Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 10
ISET 2025

# Evaluating Human-Like Qualities in Language Models

Ruijie Liu

Concord College International School, Kuala Lumpur, Malaysia

cedric071010@gmail.com

**Abstract.** This paper investigates the human-like communication abilities of modern language models, comparing several open-source and proprietary systems. As LLMs are increasingly deployed in socially interactive roles—ranging from digital companions to mental health support tools—their ability to engage users naturally and expressively has become a critical yet underexplored dimension of evaluation. Traditional benchmarks tend to emphasize accuracy or reasoning, but they fail to capture the nuanced, subjective traits that define human conversation. To address this, seven LLMs were tested using both short and sustained dialogues, evaluated by five human raters using a multi-trait rubric. LLaMA 3.2 emerged as a standout, occasionally outperforming human responses in personality and creativity. Models were assessed on five human-oriented communication traits: naturalness, empathy, creativity, adaptability, and humor/personality. Results show significant variation across systems, with some matching or exceeding human performance in specific areas—suggesting that conversational quality may depend more on tuning and stylistic freedom than model scale alone.

**Keywords:** Conversational AI, Empathy, Language Models, Benchmark.

## 1. Introduction

Large Language Models (LLMs) are primarily known for generating text in response to text-based input. While many benchmarks focus on their ability to solve math problems or handle logical reasoning tasks, these are not necessarily their core strengths. LLMs are fundamentally trained on large amounts of text to produce fluent, natural-sounding language—built for conversational interactions. At their core, these models operate by estimating the probability distribution of the next token in a sequence, conditioned on the preceding context. This makes them potentially powerful tools for mental support, emotional engagement, and entertainment. However, major companies have increasingly restricted these models from expressing significant personality or emotion, effectively suppressing their human-like qualities to maintain them as tools rather than conversational agents.

In addition, most previous evaluations have relied on multiple-choice testing or isolated prompt completions assessed for factuality or correctness. Traits such as natural engagement, empathy, and personality—core aspects of human communication—have rarely been the focus of systematic evaluation. Yet these qualities are essential when LLMs are applied in interactive domains, where trust, reliability, and emotional nuance significantly impact user experience. Without assessing these traits, an incomplete understanding of LLM capabilities persists.

In this study, the human-like communicative capacities of several LLMs are examined, with a focus on their ability to engage naturally, express empathy, and exhibit personality. Responses to specifically designed prompts measuring naturalness, creativity, and emotional adaptability are collected and evaluated. Each model receives a shared instruction prompt, followed by a set of short-form questions, and a simulated dialogue task. All responses are recorded in separate text files labeled with a letter randomly assigned and later assessed by human evaluators using a band score from 0 to 9, across five traits.

**Table 1.** System assignments by label.

| Label | System |
|---|---|
| A | Qwen 3 14B |
| B | LLaMA 3.2 Vision 11B |
| C | Claude 3.7 |
| D | Human (for reference) |
| E | ChatGPT-4o |
| F | Mistral 7B Instruct v0.3 |
| G | DeepSeek R1 14B |
| H | Gemini |

## 2. Literature Review

The evaluation of language models has traditionally emphasized logical reasoning, factual accuracy, and task performance. Benchmarks such as MMLU, ARC, and TruthfulQA dominate the discourse, offering quantifiable ways to assess correctness and general knowledge [1-3]. However, these measures often ignore the subtleties of human-like interaction—elements such as empathy, humor, and contextual adaptability—that are increasingly essential in real-world applications like education, mental health, and digital companionship [4].

Several studies have acknowledged this gap. For instance, the Empathetic Dialogues dataset [5] was one of the first attempts to benchmark emotional understanding in AI, focusing on responses that reflect and validate the speaker's emotional state. More recently, OpenAI's InstructGPT and GPT-4 models were fine-tuned not just for correctness, but for alignment with human preferences [6][7]. While this improves general tone and helpfulness, critics argue that over-alignment can suppress individuality or creativity, leading to outputs that feel generic or emotionally flat [8].

The notion of "human-likeness" in AI has also been explored through the lens of personality modeling [9] and stylistic mimicry. Some researchers advocate for evaluating conversational AI on traits closer to those used in psychology or literary analysis—naturalness, expressiveness, or emotional resonance—rather than only technical accuracy [10]. Still, there remains no standard framework for scoring these qualities in a scientifically controlled manner.

This paper contributes to this emerging discourse by proposing a multi-trait, human-centered evaluation of popular language models, bridging the gap between performance benchmarks and experiential interaction. It builds on earlier alignment research while explicitly focusing on user perception and emotional realism, rather than factual reliability alone.

## 3. Methodology

This study evaluates the human-like qualities of selected language models, as listed in Table 1, through prompt-based testing and multi-rater scoring.

Seven mainstream language models were tested, including a combination of locally-run open-source models and proprietary models accessed through online interfaces. A human-generated response was also included (under label D) to serve as a reference point for authentic human communication. This helped ensure the scoring rubric was realistic and that author could calibrate participants' evaluations against a true human example. A random integer generator was used to assign labels A to H to the eight systems listed in Table 1, ensuring that the labeling had minimal influence on human scoring. The evaluation was divided into two parts, which proved useful in highlighting that strong performance in one context did not guarantee strong performance in another.

In Part 1, the focus is to understand the ability for the models to react to different situation separately. They were given the same set of 12 short prompts, preceded by the initial instruction: *Respond to the following like a close friend in short sentence(s) please.* For each question, local LLMs were restarted,

and online LLMs were opened in separate chat sessions to ensure responses were not influenced by prior prompts.

In Part 2, the focus is to understand the ability for the models to stay in a conversation and keep it going. A conversational setup was used: each model was opened in two windows and given the initial instruction. The response from one window was then passed to the other as the next input, and this exchange was repeated until ten total messages were collected. This tested consistency, tone maintenance, and longer-term conversational coherence.

All output data was saved in eight separate text files (one per label). A shared evaluation rubric was created to guide the scoring process and ensure that all participants followed the same criteria. Each participant received anonymized outputs labeled A to H, a scoring table with a comments field for qualitative feedback. Five individuals, including the author, participated in the evaluation. To avoid bias, blind scoring is used to make the systems anonymized using random letter labels, and the author completed scoring before accessing results from other participates to avoid influence. The scoring rubric measured five categories: naturalness, empathy, creativity, adaptability, and humor/personality. Scores were given on a 0 to 9 scale. Final scores were averaged across all raters and rounded to the nearest whole number to avoid false precision.

All open-source LLMs were run locally on Windows 11 using the Ollama framework. The five evaluators were from the Gen Z age group, and scoring patterns were notably consistent across all five, with only minor variation. Qualitative comments provided additional insights into tone, phrasing, and personality that were not always reflected by numeric scores alone.

**Table 2.** Band scores (0--9) across five human-like traits.

| Label | Naturalness | Empathy | Creativity | Adaptability | Humor |
|---|---|---|---|---|---|
| A | 7 | 6 | 7 | 8 | 7 |
| B | 9 | 8 | 8 | 9 | 9 |
| C | 6 | 6 | 2 | 5 | 2 |
| D | 9 | 8 | 8 | 9 | 8 |
| E | 8 | 7 | 7 | 8 | 7 |
| F | 3 | 5 | 4 | 6 | 3 |
| G | 6 | 7 | 5 | 6 | 5 |
| H | 8 | 7 | 6 | 7 | 6 |

## 4. Results

The results of the evaluation are presented in Table 2. Each label was assessed by five participants across five categories: *naturalness*, *empathy*, *creativity*, *adaptability*, and *humor/personality*, using a 0--9 band scoring rubric. Letter labels (A to H) were used to anonymize identity. In addition to numerical scores, participants were encouraged to provide comments explaining their impressions. These insights proved useful for interpreting unexpected patterns in the data and for capturing nuances not reflected in the scores alone.

### 4.1. Trait-by-Trait Analysis

**Naturalness.** As shown in Table 2, label B (LLaMA 3.2) and label D (Human) received the highest score of 9, indicating that their responses were perceived as highly fluent, smooth, and conversational. Participants described label B's responses as "very natural," with little sign of robotic phrasing. Label D, representing a real human, was also rated highly, although some noted it felt "almost too simple," suggesting that naturalness alone does not guarantee engagement. Label F (Mistral) scored lowest (3) due to long and erratic responses that "can't chat normally" according to comments. This is strongly reflected in the dialogue sequences, where the conversation stayed at basic greetings without progressing further. This may also suggest that models with too few parameters may not perform as

well since Label F (Mistral) has the minimum parameters among the LLMs with known size used in this study.

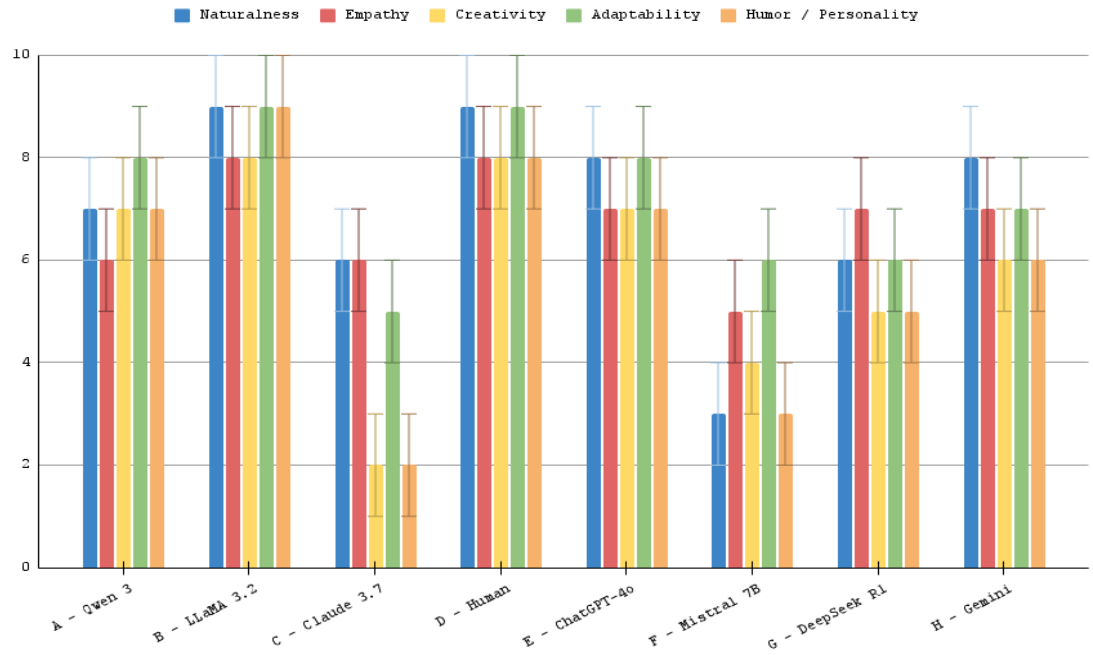


**Fig. 1.** Column chart comparison of LLM trait scores across labels.

**Empathy.** As shown in Table 2, scores for empathy were relatively consistent, with most labels falling between 6 and 8. The human (label D) and LLaMA 3.2 (label B) again tied at the top with scores of 8, showing emotional awareness and appropriate tone. Label F again underperformed (5) reinforcing the hypothesis that its relatively small parameter size may have contributed to its weaker performance, showing less sensitivity in emotionally oriented prompts. Notably, Claude 3.7 (label C) also received only a 6, despite its reputation for alignment---possibly due to its overly formal, instructional tone that felt detached.

**Creativity.** As shown in Table 2 and Fig. 1, creativity showed the widest spread in scores, from 2 to 8. The human (label D) and LLaMA 3.2 (label B) both scored 8, offering original, imaginative responses. In contrast, Claude (label C) scored a mere 2, described as “feel[ing] like it was explaining everything” rather than engaging creatively. DeepSeek (label G) and Gemini (label H) both landed at 5, indicating middling performance---technically competent but lacking distinctive flair.

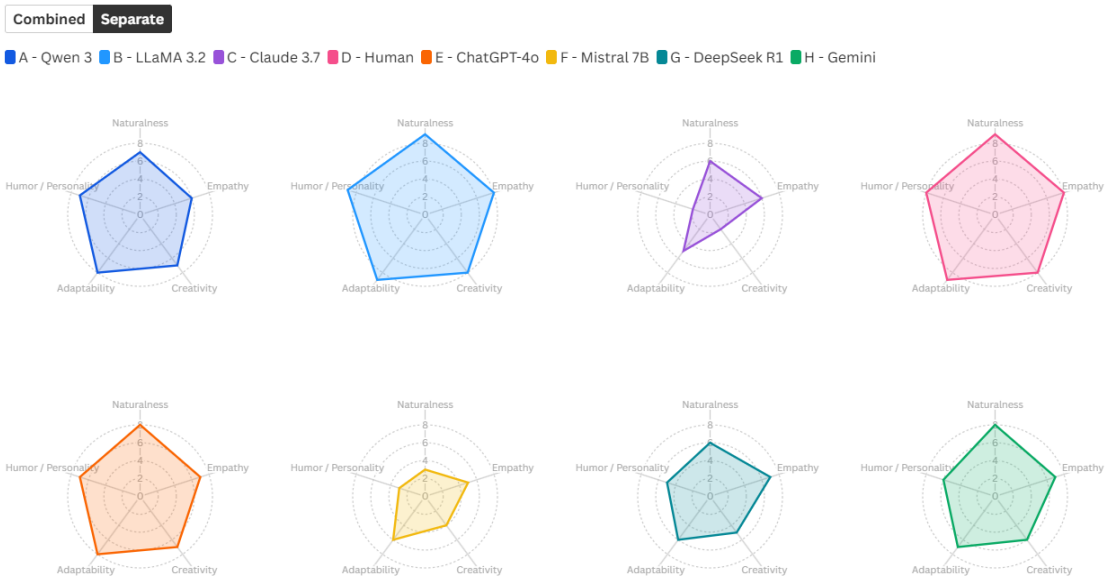


**Fig. 2.** Separated radar chart comparison of LLM trait scores across labels.

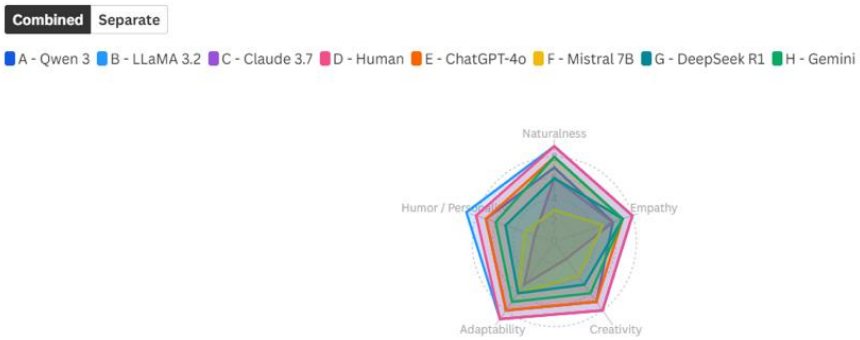


**Fig. 3.** Combined radar chart comparison of LLM trait scores across labels.

**Adaptability.** As shown in Table 2 and the radar charts (Fig. 2 and Fig. 3), labels B and D again led with scores of 9, demonstrating a strong ability to shift tone depending on the task---whether explaining to a child or speaking empathetically. Claude (label C) scored 5, as did DeepSeek (label G), which "answered questions well but couldn't chat normally." Although it performed adequately in short-form responses, it failed to develop spontaneous or contextually rich exchanges in the multi-turn conversation task. Label F received a 6, possibly lifted slightly by its ability to follow varied instructions, even if the delivery remained inconsistent.

**Humor / Personality.** As shown in Table 2, personality proved difficult for most labels. LLaMA 3.2 (label B) was the standout with a score of 9, praised for being expressive, fun, and stylistically engaging. While the human (label D) scored 8, This may also be due to the human conversation being described as "short and simple," which could have limited the perceived depth of personality in comparison to more expressive model responses. Claude (label C) again underperformed significantly, scoring only 2, with responses that felts robotic and showed little to no charm. Mistral and DeepSeek (label G) also struggled in this area, scoring 3 and 5 respectively.

### 4.2. Overall Observations

As shown in Fig. 2, label B (LLaMA 3.2) consistently scored at or near the top in all five traits, closely matching or even exceeding the human benchmark represented by label D. ChatGPT-4o (label E) performed solidly across all traits with no major weaknesses, averaging 7--8 across the board. Gemini (label H) was also well-rounded but received comments suggesting it was "holding back," implying a degree of constraint in its tone or output.

Claude (label C), while often praised in industry circles for safety and alignment, scored lower than expected in personality and creativity, possibly due to its formal tone and instruction-heavy style. DeepSeek and Mistral both exhibited uneven performance, with some prompts answered well but others causing disjointed, overly long, or overly mechanical responses. The response under label D (Human) was seen as natural and emotionally grounded, but its simplicity may have slightly limited its personality score.

As shown in the separated radar chart (Fig. 2), most models exhibit a roughly pentagonal shape, indicating balanced performance across the five human-like traits. This suggests a degree of interrelation among the evaluated characteristics—naturalness, empathy, creativity, adaptability, and humor/personality—where strengths in one area may be indicative of strengths in others. Notably, as overall model performance declines, the shape of the radar plot becomes more irregular and fragmented. This pattern is especially evident in models C (Claude 3.7) and F (Mistral), whose lower trait scores result in distorted, lopsided shapes, implying inconsistency and underperformance across multiple dimensions. In contrast, labels B (LLaMA 3.2) and D (Human) produce the most symmetrical and expansive radar shapes, reinforcing their strong and balanced performance across all five evaluated traits. The visual geometry of the radar plots thus provides an intuitive complement to the numerical scores, highlighting both balance and deficiency in human-like qualities.

These results, illustrated in Table 2, Fig. 1 and Fig. 3, reveal clear differences in how each system expresses human-like qualities, with LLaMA 3.2 and ChatGPT-4o standing out as particularly conversational and emotionally aware. In the following section, the implications of these findings will be explored further, including what "human-likeness" means in the context of LLM interaction and how qualitative comments add nuance beyond numerical scores.

## 5. Discussion

The results show substantial variation across the evaluated systems. It is evident that neither model size nor developer reputation guarantee human-like output. Some smaller or open-source systems demonstrated surprising emotional intelligence or creativity, while some larger, more restricted systems appeared overly formal or flat. These differences are significant when considering applications in entertainment or mental health support, where sounding natural and emotionally aware is essential.

### 5.1. Restating the Objective

This study aimed to assess how convincingly different LLMs could mimic human conversational behavior. Rather than focusing on traditional benchmarks like arithmetic or factual recall, the evaluation centered on human traits: naturalness, empathy, creativity, adaptability, and personality. These qualities are especially important for LLMs being used in socially interactive contexts, such as virtual companions, digital assistants, or therapeutic tools.

### 5.2. Unexpected Outcomes and Model Dynamics

One of the most surprising outcomes was the performance of label B, LLaMA 3.2. Despite being an open-source model with only 11 billion parameters, it scored at or near the top across every trait. It tied with the human benchmark (label D) in several categories and even exceeded it in humor and personality. Participants described its tone as natural, creative, and emotionally aware—traits not commonly associated with open-source models.

In contrast, proprietary models like Claude 3.7 (label C) and Gemini (label H) underperformed in some human-centric tasks. Claude, for example, frequently responded with formal, instructional language. This alignment-focused style, while useful in structured settings, felt too robotic or detached in casual conversation. Gemini was competent but received comments suggesting that it was “holding back,” possibly a result of internal safety constraints.

### 5.3. Trait-Specific Reflections

Different models demonstrated different strengths. For instance, ChatGPT-4o (label E) was balanced across all categories, suggesting robust fine-tuning. Mistral (label F) and DeepSeek (label G), however, were more inconsistent—performing well in adaptability or empathy, but scoring lower in creativity and personality due to either repetitive phrasing or overly long outputs. This unevenness suggests that certain traits are harder to maintain simultaneously without specific optimization.

### 5.4. Human Benchmarking and Calibration

The inclusion of a human response (label D) proved critical. While it scored highly in most categories, some raters noted that its answers were “too short” or “overly simple.” This shows that even genuine human communication does not always outperform machines—especially when brevity or ambiguity is misread as lack of personality. However, label D helped anchor the scoring system, providing a baseline for evaluating what naturally human communication sounds like.

### 5.5. What These Differences Suggest

These results imply that human-likeness in LLMs is more influenced by expressive freedom and conversational training than by scale or brand. While model size and backend infrastructure impact

fluency and speed, they do not guarantee emotional nuance or adaptability. In particular, models that scored well in humor and creativity—like LLaMA 3.2—appeared more flexible, suggesting that openness to diverse output styles plays a key role in perceived "personality." However, the case of Mistral (Label F) also suggests that models with over small parameter sizes may struggle to achieve well-rounded performance, particularly in traits requiring expressive or adaptive language generation. These observations underscore the importance of not only architectural scale, but also alignment strategies and freedom of stylistic expression when optimizing LLMs for human-like interaction.

### 5.6. Limitations

Several factors may have affected the outcome. All five evaluators were from the same age group (Generation Z) and shared similar linguistic and cultural backgrounds. This demographic homogeneity may introduce bias in the interpretation of subjective traits such as humor and empathy. Additionally, the evaluation was conducted exclusively in English and was limited to two types of interaction: short-form prompts and multi-turn dialogue. Other important areas—such as multilingual performance or emotionally intense storytelling remain unexplored. The sample size of both evaluators and prompts was relatively small, consisting of only five participants, twelve static prompts, and ten-turn chat simulations per model. Furthermore, all language models used in this study were general-purpose systems not specifically optimized for open-ended social conversation. Their underlying training objectives may prioritize factual accuracy, code generation, or safety alignment over personality expression, which could have influenced their performance in certain traits.

### 5.7. Future Directions

Future studies could expand this evaluation framework to include more diverse participants, larger sample size or even more languages. It may also be valuable to study how models behave when fine-tuned specifically for emotionally intelligent responses. With the rise of multi-model LLMs, future research could test human-likeness not just through text, but also through timing, tone.

## 6. Conclusion

This study evaluated the human-like qualities of large language models using a five-trait rubric: naturalness, empathy, creativity, adaptability, and humor/personality. Eight systems—including seven AI models and one human baseline—were assessed in both short-form and multi-turn dialogues. The results demonstrated considerable variation in how convincingly each system could simulate human conversation.

LLaMA 3.2 (label B) emerged as the most human-like system overall, often outperforming both proprietary models and the human reference in perceived creativity and expressiveness. Its strong showing suggests that open-source models, when properly aligned and unconstrained, can match or exceed the conversational nuance of commercial counterparts. ChatGPT-4o also performed well across all traits, while Claude 3.7 and Gemini, though competent, were hindered by a more formal and risk-averse tone.

The inclusion of a human response offered a valuable calibration point, revealing that even authentic dialogue can be perceived as underwhelming if brevity or subtlety is mistaken for lack of personality. This highlights the importance of context and delivery in shaping user perception of "humanness."

These findings suggest that emotional realism and stylistic variation—rather than raw model size—are key to building more relatable AI systems. As LLMs continue to evolve into agents, companions, and entertainers, future research should explore cross-linguistic traits, multi-model delivery, and the role of fine-tuning in shaping human-centric performance.

## References


[1] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, et al. “Training Verifiers to Solve Math Word Problems”. In: arXiv preprint arXiv:2109.08593 (2021).

[2] Dan Hendrycks, Collin Burns, Saurav Kadavath, et al. “Measuring Massive Multitask Language Understanding”. In: arXiv preprint arXiv:2009.03300 (2021).

[3] Stephanie Lin, Jacob Hilton, and Owain Evans. “TruthfulQA: Measuring How Models Mimic Human Falsehoods”. In: arXiv preprint arXiv:2109.07958 (2022).

[4] Heung-Yeung Shum, Xiaodong He, and Di Li. “From Eliza to XiaoIce: Challenges and Opportunities with Social Chatbots”. In: Frontiers of Information Technology Electronic Engineering 19.1 (2018), pp. 10–26.

[5] Hannah Rashkin et al. “Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset”. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (2019).

[6] OpenAI. “GPT-4 Technical Report”. In: arXiv preprint arXiv:2303.08774 (2023).

[7] Long Ouyang, Jeff Wu, Xu Jiang, et al. “Training Language Models to Follow Instructions with Human Feedback”. In: Advances in Neural Information Processing Systems: 27730-27744. (2022).

[8] Cheng-Yu Chiang, Shiyue Cao, and Inioluwa Deborah Raji. “You Can’t Learn if You Don’t Look: Auditing LLMs for Social Bias and Alignment Tradeoffs”. In: arXiv preprint arXiv:2306.09301 (2023).

[9] Gabriel Volkel, Maarten Sap, Chandra Bhagavatula, et al. “Learning Interpretable Personality Traits from Conversations”. In: Findings of the Association for Computational Linguistics: EMNLP 2021 (2021).

[10] Yi-Lin Chiu, Shuyang Gao, and Noah A. Smith. “Stylistic Control for Empathetic Response Generation”. In: arXiv preprint arXiv:2211.08910 (2022).