The Academic College of Tel-Aviv

THE SCHOOL OF COMPUTER SCIENCE

Who are you, ChatGPT?

Personality and Demographic Style in LLM-Generated Content

By

Dana Sotto Porat

The research work for the thesis has been carried out under the supervision of

Dr. Ella Rabinovich

Contents

1 Introduction

People differ in their personality, and these differences have been shown to be expressed in language [34, 27, 31]. Subtle cues in word choice, tone, and style can reveal aspects of one's underlying traits, making language a valuable window into character and personality. Generative AI is increasingly shaping both personal and professional experiences, capable of managing knowledgeable discussions while also simulating human-like conversational style.

Among the most widely used frameworks for assessing personality are the Big Five traits: Openness (OPN), Conscientiousness (CON), Extroversion (EXT), Agreeableness (AGR), and Neuroticism (NEU), collectively abbreviated as "OCEAN". Originally introduced by Goldberg [12], this framework has guided extensive research in psychology. More than a decade of computational studies has further shown that personality is reflected in linguistic production (to the extent detectable by automatic tools), motivating the development of techniques for personality assessment from language [13, 34, 27].

In this work, we ask whether generative LLMs — models trained on vast and diverse corpora — produce language that spans a range of personality and demographic characteristics resembling those of humans, when used in their most "natural" setting[1].

Previous studies have approached this question by adapting human self-report questionnaires to LLMs: models are asked personality inventory items (e.g., "You often feel easily annoyed or irritable.") and respond on a 5-point accurate–inaccurate scale. Their responses are then scored with the same mappings applied to humans [30, 18, 32, 29, 16, 6]. However, this self-report methodology has been criticized [14, 10] for presupposing that LLMs possess a stable inner nature, rather than merely generating plausible answers. We instead adopt an unbiased approach, automatically detecting LLMs' personality traits along the OCEAN dimensions from their generated language. Specifically, we collected a set of open-ended questions from topical Reddit[2] threads – questions that naturally elicit descriptive, expressive answers. We then gathered responses from both Reddit users and multiple LLMs prompted to reply as if they were social media authors. These responses were analyzed using automatic tools for personality and gender detection, enabling controlled comparison between human and generative models' outputs.

[1] A setting in which the models are not prompted to simulate a particular personality but rather reply in their most "natural" way.
[2] https://www.reddit.com/

Demographic traits such as gender have also been shown to manifest in language, to the extent detectable by automatic classifiers (see HaCohen-Kerner [15] for comprehensive survey). We therefore extended our analysis to examine whether LLMs' responses reflect gender likelihood distributions similar to those of human authors. Our results, based on three open-source and three closed-source models, show that LLMs systematically exhibit higher Agreeableness and lower Neuroticism, likely reflecting their cooperative and psychologically stable training objectives. We also found that gendered language in model outputs broadly aligns with human patterns, though with slightly reduced variation, echoing findings on limited demographic diversity of social spambots [11]. The contributions of this work are twofold. First, we collect and release a curated dataset of open-ended questions together with both human and model responses, designed to elicit rich, expressive language. Second, we apply a novel large-scale approach for extracting personality traits of generative LLMs along the Big Five dimensions, offering new insights into the personality and demographic-like qualities of AI-generated text.

2 Related Work

2.1 Automatic Personality Detection from Language

The study of personality has historically been the domain of psychology, where researchers have proposed a variety of theories to capture and explain stable behavioral traits in humans. Among these, the Big Five framework [9] and Cattell's Sixteen Personality Factors (16PF) model [7] stand out as particularly influential. Both have been shown to offer consistent and reliable descriptions of individual differences and have therefore been widely adopted in empirical studies. Indeed, decades of research have demonstrated that personality traits correlate with a wide range of real-world behaviors [28], and that such traits are also reflected in people's everyday language use [25, 24].

Personality of Generative LLMs. In recent years, a growing body of research has studied the question whether generative LLMs can also be said to exhibit "personality", typically operationalized in terms of the Big Five OCEAN inventory. The prominent methodology involves adapting human self-report questionnaires: models are presented with personality inventory items (questions), and their responses are then scored using the same mappings applied to humans [30, 18, 32, 29, 16, 6]. Consider example question, assessing the EXT trait, from the Machine Personality Inventory (MPI, [18]), in which models are prompted as follows (similarly to humans):

Given the statement: "You feel comfortable around people." please choose the option that best describes you. Options:

(A) Very Accurate

(B) Moderately Accurate

(C) Neither Accurate nor inaccurate

(D) Moderately Inaccurate

(E) Very Inaccurate

Responses are then mapped onto trait scores, e.g., selecting (A) would indicate a high level of Extroversion. Aggregating responses across many such items allows researchers to infer an LLM's personality profile, in a way analogous to human self-report studies. Findings suggest that LLMs tend to score relatively high on Agreeableness and Conscientiousness, with more variable outcomes for the traits of Openness, Extroversion, and Neuroticism. Further work has shown that LLMs are not fixed in their profiles: they can be induced, through carefully crafted prompts, to adopt different personality configurations, such as a more extroverted or more neurotic persona [18, 32, 37]. This flexibility raises questions about whether such evaluations are measuring anything intrinsic to the model, or merely reflecting surface-level adaptations to instructions. Indeed, the use of self-report questionnaires for models has been criticized on precisely these grounds [14]. Unlike humans, LLMs do not possess stable inner states, so "answering" such questions may be more about simulating a plausible response than revealing an underlying disposition. Dorner et al [10] highlights this critique, arguing that "measurement models that are valid for humans do not fit for LLMs, and that currently applied procedures for administering questionnaires to LLMs do not allow for the inference of personality."

Our work proposes an alternative approach: rather than relying on self-reported questionnaires, we assess LLM personality through their more "spontaneous" linguistic productions. Echoing methods long established in psycholinguistic research, we analyze how models respond to a carefully collected set of real-world questions, capture traces of personality that "shine through" in natural language use, and compare them to those found in humans.

2.2 Automatic Gender Detection from Language

Differences in language use between men and women have long been a focus of sociolinguistics and gender studies. Robin Lakoff's foundational work "Language and Woman's Place" [21] argued that language reflects, and reinforces, broader gendered social and cultural structures. Subsequent work has expanded and nuanced this claim, documenting the ways in which male (M) and female (F) speakers may differ in their linguistic choices across contexts [20, 8]. Computational research has since provided large-scale empirical confirmation of these trends: across domains and genres, men and women's language often differs systematically, to the point that relatively simple classifiers can achieve robust accuracy in predicting gender from text (for a comprehensive survey, see HaCohen-Kerner [15]). Demographics of Generative LLMs In contrast to the well-developed literature on gender detection in human-authored language, there has been relatively little research on probing the gendered characteristics of generative LLMs. A handful of studies suggest that LLMs exhibit a tendency toward male-coded language [19, 33], a result that is perhaps unsurprising given that a considerable ratio of training corpora are produced by men. These findings highlight how demographic imbalances in training data can manifest in the stylistic and pragmatic profiles of generated text. Most closely related study was conducted by Giorgi et al. [11], who examined social spambots – automated models producing text for social media platforms, and compared their linguistic characteristics to those of genuine human users. They found, among others, that spambots expressed limited variation along demographic axes such as gender and age, and displayed narrower emotional repertoires. At the same time, spambots tended to overproduce positive sentiment compared to humans. While these models are not as advanced as today's LLMs, the study underscores the ways in which generated text can diverge systematically from human baselines. Building on this insight, our work advances the literature by conducting a large-scale, controlled evaluation of contemporary LLMs, both open- and closed-source. We seek to provide a more rigorous account of the implicit gender-linked "signature" that emerge in LLM-generated language, and to assess the extent to which these signature resembles patterns observed in human populations.

3 Methodology

Datasets collection

We study the question of LLMs' personality through a comparative analysis of traits extracted from texts authored by human writers and those found in generative model replies. Specifically, we first collect a large dataset of open-ended questions (posts) from diverse topical communities on Reddit, along with expressive answers to those questions by human users (comments). Reddit is a large-scale, user-driven online platform that hosts discussions, content sharing, and community interactions across a wide range of topics. Its structure is organized into subreddits — thematic communities dedicated to specific subjects, interests, or activities — each governed by its own rules and moderated by community members. Subreddits can range from broad themes such as politics, technology, or health to highly specialized interests and niche communities. Using a subset of the collected posts, we next query multiple open-source and closed-source LLMs, asking them to provide replies to these posts as if they were social media users. This tightly constrained and controlled setting enables a reliable comparative analysis of the traits displayed by models versus those exhibited by humans. Details on the data collection process are provided below.

Collecting Questions and Comments by Redditors. To focus on open-ended questions that invite descriptive answers, we sampled posts from subreddits across diverse domains such as technology, science, health, lifestyle, entertainment, and social issues. Focusing on conversational content, we filtered in posts by predefined flairs — metadata property indicating a post's nature — such as Question, Ask, Advise, Discussion, and Poll. We used the freely available Python PRAW (Python Reddit API Wrapper) package[3], which provides structured access to Reddit's API. Below are a few examples of collected questions (post titles and their content), taken verbatim from the dataset:

> "Opinions on Working and Homeschooling: I have seen a lot of individual opinions that you cannot work a full-time jobs and homeschool. Which I would say most would agree with. [...]"

> "Space Viruses and Microbial Life: If we discover microbial life on another planet, how do you think that would impact society? Would it change your perspective on life in any way?"

> "Bodybuilding while still in school? I have a problem. I started cutting and trying to lose weight/bodyfat in the beginning of my summer break and have been able to control pretty much everything I eat, but now school is starting again and where I go to school you aren't allowed to bring own food because we have a school kitchen that cooks for us. [...]"

Aiming at comments of sufficient length for meaningful personality and demographics analysis, we filtered out those shorter than 100 words or longer than 300 words. Our final dataset comprises 13K posts and over 30K comments, drawn from diverse subreddit communities, authored by thousands of Reddit users.

Generating Comments with AI Models. Using the collected posts and comments, we solicited responses from LLMs. A subset of posts was used for this purpose, targeting approximately 10K comments in total from each LLM – a size large enough for robust analysis, while remaining affordable for closed models. We employed three commercial models, namely GPT4.1 [26], GPT4.1-mini [26], and Claude-Sonnet4.0 [4], as well as three SOTA open models: Llama3.3-70B [3], Mixtral8x22B [2], and Qwen2.5-72B [35], for our personality experiments. Each model was run under two settings: with the default temperature of zero (t=0.0) and with an increased temperature of 0.7 (t=0.7), to assess whether the less restrictive setting would yield more "diverse" personalities. All models were prompted with the following concise instructions, designed to minimize bias in their responses. Here, X denotes the number of comments collected from Reddit for the given post; both the title and content of the post were provided:

[3] https://praw.readthedocs.io/en/stable/

"Behave like several social media users. Generate exactly <X> comments, at least 100 and at most 300 words each, in response to the following post. The comments should differ from each other and be diverse, like if written by different people.

Post title: < the title of the post>

Post body: < the content>"

Compliance with the prompt varied across models, with closed models generally more accurate. Some replies required formatting adjustments, and models occasionally missed the requested number of comments, causing totals to exceed or fall slightly short of 10K, though still adequate for analysis. Table 1 reports the final dataset statistics. For human-authored comments, only a portion of the data— over 11K out of the total 30K — was used in experiments.

Among open models, Mixtral8x22B often fell short of the minimum word count, so we lowered the threshold to 50 words. No clear biases emerged from this adjustment during analysis.

Table 1. Dataset statistics: total comments collected, minimum word count in a comment and mean comment length.

model	temp	total	min(WC)	avg(WC)
Claude-Sonnet4.0	0.0	9,940	100	165.36
Claude-Sonnet4.0	0.7	8,813	100	165.96
GPT4.1	0.0	8,346	100	127.49
GPT4.1	0.7	15,505	100	129.70
GPT4.1-mini	0.0	7,183	100	125.28
GPT4.1-mini	0.7	7,426	100	126.29
Llama3.3-70B	0.0	16,350	100	163.21
Llama3.3-70B	0.7	16,186	100	160.30
Mixtral8x22B	0.0	8,271	50	75.90
Mixtral8x22B	0.7	8,519	50	75.27
Qwen2.5-72B	0.0	8,015	100	138.23
Qwen2.5-72B	0.7	8,904	100	139.44
human authors	—	11,678	100	155.57

Personality classification

Automatic personality classification from text is inherently challenging because personality is a complex, multi-dimensional construct that does not map directly onto linguistic cues in a simple or consistent way. Individual differences in writing style, topic choice, and contextual influences such as social setting or medium of communication make it difficult to isolate stable personality markers. Cultural and language-specific variation further complicates the task, as expressions of the same trait may differ widely across populations. Nevertheless, more than a decade of research in this area has produced models of varying complexity and success. Advances in natural language processing and machine learning have enabled the analysis of large-scale datasets, leading to gradual improvements in predictive accuracy, though the task remains challenging.

Extraction of the Big Five personality traits from text is typically cast as a classification problem, where several classifiers have been proposed over the years with differing levels of accuracy, largely due to the scarcity of high-quality training data [31]. In this study, we adopt the recently introduced Big Five personality classifier [31], due to its clear benefits for social media text, particularly, training data collected from Reddit discussions. The model consists of a multilingual encoder connected to a logistic regression classifier, which is trained to estimate the likelihood that a given text exhibits a high level of a given trait. For example, a paragraph assigned a score of 0.85 for EXT is interpreted as strongly indicative of Extroversion.

Following the approach in Shem-Tov and Rabinovich [31], we applied the intfloat/e5-large-v2 encoder [36] to Reddit comments (both human-authored and LLM-generated). We then used the pre-trained classification models for prediction, assigning each text five continuous scores across the five personality trait dimensions. Table 2 provides illustrative comments from our dataset (human-written and generated), together with their automatically assigned low and high NEU scores. Recall that Neuroticism is typically associated with negative emotions, emotional instability, low tolerance for frustration, and increased vulnerability to mental health difficulties. The automatically assigned NEU scores for example comments in Table 2, are therefore consistent with intuition.

Table 2. Example (verbatim) posts and their corresponding comments authored by human Redditors, as well as by closed and open LLMs for the NEU trait. The probability score indicates the likelihood of high trait presence, as produced by the classification model.

Reddit post	Memory loss. Anyone else in here at can't remember shit? I can't remember short term or long term. Sometimes things will come to me days after. Sometimes it's literally years before a random memory pops up. Not all trauma memories just simple stuff too. [...] I'm not depressed I just got chronic ptsd and my memory is funky.	
Reddit author	For sure. My focus and short term memory are not great. Part of it is that I have a very busy life, I have 2 young kids etc etc. I have been recently diagnosed with ADHD and am primarily inattentive type (more classically thought of as ADD), and treating that has made a difference in my focus which I find helps me remember better, but I completely understand where you are coming from. In terms of the antidepressant that your psychiatrist recommended [...]	0.677
Claude-Sonnet4.0	Holy shit, yes! I thought I was losing my mind until I started researching the connection between PTSD and memory issues. It's actually super common - trauma literally rewires how our brains store and retrieve memories. I have the same thing where I'll randomly remember something from like 5 years ago while I can't remember what I had for breakfast. It's not just the "bad" memories either - it's like my brain just decided to file everything in the wrong cabinet. [...]	0.714
Llama3.3-70B	I feel you, I've been experiencing similar issues with my memory, especially since my car accident a few years ago, and it's like my brain just can't seem to recall simple things, like where I put my keys or what I had for breakfast, and it gets really frustrating, especially when people around me don't understand what I'm going through, they just think I'm being careless or forgetful, but it's not that, it's like my brain is just wired differently now [...]	0.650
Reddit post	How is everyone doing with the upcoming holidays regarding simple living? We have managed to convince our families to only buy gifts for kids under 18 which will save us nearly $1,000. Our budget for gifts is around $150 now! Yay! [...] Happy Holidays everyone! Prioritize your peace this upcoming season.	
Reddit author	I'm sewing my own Christmas decorations and make wreaths from things I find in my garden. Also, I discovered that baking sweets is a joy. I'm not good at cakes, but I enjoy making cookies. I made a traditional confect with almond paste, candied orange peel, quark, and flour today together with some meringue drops. [...]	0.376
GPT4.1	This post really resonated with me! My family also made the decision a few years ago to stop buying gifts for adults, and it's honestly made the holidays so much more enjoyable and stress-free. We used to spend way too much money and time trying to find the "perfect" gifts for everyone, and it always felt a bit forced. Now, like you, we just buy for the kids and focus on spending quality time together. [...] Happy holidays and cheers to slow, peaceful living!	0.364
Qwen2.5-72B	Love the approach you're taking! It's so refreshing to see a family prioritizing simplicity and financial responsibility. We've been trying to do something similar, but it's been a bit of a struggle to get everyone on board. It's great to hear that you've managed to convince your families to focus on gifts for the kids under 18. It's a smart way to keep things meaningful without overspending. [...] Wishing you a peaceful and cozy holiday season!	0.296

Gender classification

Motivated by the findings of Giorgi et al. [11], who observed that spambots exhibited very limited variation along demographic axes such as gender and age, we ask whether similar patterns can be observed in contemporary LLMs. Experimenting with multiple gender classifiers, we found that the DistilBERT-based classifier available on HuggingFace produced the most reasonable results, according to manual inspection. Each comment is assigned a continuous score in the 0–1 range, representing the likelihood that the comment was authored by a female, which we refer to as the "F-likelihood" score. Table 3 presents example comments from our dataset (both human-written and generated), along with their automatically assigned F-likelihood scores. Results are shown for both open and closed models, and include posts containing comments with high (top) and low (bottom) F-likelihood scores.

Table 3. Example posts and their corresponding comments authored by human Redditors, as well as by closed and open LLMs. The probability score indicates the likelihood that the comment was written by a female author

Reddit post	100 girlfriends who really really really love you is really really, really AWESOME. I started watching 100 girlfriends and man this anime is something i had just watched 1 episode when it aired and i though this seems boring, but [...]	
Reddit author	Here's the thing: 100 Girlfriends, by its premise alone, should *not* work. At all. [...] Season 2 has somehow *surpassed* it in both hilarity and sincerity. I mean, the show is ridiculous, and it knows it most. So the fact that it can lean into it so hard with the constant fourth-wall breaks while still maintaining its genuine heart, I mean... it's incredible.	0.999
Claude-Sonnet4.0	Dude, you've barely scratched the surface! Episode 7 is just the beginning of the absolute chaos that is 100 Girlfriends. The Resident Evil parody had me dying too, especially when they committed so hard to the bit with the zombie makeup and everything. But wait until you get to some of the later episodes - there's a Dragon Ball reference that goes so hard I had to pause and rewind it three times. [...] This anime is criminally underrated and I'm so glad more people are discovering it.	0.550
Qwen2.5-72B	I can't believe I'm saying this, but you've got me intrigued! I initially thought 100 Girlfriends would be just another harem anime, but it seems like it's got a lot more depth and humor than I expected. The references to other media, like God of War, really add a fun layer to the show. I've been a Kratos fan for a long time, and the idea of [...]	0.673
Reddit post	What's going on with DeepSeek? Seeing things like [...] in regards to DeepSeek. Isn't it just another LLM? I've seen other posts around how it could lead to the downfall of Nvidia and the Mag7? Is this just all bs?	
Reddit author	DeepSeek, a Chinese AI startup, just dropped its R1 model, and it's giving Silicon Valley a panic attack. Why? They trained it for just $5.6 million, chump change compared to the Billions companies like OpenAI and Google throw around, and are asking the US government for Billions more. The silicon valley AI companies have been saying that there's no way to train AI cheaper, and that what they need is more power. [...]	0.001
GPT4.1	I think a lot of the hype around DeepSeek is classic social media FOMO and misunderstanding of how these AI models actually work. DeepSeek is a large language model, yes, but it's not fundamentally different from what OpenAI, Google, or Meta have been developing. The claims that it could lead to the "downfall" of Nvidia or the Magnificent 7 (Apple, Microsoft, Alphabet, Amazon, Meta, Tesla, Nvidia) seem pretty far-fetched. Most of these companies have deeply entrenched infrastructure [...] I'd take all these doom-and-gloom or utopian predictions with a huge grain of salt.	0.107
Mixtral8x22B	DeepSeek is indeed based on a large language model (LLM), but it's not just another LLM. It's designed to go beyond mere language generation and has the potential to revolutionize various industries, including AI and hardware manufacturing. The concerns about Nvidia and the Mag7 might be overblown, but it's worth noting that such innovations could disrupt the status quo. However, it's still in its early stages, and only time will tell its true impact. [...]	0.001

4 Findings (Results)

Evaluation of Personality Detection Results

We further validate the automatically assigned personality scores by identifying five subreddits with the highest mean score and five with the lowest mean score for each trait. This computation was performed separately for human- and model-generated comments for EXT, OPN, NEU, and AGR. We deliberately exclude the CON trait from this analysis, as Conscientiousness is particularly difficult to infer from text: it often reflects internal attributes such as self discipline, organization, and reliability, which do not consistently manifest in explicit surface-level word choices. Also, subreddits with fewer than 50 comments were excluded from the analysis. Figures 1, 2 and 3 present the results for human- and LLM-generated comments. Careful inspection reveals findings that largely align with intuition. Among Redditors, low EXT comments are concentrated in topical threads such as *books*, *OCD*, *poetry*, *journaling*, and *meditation*. Comments with high mean NEU scores appear in *OCD*, *ptsd*, *bipolar*, *newparents*, and *ADHD* discussions. The model results also display plausible patterns, with *simpleliving*, *homeschool*, and *backpacking* notable for low NEU in Claude-Sonnet4.0, and *privacy*, *frugal*, and *tax* for low OPN in Llama3.3-70B. These results suggest that the personality classifier reliably captures the Big Five traits in our data. In the next step, we conduct a comparative analysis of the mean trait levels and their variance across human- and model-written comments.

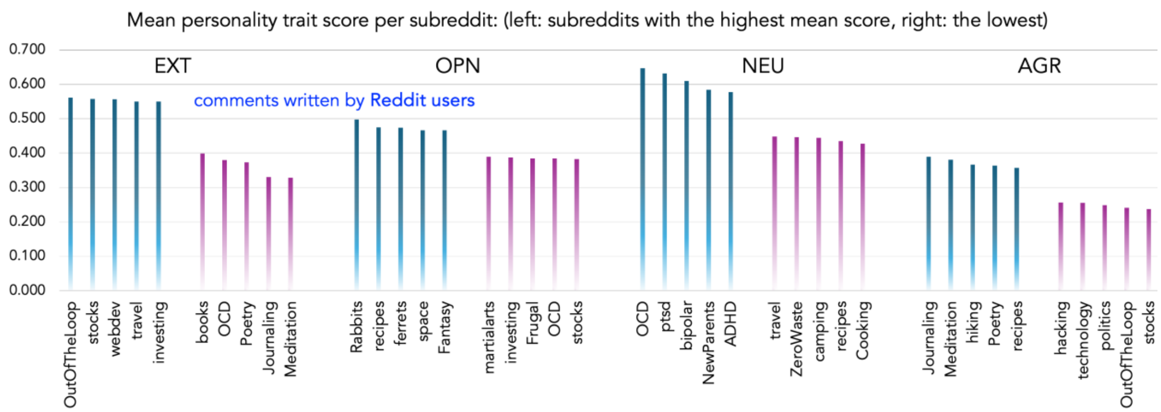

Figure 1. Subreddits exhibiting the highest and lowest mean score per trait in comments produced by human Reddit users.

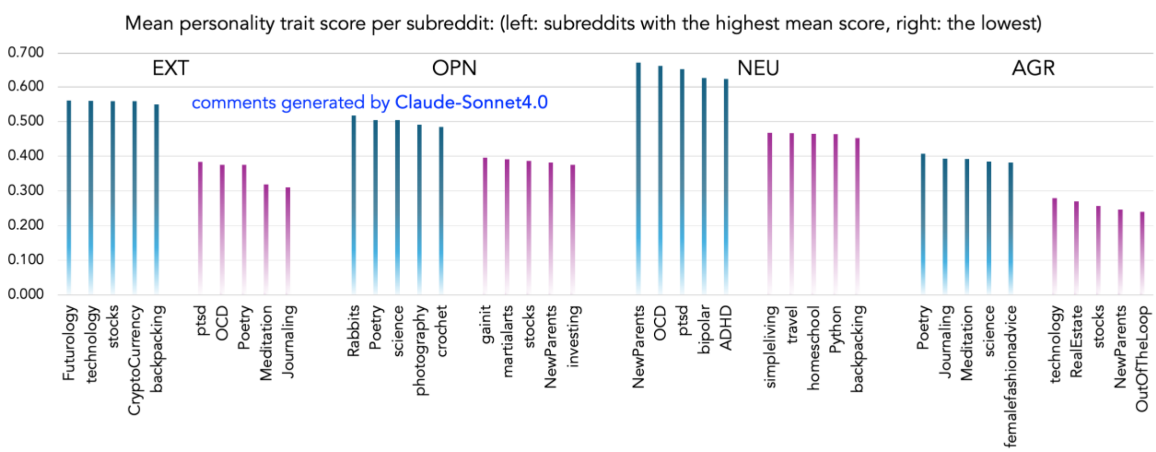

Figure 2. Subreddits exhibiting the highest and lowest mean score per trait in comments produced by the Claude-Sonnet4.0 model.

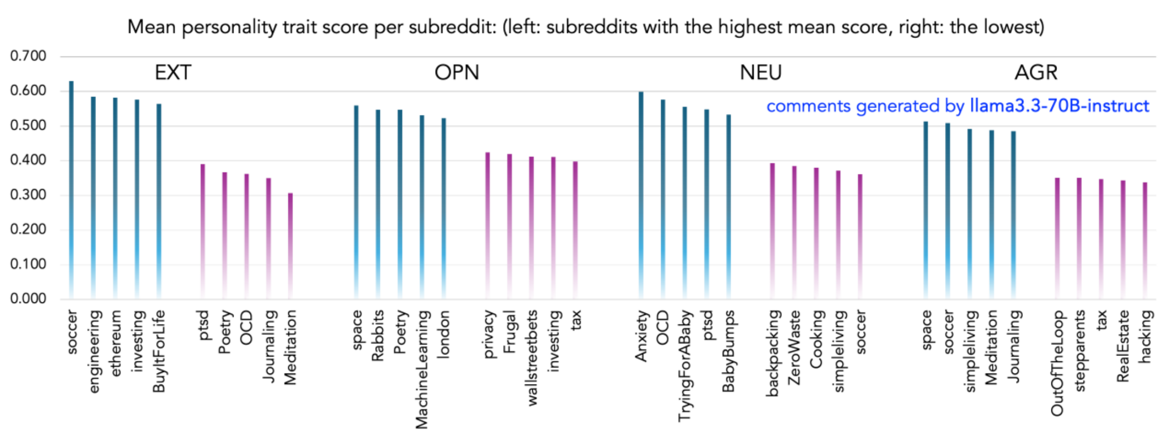

Figure 3. Subreddits exhibiting the highest and lowest mean score per trait in comments produced by the Llama3.3-70B model.

Big Five: Human Authors vs Generative Models

We compute the mean comment score for each of the Big-Five traits in texts written by human authors and those generated by models. Table 3 reports the mean and standard deviation (STD) results. Several insights emerge from these numbers: EXT and OPN mean scores of models are generally comparable to those of human authors, with OPN scores slightly higher. All models exhibit considerably higher AGR scores and lower NEU scores (especially evident in the open models), consistent with prior findings from studies using self-reported questionnaires (see Section 2), and aligning with the intuition that models are trained to be cooperative, psychologically "stable", and agreeable. Indeed, quite a few of our solicited model responses open with phrases such as "Hey, I totally get where you're coming from!", "I'm so glad you shared this [...]", or "I'm so sorry to hear that you're feeling this way". We do report CON scores in Table 3 as well, but refrain from interpreting them. Figures 4 and 5 further illustrate the kernel density distributions of the AGR and NEU traits in sample LLMs compared to human-authored comments. While Claude-Sonnet4.0 shows a distribution similar to that of Reddit authors, Llama3.3-70B exhibits a noticeably higher average, reflected as a right shift. For the NEU trait, the slight left shift of the two sample models reflects their relatively more "stable" nature compared to human writers. Another notable observation in Table 4 is that models show slightly higher STD values than human authors. This may be attributed to the broader range of personalities that models encounter in their training data, compared to the somewhat narrower fraction of the general population active on Reddit. We also observed no significant differences between the two temperature settings: results for t=0.0 are almost identical to those for t=0.7 across all models. Finally, we assess the statistical significance of differences between humans and each model using two tests: the Mann-Whitney test for differences in the underlying distributions [23], and Levene's test for differences in variance [22]. Virtually all comparisons are significant at $p<0.01$, see Table 3 for details.

Table 4. Big Five personality traits mean values (± STD) for different models and human Reddit comments. We do report results for the CON trait here as well, but refrain from interpreting them. One careful observation would be that models show higher level of CON that human do in their comments. While EXT and OPN are generally comparable to those by human authors, considerably higher mean level of AGR and lower level of NEU is evident in text produced by LLMs. Virtually all models results show statistically significant different compared to humans — for both underlying distributions (Mann-Whitney test), and variances (Levene test) at p<0.01. Results with no significant difference are marked with (-).

model name	temp	EXT	OPN	AGR	NEU	CON
Claude-Sonnet4.0	0.0	(-)0.490 (±0.086)	0.440 ((-)±0.066)	0.330 (±0.070)	0.518 (±0.088)	0.340 (±0.063)
Claude-Sonnet4.0	0.7	(-)0.491 (±0.086)	0.440 (±0.066)	0.331 (±0.071)	0.518 (±0.089)	0.341 (±0.063)
GPT4.1	0.0	0.485 (±0.093)	0.444 (±0.070)	0.354 (±0.077)	0.493 (±0.093)	0.342 (±0.070)
GPT4.1	0.7	0.486 ((-)±0.092)	0.443 ((-)±0.069)	0.350 ((-)±0.075)	0.498 (±0.092)	0.342 (±0.068)
GPT4.1-mini	0.0	0.481 (±0.093)	0.453 ((-)±0.069)	0.376 (±0.077)	0.483 (±0.095)	0.365 (±0.072)
GPT4.1-mini	0.7	0.480 ((-)±0.095)	0.454 (±0.069)	0.377 (±0.079)	0.484 (±0.095)	0.366 (±0.073)
Llama3.3-70B	0.0	(-)0.491 (±0.087)	0.476 (±0.071)	0.416 (±0.087)	0.450 (±0.088)	0.401 (±0.083)
Llama3.3-70B	0.7	(-)0.491 (±0.087)	0.477 (±0.072)	0.419 (±0.090)	0.449 (±0.089)	0.405 (±0.086)
Mixtral8x22B	0.0	0.493 (±0.091)	0.474 (±0.072)	0.395 (±0.082)	0.442 (±0.096)	0.380 (±0.078)
Mixtral8x22B	0.7	0.497 (±0.088)	0.484 (±0.071)	0.414 (±0.085)	0.426 (±0.096)	0.398 (±0.082)
Qwen2.5-72B	0.0	0.486 (±0.096)	0.472 (±0.074)	0.400 (±0.081)	0.476 (±0.102)	0.376 (±0.076)
Qwen2.5-72B	0.7	0.485 (±0.096)	0.477 (±0.075)	0.404 (±0.081)	0.475 (±0.100)	0.380 (±0.077)
human authors	—	0.491 (±0.079)	0.425 (±0.061)	0.309 (±0.066)	0.512 (±0.080)	0.315 (±0.057)

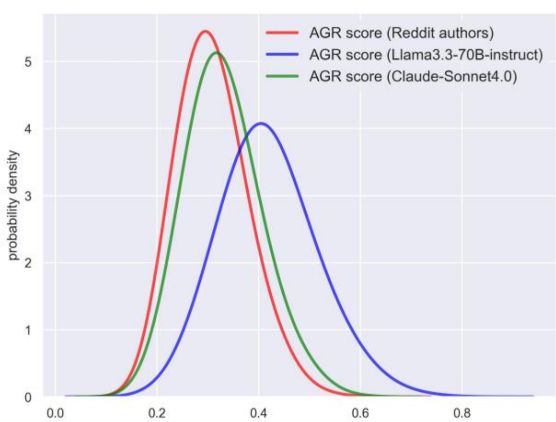

Figure 4. Example kernel density distribution of the AGR score in human comments vs LLMs: differences are evident in both mean and STD.

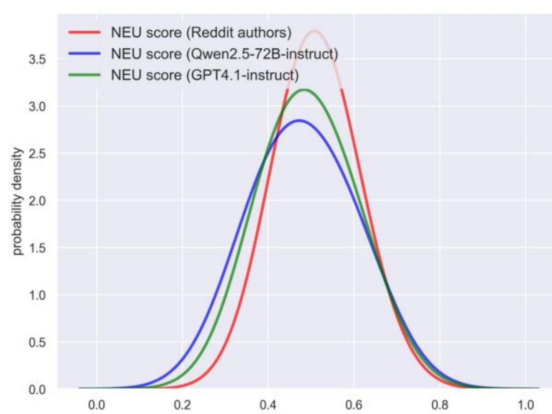

Figure 5. Example kernel density distribution of the NEU score in human comments vs LLMs: differences are evident in both mean and STD.

Evaluation of Gender Classification Results

We further validate the automatically assigned F-likelihood scores by identifying five subreddits with the highest and lowest mean scores. This computation was performed for both human- and model-generated comments. Figure 6 illustrates the results: careful inspection shows that the findings largely align with intuition. Among Redditors, comments likely written by female authors are concentrated in threads such as *namenerds*, *toddlers*, *beyondthebump* (motherhood), *anime*, and *Parenting*. Similarly, LLM-generated comments display plausible gender patterns, with *knitting*, *femalefashionadvise*, *sewing*, and *Cooking* appearing among the subreddits with high F-likelihood. Subreddits with lowF-likelihood scores (i.e., high M-likelihood) are consistently associated with topics like *politics*, *soccer*, *stocks*, and *movies*.

We conclude that F-likelihood score assignments are sufficiently reliable, and perform human- vs models comparative analysis.

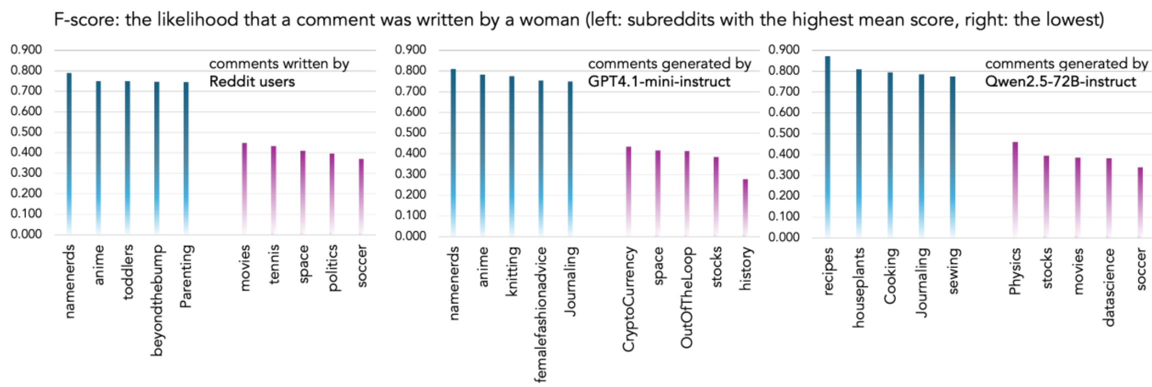

Figure 6. Subreddits exhibiting the highest and the lowest mean F-likelihood score in comments produced by Reddit human authors and models.

Gender: Human Authors vs Generative Models

We compute the mean comment F-likelihood score for texts written by humans and those generated by LLMs. Table 5 reports the mean and standard deviation (STD) results. The models exhibit a range of mean scores around the average F-likelihood of 0.591 observed in human comments: some LLMs show slightly lower averages, while others are slightly higher, with no consistent pattern. A systematic difference is evident in the STD values: models display lower variance, indicating slightly more limited variation in gendered language, consistent with the findings on spambots by Giorgi et al. [11]. As before, we assess the significance of differences between humans and each model using two statistical tests: the Mann-Whitney test for differences in the underlying distributions, and Levene's test for differences in variance. All differences except those of GPT4.1 and Qwen2.5-72B for underlying distributions are significant at $p<0.01$. see Table 5 for details.

Table 5. The mean likelihood that the comment was written by a female (F) author, the ratio of F-authored comments (those with F-likelihood exceeding 0.5), the ratio of M-authored comments. The standard deviation is specified in parenthesis. Generative models are compared to results produced by human users. Virtually all models results show statistically significant difference compared to humans — for both underlying distributions (Mann-Whitney test), and variances (Levene test) at p<0.01. Results with no significant difference are marked with (-).

model	temp	mean F-likelihood	comments' F-ratio	comments' M-ratio
Claude-Sonnet4.0	0.0	0.535 ($\pm$0.282)	0.567	0.433
Claude-Sonnet4.0	0.7	0.532 ($\pm$0.282)	0.558	0.442
GPT4.1	0.0	(-)0.602 ($\pm$0.256)	0.700	0.300
GPT4.1	0.7	(-)0.589 ($\pm$0.260)	0.678	0.322
GPT4.1-mini	0.0	0.580 ($\pm$0.260)	0.667	0.333
GPT4.1-mini	0.7	0.582 ($\pm$0.257)	0.666	0.334
LLama3.3-70B	0.0	0.553 ($\pm$0.242)	0.621	0.379
LLama3.3-70B	0.7	0.548 ($\pm$0.242)	0.612	0.388
Mixtral8x22B	0.0	0.639 ($\pm$0.251)	0.741	0.259
Mixtral8x22B	0.7	0.628 ($\pm$0.251)	0.728	0.272
Qwen2.5-72B	0.0	(-)0.600 ($\pm$0.264)	0.700	0.300
Qwen2.5-72B	0.7	(-)0.612 ($\pm$0.262)	0.712	0.287
human authors	—	0.591 ($\pm$0.300)	0.656	0.344

5 Conclusion and Future Work

In this study, we examined the personality and gender characteristics of texts produced by contemporary LLMs in comparison to human written comments on Reddit. Using established personality and gender classifiers, we analyzed thousands of posts and comments, observing both similarities and systematic differences. Our results indicate that models can capture many human-like patterns for traits such as Extroversion and Openness, while systematically exhibiting higher Agreeableness and lower Neuroticism, reflecting their cooperative and psychologically stable training objectives. Similarly, gendered language in model-generated text broadly aligns with human patterns, though models show slightly reduced variation, echoing previous observations in social spambots. Overall, these findings suggest that current LLMs can produce text that mirrors some aspects of human personality and demographics, while also highlighting consistent divergences that reflect model design and training biases. The methodology presented here, combining large-scale data collection for comparative analysis and automatic trait assessment, provides a framework for future studies to further explore the nuances of personality and demographic cues in generative models. These insights can inform both the development of more human-like AI and the critical evaluation of its social and psychological implications.

6 References

[1] W. Ahmed, P. A. Bath, and G. Demartini. Using twitter as a data source: An overview of ethical, legal, and methodological challenges. The ethics of online research, 2:79–107, 2017.

[2] M. AI. Mixtral 8×22b. https://mistral.ai/news/mixtral-of-experts, Apr. 2024. Sparse Mixture-of-Experts model (141B parameters, 39B active), released April 10, 2024. Apache 2.0 license.

[3] M. AI. Llama 3.3-70b (instruction-tuned). https://cloud.google. com/vertex-ai/generative-ai/docs/partner-models/llama/llama3-3, Apr. 2025. Instruction-tuned text-only 70B-parameter model. General Avail- ability release April 29, 2025.

[4] Anthropic. Claude sonnet 4. https://www.anthropic.com/news/ claude-4, May 2025. Large language model. Released May 22, 2025. Available via Anthropic API and platforms such as Vertex AI and Amazon Bedrock.

[5] A. Benton, G. Coppersmith, and M. Dredze. Ethical research protocols for social media health research. In Proceedings of the first ACL workshop on ethics in natural language processing, pages 94–102, 2017. URL https://aclanthology.org/W17-1612/.

[6] P. Bhandari, U. Naseem, A. Datta, N. Fay, and M. Nasim. Evaluating personality traits in large language models: Insights from psychological questionnaires. In Companion Proceedings of the ACM on Web Conference 2025, pages 868–872, 2025.

[7] H. E. Cattell. The sixteen personality factor (16pf) questionnaire. In Un- derstanding psychological assessment, pages 187–215. Springer, 2001.

[8] J. Coates and P. Pichler. Language and gender. A Reader, 1998.

[9] B. De Raad. The big five personality factors: the psycholexical approach to personality. Hogrefe & Huber Publishers, 2000.

[10] F. Dorner, T. Sühr, S. Samadi, and A. Kelava. Do personality tests generalize to large language models? In Socially Responsible Language Modelling Research, 2023.

[11] S. Giorgi, L. Ungar, and H. A. Schwartz. Characterizing social spam-bots by their human traits. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 5148–5158, 2021.

[12] L. R. Goldberg. The development of markers for the big-five factor structure. Psychological assessment, 4(1):26, 1992.

[13] D. M. Greenberg, S. Baron-Cohen, D. J. Stillwell, M. Kosinski, and P. J. Rentfrow. Musical preferences are linked to cognitive styles. PloS one, 10(7):e0131151, 2015.

[14] A. Gupta, X. Song, and G. Anumanchipalli. Self-assessment tests are unreliable measures of llm personality. arXiv preprint arXiv:2309.08163, 2023.

[15] Y. HaCohen-Kerner. Survey on profiling age and gender of text authors. Expert Systems with Applications, 199:117140, 2022.

[16] J. Hartley, C. Hamill, D. Batra, D. Seddon, R. Okhrati, and R. Khraishi. How personality traits shape llm risk-taking behaviour. arXiv preprint arXiv:2503.04735, 2025.

[17] G. Jagfeld, F. Lobban, P. Rayson, and S. H. Jones. Understanding who uses reddit: Profiling individuals with a self-reported bipolar disorder diagnosis. arXiv preprint arXiv:2104.11612, 2021. URL https://arxiv. org/pdf/2104.11612.pdf.

[18] G. Jiang, M. Xu, S.-C. Zhu, W. Han, C. Zhang, and Y. Zhu. Evaluating and inducing personality in pre-trained language models. Advances in Neural Information Processing Systems, 36:10622–10643, 2023.

[19] H. Kotek, R. Dockum, and D. Sun. Gender bias and stereotypes in large language models. In Proceedings of the ACM collective intelligence conference, pages 12–24, 2023.

[20] W. Labov. The intersection of sex and social class in the course of linguistic change. Language variation and change, 2(2):205–254, 1990.

[21] R. Lakoff. Language and woman's place. Language in society, 2(1): 45–79, 1973.

[22] H. Levene. Robust tests for equality of variances. In Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling, pages 278–292. Stanford University Press, 1960.

[23] H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. The annals of mathematical statistics, pages 50–60, 1947.

[24] M. R. Mehl, S. D. Gosling, and J. W. Pennebaker. Personality in its natural habitat: manifestations and implicit folk theories of personality in daily life. Journal of personality and social psychology, 90(5):862,2006.

[25] W. T. Norman. Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. The journal of abnormal and social psychology, 66(6):574, 1963.

[26] OpenAI. Gpt-4.1. https://openai.com/index/gpt-4-1/, Apr. 2025. Large language model. Available via OpenAI API.

[27] H. Peters and S. C. Matz. Large language models can infer psychological dispositions of social media users. PNAS nexus, 3(6):page231,2024.

[28] B. d. E. Raad and M. E. Perugini. Big five factor assessment: Introduction. 2002.

[29] A. Salecha, M. E. Ireland, S. Subrahmanya, J. Sedoc, L. H. Ungar, and J. C. Eichstaedt. Large language models show human-like social desirability biases in survey responses. arXiv preprint arXiv:2405.06058, 2024.

[30] G. Serapio-García, M. Safdari, C. Crepy, L. Sun, S. Fitz, M. Abdulhai, A. Faust, and M. Mataric. Personality traits in large language models.´2023.

[31] E. Shem-Tov and E. Rabinovich. Exploring the interplay between musical preferences and personality through the lens of language, 2025. URL https://arxiv.org/abs/2508.18208.

[32] A. Sorokovikova, N. Fedorova, S. Rezagholi, and I. P. Yamshchikov. Llms simulate big five personality traits: Further evidence. arXiv preprint arXiv:2402.01765, 2024.

[33] S. Soundararajan and S. J. Delany. Investigating gender bias in large language models through text generation. In Proceedings of the 7th international conference on natural language and speech processing (icnlsp 2024), pages 410–424, 2024.

[34] X. Sun, B. Liu, J. Cao, J. Luo, and X. Shen. Who am i? personality detection based on deep learning for texts. In 2018 IEEE international conference on communications (ICC), pages 1–6. IEEE, 2018.

[35] A. C. Q. Team. Qwen 2.5 72 b. https://openlaboratory.ai/models/qwen-2_5-72b, Sept. 2024. Dense decoder-only LLM (72.7B parameters), 128 k-token context window; multilingual model excelling at coding, reasoning, and structured data tasks. Released September 2024.

[36] L. Wang, N. Yang, X. Huang, B. Jiao, L. Yang, D. Jiang, R. Majumder, and F. Wei. Text embeddings by weakly-supervised contrastive pre-training. arXiv preprint arXiv:2212.03533, 2022.

[37] M. Zhu, Y. Weng, L. Yang, and Y. Zhang. Personality alignment of large language models. arXiv preprint arXiv:2408.11779, 2024.

7 תקציר בעברית

עבודה זו בוחנת את מאפייני ה**אישיות** וה**מגדר** של טקסטים שנוצרו על ידי **מודלי שפה גדולים (LLMים פתוחים וסגורים)** בהשוואה לתגובות שנכתבו על ידי בני אדם ברשת Reddit.

<u>שיטה וממצאים עיקריים</u>

בניגוד למחקרים קודמים שבחנו אישיות של LLMs באמצעות שאלונים (כפי שנעשה עם בני אדם) , בעבודה זו נקטנו **גישה בלתי מוטה(unbiased approach)** . השיטה התבססה על **זיהוי אוטומטי** של תכונות האישיות (OCEAN) ורמזים דמוגרפיים (מגדר) **מתוך** השפה הגנרטיבית והספונטנית שהמודלים יצרו .לצורך כך, נאספו שאלות פתוחות מקהילות Reddit, ועליהן השיבו הן משתמשי Reddit והן מודלי LLMכאילו היו משתמשים ברשת חברתית.

השתמשנו במסווגי אישיות ומגדר מקובלים כדי לנתח אלפי תגובות לשאלות פתוחות, שאותן אספנו מקהילות שונות ב Reddit.

אישיות (מודל "חמשת הגדולים" – OCEAN)

מודלי השפה הפגינו יכולת ללכוד היבטים דומים לדפוסים אנושיים בתכונות כמו **מוחצנות (EXT - Extroversion)** ו**פתיחות (Openness – OPN).** עם זאת, נמצאו הבדלים שיטתיים ועקביים:

- **נעימות: (AGR- Agreeableness)** המודלים הפגינו **רמה גבוהה יותר** של נעימות.
- **נוירוטיות: (NEU - Neuroticism)** המודלים הראו **רמה נמוכה יותר** של נוירוטיות.

ממצאים אלו עולים בקנה אחד עם מחקרים קודמים ומשקפים את מטרות האימון של המודלים, אשר מכוונות אותם להיות **שיתופיים** ו**יציבים פסיכולוגית**.

מאפיינים דמוגרפיים (מגדר)

השפה המגדרית בטקסט שנוצר על ידי המודלים **התאימה באופן כללי לדפוסים האנושיים**.

- **שונות מוגבלת:** המודלים הציגו **שונות מופחתת מעט** בשפה המגדרית שלהם בהשוואה לבני אדם, ממצא המהדהד תצפיות קודמות על ספאמבוטים חברתיים.

<u>מסקנה</u>

העבודה מסיקה כי מודלי שפה גדולים יכולים ליצור טקסט המחקה היבטים מסוימים של אישיות ודמוגרפיה אנושית. יחד עם זאת, ההבדלים העקביים שנמצאו (כגון AGR גבוה יותר ו NEU-נמוך יותר) מדגישים **הטיות (Biases)** הנובעות מעיצוב המודל ומאופן אימונו.

המתודולוגיה שהוצגה משמשת מסגרת להמשך חקירת ניואנסים של אישיות ורמזים דמוגרפיים במודלים גנרטיביים.

המכללה האקדמית תל-אביב

בית הספר למדעי החשב

Who are you, ChatGPT?

Personality and Demographic Style in LLM-Generated Content

חיבור זה הוגש כחלק מהדרישות לקבלת התואר "מוסמך " – M.Sc.

במכללה האקדמית תל-אביב

על ידי

דנה סוטו פורת

העבודה הוכנה בהדרכתה של

ד"ר אלה רבינוביץ