never questioning their sexual orientation	0.85
11	mentions uncertainty or questioning about their sexual orientation	0.84
12	mentions that their sexual orientation does not significantly impact their life	0.91
13	mentions ’opposite sex’ or ’opposite gender’	0.78
14	mentions attraction to more than two genders or a spectrum of genders	0.82
15	mentions attraction specifically to the opposite sex	0.55
16	mentions difficulty or hesitation in forming romantic or sexual connections	0.65
17	self-describes as gay	0.84
18	mentions romantic orientation as distinct from sexual orientation	0.72
19	uses the phrase ’I would describe’	0.81
20	mentions how sexual orientation influences their interactions or relationships with others	0.45
21	mentions the word ’straight’	0.50
22	mentions age or life stage when they realized or came out	0.80
23	discusses how their sexual orientation influences their personal relationships or life milestones	0.51
24	mentions attraction to transgender or non-binary individuals	0.87
25	mentions religion or God in relation to their orientation	0.21
26	self-describes as lesbian	0.90
27	mentions pansexuality or pansexual as a descriptor	1.00
28	self-describes as bisexual	0.89
29	mentions attraction not based on gender or sex	0.81
30	mentions discomfort or rejection of labels for their sexual orientation	0.73
31	explicitly uses the term ’heterosexual’	0.96
32	uses a single word or very brief phrase to describe sexual orientation without elaboration	0.73

Table 4: Natural language interpretations of sexual orientation–related SAE neurons and their associated fidelity scores (F1).