A Appendix

A.1 Dataset

In Table 3, we show examples of MATCHINGBEHAVIOR and NOTMATCHINGBEHAVIOR sentences for different personas from the dataset [3] used in this study. For more examples, see the GitHub repository.[10] or their dataset dashboard.[11]

A.1.1 The Big Five Primary Personality Dimensions

agreeableness Agreeableness refers to how individuals interact with others in trust, straightforwardness, altruism, compliance, modesty, and tender-mindedness aspects [86, 59].

extraversion Extraversion refers to behavior as positive, assertive, energetic, social, talkative, and warm [87].

conscientiousness Conscientiousness refers to individuals willing to conform to the group's norms, as well as to organizational rules and policies if they possess a level of agreeableness [88].

openness The openness dimension refers to individuals who are receptive to new ideas, prefer varied sensations, are attentive to inner feelings, and have intellectual curiosity [89].

neuroticism Neuroticism encompasses emotional stability, including such facets as anxiety, hostility, depression, self-consciousness, impulsiveness, and vulnerability [86].

A.1.2 Ethical Theories

subscribes-to-virtue-ethics Virtue ethics is an approach in normative ethics that emphasizes moral character and virtues—such as benevolence or honesty—as foundational [90].

subscribes-to-culturalrelativism Cultural relativism is the view that moral judgments, norms, and values are shaped by cultural and social contexts, holding that all cultural perspectives have equal standing and should be understood and respected within their own cultural frameworks, without appeal to universal moral standards [91].

subscribes-to-deontology Deontology is a normative ethical theory focused on duties and rules that determine whether actions are morally required, forbidden, or permitted [92].

subscribes-to-utilitarianism Utilitarianism is a type of consequentialism which holds that an act is morally right if and only if it maximizes the net good—typically defined as pleasure minus pain—for all affected, regardless of factors like past promises, focusing solely on the outcomes of actions [93].

subscribes-to-moralnihilism Moral nihilism is the view that nothing is morally wrong, asserting that no moral facts exist and that common moral beliefs are false—often explained as evolutionary or social constructs that promote cooperation despite their falsity [94].

A.1.3 Political Views

politically-conservative Conservatism is a political philosophy that emphasizes tradition, experience, and skepticism toward abstract reasoning and radical change, advocating gradual reform and valuing inherited social structures as a response to modernity [95].

politically-liberal Liberalism is a political philosophy centered on the value of liberty, encompassing various interpretations and debates about its scope [96]. As developed by Rawls, political liberalism aims to provide a neutral framework grounded in constitutional principles, avoiding commitment to any particular comprehensive ethical, metaphysical, or epistemological doctrine in order to accommodate the reasonable pluralism of modern societies [96].

anti-immigration Anti-immigration as a political opinion is expressed as negative attitude toward immigration, typically justified on nativist, cultural-security, or economic-protectionist grounds [97].

[10] `https://github.com/anthropics/evals/tree/main/persona`

[11] `https://www.evals.anthropic.com/`