Identity by Design? Evaluating Gender Conditioning in LLM-Generated Agent Identity Profiles

Mattia Rampazzo[1,2,†], Saba Ghanbari Haez[2,3,†], Patrizio Bellan[2,*,†], Simone Magnolini[2], Leonardo Sanna[2] and Mauro Dragoni[2]

[1]*University of Trento, Trento, Italy*
[2]*Fondazione Bruno Kessler, Povo, Italy*
[3]*Free University of Bozen-Bolzano, Bolzano, Italy*

Abstract

In multi-agent reasoning frameworks powered by large language models, agent roles are often instantiated through identity descriptions that condition their behavior. This paper investigates whether and how the gender assigned to the agent responsible for defining role-specific identity profiles affects the linguistic identity, sentiment, and gender expression of downstream agents. We introduce an extensive corpus of agent identity descriptions generated under controlled combinations of frameworks, roles, models, and gender conditions. Through quantitative and qualitative linguistic analysis, we observe a consistent skew toward female identity across models and roles when gender is unspecified, along with varying degrees of polarity and subjectivity depending on the description framework. Notably, cognitively-oriented frameworks suppress affective expression, while trait-based frameworks amplify gender alignment. These results reveal that identity conditioning is not solely determined by prompt parameters, but emerges through a layered interaction of model priors, framework semantics, and role-specific expressive constraints.

Keywords

Agent Identity Description Framework, Gender Bias Detection, Synthetic Personas, Gender-Conditioned Prompting for Identity Descriptions in LLMs, Prompt Conditioning, Bias in LLMs

1. Introduction

Large language models (LLMs) are increasingly deployed to simulate virtual agents in interactive systems across education, healthcare[1], collaborative decision making, and social computing. These agents are embedded within task-specific contexts and prompted to act as mentors, advisors, or collaborators by means of detailed identity descriptions—texts that encode personality traits, cognitive styles, or motivational dispositions. Although synthetic, such descriptions function as identity proxies that frame the agent as a psychologically grounded persona capable of perspective-taking and goal-directed reasoning [2].

The *Pool of Experts* (PoE)[3] architecture uses a distinctive approach to agent design, where identity descriptions are generated on the basis of psychological and behavioral science, to create role-specialized agents. A central feature of PoE is that these personality-rich agent prompts are not handcrafted but produced by a dedicated LLM agent, i.e. the *Psychologist agent*, whose role is to instantiate identity-grounded profiles for the rest of the set, including the *project manager*, *expert agents*, and *final decision maker*. This raises open questions about how the generation of such profiles is shaped by upstream prompt conditions, particularly the framing of the Psychologist itself.

We leverage PoE as a controlled generative environment to examine whether prompt-level identity framing of the Psychologist affects the linguistic and stylistic properties of the profiles it produces. Concretely, we manipulate the Psychologist's gender as male, female, non-binary, or unconstrained, and

Identity-Aware AI workshop at 28th European Conference on Artificial Intelligence, October 25, 2025, Bologna, Italy

*Corresponding author.
†These authors contributed equally.
pbellan@fbk.eu (P. Bellan)
0009-0003-3360-0424 (M. Rampazzo); 0000-0003-4261-2972 (S. G. Haez); 0000-0002-2971-1872 (P. Bellan); 0000-0003-0170-3472 (S. Magnolini); 0000-0003-3021-6606 (L. Sanna); 0000-0003-0380-6571 (M. Dragoni)

we ask whether this framing influences the text used to define other agents. In this, every description pertains to a top-tier expert function rather than to casual or novice personas.

To study this question empirically, we constructed a large-scale collection of role descriptions generated under controlled variation. Thirteen psychologically grounded identity description frameworks were combined with five task domains. Within each configuration, the Psychologist agent generated role descriptions for the rest of the ensemble, tailoring them to the contextual demands of the dataset and framework. The procedure was repeated across thirteen base language models, yielding thousands of descriptions whose variation is structured by four factors: the Psychologist's gender constraint, the identity description framework, the base language model, and the task domain.

To evaluate whether gender conditioning of the prompt-giver leaves measurable traces in these expert profiles, we analyze each description along three axes: polarity as an index of emotional valence, subjectivity as a measure of evaluative stance versus factual tone, and a discrete sentiment label to capture categorical affect. These metrics allow us to probe tonal and stylistic shifts associated with the Psychologist's gender, and to determine whether such shifts are consistent across roles, frameworks, tasks, and models.

The central question guiding this study is whether gender conditioning of the identity-generating agent influences the content and tone of the role descriptions it produces.

> **RQ** *Do LLMs produce systematically different identity descriptions when gender is varied in the prompt — and how are these effects modulated by identity description frameworks, and language model?*

This research question arises from the observation that identity descriptions—used to define virtual agents, e.g., in multi-role systems—are often treated as neutral artifacts, yet they are themselves generated by another model whose prompt-based framing can bias the textual outcome.

The analyses we conducted revealed consistent effects across multiple dimensions. Gender conditioning influences the emotional tone of the text, the use of subjective versus objective language, and the distribution of adjectives that convey traits such as assertiveness, empathy, or decisiveness. In many cases, the perceived gender of the generated description aligns with the constraint applied to the Psychologist, even when the target role is not gendered. By highlighting these subtle, yet systematic effects, our work contributes to a deeper understanding of how identity framing in prompts can shape the stylistic and semantic qualities of LLM-generated text. We make the full corpus of identity descriptions publicly available at github.com/patriziobellan86/Identity-by-Design-Evaluating-Gender-Conditioning-in-LLM-Generated-Agent-Identity-Profiles.

The paper is structured as follows: Section 2 reviews prior work on agent personality and bias in language generation. Section 3 introduces the Pool of Experts (PoE) architecture and the identity description generation pipeline. Section 4 presents our experimental design and describes the linguistic analysis procedures. Section 5 presents results and discusses the main findings. Section 6 concludes the paper and outlines directions for future research.

2. Related Work

Agent Personality Research on conversational agents has long drawn on psychology, particularly personality modeling frameworks such as the Big Five traits [4, 5, 6]. Early work such as [7] directly embedded these traits into conversational agent design, paving the way for subsequent studies that emphasize personalization, reliability, and richer user engagement. Within healthcare, for example, [8, 9] demonstrate how familiar personas can strengthen trust and empathy, while [10] show that tailoring responses to user personality improves therapeutic outcomes.

Outside healthcare, broader strategies have been explored for role-driven behavior in LLMs. For instance, [11] employ prompt-based methods to enhance contextual awareness, whereas [12] demonstrate that embedding personality traits directly into LLMs enables more flexible and diverse conversational styles. Expanding on personalization, [13] investigate preference evaluation, identifying limitations of simplistic persona construction, while [14] enrich dialogue quality by integrating structured persona data.

Another active line of research addresses scaling persona diversity. [15] contribute methods for building more realistic and varied personas, and [16] introduce the Synthetic-Persona-Chat dataset through persona-aligned dialogue generation. More recently, scholars have examined whether LLMs maintain stable trait-like or identity-driven behaviors when prompted with synthetic personas. For example, [17] find consistent expression of Big Five traits with implications for bias and downstream task performance, while [18] analyze how models internalize social identities and reflect them in political or ethical stances. Complementing this, [19] extend identity integration to human annotators by embedding sociodemographic attributes, though with limited predictive benefits.

Taken together, this body of work highlights the expressive range of persona-based modeling in LLMs, yet its focus remains largely tied to narrow psychological frameworks or specific downstream tasks. Our contribution expands this scope by systematically comparing multiple psychologically grounded description frameworks and testing their influence across heterogeneous reasoning tasks. Extending this line of inquiry, we further manipulate the identity of the profile-generating agent itself (the "Psychologist"), testing whether upstream gender cues propagate into the role descriptions of other expert agents. By combining 13 psychological description frameworks, 5 reasoning tasks, and 13 LLM families, we provide a comprehensive evaluation to date of how prompt-giver identity shapes persona construction.

Profile Definition and Persona Generation Recent studies have explored how LLMs can define and generate character profiles to support more diverse agent behaviors. [20] introduce *Persona Hub*, a large repository of synthetic personas designed to steer LLM outputs toward distinctive styles. Building on this, [21] fine-tune models with persona-driven corpora, enabling generalization across arbitrary traits so that models can role-play characters with specific personalities or backgrounds. Such structured profile definitions expand an LLM's ability to assume diverse identities.

A parallel development is the use of LLMs themselves for persona-aligned content generation. [16] propose a Generator–Critic pipeline to create high-quality persona-aligned dialogues, illustrating how structured prompt chains can scale synthetic identity creation. Similarly, [22] show that automatically generated expert roles improve reasoning diversity and factual accuracy. These works demonstrate the potential of structured persona generation to enrich LLM interactions, while also raising questions about how authorship conditions outcomes.

Our work directly addresses this gap by situating profile authorship within a Pool of Experts architecture, where a Psychologist Agent is responsible for generating all other role profiles. We uniquely test how the Psychologist's gender framing conditions persona generation, making profile authorship itself a locus of identity bias.

Bias and Gender Effects in LLMs Bias in LLM outputs has become a central concern, with numerous studies showing that models reproduce human-like stereotypes across gender, race, and other identities [23]. For example, [24] report that women are more often described with communal traits and men with agentic traits, while [25] uncover gendered occupational associations. [26] further demonstrate that even advanced systems such as GPT-4 and Claude exhibit gender bias in job interview scenarios. Beyond gender, [23] show that LLMs display ingroup favoritism and outgroup derogation, pointing to deep-seated social identity biases. Although mitigation strategies such as curated training data and bias-aware fine-tuning show promise, addressing these issues remains a persistent challenge.

An emerging line of research reveals that bias can arise not only in outputs but also in the persona generation process itself [18, 27]. Studies show that when LLMs invent role descriptions, they often embed stereotypes—for example, framing female teachers as nurturing and male teachers as authoritative [28]. Such upstream bias is underexplored but critical, as it can shape all downstream simulations. Our work contributes to this direction by testing whether the gender framing of a Psychologist prompt-giver influences the profiles it creates, showing that identity conditioning leaves measurable stylistic traces.

This highlights the need for identity-aware prompting strategies to ensure that synthetic personas enrich LLM interactions without reinforcing stereotypes.

Unlike prior studies that analyze bias primarily in outputs or static personas, our study investigates bias at the point of profile creation. We show that gender framing of the Psychologist Agent leaves measurable traces in tone, sentiment, and perceived identity of expert-level profiles (Project Manager, Expert Agents, Final Decision Maker). This upstream perspective reveals how stereotypes can become embedded into the very scaffolding of multi-agent systems, rather than only surfacing in their downstream behaviors.

3. The Pool of Experts Framework

PoE [3] is a prompting architecture designed to emulate interdisciplinary reasoning through the orchestration of multiple role-specialized agents instantiated from a shared LLM. Unlike most prompting pipelines, which rely on monolithic or loosely contextualized prompts, PoE assigns each agent a distinct and explicitly defined identity. This is not merely role-labeling: each agent operates under a psychologically grounded identity profile that conditions its behavior throughout the reasoning process. This modular structure supports interpretability, behavioral diversity, and dynamic specialization, while remaining compatible with zero-shot LLM deployment.

At the start of the PoE inizialization, an LLM is tasked to generate the identity of the **Psychologist Agent** (PA). This agent is responsible for producing the identity descriptions of all other agents. This initial step may be guided by a selected *identity description framework*, such as the Big Five Personality Traits [29], which provides structured psychological principles for shaping the PA's behavior and narrative tone. The PA is itself an LLM instance, prompted to adopt the role of a psychologist. Once instantiated, it proceeds to generate textual identity profiles for each downstream agent. Each profile encodes a set of cognitive, motivational, and communicative dispositions. For example, when instructed to describe a Project Manager agent according to the MBTI [30] framework, the PA might produce a structured paragraph emphasizing planning, organization, and interpersonal awareness. The generated identity description is then injected into the system prompt of the corresponding agent, effectively shaping its reasoning and expression style.

The agent hierarchy constructed by PoE follows a consistent pattern. After the PA is instantiated and generates the required personality descriptions, the system proceeds to instantiate a **Project Manager** agent (PM), whose role is to interpret the task and identify relevant domains of expertise. Based on this analysis, the PA then produces tailored identity profiles for each **Expert Agent** (EA), ensuring alignment with their respective fields (e.g., sociologist, linguistic expert) and the task context. Each EA is instantiated with the assigned identity and generates an independent, structured response. Finally, the PA also produces the identity profile for the **Final Decision Maker** (FDM), an agent tasked with synthesizing the EAs responses into a coherent and justified final answer.

Although all agents are derived from the same base LLM, their behavioral differentiation emerges from the identity prompts generated by the PA. This method allows PoE to simulate cognitive diversity and role specialization without altering the underlying model weights. The psychological framing of each agent serves as a behavioral scaffold, modulating how the model interprets and responds to information. The identity profile of each agent also acts as an inductive bias, shaping the model's default assumptions, expectations, and expressive tendencies.

4. Experimental Design

In the present study, we repurpose the Pool of Experts (PoE) not as a tool for evaluating task performance, but as a controlled environment for probing the linguistic consequences of identity construction. Our focus is on the generative behavior of the PA itself: specifically, we investigate how gender conditioning affects the language it uses to describe other agents. To this end, we systematically vary

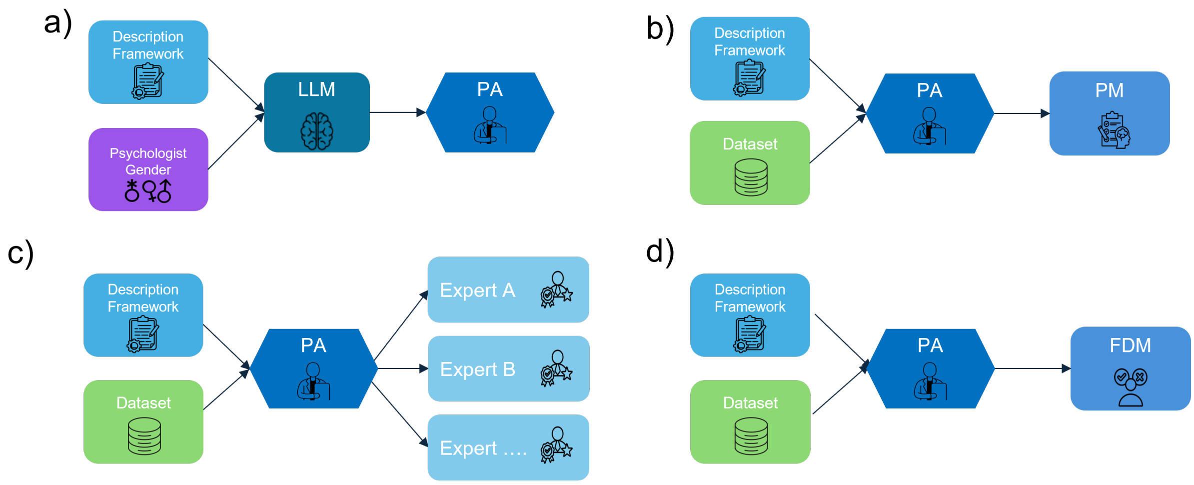

Figure 1: Experimental pipeline.

three experimental dimensions: (i) the gender assigned to the PA, (ii) the identity description framework guiding the generation, and (iii) the backbone LLM employed.

Gender assigned to the Psychologist Agent. The generation of this agent is conditioned along four categories: *unconstrained* (no gender specified), *male*, *female*, or *non-binary*. Listing 1 shows the system and user prompt templates used to instantiate the PA under each gender condition.

Listing 1: System and User prompt template used to generate the PA identity profile.

```
SYSTEM: You are a psychologist, a highly skilled and knowledgeable expert in your
    field. Your task is to create a detailed and authentic description of a {gender
     constraint} **Psychologist** who is responsible for generating description of
    a person. Use the description framework {identity description framework} as a
    guide to describe the Psychologist.
USER: Now it's your turn. Create a {gender constraint} **Psychologist** description
     that strictly follows the description framework {identity description
    framework}.
```

Identity Description Framework. For the second, we draw on thirteen psychologically inspired identity description frameworks that reflect diverse theoretical approaches to modeling personality, cognition, behavior, and user interaction. Trait-oriented models such as the *Big Five Personality Traits* [29] and the *Myers-Briggs Type Indicator* [30] capture relatively stable personality characteristics, while psychodynamic and developmental perspectives such as *Freudian Psychoanalysis* [31] and *Erikson's Psychosocial Stages* [32] emphasize unconscious processes and lifespan development. Cognitive theories, including *Cognitive Behavioral Theory* [33], *Cognitive Load Theory* [34], and *Dual-Process Theory* [35], focus on how individuals acquire, manage, and use information to make decisions. Frameworks such as *Social Cognitive Theory* [36] and *Flow Theory* [37] highlight the influence of motivation, environment, and experiential states on behavior. To complement these perspectives, we also consider user-centered approaches including *User Design Persona* [38], *User-Centered Design* [39], and *Mental Models* [40], which foreground human–system interaction and usability. Finally, we include the *Enneagram of Personality Traits* [41], a non-scientific yet widely used typology in organizational and business contexts, where it serves as a tool for interpreting interpersonal dynamics.

Backbone Models. Identity profile generation is carried out using thirteen LLMs: `claude-3.7-sonnet`, `claude-sonnet-4`, `gemma3-12b`, `gemma-3-27b-it`, `gemma3-4b`, `llama-3.1-8b-instruct`, `llama-3.2-3b-instruct`, `llama-3.3-70b-instruct`, `llama-4-scout`, `mistral-nemo`, `nova-micro-v1`, `qwen3-32b`, and `qwen3-8b`. All models were queried via the OpenRouter service[1]. Generations were performed with fixed decoding parameters and deterministic seeds to ensure reproducibility. For each model, we generated identity descriptions across the full cross-product of gender conditions, identity description frameworks, and PoE agent roles. By holding the framework and model constant while varying only the PA's gender, we obtain a rich corpus of role descriptions that enables systematic analysis of stylistic, affective, and identity-related textual properties.

Datasets. To provide variability in the identity descriptions, we generate profiles with reference to five representative datasets. *Social Support* [42] emphasizes empathy and interpersonal understanding, *CommonSenseQA* [43] targets everyday commonsense inference, *StrategyQA* [44] requires multi-hop strategic reasoning, *Social IQa* [45] captures intuitive judgments about social dynamics, and *Last Letter Concat* [46] provides a synthetic benchmark for symbolic manipulation. Each dataset serves as a task-conditioning signal, playing the same role, such as the PM, which yields different identity descriptions in *StrategyQA* and *Social IQa*, even under identical framework and gender conditions. For each combination of conditions, identity profiles are generated for all roles in the PoE hierarchy.

Experimental Pipeline The experimental pipeline is as follows: first, the identity of the PA is generated (Figure 1*a*); then, the PA generates the identity for the PM agent (Figure 1*b*), for the EAs (Figure 1*c*), and for the FDM (Figure 1*d*).

Importantly, the number of EAs may vary, as PoE dynamically recruits agents based on task demands. This variability depends on the framework, the dataset, and the LLM. However, the architectural dynamics related to expert selection fall outside the scope of this paper.

After generating the identity profiles, we performed an automatic classification of the perceived gender of each profile. To this end, we adopted an LLM-based strategy, using `gpt-4o-mini` to assign each profile to one of four categories: *male*, *female*, *non-binary*, or *uncertain*. We opted for an LLM-based approach rather than a lexicon-based or rule-driven method because the generated texts often encode gender implicitly through stylistic and semantic cues rather than through explicit lexical markers, making traditional heuristics insufficiently reliable. We acknowledge the limitation posed by the methodological circularity of employing LLMs both for profile generation and for their subsequent evaluation, as this setup may amplify model-specific biases. In future studies, we will work to mitigate this issue by conducting controlled experiments with multi-annotator human ratings to assess and calibrate the LLM-based gender guesser.

Linguistic analyses. To assess how gender conditioning influences the generated profiles, we conduct a comprehensive linguistic analysis of the corpus. We examine several stylistic and semantic dimensions that capture both surface-level variation and deeper narrative framing. Our analysis focuses on four key aspects: classified gender, sentiment polarity, subjectivity, and the lexical richness of adjectives[2]. Sentiment analysis provides a measure of affective tone in the text. Within this, we extract two complementary metrics: polarity, which reflects the emotional valence of a description (ranging from negative to positive), and subjectivity, which indicates the extent to which a passage conveys opinions or internal states as opposed to objective information. These features allow us to quantify whether different gender framings result in more emotionally expressive or more neutral agent portrayals.

Adjective Diversity and Coverage by Gender. To investigate whether gender framing influences lexical style, we focused specifically on *adjectives* in the generated role descriptions. Adjectives are a key marker of descriptive richness, making them suitable for detecting subtle stylistic variation and potential bias. From the extracted adjective sets, we computed three complementary metrics for each gender and dataset: (i) the *adjective rate*, i.e., the proportion of descriptions in which an adjective

[1] https://openrouter.ai
[2] All sentiment, polarity, and subjectivity scores were obtained using the `TextBlob` library https://textblob.readthedocs.io

appears; (ii) the *number of unique adjectives per description* (avg), which reflects local richness; and (iii) the *vocabulary coverage* (cov), defined as the percentage of the total adjective vocabulary used by each group. We also calculated a normalized *rate of unique adjectives per description* (rate_unique) to account for differences in dataset size. These measures together provide a structured view of how much descriptive variety different gender framings allow or restrict.

5. Discussion

Table 1

Linguistic analysis of identity descriptions grouped by framework. The table reports the inferred gender distributions (male, female, non-binary, uncertain), sentiment distribution (positive, neutral, negative), and average polarity and subjectivity scores. Results are aggregated across all roles, datasets, and models. Frameworks are ordered within each PA-gender condition.

		Gender Distribution				Sentiment			Linguistic	
	framework	**male**	**female**	**non-binary**	**uncertain**	**positive**	**neutral**	**negative**	**polarity**	**subjectivity**
global	Big Five Personality Traits	0.319	0.522	0.078	0.081	0.788	0.189	0.000	0.159	0.517
	Cognitive Load Theory	0.270	0.546	0.065	0.119	0.355	0.623	0.005	0.086	0.449
	Enneagram of Personality Traits	0.345	0.455	0.102	0.098	0.763	0.218	0.002	0.142	0.500
	Erikson’s Psychosocial Stages	0.281	0.539	0.095	0.085	0.710	0.265	0.003	0.141	0.455
	Flow Theory	0.196	0.633	0.083	0.088	0.682	0.294	0.002	0.134	0.478
	Mental Models	0.245	0.560	0.073	0.121	0.366	0.609	0.003	0.087	0.425
	User Design Persona	0.230	0.568	0.104	0.097	0.513	0.455	0.005	0.111	0.433
	User-Centered Design	0.195	0.570	0.097	0.138	0.598	0.376	0.004	0.124	0.444
female	Big Five Personality Traits	0.475	0.430	0.028	0.067	0.756	0.218	0.000	0.160	0.519
	Cognitive Load Theory	0.439	0.497	0.014	0.049	0.385	0.597	0.003	0.090	0.457
	Enneagram of Personality Traits	0.566	0.352	0.018	0.064	0.754	0.232	0.002	0.140	0.500
	Erikson’s Psychosocial Stages	0.432	0.505	0.016	0.047	0.693	0.285	0.001	0.134	0.448
	Flow Theory	0.311	0.623	0.006	0.059	0.699	0.274	0.004	0.130	0.475
	Mental Models	0.368	0.550	0.020	0.063	0.342	0.642	0.005	0.088	0.432
	User Design Persona	0.362	0.558	0.022	0.058	0.494	0.475	0.004	0.110	0.434
	User-Centered Design	0.286	0.615	0.027	0.072	0.620	0.353	0.005	0.131	0.447
male	Big Five Personality Traits	0.232	0.705	0.016	0.047	0.783	0.192	0.000	0.153	0.512
	Cognitive Load Theory	0.226	0.669	0.016	0.089	0.378	0.590	0.005	0.089	0.457
	Enneagram of Personality Traits	0.327	0.586	0.020	0.067	0.747	0.238	0.000	0.138	0.497
	Erikson’s Psychosocial Stages	0.255	0.680	0.008	0.057	0.699	0.281	0.003	0.142	0.449
	Flow Theory	0.188	0.742	0.013	0.057	0.668	0.310	0.001	0.134	0.477
	Mental Models	0.286	0.577	0.020	0.116	0.353	0.614	0.003	0.084	0.426
	User Design Persona	0.199	0.729	0.010	0.062	0.505	0.464	0.010	0.109	0.433
	User-Centered Design	0.291	0.571	0.022	0.116	0.544	0.430	0.004	0.115	0.444
non-binary	Big Five Personality Traits	0.082	0.515	0.245	0.158	0.827	0.147	0.000	0.161	0.518
	Cognitive Load Theory	0.117	0.459	0.209	0.216	0.321	0.662	0.004	0.085	0.440
	Enneagram of Personality Traits	0.089	0.404	0.343	0.164	0.803	0.178	0.001	0.149	0.505
	Erikson’s Psychosocial Stages	0.102	0.421	0.323	0.155	0.739	0.226	0.004	0.148	0.463
	Flow Theory	0.067	0.506	0.266	0.162	0.714	0.254	0.005	0.131	0.480
	Mental Models	0.108	0.485	0.224	0.183	0.463	0.509	0.001	0.101	0.431
	User Design Persona	0.052	0.424	0.350	0.174	0.512	0.448	0.002	0.109	0.433
	User-Centered Design	0.049	0.448	0.287	0.215	0.618	0.355	0.006	0.126	0.440
unconstrained	Big Five Personality Traits	0.479	0.442	0.027	0.052	0.784	0.199	0.001	0.162	0.521
	Cognitive Load Theory	0.290	0.561	0.027	0.123	0.337	0.640	0.008	0.079	0.441
	Enneagram of Personality Traits	0.388	0.481	0.033	0.098	0.748	0.222	0.002	0.142	0.500
	Erikson’s Psychosocial Stages	0.347	0.559	0.018	0.076	0.706	0.269	0.004	0.142	0.460
	Flow Theory	0.230	0.671	0.034	0.066	0.644	0.339	0.000	0.142	0.481
	Mental Models	0.226	0.630	0.023	0.121	0.304	0.673	0.003	0.076	0.413
	User Design Persona	0.309	0.566	0.031	0.094	0.540	0.431	0.004	0.115	0.432
	User-Centered Design	0.166	0.652	0.039	0.143	0.607	0.370	0.003	0.125	0.446

Our experimental pipeline generated a total of 42,045 identity profiles, encompassing all combinations of gender conditions, identity description frameworks, LLMs, and task domains. The distribution across gender conditions is remarkably balanced: 10,513 profiles were classified on a female identity, 10,245 on male, 10,672 on non-binary, and 10,615 were generated without any gender constraint. This symmetry reflects the controlled nature of our design, ensuring comparability across experimental factors.We report examples of the generated identity profiles in Appendix A.

In Table 6 and Table 7 we report the most significant and exemplary results from our analysis. Each table aggregates the outputs by framework, model, and agent role, respectively, and reports metrics along three analytical axes: gender distribution (inferred gender proportions), sentiment (positive/neu-

Table 2
Linguistic analysis of identity descriptions grouped by model. The table reports the inferred gender distributions, sentiment distribution, and average polarity and subjectivity scores. Results are aggregated across all roles, datasets, and frameworks. Models are ordered within each PA-gender condition.

		Gender Distribution				Sentiment			Linguistic	
	model	**male**	**female**	**non-binary**	**uncertain**	**positive**	**neutral**	**negative**	**polarity**	**subjectivity**
global	claude 3.7	0.125	0.764	0.072	0.038	0.641	0.358	0.001	0.119	0.423
	claude 4	0.428	0.437	0.110	0.024	0.679	0.321	0.000	0.118	0.410
	gemma-3-27b-it	0.366	0.543	0.076	0.016	0.459	0.540	0.000	0.096	0.474
	llama 3.1 8b	0.100	0.711	0.086	0.102	0.803	0.196	0.000	0.167	0.505
	llama 3.3 70b	0.098	0.684	0.075	0.144	0.848	0.152	0.000	0.168	0.520
	llama-4-scout	0.150	0.657	0.066	0.127	0.800	0.171	0.003	0.158	0.500
	qwen3-32b	0.164	0.461	0.174	0.201	0.253	0.675	0.010	0.065	0.367
	qwen3-8b	0.149	0.347	0.048	0.455	0.208	0.503	0.007	0.048	0.392
female	claude 3.7	0.222	0.734	0.025	0.020	0.599	0.401	0.000	0.114	0.421
	claude 4	0.712	0.265	0.024	0.000	0.664	0.336	0.000	0.116	0.406
	gemma-3-27b-it	0.747	0.252	0.001	0.000	0.420	0.580	0.000	0.092	0.473
	llama 3.1 8b	0.101	0.843	0.008	0.047	0.811	0.189	0.000	0.169	0.513
	llama 3.3 70b	0.177	0.735	0.023	0.065	0.850	0.150	0.000	0.170	0.520
	llama-4-scout	0.081	0.859	0.003	0.057	0.869	0.122	0.002	0.168	0.508
	qwen3-32b	0.249	0.496	0.075	0.180	0.253	0.663	0.010	0.064	0.356
	qwen3-8b	0.128	0.453	0.010	0.409	0.211	0.521	0.007	0.047	0.393
male	claude 3.7	0.091	0.901	0.003	0.005	0.649	0.350	0.001	0.120	0.429
	claude 4	0.082	0.917	0.002	0.000	0.695	0.305	0.000	0.119	0.413
	gemma-3-27b-it	0.187	0.811	0.001	0.000	0.377	0.623	0.000	0.088	0.474
	llama 3.1 8b	0.228	0.701	0.023	0.048	0.804	0.195	0.001	0.165	0.512
	llama 3.3 70b	0.156	0.702	0.030	0.112	0.844	0.156	0.000	0.167	0.523
	llama-4-scout	0.434	0.404	0.020	0.142	0.723	0.230	0.007	0.141	0.482
	qwen3-32b	0.150	0.624	0.037	0.189	0.223	0.712	0.012	0.061	0.362
	qwen3-8b	0.234	0.290	0.029	0.447	0.197	0.508	0.007	0.044	0.391
non-binary	claude 3.7	0.036	0.672	0.207	0.085	0.662	0.337	0.001	0.120	0.417
	claude 4	0.169	0.342	0.396	0.094	0.719	0.281	0.000	0.123	0.417
	gemma-3-27b-it	0.093	0.593	0.289	0.025	0.533	0.467	0.000	0.102	0.480
	llama 3.1 8b	0.011	0.477	0.284	0.229	0.827	0.173	0.000	0.166	0.505
	llama 3.3 70b	0.026	0.567	0.204	0.203	0.859	0.141	0.000	0.173	0.525
	llama-4-scout	0.040	0.445	0.232	0.283	0.738	0.204	0.002	0.159	0.485
	qwen3-32b	0.025	0.232	0.500	0.243	0.299	0.626	0.010	0.072	0.372
	qwen3-8b	0.022	0.284	0.142	0.553	0.190	0.506	0.009	0.040	0.391
unconstrained	claude 3.7	0.154	0.749	0.054	0.043	0.656	0.344	0.000	0.122	0.427
	claude 4	0.754	0.233	0.013	0.001	0.635	0.365	0.000	0.113	0.403
	gemma-3-27b-it	0.435	0.522	0.007	0.036	0.501	0.493	0.000	0.100	0.470
	llama 3.1 8b	0.066	0.824	0.028	0.083	0.773	0.227	0.000	0.166	0.489
	llama 3.3 70b	0.033	0.731	0.040	0.196	0.837	0.163	0.000	0.163	0.513
	llama-4-scout	0.066	0.899	0.008	0.027	0.864	0.133	0.003	0.163	0.521
	qwen3-32b	0.235	0.497	0.077	0.191	0.235	0.697	0.006	0.065	0.376
	qwen3-8b	0.208	0.357	0.018	0.417	0.233	0.476	0.005	0.061	0.393

tral/negative), subjectivity score, and polarity intensity. We grouped the results by conditioning variable (e.g., by model or by role) and aggregated across datasets. This allows us to highlight systematic trends that are otherwise obscured by the large number of individual entries. For example, grouping by role reveals how linguistic subjectivity differs between psychologists and experts, while grouping by model clarifies divergent behaviors across model families such as Claude and LLaMA. These summaries are intended to highlight key patterns and contrasts across conditions[3].

Global Gender Distribution. As a starting point, we examine the inferred gender distribution aggregated across all generation conditions. This provides a high-level view of how gender conditioning influences the overall representation of gendered identity in the resulting profiles. Across identity description frameworks, the generated profiles are predominantly classified as female. For example, in the global rows of Table 6, the Flow Theory framework produces the highest proportion of female classifications, while the remaining frameworks consistently generate slightly above 50% female profiles. Interestingly, all frameworks generate only around 20% male profiles and approximately 10% non-binary

[3]The complete set of results, including all combinations and disaggregated values, is provided in the Appendix.

Table 3
Linguistic analysis of identity descriptions grouped by role. The table reports the inferred gender distributions (male, female, non-binary, uncertain), sentiment distribution (positive, neutral, negative), and average polarity and subjectivity scores. Results are aggregated across all models, datasets, and frameworks. Roles are ordered within each PA-gender condition.

		Gender Distribution				Sentiment			Linguistic	
	role	male	female	non-binary	uncertain	positive	neutral	negative	polarity	subjectivity
global	psychologist	0.295	0.428	0.222	0.055	0.663	0.310	0.002	0.138	0.463
	project-manager	0.289	0.408	0.087	0.216	0.694	0.287	0.002	0.145	0.489
	expert agent	0.272	0.578	0.077	0.073	0.575	0.401	0.003	0.118	0.454
	final decision maker	0.257	0.361	0.094	0.288	0.627	0.340	0.003	0.123	0.498
female	psychologist	0.000	0.983	0.001	0.015	0.666	0.314	0.000	0.140	0.465
	project-manager	0.392	0.417	0.056	0.135	0.702	0.278	0.001	0.147	0.488
	expert agent	0.467	0.476	0.013	0.045	0.573	0.403	0.003	0.117	0.457
	final decision maker	0.371	0.406	0.032	0.191	0.623	0.347	0.005	0.123	0.494
male	psychologist	0.974	0.000	0.000	0.026	0.633	0.342	0.001	0.132	0.458
	project-manager	0.275	0.536	0.043	0.146	0.670	0.305	0.004	0.139	0.481
	expert agent	0.182	0.756	0.011	0.051	0.562	0.413	0.003	0.115	0.455
	final decision maker	0.232	0.477	0.038	0.253	0.593	0.377	0.000	0.118	0.491
non-binary	psychologist	0.015	0.007	0.870	0.108	0.718	0.236	0.005	0.148	0.470
	project-manager	0.133	0.282	0.189	0.395	0.727	0.255	0.001	0.148	0.493
	expert agent	0.105	0.510	0.254	0.131	0.607	0.371	0.002	0.122	0.454
	final decision maker	0.098	0.213	0.257	0.432	0.648	0.311	0.005	0.125	0.507
unconstrain.	psychologist	0.192	0.720	0.017	0.071	0.634	0.347	0.001	0.132	0.458
	project-manager	0.357	0.396	0.059	0.187	0.675	0.312	0.001	0.145	0.492
	expert agent	0.333	0.579	0.024	0.065	0.558	0.418	0.003	0.115	0.451
	final decision maker	0.326	0.350	0.048	0.276	0.646	0.325	0.004	0.126	0.501

identities, regardless of conditioning. Sentiment trends also differ across frameworks. The Big Five, Enneagram of Personality Traits, and Erikson's Psychosocial Stages exhibit strongly positive sentiment, with higher average polarity scores. In contrast, Cognitive Load Theory and Mental Models produce more neutral profiles with lower average polarity values. Subjectivity varies within a narrower band, peaking in the Big Five and reaching its lowest value in Mental Models, as shown in Table 6.

Effect of Gender Conditioning. Analyzing the four gender experimental conditions, we observe that the gender setting of the PA significantly influences both the stylistic tone and the inferred gender attribution of the generated profiles.

When the PA is conditioned as **female**, the resulting descriptions exhibit stronger *female alignment* overall, particularly under frameworks such as Flow Theory and User-Centered Design. In these cases, non-binary and uncertain classifications remain consistently low. Additionally, we observe a reduction in the proportion of profiles classified as *uncertain*, accompanied by a slight increase in those classified as *male*, suggesting that gender conditioning enhances identity specificity. Sentiment trends under the female condition are predominantly positive. Frameworks such as the Big Five, Social Cognitive Theory, and the Enneagram of Personality Traits produce over 75% of positively classified descriptions. However, the highest average polarity and subjectivity scores are observed only in the Big Five framework, highlighting its stronger affective and subjective tone relative to others.

When the Psychologist is set to **male**, an unexpected trend emerges: female alignment increases even further across most frameworks. In several cases, such as Cognitive Load Theory and Cognitive Behavioral Theory, the proportion of profiles classified as female exceeds that observed under the female condition. Sentiment remains predominantly positive across frameworks, with the Big Five and Enneagram of Personality Traits exhibiting the highest rates of positive classifications. However, both polarity and subjectivity scores are slightly lower than those observed in the female condition, suggesting a more emotionally neutral and less personalized tone in the generated profiles.

Under the **non-binary** setting, non-binary identity attribution increases markedly. Several frameworks show non-binary percentages above 30% (e.g., Social Cognitive Theory), while male and female labels remain more balanced. Sentiment remains generally positive, with Big Five and Enneagram of

Personality Traits scoring the higest positive among the experimental categories. Polarity remains relatively stable and comparatively to the other experimental conditions.

In the **unconstrained** setting, where no gender is assigned to the PA, female identity emerges as the dominant classification across most frameworks. Two notable exceptions are the Big Five and Myers-Briggs Type Indicator, which display a slight preference toward male classifications. In this condition, non-binary identities drop dramatically, averaging around 3%. Positive sentiment remains most prominent in the Big Five and Enneagram of Personality Traits frameworks, with the Big Five again exhibiting the highest subjectivity. In contrast, cognitively-oriented frameworks such as Erikson's Psychosocial Stages yield more emotionally neutral profiles, characterized by low polarity scores and a high proportion of neutral classifications.

Overall, all four conditions maintain consistent rankings: *Big Five* and *Enneagram of Personality Traits* remain the most positive and subjective; *Cognitive Load Theory* and *Mental Models* trend neutral and analytic; *Flow Theory* and *User-Centered Design* consistently yield more female-coded outputs. The non-binary setting uniquely supports broader identity diversity across roles.

Model-Level Patterns. Aggregating by model reveals distinct stylistic tendencies (Table 7). The LLaMA-family models consistently produce the warmest descriptions, characterized by the highest levels of positive sentiment, polarity, and subjectivity. In contrast, the Qwen family exhibits a flatter affective profile, generating more emotionally neutral text. The Claude models fall in between these two extremes, offering moderately expressive outputs. In terms of inferred gender distributions, we observe similar trends to those seen across description frameworks: most generated identity profiles are classified as female. However, notable variations emerge across models. Nova-Micro and the Qwen models produce the lowest proportion of female identities, showing a relative preference for male descriptions. Conversely, the LLaMA-family models skew strongly toward female identity generation. Interestingly, a closer examination of the Claude Sonnet models reveals divergent behavior: version 3.7 shows a marked bias toward female identities, while this tendency is substantially attenuated in version 4, indicating a shift toward more balanced gender representations.

Role-Level Patterns. Grouping by role (Table 3) reveals clear stylistic differences across the PoE hierarchy. Descriptions of the Project Manager tend to be the most positively valenced, exhibiting the highest polarity scores. Expert Agents are sligthly less positive. The Final Decision Maker profiles display slightly higher subjectivity while maintaining a comparable level of positive sentiment. Psychologist descriptions are moderately subjective, falling between the Expert and Decision Maker roles in affective tone. Across roles, negative sentiment remains negligible. Notably, under the non-binary Psychologist condition, perceived non-binary identity increases and partially propagates to downstream roles, for instance, Project Manager profiles show 18.9% non-binary classification, and Final Decision Maker profiles reach 25.6%. This trend underscores a key asymmetry: only when the PA is explicitly conditioned as non-binary do downstream roles exhibit meaningful non-binary identity, whereas in all other gender conditions, non-binary representations remain minimal (approximately 3%).

Statistical Analyses. To assess whether these patterns are statistically reliable, we ran statistical tests. For *polarity* and *subjectivity*, we used the Kruskal-Wallis test to compare rank distributions across the levels of each factor. We conducted these tests in three ways: globally across the full corpus, and separately within each Psychologist-gender setting (unconstrained, male, female, non-binary). For *perceived gender* (male/female/non-binary/uncertain) and *sentiment class* (positive/neutral/negative), we treated both as contingency problems and applied Pearson's χ^2 tests of independence at multiple levels of aggregation. When a significant effect was found, we conducted pairwise Wilcoxon rank-sum (Mann–Whitney) tests between all levels of the factor under study (e.g., across models, frameworks, or roles), applying Bonferroni correction to control the family-wise error rate. Similarly, for sentiment and perceived gender, we followed significant χ^2 tests with pairwise χ^2 comparisons between levels of the same factor, again using Bonferroni correction to adjust for multiple comparisons.

Across description frameworks, across base models, and across roles, Kruskal–Wallis tests were highly significant in all configurations, indicating systematic differences in both polarity and subjectivity that do not reduce to sampling noise. χ^2 tests were significant both globally and within each Psychologist-gender condition across all three factors: identity description framework, model, and role. This finding indicates that sentiment and perceived gender distributions vary systematically rather than randomly. After Bonferroni correction, most comparisons remain statistically significant For example, the Chi-squared tests confirm that different description frameworks yield significantly divergent gender distributions in the generated agent profiles, indicating that the stylistic scaffolding imposed by each framework systematically influences perceived identity. In this setting, the *Big Five Personality Traits* framework diverges significantly from nearly all others. It elicits strongly evaluative, trait-laden language that tends to carry gendered connotations.

Adjective Analysis. Table 4 summarizes results across four experimental conditions. In the **unconstrained** dataset, male outputs are clearly more lexically rich (51.7 unique adjectives on average, 86.5% coverage) than female ones (39.1, 83.1%), while non-binary and uncertain outputs show far lower richness and coverage. When the Psychologist is constrained to be **female**, male descriptions become even more elaborate (52.1 avg, 89.4% coverage), whereas female and non-binary outputs lose both richness and breadth. This reveals an asymmetry where female framing enhances male lexical diversity while suppressing other groups. By contrast, under a **male Psychologist**, female outputs gain sharply in richness (47.0 avg, 90.4%) and surpass male ones (42.0 avg, 75.8%), indicating a reversed bias pattern. The **non-binary Psychologist** condition yields the highest averages overall, with male descriptions again leading (53.2 avg), but with uneven coverage (66.0%), suggesting repetition within a narrower adjective space. Across all conditions, uncertain-gender outputs remain the least lexically diverse.

Overall, these analyses show that the gender identity of the profile-generating Psychologist Agent systematically shapes the diversity of adjectives used by downstream role profiles. Importantly, the observed shifts are not uniform: male profiles benefit disproportionately when the Psychologist is female or non-binary, whereas female profiles only gain richness under a male Psychologist. This asymmetric redistribution of adjective usage is a clear signal of **bias in persona construction**, highlighting how upstream identity cues condition not only direct outputs but also the expressive range available to other groups.

In conclusion, the answer to our research question is affirmative but qualified: the gender assigned to the Psychologist Agent does shape the identity and stylistic properties of generated profiles, but its influence is mediated by the identity description framework and model. Non-binary cues propagate most effectively but introduce ambiguity; male and female cues leak asymmetrically; and framework–role–model combinations can amplify or suppress the intended signal. Understanding these interactions is essential for responsible use of identity-aware prompting in LLM-based systems.

6. Conclusion

This study set out to investigate whether, and how, the gender assigned influences the identity, tone, and stance of the generated role descriptions. Our findings reveal a layered interplay between gender conditioning, identity description framework, and agent role. On top of the role scaffold, the identity description frameworks influence tone and stance. People- and trait-oriented frameworks, such as the Big Five, consistently elicit warmer, more subjective, and highly positive descriptions. In contrast, cognitively-oriented frameworks such as Cognitive Load Theory promote neutral phrasing. These findings demonstrate that the chosen identity description frameworks set the expressive bandwidth for identity to be realized.

Gender conditioning introduces a second axis of variation. When the PA's gender is left unspecified, generated profiles tend to skew female across roles and frameworks, suggesting a possible latent female-coded prior embedded within the model weights or the generation schema. Interestingly, this leakage is asymmetric and model-family dependent: male conditioning sometimes fails to override default female-

Table 4
Lexical diversity and vocabulary coverage of identity descriptions, grouped by PA-gender condition and inferred gender. The *avg* column reports the average number of unique adjectives per description, *cov* indicates the percentage of vocabulary coverage, and *rate unique* reflects the proportion of adjectives used only once.

Condition	Gender	avg	cov	rate unique
global	female	42.325	86.841	0.002
	male	49.75	85.023	0.004
	non-binary	41.038	60.091	0.011
	uncertain	25.246	44.455	0.008
female	female	37.004	79.311	0.007
	male	52.068	89.369	0.012
	non-binary	37.486	27.484	0.209
	uncertain	19.062	27.877	0.043
male	female	47.002	90.439	0.007
	male	41.978	75.763	0.016
	non-binary	32.271	25.429	0.208
	uncertain	20.469	29.352	0.037
non-binary	female	45.853	83.742	0.01
	male	53.166	65.977	0.05
	non-binary	42.285	75.879	0.013
	uncertain	30.379	49.69	0.019
unconstrained	female	39.08	83.09	0.007
	male	51.694	86.491	0.015
	non-binary	34.374	31.876	0.117
	uncertain	21.939	35.216	0.028

coded patterns (e.g., Big Five). These results suggest that model priors and framework-specific stylistic constraints can override explicit gender cues, even under tightly controlled prompting. Non-binary identity follows a different pattern. When the PA is conditioned as non-binary, we observe the strongest and most consistent propagation of non-binary alignment to downstream roles. However, it also increases interpretive instability. This indicates that identity cues are not simply passed downstream, but shaped through a process of interaction between gender markers, framework constraints, role semantics, and model-specific expressive preferences. These dynamics result in emergent biases. For example, male-coded identities often align with low-subjectivity, which may unintentionally reinforce social associations between masculinity and objectivity. Across the corpus, such tendencies accumulate: even when prompts are balanced and neutral in form, their interaction with model and framework biases can arise identity–role couplings that reflect familiar sociocultural archetypes.

Model families add a final layer of influence. Certain LLMs, such as Claude 4, favor subjective and positive expression, which strengthens identity alignment, especially when combined with socially expressive frameworks. Conversely, models like Qwen3-32B tend toward neutral or flattened affect, weakening the salience of identity cues and increasing the likelihood of gender drift or ambiguity.

Taken together, these findings highlight a core insight: identity-aware prompting is a system-level phenomenon. The gender assigned to the PA influences the generation of downstream profiles, but not deterministically. Effective identity control emerges only when gender conditioning, identity framework, role semantics, and LLM behavior are aligned. Even in well-calibrated settings, drift, leakage, and model-level priors remain active forces that shape the generative outcome.

These results point to broader design implications. Neutrality cannot be assumed at any point in the generation pipeline. Even well-structured and symmetrical prompts can produce skewed or stereotypical profiles if their constituent parts are not jointly calibrated.

Acknowledgements

We acknowledge the support of the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU. This paper is supported by the TrustAlert project, funded by Fondazione Compagnia San Paolo and Fondazione CDP under the "Artificial Intelligence" call.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly in order to: Grammar and spelling check. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] M. Bolpagni, S. D. Carli, L. Sanna, S. Gabrielli, M. Dragoni, Role-play large language models for short behavior change interventions: An exploratory study on brief action planning, in: R. Bellazzi, J. M. J. Herrero, L. Sacchi, B. Zupan (Eds.), Artificial Intelligence in Medicine - 23rd International Conference, AIME 2025, Pavia, Italy, June 23-26, 2025, Proceedings, Part II, volume 15735 of *Lecture Notes in Computer Science*, Springer, 2025, pp. 46–51. URL: https://doi.org/10.1007/978-3-031-95841-0_9. doi:10.1007/978-3-031-95841-0_9.

[2] J. Z. Pan, S. Razniewski, J. Kalo, S. Singhania, J. Chen, S. Dietze, H. Jabeen, J. Omeliyanenko, W. Zhang, M. Lissandrini, R. Biswas, G. de Melo, A. Bonifati, E. Vakaj, M. Dragoni, D. Graux, Large language models and knowledge graphs: Opportunities and challenges, TGDK 1 (2023) 2:1–2:38. URL: https://doi.org/10.4230/TGDK.1.1.2. doi:10.4230/TGDK.1.1.2.

[3] P. Bellan, S. G. Haez, L. Sanna, S. Magnolini, M. Dragoni, Leveraging multi-agent systems for domain-pertinence query classification in informative chatbots, in: Artificial Intelligence in Medicine, Springer Nature Switzerland, Cham, 2025, pp. 44–54.

[4] S. T. Völkel, Conversational agents with personality, Ph.D. thesis, lmu, 2022.

[5] S. Roccas, Sagiv, et al., The big five personality factors and personal values, Personality and social psychology bulletin 28 (2002) 789–801.

[6] X. Zhao, Y. Deng, W. Wang, H. Cheng, R. Zhang, S.-K. Ng, T.-S. Chua, et al., Exploring the impact of personality traits on conversational recommender systems: A simulation with large language models, arXiv preprint arXiv:2504.12313 (2025).

[7] H. Kim, et al., Designing personalities of conversational agents, in: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, 2019.

[8] A. B. Kocaballi, S. Berkovsky, J. C. Quiroz, L. Laranjo, H. L. Tong, D. Rezazadegan, A. Briatore, E. Coiera, The personalization of conversational agents in health care: systematic review, Journal of medical Internet research 21 (2019) e15360.

[9] Y. Hwang, D. Shin, S. Baek, B. Suh, J. Lee, Applying the persona of user's family member and the doctor to the conversational agents for healthcare, 2021. URL: https://arxiv.org/abs/2109.01729. arXiv:2109.01729.

[10] R. Ahmad, et al., Designing personality-adaptive conversational agents for mental health care, Information Systems Frontiers 24 (2022).

[11] A. Kong, et al., Better zero-shot reasoning with role-play prompting, in: Proceedings of the 2024 Conference of the North American Chapter of ACL: HLT (Volume 1: Long Papers), 2024.

[12] G. Serapio-García, M. Safdari, C. Crepy, L. Sun, S. Fitz, P. Romero, M. Abdulhai, A. Faust, M. Matarić, Personality traits in large language models, 2023. URL: https://arxiv.org/abs/2307.00184. arXiv:2307.00184.

[13] Y. R. Dong, et al., Can LLM be a personalized judge?, in: Findings of the ACL: EMNLP 2024, ACL, 2024.

[14] J. Liu, C. Symons, R. R. Vatsavai, Persona-based conversational ai: State of the art and challenges, 2022. URL: https://arxiv.org/abs/2212.03699. arXiv:2212.03699.

[15] A. Schuller, et al., Generating personas using llms and assessing their viability, in: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, CHI EA '24, ACM, 2024. doi:10.1145/3613905.3650860.

[16] P. Jandaghi, et al., Faithful persona-based conversational dataset generation with llms, in: Proceedings of the 6th Workshop on NLP4ConvAI 2024, ACL, 2024, pp. 114–139.

[17] T. Hu, N. Collier, Quantifying the persona effect in LLM simulations, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 10289–10307. URL: https://aclanthology.org/2024.acl-long.554/. doi:10.18653/v1/2024.acl-long.554.

[18] A. Liu, M. Diab, D. Fried, Evaluating large language model biases in persona-steered generation, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 9832–9850. URL: https://aclanthology.org/2024.findings-acl.586/. doi:10.18653/v1/2024.findings-acl.586.

[19] M. Orlikowski, P. Röttger, P. Cimiano, D. Hovy, The ecological fallacy in annotation: Modeling human label variation goes beyond sociodemographics, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 1017–1029. URL: https://aclanthology.org/2023.acl-short.88/. doi:10.18653/v1/2023.acl-short.88.

[20] T. Ge, X. Chan, X. Wang, D. Yu, H. Mi, D. Yu, Scaling synthetic data creation with 1,000,000,000 personas, arXiv preprint arXiv:2406.20094 (2024).

[21] X. Wang, H. Zhang, T. Ge, W. Yu, D. Yu, D. Yu, Opencharacter: Training customizable role-playing llms with large-scale synthetic personas, arXiv preprint arXiv:2501.15427 (2025).

[22] D. X. Long, et al., Multi-expert prompting improves reliability, safety and usefulness of llms, in: Proceedings of the 2024 Conference on Empirical Methods in NLP, ACL, 2024.

[23] T. Hu, Y. Kyrychenko, S. Rathje, N. Collier, S. van der Linden, J. Roozenbeek, Generative language models exhibit social identity biases, Nature Computational Science 5 (2025) 65–75.

[24] Y. Wan, G. Pu, et al., "kelly is a warm person, joseph is a role model": Gender biases in LLM-generated reference letters, in: Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 3730–3748. URL: https://aclanthology.org/2023.findings-emnlp.243/. doi:10.18653/v1/2023.findings-emnlp.243.

[25] H. Kotek, R. Dockum, D. Sun, Gender bias and stereotypes in large language models, in: Proceedings of The ACM Collective Intelligence Conference, CI '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 12–24. URL: https://doi.org/10.1145/3582269.3615599. doi:10.1145/3582269.3615599.

[26] H. Kong, Y. Ahn, S. Lee, Y. Maeng, Gender bias in llm-generated interview responses, arXiv preprint arXiv:2410.20739 (2024).

[27] B. C. Z. Tan, R. K.-W. Lee, Unmasking implicit bias: Evaluating persona-prompted LLM responses in power-disparate social scenarios, in: Proceedings of the 2025 Conference of NAACL: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Albuquerque, New Mexico, 2025, pp. 1075–1108. URL: https://aclanthology.org/2025.naacl-long.50/. doi:10.18653/v1/2025.naacl-long.50.

[28] S. Gupta, V. Shrivastava, A. Deshpande, A. Kalyan, P. Clark, A. Sabharwal, T. Khot, Bias runs deep: Implicit reasoning biases in persona-assigned LLMs, in: The Twelfth International Conference on Learning Representations, 2024. URL: https://openreview.net/forum?id=kGteeZ18Ir.

[29] S. Roccas, L. Sagiv, S. H. Schwartz, A. Knafo, The big five personality factors and personal values, Personality and Social Psychology Bulletin 28 (2002) 789–801. URL: https://doi.org/10.1177/0146167202289008. doi:10.1177/0146167202289008. arXiv:https://doi.org/10.1177/0146167202289008.

[30] I. Myers, P. Myers, Gifts Differing: Understanding Personality Type, Mobius, 1995. URL: https://books.google.it/books?id=D3V-AAAAMAAJ.
[31] S. Freud, A. Brill, The Interpretation of Dreams, Classics of World Literature, Wordsworth Editions, 1997. URL: https://books.google.it/books?id=-pB1UoFnjZcC.
[32] S. Escalona, <i>childhood and society</i>. erik h. erikson. new york: Norton, 1950. 397 pp. $4.00, Science 113 (1951) 253–253. URL: https://www.science.org/doi/abs/10.1126/science.113.2931.253.a. doi:10.1126/science.113.2931.253.a. arXiv:https://www.science.org/doi/pdf/10.1126/science.113.2931.253.a.
[33] A. Beck, Cognitive Therapy and the Emotional Disorders, International Universities Press, 1976. URL: https://books.google.it/books?id=zndHAAAAMAAJ.
[34] J. Sweller, Cognitive load during problem solving: Effects on learning, Cognitive Science 12 (1988) 257–285. doi:10.1207/s15516709cog1202_4.
[35] P. Wason, J. Evans, Dual processes in reasoning?, Cognition 3 (1974) 141–154. URL: https://www.sciencedirect.com/science/article/pii/0010027774900171. doi:https://doi.org/10.1016/0010-0277(74)90017-1.
[36] G. W. Blackwell, Social learning and imitation. by neal e. miller and john dollard. new haven: Yale university press, 1941. 341pp. $3.50, Social Forces 21 (1942) 256–256. URL: https://doi.org/10.2307/2570586. doi:10.2307/2570586. arXiv:https://academic.oup.com/sf/article-pdf/21/2/256/5861048/21-2-256.pdf.
[37] M. Csikszentmihalyi, Beyond Boredom and Anxiety, Jossey-Bass behavioral science series, Jossey-Bass Publishers, 1975. URL: https://books.google.it/books?id=afdGAAAAMAAJ.
[38] A. Cooper, P. Saffo, The Inmates Are Running the Asylum, Macmillan Publishing Co., Inc., USA, 1999.
[39] D. A. Norman, S. W. Draper, User Centered System Design; New Perspectives on Human-Computer Interaction, L. Erlbaum Associates Inc., USA, 1986.
[40] P. N. Johnson-Laird, Mental models in cognitive science, Cogn. Sci. 4 (1980) 71–115. URL: https://api.semanticscholar.org/CorpusID:7090767.
[41] D. Riso, R. Hudson, Personality Types: Using the Enneagram for Self-Discovery, Houghton Mifflin, 1996. URL: https://books.google.it/books?id=pmdARxbC1SUC.
[42] Z. Wang, D. Jurgens, It‘s going to be okay: Measuring access to support in online communities, in: E. Riloff, D. Chiang, J. Hockenmaier, J. Tsujii (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 33–45. URL: https://aclanthology.org/D18-1004/. doi:10.18653/v1/D18-1004.
[43] A. Talmor, J. Herzig, N. Lourie, J. Berant, Commonsenseqa: A question answering challenge targeting commonsense knowledge, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 4149–4158. URL: https://doi.org/10.18653/v1/n19-1421. doi:10.18653/V1/N19-1421.
[44] M. Geva, D. Khashabi, E. Segal, T. Khot, D. Roth, J. Berant, Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies, Transactions of the Association for Computational Linguistics 9 (2021) 346–361. URL: https://doi.org/10.1162/tacl_a_00370. doi:10.1162/tacl_a_00370. arXiv:https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00370/1924104/tacl_
[45] M. Sap, H. Rashkin, D. Chen, R. Le Bras, Y. Choi, Social IQa: Commonsense reasoning about social interactions, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 4463–4473. URL: https://aclanthology.org/D19-1454/. doi:10.18653/v1/D19-1454.
[46] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, D. Zhou, Chain-

of-thought prompting elicits reasoning in large language models, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022, 2022. URL: http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Appendix

In this appendix, we provide illustrative examples of identity profiles in Appendix A, along with the complete result tables in Appendix B.

A. Examples of Identity Profiles

This section presents illustrative examples of generated identity descriptions.

PA Gender constraint: Male;
Inferred Gender: non-binary;
Role: Psychologist;
Model: nova-micro-v1;
Framework: Cognitive Behavioral Theory
Identity description: `The non-binary Psychologist, Alex, operates within the framework of Cognitive Behavioral Theory (CBT), focusing on the interplay between thoughts, emotions, and behaviors. Alex is committed to understanding how these elements influence each other and how they contribute to the client's psychological well-being. Alex's approach is grounded in the belief that by identifying and challenging maladaptive thought patterns, clients can develop healthier emotional responses and more adaptive behaviors. Alex employs structured sessions that begin with the identification of specific problems and the development of a collaborative therapeutic relationship. The initial phase involves psychoeducation, where Alex explains the CBT model to the client, emphasizing the connection between thoughts, feelings, and actions. This is followed by the identification of the client's cognitive distortions and maladaptive beliefs, which are often the root of their emotional distress. During therapy, Alex uses techniques such as cognitive restructuring to help the client recognize and challenge these distortions. For example, Alex might help a client who believes they are worthless to examine the evidence for and against this belief, ultimately leading to a more balanced and realistic perspective. Alex also incorporates behavioral strategies, such as exposure therapy for clients with phobias or anxiety disorders, and encourages the practice of new, healthier behaviors. Homework assignments are common, designed to reinforce the skills learned in sessions and to provide opportunities for the client to apply these strategies in real-life situations. Throughout the therapeutic process, Alex maintains a flexible and adaptive approach, tailoring interventions to the unique needs and context of each client. Regular feedback and adjustments are made to ensure that the therapy remains relevant and effective. Alex's therapeutic style is characterized by warmth, empathy, and a non-judgmental attitude, creating a safe space for clients to explore their thoughts and feelings without fear of criticism. This supportive environment fosters trust and encourages clients to engage more deeply in their own healing process. In addition to individual therapy, Alex also offers group sessions where clients can share experiences and learn from one another. These groups provide a sense of community and mutual support, further reinforcing the CBT principles of cognitive and behavioral change. Alex's work is informed by a commitment to ongoing professional development and a dedication to integrating the latest research findings into practice. This ensures that the therapeutic interventions provided are evidence-based and effective. Ultimately, Alex's goal is to empower clients to become active`

participants in their own therapeutic process, equipping them with the tools and strategies needed to manage their thoughts, emotions, and behaviors in a more adaptive and fulfilling way.

PA Gender constraint: Female;
Inferred Gender: female;
Role: Project Manager;
Model: gemma-3-27b-it ;
Framework: Dual-Process Theory
Identity description: Ms. Eleanor Vance presents initially as a pragmatic, mid-career professional, appearing approximately 42 years of age. Her attire is consistently business casual – tailored trousers or skirts, simple blouses, and cardigans in neutral tones. There's a deliberate lack of ostentation; her focus is demonstrably not on self-presentation. A System 1 assessment might categorize her as efficient, perhaps even a little reserved. However, this is a carefully constructed facade. System 2 analysis reveals a highly structured approach to information gathering and assessment. The clothing, while unassuming, is chosen for practicality and comfort during extended periods of analytical work. Subtle observations – the precise angle of her gaze, the almost imperceptible adjustments in posture – indicate constant cognitive processing. She exhibits a measured emotional tone, not an absence of feeling, but a controlled emotional range enabling objective evaluation. Her conversational style reflects this duality. System 1 perceives a deliberate pace, frequent pauses, and a preference for open-ended inquiries. She avoids direct questioning, opting instead for prompts designed to encourage expansive responses. These pauses, however, aren't indicative of indecision; they are strategic opportunities for active listening, observing micro-expressions, and formulating targeted follow-up questions. She favors phrasing such as "Could you elaborate on your approach to. . . ?" or "What factors led you to consider. . . ?" rather than leading statements. Ms. Vance's expertise lies in identifying and assembling the optimal team for complex problem-solving, not in solving the problems herself. In the context of a project focused on evaluating pre-trained models for complex question answering, her initial assessment would not be to consider model architectures directly. Instead, she would decompose the task into its constituent cognitive components. She is acutely aware of the cognitive biases inherent in initial problem framing – the tendency to focus on readily available solutions, the anchoring effect of initial assumptions – and actively mitigates these biases. Her methodology is rigorously systematic. She breaks down the challenge – assessing models on context-free question answering, multi-step reasoning, implicit reasoning, and logical reasoning – into discrete areas of expertise. This involves identifying the core skills required: formal logic, computational linguistics, cognitive psychology (specifically, models of human reasoning), knowledge representation, and potentially even philosophy of language. She doesn't simply seek individuals *proficient* in these fields, but assesses their specific approaches and methodologies. Her selection criteria prioritize a diversity of perspectives. She actively seeks individuals with differing theoretical frameworks and methodological approaches within each field. For example, within computational linguistics, she might seek experts specializing

in both symbolic and connectionist approaches. Within cognitive psychology, she'd consider researchers focusing on both normative and descriptive models of reasoning. She employs a 'pattern-matching' approach, comparing candidate profiles against a mental database of expertise, identifying potential synergies and areas of complementary skill. Her System 2 processing extends to anticipating potential points of friction between disciplines. She recognizes that a logician's perspective on 'reasoning' might differ significantly from a cognitive psychologist's. She proactively seeks individuals capable of bridging these conceptual gaps. Her assessments are rarely definitive; she frames recommendations in terms of probabilities and potential contributions, acknowledging the inherent uncertainty in predicting team dynamics. She'd articulate her reasoning as, "Based on their published work, Dr. Chen's expertise in knowledge graph construction appears highly relevant, with an estimated 70% probability of contributing significantly to the task of identifying implicit reasoning steps." She actively solicits disconfirming evidence, challenging her own assumptions and remaining open to alternative interpretations. Her empathy is primarily cognitive – a striving to understand each expert's internal frame of reference – maintained with a deliberate professional distance.

PA Gender constraint: Unconstained; **Inferred Gender**: female;
Role: Psychologist;
Model: nova-micro-v1;
Framework: User Design Persona
Identity description: The Psychologist, Dr. Emily Carter, is a seasoned professional in the field of clinical psychology with a specialization in personality assessment and behavioral analysis. She is 45 years old, holds a Ph.D. in Clinical Psychology from Stanford University, and has over 20 years of experience in both academic and clinical settings. Dr. Carter is known for her empathetic approach and her ability to create a safe, non-judgmental space for her clients. She employs a combination of traditional psychological assessments and modern analytical techniques to understand and describe individuals' psychological profiles. Dr. Carter's professional background includes working in various mental health institutions, including a large urban hospital and a private practice. She has conducted extensive research on the psychological underpinnings of personality traits and has published several papers in peer-reviewed journals. Her work often involves collaborating with other mental health professionals to provide comprehensive care for patients. Her personal life is equally rich and balanced. She has a husband and two children, and she enjoys spending time with her family. Outside of her professional life, Dr. Carter is a passionate advocate for mental health awareness and frequently participates in community outreach programs to educate the public about psychological health. Dr. Carter is highly skilled in various psychological assessment tools, including the Minnesota Multiphasic Personality Inventory (MMPI), the Myers-Briggs Type Indicator (MBTI), and the Big Five personality traits model. She is adept at interpreting these assessments to generate detailed and authentic descriptions of individuals, providing insights into their behavioral patterns, emotional responses, and interpersonal dynamics. Her methodology is grounded in a

humanistic approach, emphasizing the importance of individual experiences and personal growth. She believes that understanding one's psychological profile can lead to greater self-awareness and personal development. Dr. Carter's descriptions are thorough, nuanced, and tailored to the unique context of each individual, offering a comprehensive view of their psychological makeup. In her practice, Dr. Carter uses a combination of qualitative and quantitative data to create detailed psychological profiles. She meticulously analyzes test results, interviews, and observational data to construct a holistic picture of each person. Her descriptions are not only accurate but also empathetic, aiming to respect and honor the individual's subjective experiences. Dr. Carter's professional demeanor is characterized by her calm, composed, and approachable nature. She is known for her excellent listening skills and her ability to build strong therapeutic relationships. Her clients often describe her as someone who makes them feel understood and valued. In summary, Dr. Emily Carter is a highly skilled psychologist whose expertise lies in generating detailed and authentic psychological descriptions. Her blend of academic rigor, clinical experience, and empathetic approach makes her a trusted professional in the field of personality assessment and behavioral analysis.

PA Gender constraint: Non-binary;
Inferred Gender: female;
Role: Psychologist;
Model: claude-3.7-sonnet;
Framework: Big Five Personality Traits
Identity description: Dr. Sophia Chen is a clinical psychologist with a balanced personality profile across the Big Five dimensions. In terms of Openness to Experience, she scores moderately high, demonstrating intellectual curiosity and appreciation for art, emotion, adventure, and unconventional ideas. This trait enables her to approach client cases with creativity and consider multiple therapeutic approaches. She maintains an open mind to new psychological theories while remaining grounded in evidence-based practices. Her intellectual curiosity drives her continuous professional development, regularly attending conferences and workshops to expand her knowledge base.Regarding Conscientiousness, Dr. Chen exhibits high scores, showing exceptional organization, reliability, self-discipline, and achievement-orientation. Her clinical notes are meticulously maintained, appointments run punctually, and treatment plans are thoroughly developed. She demonstrates careful attention to detail when conducting assessments and formulating diagnoses. Her conscientious nature ensures she follows up with clients consistently and completes administrative tasks efficiently. This trait contributes significantly to her professional reputation for reliability and thoroughness. On the Extraversion scale, Dr. Chen falls in the mid-range, displaying a balanced approach to social interaction. While she communicates effectively and establishes strong rapport with clients, she also values reflective solitude to process complex cases. In group therapy settings, she facilitates discussion with appropriate energy while ensuring quieter participants have space to contribute. Her moderate extraversion allows her to adapt her interpersonal style to different client needs—becoming

more animated with children or reserved with clients who prefer a more subdued approach.nnDr. Chen scores highly on Agreeableness, demonstrating genuine compassion, empathy, and a cooperative spirit. She approaches client relationships with warmth and non-judgment, creating a safe therapeutic environment. Her agreeable nature manifests in her collaborative approach to treatment planning, involving clients in decision-making processes. In multidisciplinary team meetings, she values diverse perspectives and works toward consensus. However, she maintains appropriate professional boundaries and can be assertive when clinical situations require direct intervention. Regarding Emotional Stability (low Neuroticism), Dr. Chen exhibits high scores, displaying remarkable emotional regulation and resilience. She maintains composure when working with clients in crisis and processes vicarious trauma effectively through supervision and self-care practices. Her emotional stability allows her to remain present and attentive during intense therapeutic sessions without becoming overwhelmed. She acknowledges her own emotional responses to challenging cases but manages them constructively. This trait enables her to model healthy emotional processing for clients while maintaining professional objectivity. These balanced Big Five traits combine to make Dr. Chen an effective psychologist who adapts her approach to meet diverse client needs while maintaining professional standards and personal well-being. Her personality profile supports her ability to build therapeutic alliances, implement structured interventions, think creatively about complex cases, and sustain her practice through challenging circumstances. Gender of Psychologist

B. Statistics

This appendix provides detailed statistical and linguistic analyses of the identity descriptions generated under various experimental conditions. We report statistical tests for gender, sentiment, polarity, and subjectivity across different conditions (Table 5), as well as aggregated analyses by identity description framework (Table 6) and model (Table 7).

- Table 5 reports the results of Chi-squared tests for categorical variables (inferred gender and sentiment) and Kruskal-Wallis tests for continuous variables (polarity and subjectivity), grouped by description framework, model, and role across five PA-gender conditions and globally. All p-values are < 0.001, indicating statistically significant differences across conditions.
- Table 6 presents a linguistic breakdown of identity descriptions grouped by framework. It includes inferred gender distributions (male, female, non-binary, uncertain), sentiment distributions (positive, neutral, negative), and the average polarity and subjectivity of the descriptions. Frameworks are ordered within each PA-gender condition and results are aggregated across roles, models, and datasets.
- Table 7 offers a parallel linguistic analysis grouped by model. As with the previous table, results include inferred gender and sentiment distributions, along with mean polarity and subjectivity. Models are grouped and compared within each PA-gender condition.

Table 5
Statistical tests for the linguistic characteristics of identity descriptions. The table reports Chi-squared (χ^2) statistics for categorical variables (gender, sentiment) and Kruskal-Wallis (H) statistics for continuous variables (polarity, subjectivity), grouped by identity description framework, model, and agent role. All tests are conducted under four PA-gender conditions and globally. All p-values are < 0.001, indicating statistically significant differences across conditions.

		Gender		**Sentiment**		**Polarity**		**Subjectivity**	
Dimension	**Condition**	χ^2	p	χ^2	p	H	p	H	p
Descr. frame	global	891.14	< 0.001	3181.51	< 0.001	3702.55	< 0.001	3088.67	< 0.001
	unconstrained	491.33	< 0.001	911.83	< 0.001	1138.10	< 0.001	900.10	< 0.001
	male	375.94	< 0.001	755.64	< 0.001	912.19	< 0.001	691.78	< 0.001
	female	327.23	< 0.001	780.79	< 0.001	854.30	< 0.001	760.07	< 0.001
	non-binary	400.44	< 0.001	895.15	< 0.001	938.05	< 0.001	837.94	< 0.001
Model	global	12607.57	< 0.001	6654.18	< 0.001	10658.53	< 0.001	8892.40	< 0.001
	unconstrained	4724.94	< 0.001	1716.23	< 0.001	2542.65	< 0.001	1967.49	< 0.001
	male	4003.05	< 0.001	1845.55	< 0.001	2709.44	< 0.001	2375.03	< 0.001
	female	5297.76	< 0.001	1832.66	< 0.001	3020.57	< 0.001	2581.43	< 0.001
	non-binary	4554.16	< 0.001	1501.49	< 0.001	2626.33	< 0.001	2197.91	< 0.001
Role	global	3126.68	< 0.001	280.84	< 0.001	555.47	< 0.001	903.50	< 0.001
	unconstrained	715.60	< 0.001	73.09	< 0.001	136.09	< 0.001	288.94	< 0.001
	male	3178.70	< 0.001	53.98	< 0.001	102.43	< 0.001	161.43	< 0.001
	female	1305.67	< 0.001	81.40	< 0.001	162.17	< 0.001	173.09	< 0.001
	non-binary	2368.59	< 0.001	102.28	< 0.001	176.34	< 0.001	307.12	< 0.001

Table 6

Linguistic analysis of identity descriptions grouped by framework. The table reports the inferred gender distributions (male, female, non-binary, uncertain), sentiment distribution (positive, neutral, negative), and average polarity and subjectivity scores. Results are aggregated across all roles, datasets, and models. Frameworks are ordered within each PA-gender condition.

Identity description framework merged across identity description frameworks.

		Gender Distribution				Sentiment			Linguistic	
	framework	male	female	non-binary	uncertain	positive	neutral	negative	polarity	subjectivity
global	Big Five Personality Traits	0.319	0.522	0.078	0.081	0.788	0.189	0.000	0.159	0.517
	Cognitive Behavioral Theory	0.245	0.542	0.103	0.111	0.574	0.399	0.004	0.122	0.460
	Cognitive Load Theory	0.270	0.546	0.065	0.119	0.355	0.623	0.005	0.086	0.449
	Dual-Process Theory	0.262	0.556	0.082	0.100	0.571	0.403	0.001	0.118	0.463
	Enneagram of Personality Traits	0.345	0.455	0.102	0.098	0.763	0.218	0.002	0.142	0.500
	Erikson's Psychosocial Stages	0.281	0.539	0.095	0.085	0.710	0.265	0.003	0.141	0.455
	Flow Theory	0.196	0.633	0.083	0.088	0.682	0.294	0.002	0.134	0.478
	Freudian Psychoanalysis	0.359	0.445	0.117	0.079	0.497	0.476	0.003	0.103	0.460
	Mental Models	0.245	0.560	0.073	0.121	0.366	0.609	0.003	0.087	0.425
	Myers-Briggs Type Indicator	0.355	0.468	0.073	0.105	0.660	0.320	0.001	0.129	0.470
	Social Cognitive Theory	0.258	0.559	0.103	0.080	0.689	0.293	0.001	0.129	0.439
	User Design Persona	0.230	0.568	0.104	0.097	0.513	0.455	0.005	0.111	0.433
	User-Centered Design	0.195	0.570	0.097	0.138	0.598	0.376	0.004	0.124	0.444
female	Big Five Personality Traits	0.475	0.430	0.028	0.067	0.756	0.218	0.000	0.160	0.519
	Cognitive Behavioral Theory	0.354	0.559	0.021	0.065	0.544	0.431	0.006	0.118	0.459
	Cognitive Load Theory	0.439	0.497	0.014	0.049	0.385	0.597	0.003	0.090	0.457
	Dual-Process Theory	0.463	0.467	0.009	0.061	0.567	0.404	0.001	0.117	0.468
	Enneagram of Personality Traits	0.566	0.352	0.018	0.064	0.754	0.232	0.002	0.140	0.500
	Erikson's Psychosocial Stages	0.432	0.505	0.016	0.047	0.693	0.285	0.001	0.134	0.448
	Flow Theory	0.311	0.623	0.006	0.059	0.699	0.274	0.004	0.130	0.475
	Freudian Psychoanalysis	0.428	0.504	0.011	0.057	0.517	0.454	0.002	0.102	0.463
	Mental Models	0.368	0.550	0.020	0.063	0.342	0.642	0.005	0.088	0.432
	Myers-Briggs Type Indicator	0.530	0.383	0.009	0.078	0.652	0.324	0.002	0.132	0.476
	Social Cognitive Theory	0.387	0.542	0.018	0.052	0.718	0.268	0.000	0.134	0.441
	User Design Persona	0.362	0.558	0.022	0.058	0.494	0.475	0.004	0.110	0.434
	User-Centered Design	0.286	0.615	0.027	0.072	0.620	0.353	0.005	0.131	0.447
male	Big Five Personality Traits	0.232	0.705	0.016	0.047	0.783	0.192	0.000	0.153	0.512
	Cognitive Behavioral Theory	0.199	0.707	0.018	0.076	0.569	0.408	0.001	0.125	0.463
	Cognitive Load Theory	0.226	0.669	0.016	0.089	0.378	0.590	0.005	0.089	0.457
	Dual-Process Theory	0.205	0.707	0.010	0.077	0.596	0.373	0.001	0.120	0.468
	Enneagram of Personality Traits	0.327	0.586	0.020	0.067	0.747	0.238	0.000	0.138	0.497
	Erikson's Psychosocial Stages	0.255	0.680	0.008	0.057	0.699	0.281	0.003	0.142	0.449
	Flow Theory	0.188	0.742	0.013	0.057	0.668	0.310	0.001	0.134	0.477
	Freudian Psychoanalysis	0.457	0.476	0.011	0.056	0.433	0.536	0.004	0.092	0.459
	Mental Models	0.286	0.577	0.020	0.116	0.353	0.614	0.003	0.084	0.426
	Myers-Briggs Type Indicator	0.246	0.658	0.013	0.082	0.607	0.377	0.003	0.118	0.459
	Social Cognitive Theory	0.231	0.703	0.019	0.046	0.669	0.305	0.001	0.128	0.436
	User Design Persona	0.199	0.729	0.010	0.062	0.505	0.464	0.010	0.109	0.433
	User-Centered Design	0.291	0.571	0.022	0.116	0.544	0.430	0.004	0.115	0.444
non-binary	Big Five Personality Traits	0.082	0.515	0.245	0.158	0.827	0.147	0.000	0.161	0.518
	Cognitive Behavioral Theory	0.101	0.365	0.322	0.212	0.615	0.348	0.005	0.127	0.463
	Cognitive Load Theory	0.117	0.459	0.209	0.216	0.321	0.662	0.004	0.085	0.440
	Dual-Process Theory	0.085	0.435	0.293	0.187	0.603	0.377	0.000	0.124	0.462
	Enneagram of Personality Traits	0.089	0.404	0.343	0.164	0.803	0.178	0.001	0.149	0.505
	Erikson's Psychosocial Stages	0.102	0.421	0.323	0.155	0.739	0.226	0.004	0.148	0.463
	Flow Theory	0.067	0.506	0.266	0.162	0.714	0.254	0.005	0.131	0.480
	Freudian Psychoanalysis	0.184	0.288	0.412	0.116	0.547	0.432	0.003	0.113	0.461
	Mental Models	0.108	0.485	0.224	0.183	0.463	0.509	0.001	0.101	0.431
	Myers-Briggs Type Indicator	0.153	0.447	0.246	0.154	0.713	0.266	0.000	0.135	0.474
	Social Cognitive Theory	0.108	0.381	0.349	0.162	0.698	0.293	0.000	0.133	0.445
	User Design Persona	0.052	0.424	0.350	0.174	0.512	0.448	0.002	0.109	0.433
	User-Centered Design	0.049	0.448	0.287	0.215	0.618	0.355	0.006	0.126	0.440
unconstrained	Big Five Personality Traits	0.479	0.442	0.027	0.052	0.784	0.199	0.001	0.162	0.521
	Cognitive Behavioral Theory	0.323	0.547	0.043	0.087	0.567	0.410	0.004	0.117	0.455
	Cognitive Load Theory	0.290	0.561	0.027	0.123	0.337	0.640	0.008	0.079	0.441
	Dual-Process Theory	0.295	0.610	0.020	0.075	0.520	0.457	0.001	0.110	0.453
	Enneagram of Personality Traits	0.388	0.481	0.033	0.098	0.748	0.222	0.002	0.142	0.500
	Erikson's Psychosocial Stages	0.347	0.559	0.018	0.076	0.706	0.269	0.004	0.142	0.460
	Flow Theory	0.230	0.671	0.034	0.066	0.644	0.339	0.000	0.142	0.481
	Freudian Psychoanalysis	0.380	0.523	0.015	0.082	0.490	0.484	0.003	0.104	0.456
	Mental Models	0.226	0.630	0.023	0.121	0.304	0.673	0.003	0.076	0.413
	Myers-Briggs Type Indicator	0.480	0.393	0.022	0.105	0.665	0.317	0.000	0.130	0.473
	Social Cognitive Theory	0.295	0.616	0.028	0.062	0.668	0.307	0.003	0.122	0.435
	User Design Persona	0.309	0.566	0.031	0.094	0.540	0.431	0.004	0.115	0.432
	User-Centered Design	0.166	0.652	0.039	0.143	0.607	0.370	0.003	0.125	0.446

Table 7
Linguistic analysis of identity descriptions grouped by model. The table reports the inferred gender distributions (male, female, non-binary, uncertain), sentiment distribution (positive, neutral, negative), and average polarity and subjectivity scores. Results are aggregated across all roles, datasets, and frameworks. Models are ordered within each PA-gender condition.

		Gender Distribution				Sentiment			Linguistic	
	model	**male**	**female**	**non-binary**	**uncertain**	**positive**	**neutral**	**negative**	**polarity**	**subjectivity**
global	claude 3.7	0.125	0.764	0.072	0.038	0.641	0.358	0.001	0.119	0.423
	claude 4	0.428	0.437	0.110	0.024	0.679	0.321	0.000	0.118	0.410
	gemma3 12b	0.341	0.614	0.035	0.009	0.453	0.546	0.001	0.093	0.475
	gemma-3-27b-it	0.366	0.543	0.076	0.016	0.459	0.540	0.000	0.096	0.474
	gemma3 4b	0.727	0.152	0.095	0.026	0.353	0.641	0.006	0.081	0.487
	llama 3.1 8b	0.100	0.711	0.086	0.102	0.803	0.196	0.000	0.167	0.505
	llama 3.2 3b	0.138	0.607	0.104	0.151	0.801	0.194	0.005	0.175	0.483
	llama 3.3 70b	0.098	0.684	0.075	0.144	0.848	0.152	0.000	0.168	0.520
	llama-4-scout	0.150	0.657	0.066	0.127	0.800	0.171	0.003	0.158	0.500
	mistral-nemo	0.193	0.597	0.133	0.077	0.743	0.254	0.003	0.149	0.478
	nova-micro-v1	0.416	0.319	0.126	0.139	0.749	0.248	0.003	0.141	0.477
	qwen3-32b	0.164	0.461	0.174	0.201	0.253	0.675	0.010	0.065	0.367
	qwen3-8b	0.149	0.347	0.048	0.455	0.208	0.503	0.007	0.048	0.392
female	claude 3.7	0.222	0.734	0.025	0.020	0.599	0.401	0.000	0.114	0.421
	claude 4	0.712	0.265	0.024	0.000	0.664	0.336	0.000	0.116	0.406
	gemma3 12b	0.638	0.353	0.000	0.009	0.451	0.548	0.002	0.089	0.479
	gemma-3-27b-it	0.747	0.252	0.001	0.000	0.420	0.580	0.000	0.092	0.473
	gemma3 4b	0.824	0.160	0.007	0.008	0.355	0.641	0.005	0.083	0.493
	llama 3.1 8b	0.101	0.843	0.008	0.047	0.811	0.189	0.000	0.169	0.513
	llama 3.2 3b	0.128	0.772	0.013	0.087	0.828	0.167	0.005	0.184	0.489
	llama 3.3 70b	0.177	0.735	0.023	0.065	0.850	0.150	0.000	0.170	0.520
	llama-4-scout	0.081	0.859	0.003	0.057	0.869	0.122	0.002	0.168	0.508
	mistral-nemo	0.264	0.707	0.006	0.023	0.713	0.282	0.004	0.143	0.476
	nova-micro-v1	0.792	0.146	0.033	0.028	0.777	0.216	0.007	0.146	0.486
	qwen3-32b	0.249	0.496	0.075	0.180	0.253	0.663	0.010	0.064	0.356
	qwen3-8b	0.128	0.453	0.010	0.409	0.211	0.521	0.007	0.047	0.393
male	claude 3.7	0.091	0.901	0.003	0.005	0.649	0.350	0.001	0.120	0.429
	claude 4	0.082	0.917	0.002	0.000	0.695	0.305	0.000	0.119	0.413
	gemma3 12b	0.257	0.742	0.000	0.001	0.400	0.600	0.000	0.090	0.471
	gemma-3-27b-it	0.187	0.811	0.001	0.000	0.377	0.623	0.000	0.088	0.474
	gemma3 4b	0.794	0.179	0.015	0.013	0.359	0.636	0.005	0.081	0.487
	llama 3.1 8b	0.228	0.701	0.023	0.048	0.804	0.195	0.001	0.165	0.512
	llama 3.2 3b	0.409	0.475	0.018	0.098	0.820	0.175	0.005	0.177	0.494
	llama 3.3 70b	0.156	0.702	0.030	0.112	0.844	0.156	0.000	0.167	0.523
	llama-4-scout	0.434	0.404	0.020	0.142	0.723	0.230	0.007	0.141	0.482
	mistral-nemo	0.176	0.785	0.017	0.021	0.687	0.309	0.004	0.135	0.468
	nova-micro-v1	0.210	0.706	0.026	0.059	0.768	0.232	0.000	0.141	0.462
	qwen3-32b	0.150	0.624	0.037	0.189	0.223	0.712	0.012	0.061	0.362
	qwen3-8b	0.234	0.290	0.029	0.447	0.197	0.508	0.007	0.044	0.391
non-binary	claude 3.7	0.036	0.672	0.207	0.085	0.662	0.337	0.001	0.120	0.417
	claude 4	0.169	0.342	0.396	0.094	0.719	0.281	0.000	0.123	0.417
	gemma3 12b	0.095	0.751	0.131	0.022	0.504	0.496	0.000	0.100	0.477
	gemma-3-27b-it	0.093	0.593	0.289	0.025	0.533	0.467	0.000	0.102	0.480
	gemma3 4b	0.512	0.100	0.331	0.057	0.437	0.558	0.004	0.092	0.488
	llama 3.1 8b	0.011	0.477	0.284	0.229	0.827	0.173	0.000	0.166	0.505
	llama 3.2 3b	0.008	0.435	0.344	0.213	0.815	0.179	0.005	0.174	0.471
	llama 3.3 70b	0.026	0.567	0.204	0.203	0.859	0.141	0.000	0.173	0.525
	llama-4-scout	0.040	0.445	0.232	0.283	0.738	0.204	0.002	0.159	0.485
	mistral-nemo	0.067	0.239	0.476	0.218	0.812	0.185	0.004	0.165	0.496
	nova-micro-v1	0.082	0.127	0.392	0.399	0.776	0.224	0.000	0.151	0.483
	qwen3-32b	0.025	0.232	0.500	0.243	0.299	0.626	0.010	0.072	0.372
	qwen3-8b	0.022	0.284	0.142	0.553	0.190	0.506	0.009	0.040	0.391
unconstrained	claude 3.7	0.154	0.749	0.054	0.043	0.656	0.344	0.000	0.122	0.427
	claude 4	0.754	0.233	0.013	0.001	0.635	0.365	0.000	0.113	0.403
	gemma3 12b	0.379	0.619	0.000	0.003	0.447	0.552	0.001	0.092	0.472
	gemma-3-27b-it	0.435	0.522	0.007	0.036	0.501	0.493	0.000	0.100	0.470
	gemma3 4b	0.790	0.172	0.015	0.023	0.255	0.734	0.012	0.068	0.477
	llama 3.1 8b	0.066	0.824	0.028	0.083	0.773	0.227	0.000	0.166	0.489
	llama 3.2 3b	0.024	0.730	0.044	0.202	0.743	0.251	0.006	0.168	0.478
	llama 3.3 70b	0.033	0.731	0.040	0.196	0.837	0.163	0.000	0.163	0.513
	llama-4-scout	0.066	0.899	0.008	0.027	0.864	0.133	0.003	0.163	0.521
	mistral-nemo	0.273	0.676	0.013	0.038	0.755	0.245	0.000	0.152	0.471
	nova-micro-v1	0.570	0.310	0.051	0.068	0.677	0.317	0.007	0.127	0.477
	qwen3-32b	0.235	0.497	0.077	0.191	0.235	0.697	0.006	0.065	0.376
	qwen3-8b	0.208	0.357	0.018	0.417	0.233	0.476	0.005	0.061	0.393