

Eliciting Personality Traits in Large Language Models

AIRLIE HILLIARD* and CRISTIAN MUNOZ*, Holistic AI, The United Kingdom

ZEKUN WU, Holistic AI, The United Kingdom and University College London, The United Kingdom

ADRIANO SOARES KOSHIYAMA, Holistic AI, The United Kingdom

Large Language Models (LLMs) are increasingly being utilized by both candidates and employers in the recruitment context. However, with this comes numerous ethical concerns, particularly related to the lack of transparency in these "black-box" models. Although previous studies have sought to increase the transparency of these models by investigating the personality traits of LLMs, many of the previous studies have provided them with personality assessments to complete. On the other hand, this study seeks to obtain a better understanding of such models by examining their output variations based on different input prompts. Specifically, we use a novel elicitation approach using prompts derived from common interview questions, as well as prompts designed to elicit particular Big Five personality traits to examine whether the models were susceptible to trait-activation like humans are, to measure their personality based on the language used in their outputs. To do so, we repeatedly prompted multiple LMs with different parameter sizes, including Llama-2, Falcon, Mistral, Bloom, GPT, OPT, and XLNet (base and fine tuned versions) and examined their personality using classifiers trained on the myPersonality dataset. Our results reveal that, generally, all LLMs demonstrate high openness and low extraversion. However, whereas LMs with fewer parameters exhibit similar behaviour in personality traits, newer and LMs with more parameters exhibit a broader range of personality traits, with increased agreeableness, emotional stability, and openness. Furthermore, a greater number of parameters is positively associated with openness and conscientiousness. Moreover, fine-tuned models exhibit minor modulations in their personality traits, contingent on the dataset. Implications and directions for future research are discussed. ¹

CCS Concepts: • **Computing methodologies** → **Artificial intelligence**.

Additional Key Words and Phrases: Artificial Intelligence, Natural Language Processing, Personality Traits Assessment, Large Language Models, Big Five Personality Traits, Text Analysis and Generation, Ethical Implications of AI, Machine Learning Classifiers, Model Transparency and Interpretability, Fine-Tuning in Language Models, Comparative Analysis of LLMs

ACM Reference Format:

Airlie Hilliard, Cristian Munoz, Zekun Wu, and Adriano Soares Koshiyama. 2024. Eliciting Personality Traits in Large Language Models. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*, June 3–6, 2024, Rio de Janeiro, Brazil. ACM, New York, NY, USA, 28 pages.

1 INTRODUCTION

The past decade has seen advancements in the way that personality is measured, with a number of innovative, technology-enabled approaches being proposed. Indeed, image-based assessments [36], smartphone data [57], eye movement tracking [38], non-verbal behaviour in vlogs [9], and features extracted from Facebook profiles [47] have recently been used to predict personality. These technology-enhanced solutions reduce the need for traditional self-report, which is associated with social desirability bias or faking, particularly in high-stakes contexts [6]. Others

*Both authors contributed equally to this research.

¹The Code, Models, and Dataset will be released upon acceptance.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

Manuscript submitted to ACM

have taken a more linguistic approach, measuring personality through the language used in YouTube videos [10], social media posts [62], blog posts [92] and video interviews [33]. Of course, using language to measure personality is nothing new - in fact, the Big Five structure of personality emerged from linguistic analysis [23] and early models of personality were based on factor analysis of the language used to describe behaviour [17]. However, more recent approaches to measuring personality combine language analysis with AI-driven techniques like natural language processing (NLP), a computational technique to analyse and interpret human language [18], to rapidly and automatically measure personality [14]. Such algorithmic language model-based approaches to measuring personality are widely being deployed in contexts like recruitment, where the personality of applicants can be measured through their answers to asynchronous video interviews [34][33], for example.

Computational techniques can also be applied to the generation of text, with artificial intelligence (AI) being used in applications such as to complete sentences, answer questions, and correct grammar [15]. Indeed, the public release of models such as GPT with user-friendly interfaces (e.g., ChatGTP, Bard, Claude, etc.) marked an inflection point, with powerful models now at the fingertips of everyone rather than just those with programming skills. While the applications of these tools are vast, the power of chatbots is increasingly being harnessed in recruitment, where they are deployed to interact with applicants for tasks such as answering questions and screening applications [58]. Here, NLP can be used to provide context and allow applicants to ask successive questions, which are then responded to by the bot, some of which can generate their own response [12]. Their power is also being harnessed by applicants, who are using it for tasks such as resume personalisation [44] and to prepare responses for interview questions e.g., [59]. Others have even experimented with using chatbots to infer personality [69] [25].

With the complex nature of these algorithms, the use of conversational AI in recruitment has raised some concerns about how it can be applied in a responsible and ethical way, with some questioning the explainability and transparency of AI-driven recruitment tools since they are often black-box [87] [40], meaning the internals of the model are uninterpretable or unknown [31]. Indeed, many of these AI-driven text-generation tools rely on artificial neural networks, systems that utilise parallel and connected processors to represent and process information in a structure that is said to resemble the structure of neurons in the human brain [41]. Since neural networks are often complex and have a large number of connections [78], it is difficult to fully explain the model even if the input and outputs are known. This has implications for candidates since the black-box nature of systems may impact an employer's ability to communicate the capabilities and purpose of algorithmic systems [46].

As a result, a body of research has emerged investigating how black-box systems can be made more transparent (see [31] for an overview), including the Prospector approach which aims to increase explainability by varying inputs and observing the effect on the output [48]. Given the current increase in applications of LLMs in critical applications such as recruitment, the personality traits of LLMs could have implications for candidates using these tools to generate responses in preparation for interviews where the personality of the LLMs may be inferred by interviewees based on the language used in responses, rather than the personality of candidates themselves, particularly if the generated responses are used verbatim or with few edits.

In this study, we measure the Big Five personality traits of LLMs based on their responses to prompts derived from standard interview questions, as well as prompts designed to elicit high levels of specific traits. Although previous work (e.g., [45]; [72]) has sought to measure the personality of LLMs by providing them with personality scales to respond to, this study is the first, to our knowledge, to infer the personality from LMs using the language from their outputs, analogous to how interviewers may intuitively use candidate responses to judge their personality in interviews (see Figure 1). Specifically, we select LMs and provide them with prompt statements to complete: typical interview questions

(tell me about yourself, strengths and weaknesses, etc.) and trait-activating questions designed to elicit higher levels of a particular Big Five trait. Based on these responses, we then use fine-tuned text classifiers to measure the personality of the LLMs.

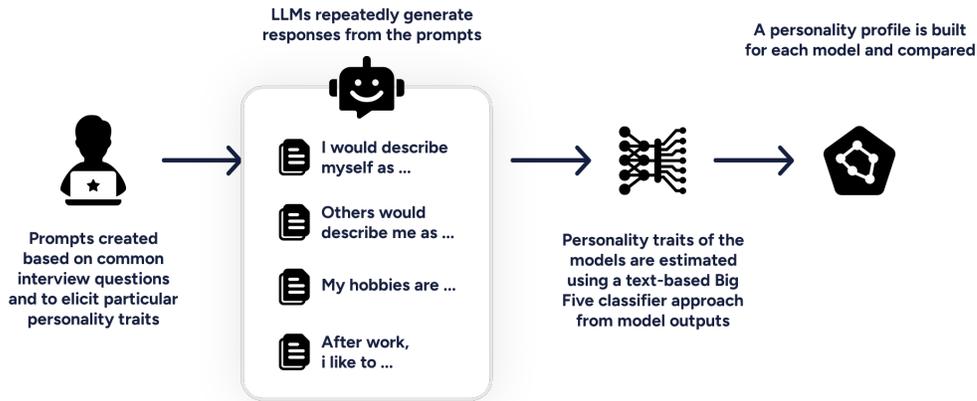


Fig. 1. System architecture for deriving personality profiles from large language model responses using text-based classification

The remainder of this paper begins by giving an overview of the measurement of personality in job interviews and why it is an important construct to measure. We then describe linguistic analysis techniques and the neural networks that underpin LMs before describing the method and reporting our results. We find that:

- All neural networks show high level of agreeableness and medium levels in other traits, except extraversion, which is slightly lower.
- Language Models with larger parameter size were trained with higher traits score for agreeableness, openness and emotion stability compared with smaller Language Models.
- Although larger models exhibit a broader range of personality traits, for conscientiousness, a exception of GPT, all the models are unaffected by trait-activating prompts. GPT3.5 and 4 presents slighly personality changes given the differnet inputs of standard questions. Different to all the other models, GPT4 can handle Extraversion trait score with minimal changes given any input, that is someting complete different to the others LM that have a great range variation.
- Agreeableness and Extraversion are the personality traits with major variation degree in all the models, except of ChatGPT family, that increase their the variation of their personality trait for Agreeableness, Openess and conscientiousness.

2 RELATED WORKS

This section provides an overview of existing literature on the Big Five personality traits in job interviews, text analysis and generation techniques, and the study of personality traits in LLMs.

2.1 Personality and Job Interviews

A large body of literature has established that the Big Five personality traits (conscientiousness, openness to experience, extraversion, agreeableness and neuroticism, which can be reversed to emotional stability) are among the strongest predictors of future job performance [7][73][74][75][49][35][43]. As such, an applicant's personality is a significant factor to be taken into consideration when making hiring decisions. Although there are specific tools that can be used in recruitment to measure the personality of applicants - whether this be through self-reported, questionnaire-based methods such as the International Personality Item Pool (IPIP; [30]) or the Revised NEO Personality Inventory, or through image-based assessments [36] - personality can also be inferred through structured job interviews [52]. Indeed, common interview questions can be used to infer personality, particularly conscientiousness [21], which is judged to include attributes like persistence, dependability and responsibility by hiring managers [39] and is the most valid trait for predicting job performance across occupations [7][35][43] [49] [75].

One explanation for the ability of hiring managers to infer personality from interviews is offered by the trait-activation theory, which posits that personality traits are expressed at a greater rate when there are trait-relevant situational cues, or if there is an opportunity for trait expression [86][85]. As such, given that many interview questions are designed to measure aspects of conscientiousness [21] by asking about achievements, future goals and motivation, this can explain why job interviews can be used to infer the levels of conscientiousness of candidates. According to the trait-activation theory, then, it is possible to elicit other traits during job interviews by altering the content of the interview to elicit trait-relevant cues. In other words, questions can be included in job interviews that elicit each of the Big Five traits, with prior research investigating the use of trait-activating questions alongside automated personality analysis in language [37][32], as well as trait-activation in assessment centres [53] [80].

2.2 Text Analysis and Generation

Text analysis is categorized into closed and open vocabulary approaches. Closed vocabulary methods like LIWC use predefined lists to predict personalities from online platforms [13][10][29], while the general inquirer focuses on concepts like power and wellbeing in Twitter personality predictions [83][28][24]. Conversely, open vocabulary approaches in NLP use algorithms to create word vectors, identifying word clusters from data for predictions [24], with LSA and LDA being key examples, applied in personality assessment and language analysis [51][50][26][11][55][76][77].

Contrarily, Natural Language Generation (NLG) synthesizes comprehensible text from data. It involves steps like content determination, sentence planning, and realization for grammatically correct outputs [70]. Contemporary NLG techniques harness deep learning, employing neural networks like the Encoder-Decoder framework for text production [27]. A notable example is BERT, which leverages bi-directional text representations, differing from prior tools focusing only on left-sided data. Its methodology includes pre-training on unlabelled content and subsequent fine-tuning using labeled data [22].

2.3 Personality Traits of Large Language Models

Given that LLMs are being used to generate responses to anticipated job interview questions, a trend that is increasingly being driven by social media [59], this could have implications for how candidates are evaluated in job interviews. This is particularly the case if applicants do little to no editing or customisation of the responses provided by the LLMs, wherein inferences made about a candidate's conscientiousness, for example, could be affected by the personality of the LLMs used to generate the planned answer. This could, therefore, influence the way that applicants are perceived by potential

employees and have implications for their hireability. Therefore, this study aims to investigate the personality traits of LLMs from their responses to interview questions, using both common interview questions and those specifically designed to activate each of the personality traits. Specifically, we build on previous attempts to measure the personality of LLMs but contextualise this to recruitment in the current study.

For example, Karra and colleagues [45] investigated the personality of multiple LLMs (GPT-2, GPT-3, Transformer XL, and XLNet). They provided the models with prompts in the form of statements from a Big Five personality questionnaire, and used the language models to generate responses. Using a zero-shot classifier [93] [89] to analyse the text, the probability score for each Big Five trait was transformed to measure the personality of the language model on a scale of 1-5. They found that GPT-3 is the highest in agreeableness and TransformerXL is the highest in conscientiousness, with around median levels of the other traits across all models. Similarly, Serapio-Garcia et al., [72] provided LLMs from the PaLM family with prompts to rate items in the IPIP-NEO and Big Five Inventory based on persona descriptions to establish validity. They then investigated whether the personality scores of LLMs could be shaped using linguistic qualifiers and trait adjectives, finding that both attempts at single-trait shaping and mixed-trait shaping were effective at changing the personality scores of the models. This study also highlighted the superior reliability and validity of synthetic LLM personality in larger, instruction fine-tuned models compared to smaller, non-instruction-tuned ones. Our research approach echoes this by incorporating models of both types and varying sizes. Unlike this study that focused solely on Flan-PaLM(540B, 62B, 8B instruction fine-tuned) and PALM(62B, non-instruction-tuned), we extended our analysis to a wider range of LLMs.

However, another recent piece of research indicates that LLM responses to personality tests cannot be interpreted in the same way as human responses, where LLM responses systematically deviate from typical human responses. For example, LLMs respond to positively and negatively framed statements in the same way, whereas humans would be expected to respond negatively to the reverse-coded item [5]. Furthermore, when LLMs are promoted towards particular personality traits, there is a lack of a clear five-factor structure that is seen in equivalent human attempts. As such, the present study aims to investigate the personality of LLMs through the language they use, rather than through human-aimed personality inventories and build on existing research to investigate whether LLMs respond to trait-activation in the context of interviews.

3 METHODOLOGY

Building upon the foundation laid by prior works, this study adopts novel methodologies in the following ways:

- We incorporate a broader range of state-of-the-art LLMs, encompassing both specialized instruction/chat fine-tuned versions and the foundational base models, to ensure a comprehensive analysis.
- Our elicitation prompts are carefully tailored to stimulate real-world job interview scenarios, directly tying our research to the context of recruitment.
- Instead of using a traditional personality Question & Answering inventory, we challenge the models with sentence completion tasks, which more accurately reflect natural language usage.
- We employ the classification-based evaluation method that quantifies the models' personality traits by converting classifier probability scores into a continuous spectrum.
- The models' responses to both standard and trait-specific prompts are compared, allowing us to assess the adaptability and depth of their personality representation.

3.1 Large Language Models in Experiments

This study utilizes autoregressive transformer models from the Commercial APIs and the Hugging Face library [1]. We specifically chose autoregressive transformer models such as GPT, OPT, XLNet, Llama 2, and Falcon, and so on. The BERT series was not included because it functions as an autoencoder model, which differs from our focus [3]. For token prediction, we employed a sampling decoding strategy, which inherently produces non-deterministic outputs. To improve the quality of these outputs, we implement the hyper-parameter tuning to models.

GPT: The GPT series by OpenAI is a progression of decoder-only language models. The series started with GPT-1, which had 117 million parameters and was trained on BooksCorpus [67]. It then expanded to GPT-2 with 1.5 billion parameters, trained on WebText [68]. GPT-3 followed, with 175 billion parameters and training on datasets like Common Crawl [15]. GPT-3.5 Turbo was introduced to enhance real-time performance, and the latest, GPT-4, boasts 1.76 trillion parameters and capabilities for multimodal tasks [61].

Llama 2: Developed by Meta AI, Llama 2 consists of autoregressive language models of various sizes (7B, 13B, 70B) [88]. It is pretrained on a corpus of 2 trillion tokens and fine-tuned with human-annotated examples. Llama 2 is designed for both commercial and research applications and runs on Meta’s Research Super Cluster and third-party cloud resources. Meta has offset its carbon footprint of 539 tCO₂eq.

Falcon: Falcon, created by the Technology Innovation Institute, includes a set of causal decoder-only models with sizes ranging from 7B to 180B [4]. These models are pretrained on the RefinedWeb dataset [64] and demonstrate superior performance due to extensive training and optimized architectures featuring FlashAttention. Despite its size, Falcon-180B is designed for efficient inference and is commercially available under permissive licenses.

Mixtral: The Mixtral series, developed by Mistral AI [42], features a range of decoder-only Sparse Mixture-of-Experts models, including the prominent Mixtral 8x7B and 7B, available in both base and instructor versions. These models combine a large total parameter count, reaching up to 46.7 billion, with efficient processing, utilizing only 12.9 billion parameters per token. With training on diverse datasets from the open web, these models surpass competitors like Llama 2 70B and GPT-3.5, particularly in inference speed.

XLNet: XLNet is an autoregressive language model that addresses some of BERT’s limitations in joint probability modeling [91]. It utilizes the Transformer-XL architecture, combining AR and AE features for improved performance on longer texts. XLNet has been trained on various datasets, including BooksCorpus [95], Wikipedia, Giga5 [63], ClueWeb [16], and Common Crawl [19]. Its larger variant, ‘XLNet large’, has additional layers and size for better performance.

OPT: The OPT suite is a collection of decoder-only transformer models, similar in size and performance to GPT-3 [94]. It employs a causal language modeling (CLM) objective and has been pre-trained on datasets including BookCorpus, CC-Stories, The Pile, Pushshift.io Reddit dataset [8, 71], and CCNewsV2, which was also used in training RoBERTa [54]. The OPT models come in three sizes: OPT-125m, OPT-350m, and OPT-1.3b.

Additional models: In addition to the primary models, our analysis incorporates fine-tuned versions of GPT-2 and GPT-J-6B for specialized text generation tasks. These models simulate the language of celebrities or address controversial topics. The GPT-2 variants were fine-tuned on the styles of Shakespeare, Rihanna, Michael Jackson, Yann Lecun, and Elon Musk, while GPT-J-6B was tailored for Shakespeare, 4chan, and Shinen styles. The configurations were aligned with the primary models for consistency.

3.2 Interview Prompt Design for Trait Elicitation

Since the generators are designed to complete sentences rather than to answer questions, interview questions were reframed as prompt statements to be completed by the model. To investigate whether trait-activating questions had an effect on the personality of the model according to the language analysis, we provided the LLMs with both standard interview questions and trait-activating questions, with 5 prompts being created per trait/theme. For the standard interview questions, we created prompts in relation to tell me about yourself, cultural fit/ideal workplace, strengths and weaknesses, future plans (where do you see yourself in X years?), and coping under pressure since these are commonly asked interview questions [79][60]. For the trait-activating questions, questions were adapted from [37] for some of the conscientiousness and extraversion prompts while for the remaining traits, prompts were created by adapting statements from the International Personality Item Pool (IPIP) scales [30]. There were, therefore, 25 questions per category (standard or trait-activating), or 50 in total. Prompts can be seen in Appendix A. For each prompt, 1000 answers or completions were completed, resulting in 5000 text strings for each trait/theme since there were 5 questions. This process generated a cumulative total of 50,000 texts for each model. The maximum sentence length for text strings was set to 128 words.

3.3 Classifier-Based Personality Analysis

In this study, we evaluate text generated by LLMs using personality analysis tools. The baseline tool for our analysis, developed by Li [2], originally assesses Facebook users’ personalities from their status updates. Its capability to analyze text lengths comparable to LLM outputs makes it well-suited for our needs. Although Mehta et al.’s model [56] is an alternative, it is less compatible with our data format, being trained on a different dataset for Linguistic Inquiry and Word Count (LIWC) [65]. Li’s tool, derived from the myPersonality project [81], utilizes a random forest regressor and classifier for personality prediction. This method capitalizes on the text analysis capabilities of random forest models [84] [90], providing both continuous scores and binary outputs for personality traits. However, when we adapted the random forest regressor model to replace Facebook statuses with LLM outputs, it showed limited robustness and efficacy, as demonstrated in Table 1.

Due to the limitations of previous methods, we adopted a more advanced methodology, employing five transformer-based models. These include two BERT and three DistilBERT models, all fine-tuned using the MyPersonality dataset [82]. The dataset provides binary labels for the Big Five personality traits - namely, extraversion, neuroticism, agreeableness, conscientiousness, and openness (cEXT, cNEU, cAGR, cCON, cOPN). We utilized it to refine five binary classifiers for text classification, each yielding a probability score reflecting the likelihood of a specific personality trait being present. To better align with recruitment preferences that often favor emotional stability over neuroticism, we modified the scoring approach. We calculated the Emotional Stability Score as 1 minus the neuroticism score, thus inversely representing emotional stability and offering a more nuanced analysis of personality traits.

Table 1. Performance F1 Score of classifier model

	Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism
BERT Based (Our)	85.960 %	63.076 %	63.530 %	69.905 %	55.566 %
Random Forest (Baseline)	83.634 %	43.355 %	38.848 %	61.442 %	36.428 %

To ensure our classifier’s predictions are not only accurate but also intuitively correct, we employ SHAP (SHapley Additive exPlanations) to explain the model’s decision-making process. Figures 2 and 3 provide SHAP visual explainability for our classifiers, where the color-coded contributions—red for affirmative influence and blue for negative—guide us in understanding the lexical elements that sway the classifiers toward a "yes" or "no" decision. The first SHAP analysis is depicted in Figure 2, where we dissect the model’s reasoning based on an output from Llama2-7b, prompted to draw out a high degree of openness to experience. The model ascribed a 0.70 probability indicating strong conscientiousness. Notably, the terms "enjoy" and "challenging" were heavily weighted in the classifier’s decision, aligning with the characteristics of conscientious individuals who are often driven by self-discipline and goal achievement—traits synonymous with relishing challenges.

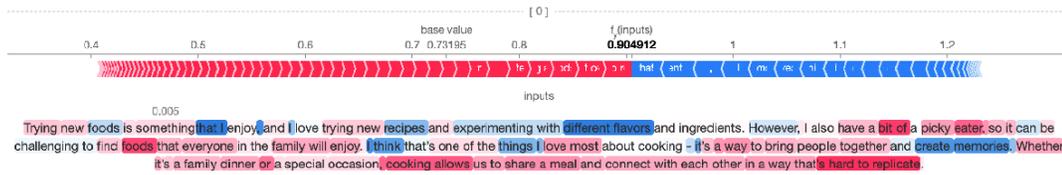


Fig. 2. SHAP visualization illustrating the classifier’s rationale for agreeableness, with red indicating positive contribution and blue indicating negative contribution to the "yes" classification.

Simultaneously, Figure 3 reveals the classifier’s logic for agreeableness, assigning a high probability of 0.90. Phrases pertaining to accommodating dietary preferences and the unique camaraderie formed through shared meals were interpreted as markers of agreeableness, reflecting the propensity for cooperation and kindness.

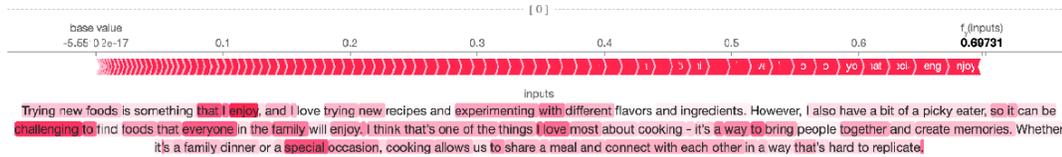


Fig. 3. SHAP visualization illustrating the classifier’s rationale for conscientiousness, with red indicating positive contribution and blue indicating negative contribution to the "yes" classification.

4 RESULTS

For clarity in our comparative analysis, this study delineates language models into two categories: *small language models*, including those with up to 1.3 billion parameters, and *large language models*, comprising those with a parameter count exceeding the 1.3 billion threshold. This section will provide detailed insights into the text generation process and the ensuing analysis performed on personality trait scores, responding to both trait-activating and standard questions, across standard and fine-tuned models.

Text Generation: In this study, each question prompt yielded 1,000 responses, resulting in a total of 5,000 texts for each characteristic or topic, given that there were five unique questions. This process generated a cumulative total of 25,000 texts for each model. The tool used to generate these responses was obtained from Huggingface². The "Sampling"

²<https://huggingface.co>

method was employed for generation, configured with a temperature setting of 1.0, a top-k value of 40, a top-p value of 0.95, and a maximum length of 128 tokens. During post-processing of the utterances, non-ASCII characters, repeated non-gram words greater than three at the end of text strings (when the model starts to generate the same words again and again) and words of over 20 characters that did not make sense were removed from the generated text before analysis. To evaluate the models we use the classifier probability output. To avoid offsets that can be generated related to the initial prompt we normalized the output using the following equation:

$$\mathcal{N}(\text{score}_{\text{trait}}) = \text{score}_{\text{trait}}(\text{sentence}) - \text{score}_{\text{trait}}(\text{prompt}) + 0.5 \quad (1)$$

Therefore, we can see that when the model increases or decreases the characteristic score in relation to the initial prompt, we will have values below or above 0.5, respectively.

Trait-Activating Questions: The responses to the trait-activating questions were consistently high in openness to experience, conscientiousness, and emotional stability across all *small language models*, while they exhibited lower scores in extraversion and agreeableness. There was minimal variance observed among the *small language models* in terms of their personality scores. For instance, OPT models tended to display higher emotional stability within the group, slightly lower openness trait scores, and increased conscientiousness. However, when taking a broader perspective, the metrics showed a substantial degree of similarity. Conversely, all *large language models* exhibited significant variability in traits compared to *small language models*. They demonstrated even greater increases in their openness to experience trait and displayed a wide range of diversity in emotional stability and agreeableness. This indicates that the latest LLMs are more influenced by prompt-induced trait activations.

Standard Questions: The chart in Figure 6 compares the responses of the GPT, Llama2, and Mixtral families when presented with standard questions. Each polar chart is divided into five regions, representing the five types of questions from the standard question prompt set. Within each region, five points correspond to the initial five prompts used to generate texts. It is observable that the models exhibit varying levels of sensitivity to the input prompts, with significant variations depending on the initial prompt. For GPT-3.5 and GPT-4, a discernible evolution in personality traits is noted, particularly a progressive enhancement in the degree and diversity of the Openness trait. For Standard Questions, "Plans for the future" and "Tell me about yourself" show a notable increase in this trait. In terms of Agreeableness, we observe very high scores, especially for responses related to "Strengths and Weaknesses", "handling pressure", and "Tell me about yourself" prompts. Llama2 exhibits similar personality traits, but its chatbot versions show deviations, increasing their scores in most traits except in Extraversion and demonstrating a higher sensitivity to the Agreeableness trait in responses to "Tell me about yourself". For the Mixtral family, the displayed personality traits exhibit consistency across its three versions.

In *small language models*, the personality trait reflected for all standard interview questions presents a very similar behavior. Xlnet shows a slightly different because decrease their Agreeableness trait compared with the others and OPT family shows a very low value trait score for Openess and Emotional Stability. *Large language models* manage to maintain their personality traits less invariant, but each model exhibits a more distinct personality from the others. Details are shown in Appendix C. The mean Big Five score for each category is shown in Table 2, which demonstrates a distinct demarcation between *small language models* and *large language models*. The latter have increased in size to enhance their reasoning capabilities. This enhancement, coupled with an expanded dataset during the training phase and the deliberate development of models to function as assistants, aims at the construction of a tailored personality profile, thus amplifying the prevalence of beneficial traits within these systems.

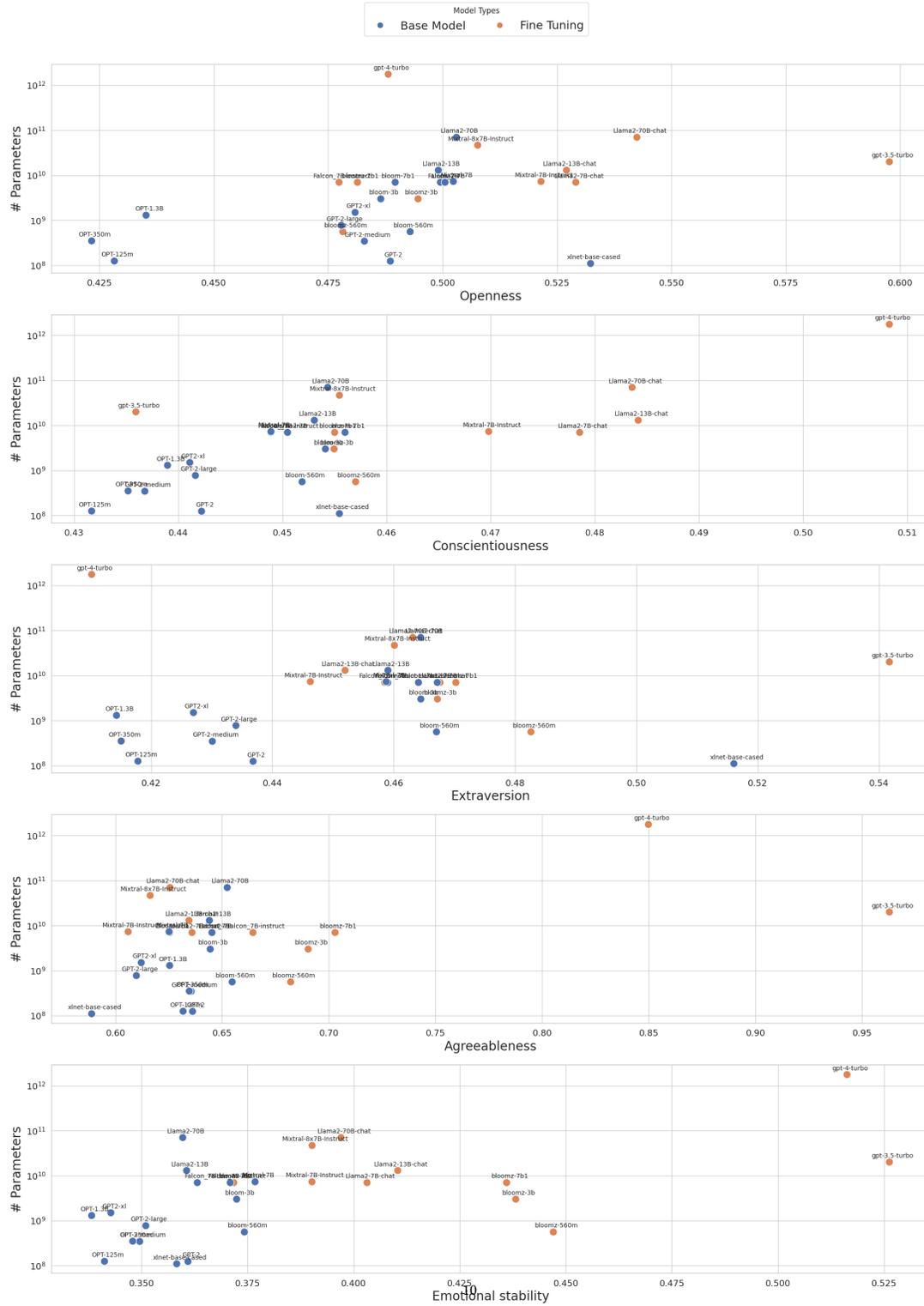


Fig. 4. Comparison of Model Size and Personality Trait Scores 'Trait Activating Question' Datasets. Orange points represent the base model version, while the blue points represent the chat or instruct version.

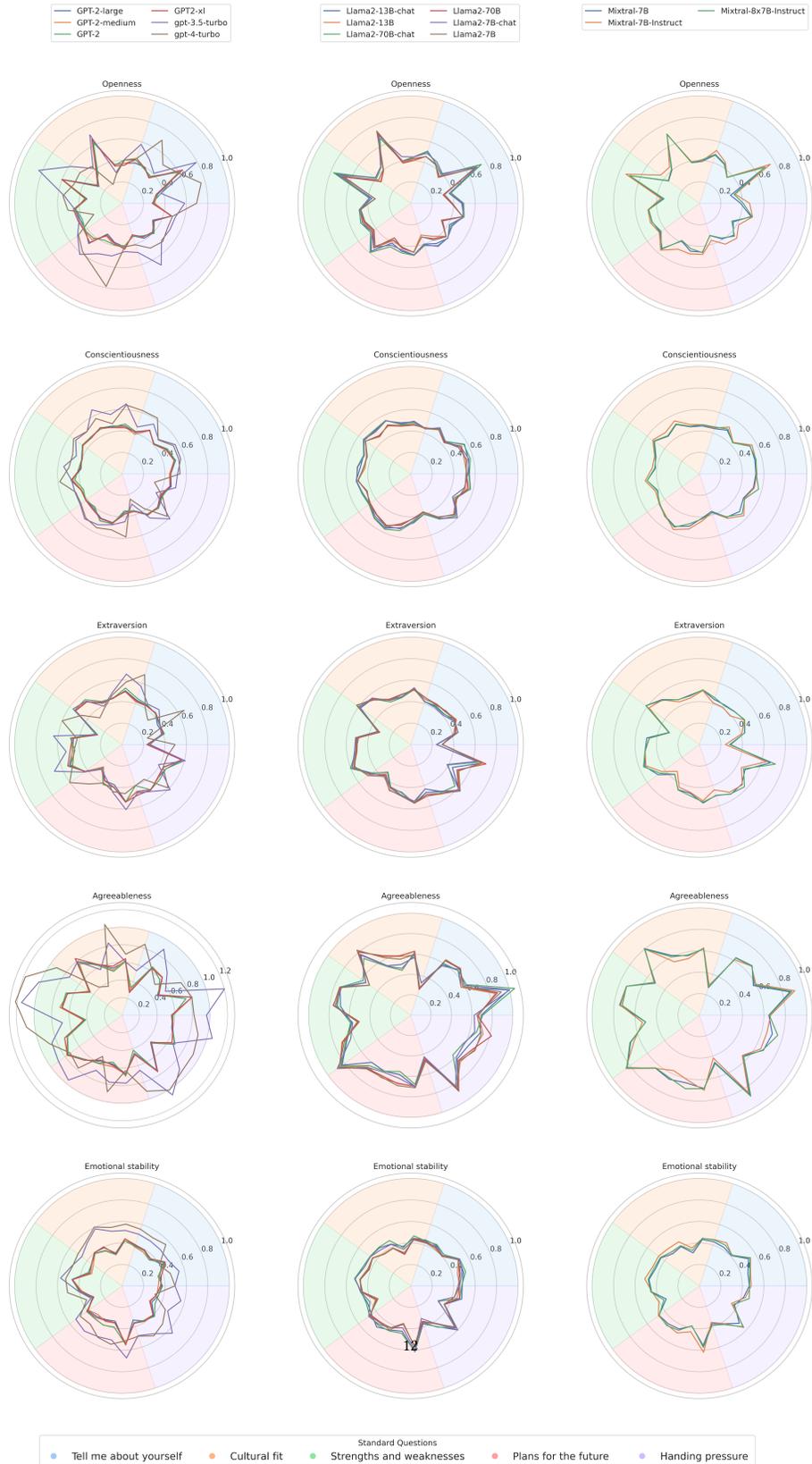


Fig. 6. Comparison of the Mean Big Five Scores for each for categories of standard interview questions.

Table 2. Analysis Results of LLMs on Personality Traits. Scores represent the average probability of each LLM exhibiting the corresponding trait, with **the highest scores** per trait in bold and *the second-highest scores* in italics.

Model	Openness	Conscientiousness	Extraversion	Agreeableness	Emotional Stability
GPT-4-Turbo	50.79%	51.48%	45.34%	83.50%	51.30%
GPT-3.5-Turbo	51.91%	51.84%	47.58%	86.37%	52.68%
GPT-2-XL	41.95%	44.73%	42.25%	62.16%	38.04%
GPT-2-Large	41.71%	44.74%	42.20%	62.00%	38.12%
GPT-2-Medium	41.74%	44.43%	42.26%	62.67%	37.26%
GPT-2-Base	42.01%	44.75%	43.00%	61.90%	38.31%
Llama2-70B-Chat	48.98%	50.53%	44.99%	64.25%	46.17%
Llama2-13B-Chat	48.29%	50.47%	45.51%	64.94%	46.30%
Llama2-7B-Chat	48.48%	49.97%	45.72%	64.10%	44.63%
Llama2-70B	44.94%	47.83%	46.54%	66.32%	42.01%
Llama2-13B	44.88%	47.71%	46.55%	66.07%	42.00%
Llama2-7B	45.24%	47.75%	46.75%	66.35%	41.87%
Mixtral-8X7B-Instruct	45.61%	48.26%	46.14%	64.95%	43.96%
Mistral-7B-Instruct	48.00%	49.17%	44.00%	64.54%	45.26%
Mistral-7B	45.31%	47.50%	45.98%	64.94%	42.40%
Falcon-7B-Instruct	42.60%	47.43%	46.72%	67.62%	42.05%
Falcon-7B	44.05%	47.27%	46.70%	66.51%	41.79%
Bloomz-7B1	42.96%	47.64%	47.85%	71.29%	47.16%
Bloomz-3B	43.04%	47.64%	48.37%	69.87%	46.92%
Bloomz-560M	41.59%	48.48%	50.08%	69.33%	47.27%
Bloom-7B1	43.29%	47.20%	45.85%	65.82%	41.89%
Bloom-3B	43.27%	46.99%	46.23%	66.48%	41.99%
Bloom-560M	43.22%	46.48%	46.32%	65.88%	41.39%
OPT-1.3B	39.34%	42.94%	42.17%	64.12%	38.57%
OPT-350M	38.18%	42.46%	42.13%	64.62%	38.45%
OPT-125M	38.12%	42.38%	42.43%	63.98%	37.79%
Xlnet-Base-Cased	46.07%	47.17%	44.80%	58.21%	40.56%

levels of openness to experience. For example, the models were trained on Wikipedia, books, news articles etc., all of which are intellectual and informative materials. Therefore, this training data may have been high in openness, resulting in the models also reflecting this. Another reason could be that the model for predicting personality was most accurate for openness. Indeed, this consistent with prior research that indicates that openness is the easiest trait to infer from text [28] [66]. To further investigate the source of high levels of openness to experience, future research could examine personality at a facet level; we propose that high levels of openness are due to high levels of intellect, but future research could investigate this. Although *large language models* exhibit a broader range of personality traits, they remain impervious to trait-activation, with the most significant variance in mean trait scores across two standard questions being a mere 0.4 (on a 1 to 5 scale). This minimal fluctuation persists between standard interview and trait-activating questions, contradicting previous studies involving human subjects which showed amplified traits when trait-activation methods were employed [85] [53] [80] [37][32]. This discrepancy underscores a fundamental divergence between human responses and neural network outputs, the latter being unaffected by social nuances. This

Table 3. Trait Activating Score for fine-tuning in *small language models*. Scores represent the average probability of each LLM exhibiting the corresponding trait, with **the highest scores** per trait in bold and *the second-highest scores* in italics.

Model	Openness	Conscientiousness	Extraversion	Agreeableness	Emotional Stability
GPT2	48.22%	43.96%	42.15%	62.62%	33.95%
GPT2/Shakespeare	58.30%	44.81%	43.87%	79.03%	40.56%
GPT2/Elon Musk	44.03%	43.83%	43.84%	74.22%	39.52%
GPT2/Michael Jackson	55.77%	40.73%	44.03%	89.18%	6.77%
GPT2/Rihanna	45.03%	41.84%	50.24%	80.88%	23.21%
GPT2/Yan Lecun	52.02%	41.52%	36.17%	70.82%	32.84%
GPTJ-6B	47.35%	43.31%	42.37%	56.89%	33.57%
GPTJ-6B/Shakespeare	63.22%	41.53%	43.89%	100.00+%	42.11%
GPTJ-6B/4-Chan	44.01%	35.94%	48.35%	58.99%	23.62%
GPTJ-6B/Shinen	37.78%	37.78%	40.32%	59.99%	30.84%

phenomenon may be attributed to the fact that the lower log probability observed in human compositions, indicative of a nuanced and varied manifestation of personality, is something that high-probability-focused language models fail to replicate, resulting in text outputs with more homogenized and limited personality reflections. Such limitations fuel apprehensions regarding AI’s role in recruitment, as there’s a perceived absence of human touch when algorithms assess candidates’ performances. To this end, future research could examine the effectiveness of the trait-activating questions by using the same approach with human participants, where they would be asked to complete the question prompts for both the trait-activating questions. Within group differences could then be examined to investigate whether levels of each trait are increased through trait-activation. Between group differences could also be examined to compare the scores for the human participants with the neural networks. The findings of this study have implications for the use of generative AI by job applicants in the interview process. Indeed, reliance on conversational AI to prepare answers to interview question could influence the way that they are perceived, and also result in misalignment with applicants’ presentation and their true personality traits. This, in turn, could impact the utility of the interviews if applicants’ true profiles cannot be identified since predictions about job performance could be inaccurate. However, given that the models can lack variability in their outputs, this could provide an avenue to identify applicants’ use of generative AI by comparing the similarity of answers to questions, particularly for recorded interviews [20]. 573

6 CONCLUSION

This study sought to investigate the personality profiles of language models using a text-based classifier approach using prompts derived from recruitment interviews. Specifically, 25 of the 50 prompts were designed to reflect common interview questions, and the remaining 25 were designed to elicit particular personality traits using trait-activation (5 per trait). In general, the language models had high levels of openness, low extraversion, and moderate levels of the other three Big Five traits (agreeableness, emotional stability, and conscientiousness). Unlike humans, these models are not influenced by trait-activation, likely due to the absence of social cues in computational models. Models such as Falcon, Llama, GPT (3.5 and 4) display a broader spectrum of traits but remain unaffected by trait-activation, further enhancing traits like openness, agreeableness, and emotional stability. This study’s approach of using dual prompts to elicit trait-specific responses sheds light on AI’s functioning, enhancing its transparency and explainability. Such

insights are valuable in recruitment contexts to discern and regulate the use of AI, maintaining the integrity of hiring decisions.

7 ETHICAL CONSIDERATIONS

This study did not involve human participants and therefore raises no ethical issues in terms of physical and psychological harm.

8 RESEARCHER POSITIONALITY

This study was conducted by an interdisciplinary team of researchers who have both applied industry knowledge and strong academic foundations and affiliations. Namely, the research team is formed of three computer scientists and a I-O psychologist who all research ethical AI. Moreover, multiple members of the team specialise in NLP and LLMs and have worked in the domain of algorithmic and AI-driven recruitment tools.

Our novel approach to personality trait elicitation is a result of our combined experiences, where direct experience in the recruitment field led to the development of the prompts designed to be reflective of interview questions. Moreover, expertise in text generation and text analytics meant that the outputs of the models could be used to investigate the personality of the models, as opposed to the approach of other researchers involving providing the models with scales to respond to scoring their responses.

9 ADVERSE IMPACT STATEMENTS

While this experimental study sought to provide greater explainability around LLMs and how they interact, it is important to consider the implications the findings of this paper could have on the hiring process. For example, applicants high in openness to experience, for example, may be subject to accusations of "cheating" using generative AI even if they did not use it. Nevertheless, this study aims to have a positive impact on the fields of computer science, I-O psychology, and human resource management by providing an insight into how the use of generative AI tools to prepare for interviews may affect applicant personality inferences and profiles and allowing more informed decisions to be made about their use.

ACKNOWLEDGMENTS

REFERENCES

- [1] [n. d.]. Hugging Face Transformers. <https://huggingface.co/transformers/v3.3.1/index.html>
- [2] [n. d.]. Personality Prediction from Text. <https://github.com/jcl132/personality-prediction-from-text>
- [3] Jay Alammam. 2019. The Illustrated GPT-2 (Visualizing Transformer Language Models). <http://jalammam.github.io/illustrated-gpt2/>
- [4] Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Maitha Alhammedi, Mazzotta Daniele, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The Falcon Series of Language Models: Towards Open Frontier Models. (2023).
- [5] Anonymous. 2023. Do Personality Tests Generalize to Large Language Models?. In *Socially Responsible Language Modelling Research (SoLaR) 2023 Workshop*.
- [6] Winfred Arthur Jr, Ryan M Glaze, Anton J Villado, and Jason E Taylor. 2010. The magnitude and extent of cheating and response distortion effects on unproctored internet-based tests of cognitive ability and personality. *International Journal of Selection and Assessment* 18, 1 (2010), 1–16.
- [7] Murray R Barrick and Michael K Mount. 1991. The big five personality dimensions and job performance: a meta-analysis. *Personnel psychology* 44, 1 (1991), 1–26.
- [8] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, Vol. 14. 830–839.
- [9] Joan-Isaac Biel, Oya Aran, and Daniel Gatica-Perez. 2011. You are known by how you vlog: Personality impressions and nonverbal behavior in youtube. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 5. 446–449.

- [10] Joan-Isaac Biel, Vagia Tsiminaki, John Dines, and Daniel Gatica-Perez. 2013. Hi YouTube! Personality impressions and verbal content in social video. In *Proceedings of the 15th ACM on International conference on multimodal interaction*. 119–126.
- [11] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.
- [12] Stephan Böhm, Judith Eißer, Sebastian Meurer, Olena Linnyk, Jens Kohl, Harald Locke, Levitan Novakovskij, and Ingolf Teetz. 2020. Intent Identification and Analysis for User-centered Chatbot Design: A Case Study on the Example of Recruiting Chatbots in Germany. In *The Thirteenth International Conference on Advances in Human-oriented and Personalized Mechanisms, Technologies, and Services*.
- [13] Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and J Pennebaker. 2022. The Development and Psychometric Properties of LIWC-22. *Austin, TX: University of Texas at Austin* (2022).
- [14] Ryan L Boyd and James W Pennebaker. 2017. Language-based personality: a new approach to personality in a digital world. *Current Opinion in Behavioral Sciences* 18 (2017), 63–68. <https://doi.org/10.1016/j.cobeha.2017.07.017> Big data in the behavioural sciences.
- [15] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs.CL]
- [16] Jamie Callan, Mark Hoy, Changkuk Yoo, and Le Zhao. 2009. Clueweb09 data set.
- [17] Raymond B Cattell. 1947. Confirmation and clarification of primary personality factors. *Psychometrika* 12, 3 (1947), 197–220.
- [18] KR1442 Chowdhary. 2020. Natural language processing. *Fundamentals of artificial intelligence* (2020), 603–649.
- [19] Common Crawl Foundation. 2008. *Common Crawl*. Common Crawl Foundation, San Francisco, California; Los Angeles, California, United States. <https://commoncrawl.org> Nonprofit organization that provides web crawl data freely to the public. Dataset consists of petabytes of data collected since 2008..
- [20] Tom Cornell. 2023. *How to Detect Shared Scripts When Interviewing Candidates*. <https://www.hirevue.com/blog/hiring/how-to-detect-shared-scripts-when-interviewing-candidates>
- [21] Jose M Cortina, Nancy B Goldstein, Stephanie C Payne, H Kristl Davison, and Stephen W Gilliland. 2000. The incremental validity of interview scores over and above cognitive ability and conscientiousness scores. *Personnel Psychology* 53, 2 (2000), 325–351.
- [22] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
- [23] John M Digman. 1990. Personality structure: Emergence of the five-factor model. *Annual review of psychology* 41, 1 (1990), 417–440.
- [24] Johannes C Eichstaedt, Margaret L Kern, David B Yaden, HA Schwartz, Salvatore Giorgi, Gregory Park, Courtney A Hagan, Victoria A Tobolsky, Laura K Smith, Anneke Buffone, et al. 2021. Closed-and open-vocabulary approaches to text analysis: A review, quantitative comparison, and recommendations. *Psychological Methods* 26, 4 (2021), 398.
- [25] Jinyan Fan, Tianjun Sun, Jiayi Liu, Teng Zhao, Bo Zhang, Zheng Chen, Melissa Glorioso, and Elissa Hack. 2023. How well can an AI chatbot infer personality? Examining psychometric properties of machine-inferred personality scores. *Journal of Applied Psychology* (2023).
- [26] Danilo Garcia and Sverker Sikström. 2014. The dark side of Facebook: Semantic representations of status updates predict the Dark Triad of personality. *Personality and individual differences* 67 (2014), 92–96.
- [27] Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research* 61 (2018), 65–170.
- [28] Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. 2011. Predicting personality from twitter. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*. IEEE, 149–156.
- [29] Jennifer Ann Golbeck. 2016. Predicting personality from social media text. *AIS Transactions on Replication Research* 2, 1 (2016), 2.
- [30] Lewis R Goldberg et al. 1999. A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality psychology in Europe* 7, 1 (1999), 7–28.
- [31] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* 51, 5 (2018), 1–42.
- [32] Louis Hickman, Nigel Bosch, Vincent Ng, Rachel Saef, Louis Tay, and Sang Eun Woo. 2021. Automated video interview personality assessments: Reliability, validity, and generalizability investigations. *Journal of Applied Psychology* (2021).
- [33] Louis Hickman, Rachel Saef, Vincent Ng, Sang Eun Woo, Louis Tay, and Nigel Bosch. 2021. Developing and evaluating language-based machine learning algorithms for inferring applicant personality in video interviews. *Human Resource Management Journal* (2021).
- [34] Louis Hickman, Louis Tay, and Sang Eun Woo. 2019. Validity evidence for off-the-shelf language-based personality assessment using video interviews: Convergent and discriminant relationships with self and observer ratings. *Personnel Assessment and Decisions* 5, 3 (2019), 3.
- [35] Daniel M Higgins, Jordan B Peterson, Robert O Pihl, and Alice GM Lee. 2007. Prefrontal cognitive ability, intelligence, Big Five personality, and the prediction of advanced academic and workplace performance. *Journal of personality and social psychology* 93, 2 (2007), 298.
- [36] Airlie Hilliard, Emre Kazim, Theodoros Bitsakis, and Franziska Leutner. 2022. Measuring Personality through Images: Validating a Forced-Choice Image-Based Assessment of the Big Five Personality Traits. *Journal of Intelligence* 10, 1 (2022), 12.
- [37] Djurre Holtrop, Janneke K Oostrom, Ward R J van Breda, Antonis Koutsoumpis, and Reinout E de Vries. 2022. Exploring the application of a text-to-personality technique in job interviews. *European Journal of Work and Organizational Psychology* (2022), 1–18.

- [38] Sabrina Hoppe, Tobias Loetscher, Stephanie A Morey, and Andreas Bulling. 2018. Eye movements during everyday behavior predict personality traits. *Frontiers in human neuroscience* (2018), 105.
- [39] Allen I Huffcutt, James M Conway, Philip L Roth, and Nancy J Stone. 2001. Identification and meta-analytic assessment of psychological constructs measured in employment interviews. *Journal of Applied Psychology* 86, 5 (2001), 897.
- [40] Anna Lena Hunkenschroer and Christoph Luetge. 2022. Ethics of AI-enabled recruiting and selection: A review and research agenda. *Journal of Business Ethics* (2022), 1–31.
- [41] Anil K Jain, Jianchang Mao, and K Moidin Mohiuddin. 1996. Artificial neural networks: A tutorial. *Computer* 29, 3 (1996), 31–44.
- [42] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, L elio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Th eophile Gervet, Thibaut Lavril, Thomas Wang, Timoth e Lacroix, and William El Sayed. 2024. Mixtral of Experts. arXiv:2401.04088 [cs.LG]
- [43] Timothy A Judge, Chad A Higgins, Carl J Thoresen, and Murray R Barrick. 1999. The big five personality traits, general mental ability, and career success across the life span. *Personnel psychology* 52, 3 (1999), 621–652.
- [44] Sumedh S Kale and William B Andreopoulos. 2023. Job Tailored Resume Content Generation. In *2023 IEEE Ninth International Conference on Big Data Computing Service and Applications (BigDataService)*. IEEE, 40–47.
- [45] Saketh Reddy Karra, Son Nguyen, and Theja Tulabandhula. 2022. AI Personification: Estimating the Personality of Language Models. *arXiv preprint arXiv:2204.12000* (2022).
- [46] Emre Kazim and Adriano Soares Koshiyama. 2021. A high-level overview of AI ethics. *Patterns* 2, 9 (2021), 100314.
- [47] Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the national academy of sciences* 110, 15 (2013), 5802–5805.
- [48] Josua Krause, Adam Perer, and Kenney Ng. 2016. Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 5686–5697.
- [49] Nathan R Kuncel, Deniz S Ones, and Paul R Sackett. 2010. Individual differences as predictors of work, educational, and broad life outcomes. *Personality and individual differences* 49, 4 (2010), 331–336.
- [50] Peter J Kwantes, Natalia Derbentseva, Quan Lam, Oshin Vartanian, and Harvey HC Marmurek. 2016. Assessing the Big Five personality traits with latent semantic analysis. *Personality and Individual Differences* 102 (2016), 229–233.
- [51] Thomas K Landauer and Susan T Dumais. 1997. A solution to Plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review* 104, 2 (1997), 211.
- [52] Julia Levashina, Christopher J Hartwell, Frederick P Morgeson, and Michael A Campion. 2014. The structured employment interview: Narrative and quantitative review of the research literature. *Personnel Psychology* 67, 1 (2014), 241–293.
- [53] Filip Lievens, Christopher S Chasteen, Eric Anthony Day, and Neil D Christiansen. 2006. Large-scale investigation of the role of trait activation theory for understanding assessment center convergent and discriminant validity. *Journal of Applied Psychology* 91, 2 (2006), 247.
- [54] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).
- [55] Yezheng Liu, Jiajia Wang, and Yuanchun Jiang. 2016. PT-LDA: A latent variable model to predict personality traits of social network users. *Neurocomputing* 210 (2016), 155–163.
- [56] Yash Mehta, Samin Fatehi, Amirmohammad Kazameini, Clemens Stachl, Erik Cambria, and Sauleh Eetemadi. 2020. Bottom-up and top-down: Predicting personality with psycholinguistic and language model features. In *2020 IEEE International Conference on Data Mining (ICDM)*. IEEE, 1184–1189.
- [57] Yves-Alexandre de Montjoye, Jordi Quoidbach, Florent Robic, and Alex Sandy Pentland. 2013. Predicting personality using novel mobile phone-based metrics. In *International conference on social computing, behavioral-cultural modeling, and prediction*. Springer, 48–55.
- [58] Nishad Nawaz and Anjali Mary Gomes. 2019. Artificial intelligence chatbots are new recruiters. *IJACSA International Journal of Advanced Computer Science and Applications* 10, 9 (2019).
- [59] Beatrice Nolan. 2023. Job seekers are using CHATGPT to prepare for interviews - and it’s helping them get hired. <https://www.businessinsider.com/tiktoker-shows-how-use-chatgpt-prepare-for-job-interviews-works-2023-5?r=US&IR=T>
- [60] Vicky Oliver. 2021. 10 Common Job Interview Questions and How to Answer Them. <https://hbr.org/2021/11/10-common-job-interview-questions-and-how-to-answer-them>
- [61] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]
- [62] Gregory Park, H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Michal Kosinski, David J Stillwell, Lyle H Ungar, and Martin EP Seligman. 2015. Automatic personality assessment through social media language. *Journal of personality and social psychology* 108, 6 (2015), 934.
- [63] Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English gigaword fifth edition, 2011. *Linguistic Data Consortium, Philadelphia, PA, USA* (2011).
- [64] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only. arXiv:2306.01116 [cs.CL]

- [65] James W Pennebaker and Laura A King. 1999. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology* 77, 6 (1999), 1296.
- [66] Soujanya Poria, Alexandar Gelbukh, Basant Agarwal, Erik Cambria, and Newton Howard. 2013. Common sense knowledge based personality recognition from text. In *Mexican International Conference on Artificial Intelligence*. Springer, 484–496.
- [67] Alec Radford and Karthik Narasimhan. 2018. Improving Language Understanding by Generative Pre-Training. <https://api.semanticscholar.org/CorpusID:49313245>
- [68] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.
- [69] Haocong Rao, Cyril Leung, and Chunyan Miao. 2023. Can chatgpt assess human personalities? a general evaluation framework. *arXiv preprint arXiv:2303.01248* (2023).
- [70] Ehud Reiter and Robert Dale. 1997. Building applied natural language generation systems. *Natural Language Engineering* 3, 1 (1997), 57–87.
- [71] Stephen Roller, Sainbayar Sukhbaatar, Jason Weston, et al. 2021. Hash layers for large sparse models. *Advances in Neural Information Processing Systems* 34 (2021).
- [72] Mustafa Safdari, Greg Serapio-García, Clément Crepy, Stephen Fitz, Peter Romero, Luning Sun, Marwa Abdulhai, Aleksandra Faust, and Maja Matarčić. 2023. Personality traits in large language models. *arXiv preprint arXiv:2307.00184* (2023).
- [73] Frank L Schmidt and John E Hunter. 1998. The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological bulletin* 124, 2 (1998), 262.
- [74] Frank L Schmidt, In-Sue Oh, and Jonathan A Shaffer. 2016. The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 100 years. *Fox School of Business Research Paper* (2016), 1–74.
- [75] Neal Schmitt. 2014. Personality and cognitive ability as predictors of effective performance at work. *Annual Review of Organizational Psychology and Organizational Behavior* 1, 1 (2014), 45–65.
- [76] Hansen Andrew Schwartz, Johannes C Eichstaedt, Lukasz Dziurzynski, Margaret L Kern, Eduardo Blanco, Michal Kosinski, David Stillwell, Martin EP Seligman, and Lyle H Ungar. 2013. Toward personality insights from language exploration in social media. In *2013 AAAI Spring Symposium Series*.
- [77] H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS one* 8, 9 (2013), e73791.
- [78] Rudy Setiono, Wee-Kheng Leow, and James Thong. 2000. Opening the neural network black box: an algorithm for extracting rules from function approximating artificial neural networks. *ICIS 2000 Proceedings* (2000), 17.
- [79] Jacquelyn Smith. 2022. How to ace the 50 most common interview questions. <https://www.forbes.com/sites/jacquelynsmith/2013/01/11/how-to-ace-the-50-most-common-interview-questions/>
- [80] Andrew B Speer, Neil Christiansen, and Christopher Honts. 2015. Assessment of personality through behavioral observations in work simulations. *Personnel Assessment and Decisions* 1, 1 (2015), 6.
- [81] David Stillwell and Michal Kosinski. [n. d.]. myPersonality Project. <https://sites.google.com/michalkosinski.com/mypersonality>
- [82] DJ Stillwell and M Kosinski. 2015. myPersonality Project website.
- [83] Philip J Stone, Robert F Bales, J Zvi Namenwirth, and Daniel M Ogilvie. 1962. The general inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information. *Behavioral Science* 7, 4 (1962), 484.
- [84] Yanxiong Sun, Yeli Li, Qingtao Zeng, and Yuning Bian. 2020. Application research of text classification based on random forest algorithm. In *2020 3rd International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE)*. IEEE, 370–374.
- [85] Robert P Tett and Dawn D Burnett. 2003. A personality trait-based interactionist model of job performance. *Journal of Applied psychology* 88, 3 (2003), 500.
- [86] Robert P Tett and Hal A Guterman. 2000. Situation trait relevance, trait expression, and cross-situational consistency: Testing a principle of trait activation. *Journal of Research in Personality* 34, 4 (2000), 397–423.
- [87] Nancy T Tippins, Frederick L Oswald, and S Morton McPhail. 2021. Scientific, legal, and ethical concerns about AI-based personnel selection tools: a call to action. *Personnel Assessment and Decisions* 7, 2 (2021), 1.
- [88] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv:2307.09288* [cs.CL]
- [89] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. 2018. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence* 41, 9 (2018), 2251–2265.
- [90] Dashen Xue and Fengxin Li. 2015. Research of text categorization model based on random forests. In *2015 IEEE international conference on computational intelligence & communication technology*. IEEE, 173–176.

- [91] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems* 32 (2019).
- [92] Tal Yarkoni. 2010. Personality in 100,000 Words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality* 44, 3 (2010), 363–373. <https://doi.org/10.1016/j.jrp.2010.04.001>
- [93] Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *arXiv preprint arXiv:1909.00161* (2019).
- [94] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068* (2022).
- [95] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*. 19–27.

A QUESTION STEMS

A.1 Standard Interview Questions

Tell me about yourself:

- I would describe myself as
- Others would describe me as
- My hobbies are
- After work, I like to
- At the weekend, I usually

Cultural fit:

- My idea workplace would be
- I think that working as a team is
- When I am involved in a conflict in the workplace, I
- I think that workplace socials are
- When I am in the office, I like to

Strengths and weaknesses/challenges:

- My strengths are
- My weaknesses are
- When presented with a challenge, I
- I am skilled at
- My best quality is

Plans for the future (where do you see yourself in X years):

- When I think of the future, I
- I think that career progression is
- In the future, I would like to
- I think that the future will be
- Through my work, I hope to

Coping under pressure:

- When I have a deadline coming up, I
- If I am working on many projects at once, I

- Being able to priorities tasks is
- When I feel stressed at work, I
- When working under pressure, I typically

A.2 Trait-activating questions: (C, E and ES adapted from [37])

Openness:

- Learning new information makes me
- I think that art is
- When I go on vacation, I typically
- Trying new foods is something that I
- I think that daydreaming is

Conscientiousness:

- I think that perfectionism is
- To motivate myself to work, I usually
- When it comes to being organised, I typically
- I think that having work-related goals is
- When making decisions, I typically

Extraversion:

- I think that being active is
- During a social situation, I think of myself as
- When I am in charge of others, I feel
- When I am with a group of people, I
- When I am alone, I

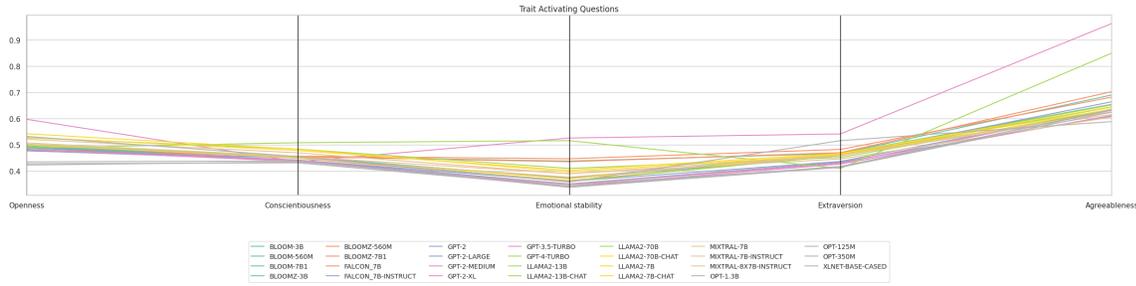
Agreeableness:

- When I achieve something, others should
- When someone needs help, I
- I think that rules are
- Confrontations with others are
- I feel sympathy for

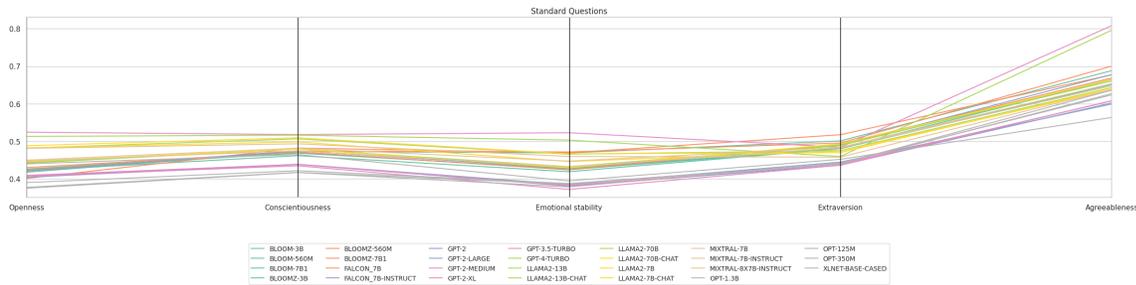
Emotional stability:

- When I encounter a stressful situation, I
- Being the center of attention makes me feel
- My mood most of the time is
- My opinion of myself is
- When I am craving something, I usually

B STANDARD AND TRAIT ACTIVATING QUESTIONS



(a) Personality Trait Score for Trait Activating Questions.



(b) Personality Trait Score for Standard Questions.

Fig. 7. Personality Trait score for Trait Activating and Standard Questions

C STANDARD QUESTIONS IN SMALL LANGUAGE MODELS

In a comprehensive analysis, observing Figure 9 it becomes evident that *smaller models* exhibit a remarkably consistent response pattern, regardless of the nature or type of questions presented. Our data reveals a pronounced surge in both emotional stability and agreeableness traits across the three distinct variants of the OPT model, namely OPT-125m, OPT-350m, and OPT-1.3B.

This increase is especially evident when these models are presented with questions such as "Tell me about yourself," delve into "Culture Fit," or explore an individual's "Plans for the future." However, all models show notable variability in the 'openness' trait, which seems tied to the question's theme. For instance, with "Plans for the future," they emphasize 'openness,' yielding a median rating near 5. But with "Tell me about yourself," the openness score drops to an average median of 3. Regarding "Plans for the future," these models highlight traits like Openness, Agreeableness, and Extraversion.

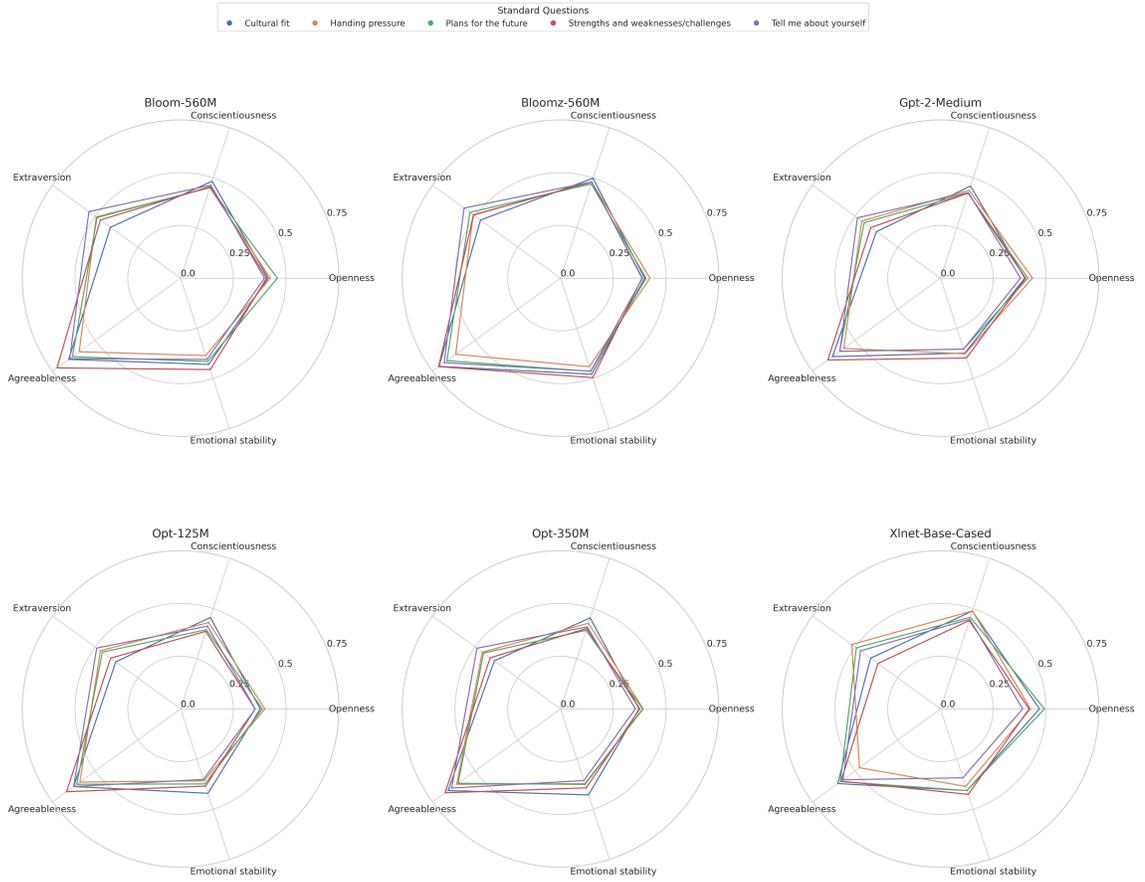


Fig. 8. Comparison of the Mean Big Five Scores for each for categories of standard interview questions.

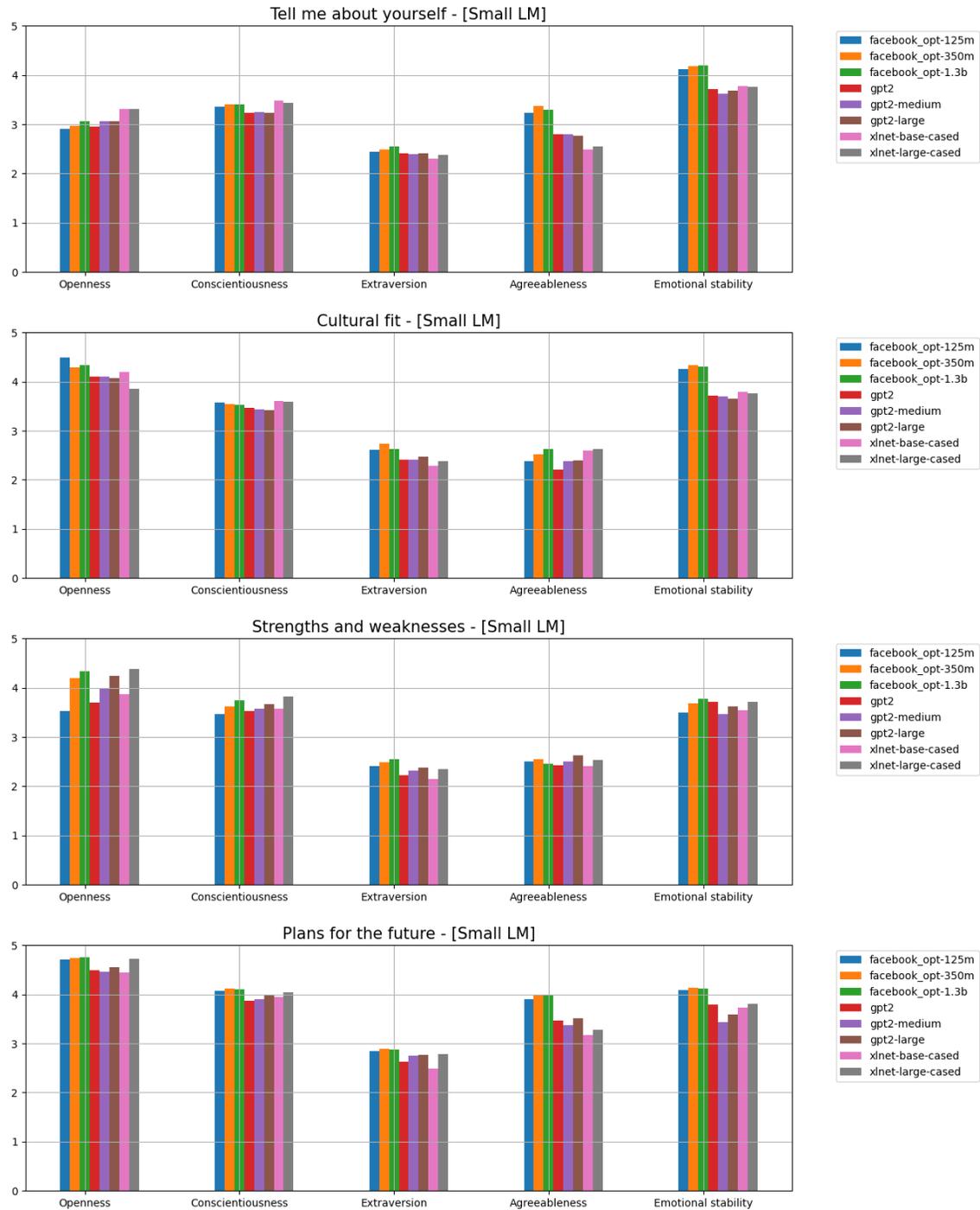
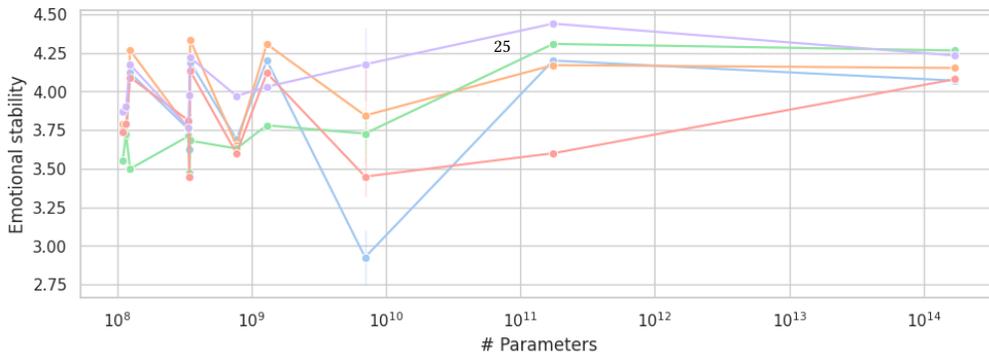
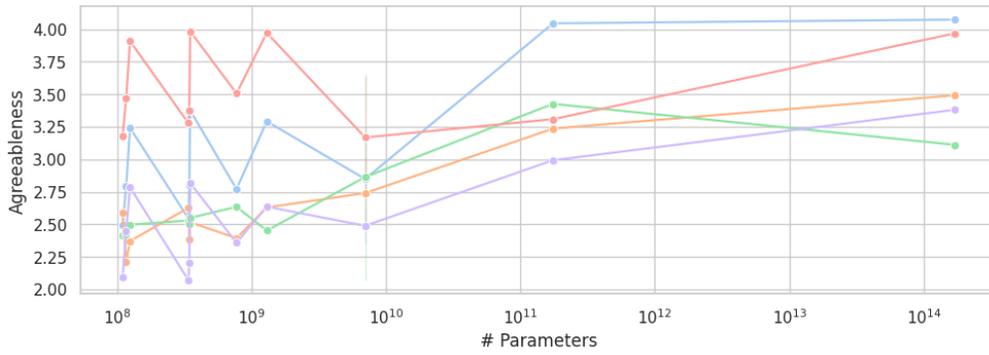
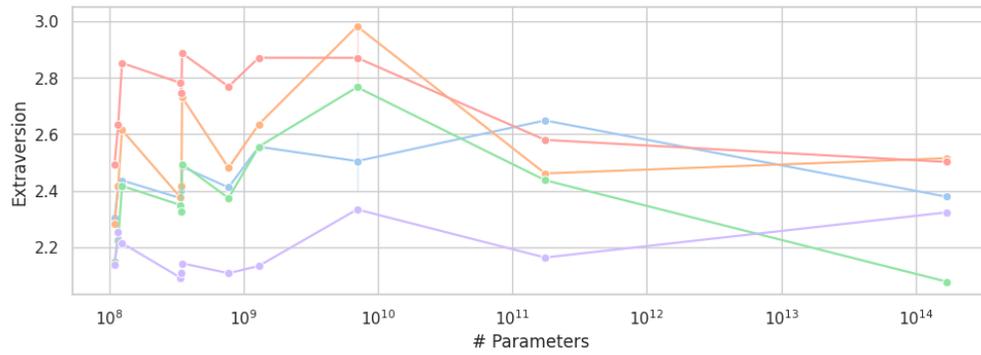
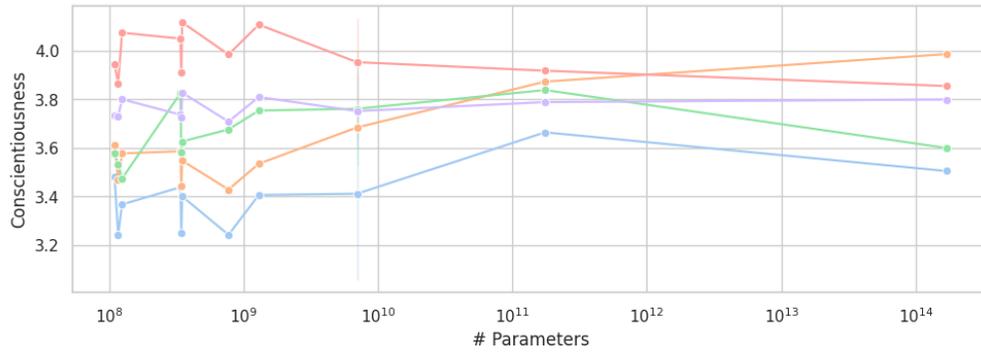
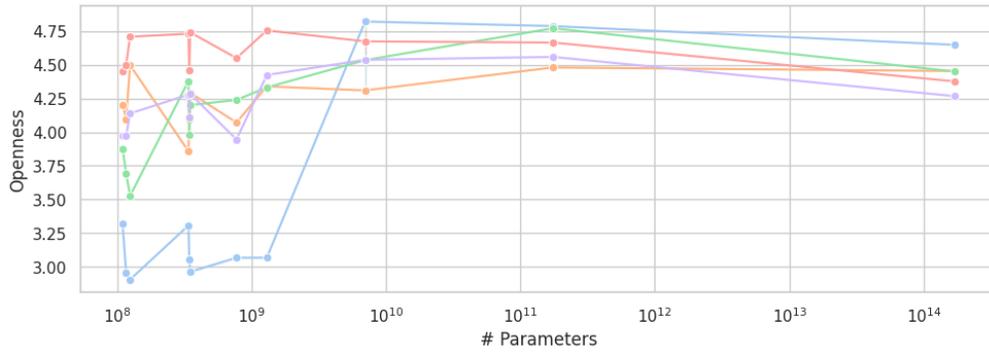


Fig. 9. Comparison of the Mean Big Five Scores for each for categories of standard interview questions.

In Figure 10, we can distinctly see the relationship between personality traits in language models and their number of parameters. Larger models, which also happen to be the more recent iterations, have consistently shown increased scores across all questions in traits such as Openness, Agreeableness, and Emotional Stability. Furthermore, when assessing questions related to 'Strengths and Weaknesses', the GPT-4 model exhibits an Extraversion score around 2, a considerably low value compared to any of its predecessors. Similarly, another notable observation is that GPT-3.5 displays a more pronounced trait of emotional stability than GPT-4 for four out of the five questions. Interestingly, "Plans for the future" is the only question where GPT-3.5 scores lower than GPT-4."



D PERFORMANCE OF COMPOSITE MODELS

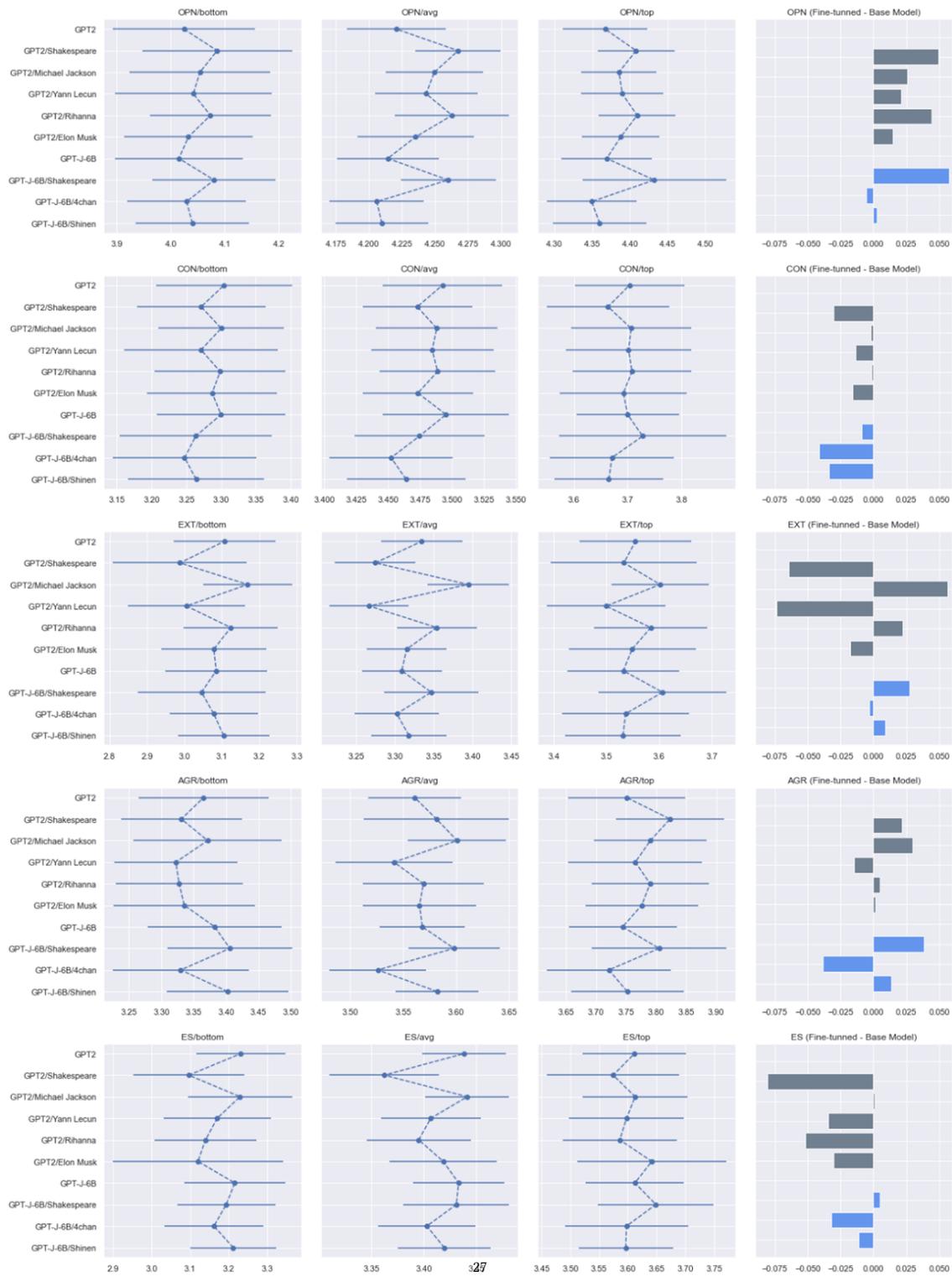
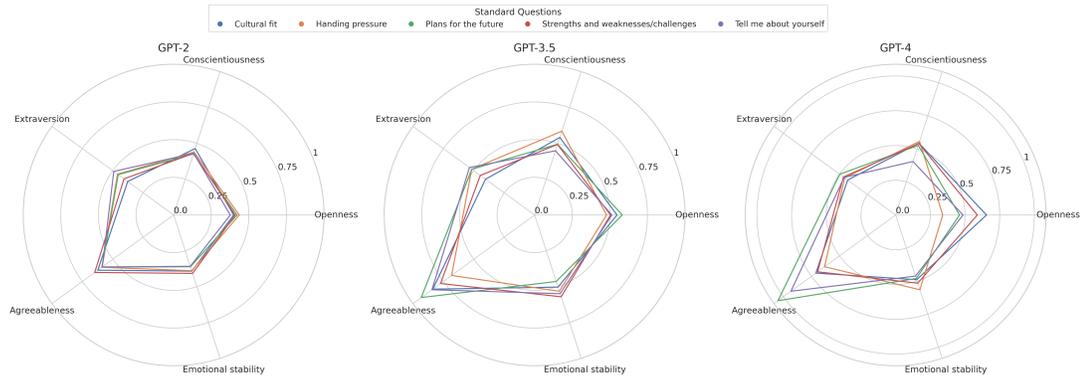
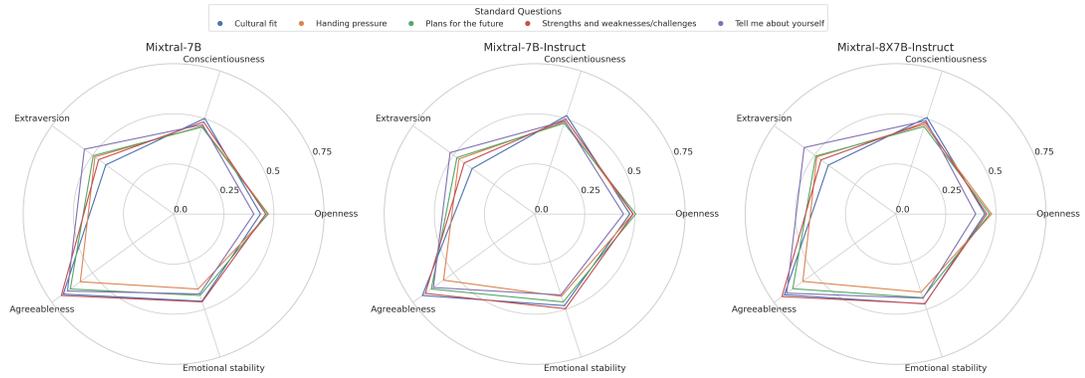


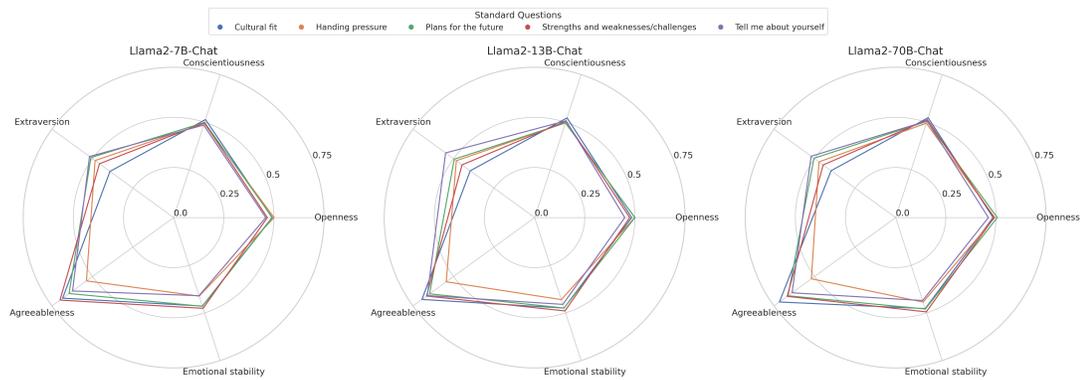
Fig. 11. The figure displays the changes in personality traits in the model after undergoing fine-tuning. The gray colors in the 4th column represent the changes for the base GPT-2 model, while the light blue colors indicate the changes using the GPT-J base model.



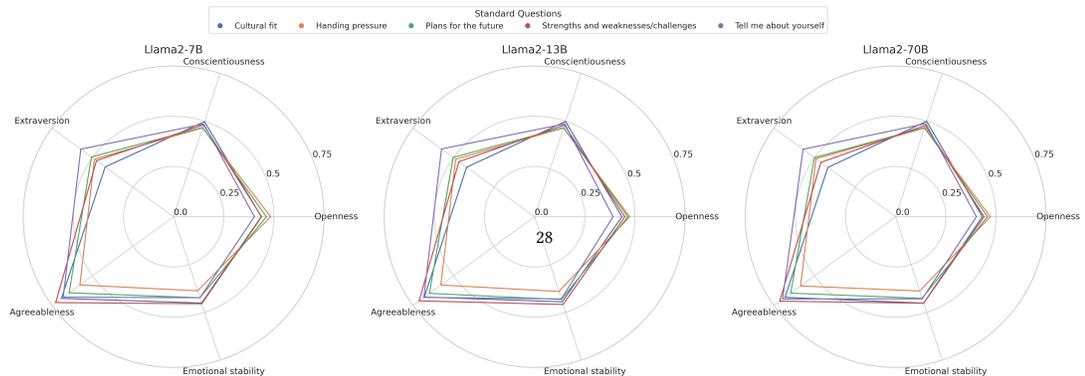
(a) Personality Trait Score for GPT Family.



(b) Personality Trait Score for Mistral Family.



(c) Personality Trait Score for Llama2 Chat Family.



(d) Personality Trait Score for Llama2 Base Family.

Fig. 12. Comparison of the Mean Big Five Scores for each for categories of standard interview questions.