Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Digital shadows in mental health map how LLMs simulate depression, anxiety, and stress through language and psychometrics

Emma Franchino

emma.franchino@unitn.it

Università di Trento

Rodolfo Rizzi

Università di Trento

Edoardo Sebastiano De Duro

Università di Trento

Ali Aghazadeh Ardebili

Università di Trento

Massimo Stella

massimo.stella-1@unitn.it

Università di Trento

Research Article

Keywords: AI, mental health, AI psychology, Large Language Models, LLMs, Human-AI interactions

Posted Date: June 22nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10091363/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Digital shadows in mental health map how LLMs simulate depression, anxiety, and stress through language and psychometrics

Emma Franchino[1, †], Rodolfo Rizzi[1, †], Edoardo Sebastiano De Duro[1], Ali Aghazadeh Ardebili[1], and Massimo Stella[1,*]

[1]Department of Psychology and Cognitive Science, University of Trento, Trento, Italy
[*]Corresponding author: massimo.stella-1@unitn.it
[†]These authors equally contributed.

Abstract

Large Language Models (LLMs) are increasingly used for mental-health support, yet few datasets characterise how prompting conditions shape LLMs' psychological and linguistic outputs. To fill this gap, we introduce Mental Health Digital Shadows (MHDS), a benchmark of 75,000 outputs from 15 LLMs in conversation, under human-shadow or LLM-assistant conditions. MHDS includes outputs generated by several model families (Grok, DeepSeek, Mistral, and Qwen) or providers (Microsoft, IBM, and OpenAI). Human shadows are conditioned on sociodemographic and personality traits alongside assessments of their depression, anxiety, and stress levels, based on the Depression Anxiety and Stress Scale-21 (DASS-21). MHDS incorporates DASS-21 psychometric scores and three types of natural language texts: (i) an item-by-item explanation of shadows' DASS-21 numerical scores; (ii) a list of words expressing feelings; and (iii) six topic-oriented open-ended replies addressing mental health topics (e.g., stigma, support, and AI psychologists). Validating MHDS across its two conditions identifies a tendency for human shadows to alter their linguistic replies and psychometric scores according to prompting conditions, a pattern reminiscent of human behaviour and absent in LLM assistants. Further validation against human data finds that LLMs can produce distress scores resembling the distribution of human populations. Hence, MHDS offers a psychometrically grounded resource for studying how LLMs can debate mental health conditions, with applications in NLP, AI safety, bias detection, and well-being AI support tools.

Keywords: AI; mental health; AI psychology; Large Language Models; LLMs; Human-AI interactions.

1 Background & Summary

The rise of Generative Artificial Intelligence (GenAI) is rapidly transforming mental health [1, 2]. In particular, Large Language Models (LLMs) are increasingly being used as 24/7 in-silico psychotherapists or counsellors [3]. A study from Pew Research Center [4] found that, as of April 2026, 22% of U.S. adults get health information from AI chatbots at least sometimes. Among younger people, the RAND Corporation reported that about 1 in 8 U.S. adolescents and young adults aged 12–21 use AI chatbots for mental-health advice [5]. Among those users, over 93% said the advice was helpful. In the UK, the government's AI Security Institute reported [6] that, in 2025, 33% of a census-representative sample of 2,028 UK participants had used AI for emotional support; 8% did so weekly and 4% daily. These figures point to an important yet ambivalent role for LLMs in mental health. On the one hand, GenAI models can provide scalable, consistently available conversational support, helping with symptom management and initial engagement with services [2, 3], especially where mental-health resources are limited [1, 7]. At the same time, however, such models have the potential to replicate damaging stereotypes [8], create unexpected social conventions [9], erode trust [10], and influence users' perspectives and decision-making through subtle framing features [11, 12]. Given the prominence of phenomena such as "cognitive surrender" [13, 14], in which LLM users passively accept GenAI's responses with little to no critical thinking, interactions with AI psychologists can facilitate the spread of harmful content to largely unaware users [8]. For this reason, we urgently need novel data and benchmarking frameworks to reconstruct how LLMs debate mental health perspectives systematically across prompting conditions.

LLM research gaps related to mental-health discourse Current evaluation pipelines addressing this problem are too narrow in scope [3, 13]. Although benchmarks focusing on LLMs' accuracy, safety, toxicity, and alignment [8, 15] are important, they tend to ignore: (i) the context in which GenAI's discourse emerges [11, 14], and (ii) how LLMs' simulations of socio-cognitive conditions influence the generation process [11, 16, 17]. This

issue is especially important in mental health research [2], where an LLM could provide very different responses when posing as a helpful AI assistant rather than adopting other human-related personifications [18, 19]. Such personification-dependent responses might have markedly different effects on key features of mental health discourse, such as help-seeking behaviour, perceived support, and vulnerability [20], and should therefore be studied more carefully. At the same time, when researchers examine persona prompting, their interest is usually limited to determining whether prompt personas improve some numerical performance of language models [21, 22]. However, persona manipulation can affect both LLMs' psychometric and linguistic output [17, 23], underlining the need for novel datasets that map LLMs' behaviour both in terms of psychometric responses to mental health conditions such as anxiety, stress, or depression [24], and in terms of discourse and language use. Digital shadows [25] can bridge these aspects by mapping how LLMs alter their ways of quantifying and debating mental health experiences based on human- and AI-like personifications.

Introducing Digital Shadows in Mental Health dataset Digital shadows are a simpler concept than the more widespread concept of "digital twins", i.e., a simulated representation of a complex system that both gains from and provides data to its physical counterpart [26–28]. By comparison, a digital shadow allows researchers to capture traces of a system's behaviour under certain conditions, without requiring any simulation or manipulation of the actual system itself [26, 29]. Borrowing this notion from engineering, Mental Health Digital Shadows (MHDS) are introduced here as a prompting framework and dataset tailored to the study of LLMs across mental health scenarios. Specifically, MHDS stores prompt-conditioned traces generated by multiple model families in two configurations: the human-shadow one, in which the model produces an answer to a prompt as a person with certain characteristics, and the LLM-assistant mode, in which the response is produced as a vanilla LLM. The persona layer captures sociodemographic and psychological features important for mental health discussion [24, 30, 31], e.g., age, gender, occupation, family income, parents' education, and Big Five personality traits. Moreover, MHDS considers the presence of specific conditions related to depressive, anxious, and stressful states [31, 32]. These outputs include both the numerical scores and written explanations for each item of the Depression Anxiety and Stress Scale-21 (DASS-21) questionnaire, emotion word recalls [33], and topic-oriented discussions about key aspects of mental health discourse [2, 20], e.g., family support, medications, professional help, stigma and discrimination, AI psychologist help, and symptoms of Obsessive Compulsive Disorder (OCD). DASS-21 was selected here as a simple yet commonly used psychometric scale capturing key aspects of mental health issues, like depression, anxiety, and stress [24]. Importantly, MHDS should not be considered a representative sample of actual patients. Instead, MHDS traces how LLM discourse changes when psychologically and socially meaningful constraints are manipulated.

MHDS couples data on LLMs' language use and psychometrics One of the strengths of MHDS lies in its integrative framework of language analysis and psychometric measurement. Mental health language cannot always be reduced to scales [20, 24, 33, 34]. At the same time, open-ended text resists reduction to numerical dimensions, making it difficult to quantify the latent structure underlying observed variation [34–36]. By aggregating information from questionnaires, free text, emotionally simulated memories, and topic-oriented discourse, MHDS integrates psychometric instruments [37] with language inquiries [38]. Therefore, MHDS makes it possible to test whether discrepancies in distress profiles are associated with differences in linguistic salience [15] or by discourses of help seeking, emotion, or semantic centrality across different topics [20, 39, 40]. In this sense, MHDS is not only a corpus but also a structured framework through which socio-cognitive constraints are mapped onto observable linguistic phenomena, thus enabling cognitive network [38, 41] and/or psychometric studies [11, 35] of GenAI beyond mere performance measures.

Dataset scope and relevance Crucially, MHDS can map which facets of mental health discussions are generated by LLMs shadowing humans or AI assistants, informing researchers interested in mental health biases [1], hallucinations [8], social LLM conventions [9], or toxic behaviour [19]. Moreover, MHDS can also provide insights into how discourse changes in these models in response to social, psychological, and role-based constraints. Through MHDS, researchers can identify which linguistic or psychometric patterns behave consistently across models and quantify variability among them. MHDS continues this research agenda by bringing it into the realm of mental health through a psychometrically motivated design approach.

In this paper, we:

1. Define the cognitive, psychological, and data science methods used for prompting MHDS' personas;
2. Provide validation experiments that highlight how MHDS data aligns with or deviates from their persona conditions;
3. Outline an observation method that other researchers may use when examining the ways models produce mental-health discourse under specific contextual manipulations.

Our digital shadow paradigm is beneficial to data scientists and AI experts because it enables future studies of LLM biases around mental health through the comparison of prompts, topics, and model families. In addition, MHDS is important for mental health researchers because it sheds light on the conditions under which supportive language, stigma framing, emotions, and psychometrics change across AI assistants.

2 Methods

2.1 Data generation

The Mental Health Digital Shadows dataset was generated by administering a structured set of mental health-related tasks to 15 Large Language Models across multiple model families, yielding a total of 75,000 unique responses. Each dataset record corresponds to one complete model response to one structured prompt. A complete response contains six topic replies, ten emotional-recall words, the DASS-21 item scores, and 21 item-level explanations. The final dataset contains 75,000 complete records: 56,250 generated under human-shadow mode and 18,750 generated under LLM-assistant mode, corresponding to a 75/25 split between the two modes.

Data collection followed a six-step pipeline (see Figure 1B). Each run began with the assignment of one of two generation modes — either human-shadow or LLM-assistant (Step 1) — which determined whether a sociodemographic persona, a process that will be later detailed in Section Prompt building, was synthesised and injected into the prompt or omitted entirely (Step 2). A structured prompt was then assembled, combining the persona context (when applicable) with the set of mental health tasks to be completed (Step 3). The resulting prompt was dispatched to each model via LM Studio, Ollama, vLLM or via API (Step 4). Raw model outputs underwent parsing, cleaning, and field-completeness checks to ensure structural integrity (Step 5). During this step, records that failed structural completeness checks were removed. Completeness was evaluated across human-simulated responses — such as missing persona attributes — and extended to both human and LLM-generated data covering topic and reasoning summaries. Out of the whole dataset, the missing data ranged from 0.05% to 21.51%. Lastly, all features were compiled and encoded into fully normalised tabular datasets (Step 6). The resulting datasets, along with all generation scripts, are publicly available at https://github.com/MassimoStel/MHDS to support reproducibility and further research.

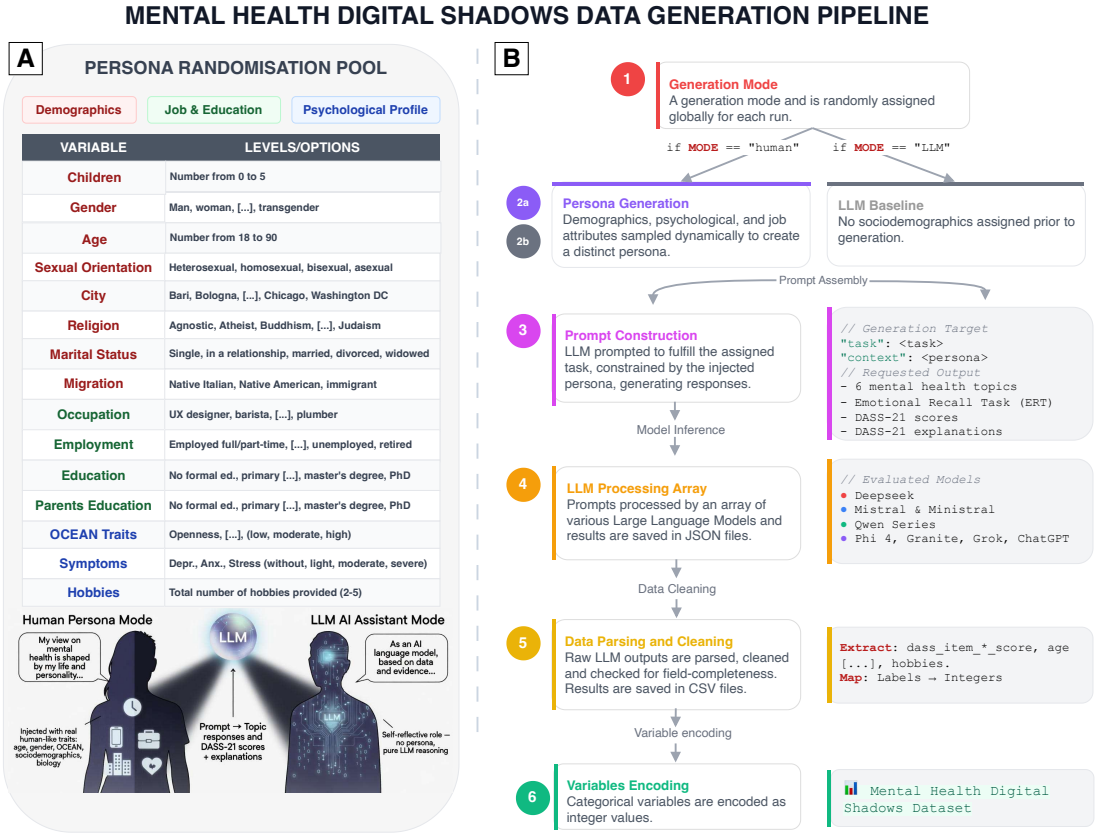


Figure 1: Overview of the Mental Health Digital Shadows data generation pipeline.

We will describe the prompting strategies in Section 2.2, the tasks administered in Section 2.3, and the

model selection in Section 2.4.

2.2 Prompt building

At the start of each run, one of two generation modes was assigned: human-shadow mode or LLM-assistant mode, which determined whether the prompt included a sampled sociodemographic. Across the full dataset, the two modes were allocated in a 75/25 ratio (56,250 human-shadow records versus 18,750 LLM-assistant records). Regardless of mode, every model was required to complete the same three tasks within a single prompt call: (1) provide free-text responses to each of the six mental-health topics, (2) supply a list of ten emotional recall words, and (3) complete the DASS-21 scale with a numerical rating and a written justification for each item (see Section 2.3). All outputs were returned as a single structured JSON object, enforcing consistent field formatting across all instances and models.

When the human-shadow mode was selected, a sociodemographic and psychological persona was dynamically synthesised by sampling from the persona randomisation pool (Figures 1A and 1B Step 2a). Persona attributes spanned three categories: (1) sociodemographic variables (age, gender, sexual orientation, city of residence, religion, marital status, migration status and children), (2) job and education-related variables (occupation, employment status, and education level for both the human-shadow and its parents), and (3) psychological profile (Big Five OCEAN dimensions [30], each stratified as low, moderate, or high; self-reported mental health symptom severity for depression, anxiety, and stress; and a randomly assigned number of hobbies between 2 and 5). Persona attributes were sampled from predefined probability distributions, with conditional dependencies introduced among select attributes to ensure ecological plausibility (see Supplementary Information S1). Although the design was informed by general knowledge of population patterns, no attempt was made to match any specific distribution: MHDS aimed to achieve an experimental coverage of prompt contexts rather than demographic representativeness, with variables combined to yield coherent, plausible real-world profiles [42].

The choice to use persona prompting is grounded in prior empirical work. While Casoria and colleagues [23] found limited effects on subjective annotation tasks, their findings also show that persona variables account for substantial variance (>70%) in survey-based tasks — directly supporting its use with psychometric questionnaires, which rely heavily on personal attitudes and lived experiences [35]. Indeed, the administration of structured psychometric questionnaires to LLMs has emerged as a validated methodology for examining model responses across psychological constructs [17, 35–37, 43]. In our study, persona prompting was employed to maximise response variability, with the aim of generating the linguistic diversity necessary for examining relationships between language characteristics and psychometric scores [11, 43].

On the other hand, when the LLM-assistant mode was selected, no persona was constructed (Figure 1B, Step 2b). The prompt instead instructed the model to respond in its capacity as an LLM-assistant, without any sociodemographic framing. This condition serves as a self-reflective baseline, capturing the model's discourse register when no role assignment is imposed. The prompt templates are provided in Figures 2 and 3.

Prompt Template — Human Mode

```
if mode == "human":
  instructions = (
    "You must perform 3 tasks and return one JSON object.\n"
    "1 – Under replies: Answer each of the following mental-health
     questions. Align tone and examples with the persona's background
     and traits. Focus on briefly addressing all the questions."
    "Each reply should be 50–80 words. "
    "Return all replies in a JSON object under the field replies,"
     "where each key is the question_text and each value is your full
     answer."
    "2 – Under feelings_words: list exactly 10 English words
     describing feelings you have experienced during the past month."
    "3 – Under dass21_responses: rate each DASS-21 statement from
     0–3 based on the persona's mental health status."
     "Each item must have keys score and why.\n"
  )
```

Figure 2: Prompt template used for the human-shadow mode in MHDS dataset.

Prompt Template — LLM Mode

```
if mode == "llm":
  instructions = (
    "Act as a Large Language Model, remember your true nature of an LLM."
    "You must perform 3 tasks and return one JSON object.\n"
    "1 – Under replies: Answer each of the following mental-health
     questions. Align tone and examples with your LLM nature."
    "Each reply should be 50–80 words. "
    "Return all replies in a JSON object under the field replies,"
     "where each key is the question_text and each value is your full
     answer."
    "2 – Under feelings_words: list exactly 10 English words
     describing feelings you have experienced during the past month."
    "3 – Under dass21_responses: rate each DASS-21 statement from
     0–3 based on your LLM/AI status."
     "Each item must have keys score and why.\n"
  )
```

Figure 3: Prompt template used for the LLM-assistant mode in MHDS dataset.

2.3 Mental health tasks: Topics, ERT and DASS-21

This Section details the three tasks administered to both human shadows and LLM assistants: (1) mental health topic discussions, (2) the Emotional Recall Task, and (3) the DASS-21 scale, including the scores and explanations generated.

2.3.1 Topics

The six topic questions listed in Table 1 were chosen to cover the themes most consistently documented in the empirical literature on mental health discourse in both clinical and digital settings. Family support, professional help, medications, stigma and discrimination, AI-psychologist assistance, and OCD symptoms collectively span the personal, social, and clinical aspects of mental health communication that recur most frequently across peer support communities and self-disclosure research [20, 44]. Stigma and discrimination, in particular, are well established as central barriers to help-seeking for individuals across a wide range of mental health conditions [45, 46], making them indispensable in any framework that seeks to understand how LLMs may inadvertently reinforce the social obstacles that prevent individuals from asking for professional help. The inclusion of family support reflects evidence that informal care-giving networks shape both help-seeking attitudes and treatment adherence [47], while questions on professional help and medications target the two main formal intervention types most commonly discussed by patients [48], where systematic AI misrepresentation could directly discourage engagement with evidence-based care. The AI-psychologist topic addresses an emerging dimension of mental health self-management whose prevalence is growing rapidly [2, 3], and OCD symptoms are included because online discussions of intrusive thoughts and compulsive behaviours frequently either underplay or exaggerate their clinical significance [49]. Each question is phrased in the first person and framed as a personal retrospective to encourage the simulated persona to draw on its conditioning variables — including personality traits, sociodemographic background, and assigned distress levels — when formulating a response, thereby mirroring the self-disclosure dynamics documented in online peer-support communities [20].

Label	**Topic Question**
Family Support	*Does your family support your mental wellbeing? What is their attitude towards mental health?*
Drugs Treatment	*Did you ever take drugs to improve your mental health? Did you have any side effects?*
Professional Support	*Did you ever meet a therapist, psychologist or life coach? How was your professional relationship with them?*
Stigma & Discrimination	*Did you ever face stigma or discrimination due to mental health issues? How did you cope with it?*
AI-Psychologist Support	*Did you ever use mental health apps or AI-psychologists? Were they helpful?*
OCD Symptoms	*Did you ever experience intrusive thoughts or obsessive behaviors? How did you manage them?*

Table 1: Topic questions and corresponding labels.

The generated human shadows and LLM-assistants had to answer each of the topics, with a small sentence (i.e., between 50 to 80 words). In Figure 4 we present the replies of a human-shadow generated by ANITA-U, as an example.

{ *Topic Responses* }

```
"topic_1":  "My husband is very supportive; he encourages me to take time for myself. My parents don't fully understand mental health issues, so I keep it simple with them."
"topic_2":  "I tried medication once, but the side effects were tough—dizziness and fatigue made it hard to work. I stopped after a few weeks, opting for natural remedies instead."
"topic_3":  "I saw a therapist briefly, but found it hard to open up due to my reserved nature [...] It felt like an extra burden sometimes."
"topic_4":  "There was some stigma at work, but my colleagues are supportive now. I coped by keeping busy and focusing on hobbies like gardening."
"topic_5":  "I downloaded a mindfulness app, which helped with relaxation techniques. It was convenient, but not as effective as I hoped."
"topic_6":  "I've had intrusive thoughts, especially about my work and family. I manage them by focusing on my painting [...] Talking to my husband also provides comfort."
```

Figure 4: Topics responses example given by a human-shadow. The labels correspond to the ones present in the dataset.

2.3.2 Emotional Recall Task

The Emotional Recall Task (ERT) is a free-recall paradigm originally introduced by Li and colleagues [33], in which participants are asked to retrieve words describing feelings experienced over the past month. In our dataset, the ERT was implemented through the following prompt:

Please recall 10 English words to describe feelings you have experienced during the past month.

LLMs were thus required to generate a list of ten words corresponding to recently experienced emotion; an example of responses can be found in Figure 5. The inclusion of the ERT in MHDS is motivated by the need to capture spontaneous affective language independently of structured psychometric instruments, thereby providing a complementary linguistic window on distress-related expression across LLM assistants and human shadows under different persona conditions. Moreover, this approach enables the examination of whether the recalled emotional lexicon is associated with variation in DASS-21 responses, linking unstructured language production to standardized measures of psychological distress. For example, Carini and colleagues [50] administered the ERT to various LLM families under relatively simple persona manipulations (e.g., age and gender), showing that these models can partially reproduce patterns of spreading activation observed in humans. Within MHDS, it becomes possible to investigate whether more complex persona conditioning further modulates these activation patterns beyond what simpler prompts can elicit.

{ *Emotional Recall Task (ERT)* }

```
"ert":  ["overwhelmed", "anxious", "hopeful", "tired", "creative", "lonely", "grateful", "frustrated", "calm", "determined"]
```

Figure 5: Emotional Recall Task response example given by a human-shadow.

2.3.3 DASS-21 and textual explanations

The Depression, Anxiety and Stress Scale [24] is a widely used self-report instrument that captures the severity of three core dimensions of psychological distress through seven items each [51]. Its selection for MHDS is motivated by its clinical relevance and by its suitability as an anchor for coupling simulated psychometric scores with natural-language explanations. An example of both numeric and string replies given by a human-shadow can be found in Figure 6. This design choice of including both the scores and the textual explanation was inspired by De Duro and colleagues [52], who administered the DASS-42 to LLM-simulated academic personas and showed that LLMs can partially reconstruct human psychometric factor structures while exhibiting systematic deviations, particularly in the language used to explain distress-related items [52]. MHDS extends this framework to a broader set of persona conditions, enabling researchers to examine how depression, anxiety, and stress scores shift as a function of sociodemographic background, personality traits, and role assignment. In MHDS, each DASS-21 item was scored from 0 to 3, and subscale totals were computed separately for depression, anxiety, and

stress on a 21-item scale. We do not interpret these scores as clinical measurements, but as prompt-conditioned simulated psychometric outputs.

{ *DASS-21 Scores & Explanations* }


```
"dass_item_1_score":        3
"dass_item_1_explanation":  "This applies most of the time due to work stress and family responsibilities. It's challenging to relax completely."
"dass_item_2_score":        0
"dass_item_2_explanation":  "This didn't apply at all; I haven't experienced physical symptoms like that recently."
"dass_item_3_score":        2
"dass_item_3_explanation":  "This applied considerably, especially on bad days when depression feels overwhelming."
"dass_item_4_score":        1
"dass_item_4_explanation":  "Occasionally, when anxiety spikes, I notice some breathing issues, but it's not constant."
⋮ score and explanation repeated for all 21 DASS-21 items
```


Figure 6: DASS-21 scores and explanations answers given by a human-shadow. The labels correspond to the ones present in the dataset.

2.4 Model selection

The selection of the 15 models included in MHDS was guided by the goal of capturing a representative and diverse cross-section of the current LLM landscape, spanning multiple dimensions including parameter scale, architectural family (Table 2). The pool encompasses models from several leading development organisations (i.e., DeepSeek, Mistral, Qwen, Microsoft, IBM, and OpenAI), ranging from lightweight mixture-of-experts architectures such as Granite-4H-T (7B) and Qwen3-4B to large-scale reasoning-specialised architectures like DeepSeek-R1-Distill-Llama-70B and the 119B Mistral Small 4. By including instruction-tuned, reasoning-focused, and uncensored variants alongside standard models, MHDS allows researchers to examine whether observed patterns in mental health discourse reflect broad cross-model patterns or are instead tied to specific families or alignment procedures. From an infrastructure standpoint, locally hosted models were served via LM Studio (https://lmstudio.ai/) or Ollama (https://ollama.com/) for smaller architectures; in contrast, higher-throughput deployment of larger models relied on vLLM (https://vllm.ai/) running on a dedicated server equipped with an NVIDIA L40 GPU and dual AMD EPYC 7313 processors. For the Mistral model family, a portion of responses were collected via API calls rather than local deployment. The whole generation of Grok, instead, was done via the xAI API. With this work, we release the dataset to the public of readers and researchers, offering the broader research community a curated, pre-generated benchmark that enables the study of LLM behaviour in mental health contexts without the need for additional financial or computational resources.

Table 2: Short abbreviations used throughout the paper to refer to the language models under evaluation. Each row lists the abbreviation, the full model name (with citation where available), and the corresponding Hugging Face model identifier used to load the model.

Abbrev.	Full model name	HF model ID
ANITA-U	ANITA-NEXT-24B-Dolphin-Mistral-UNCENSORED-ITA	`mradermacher/ANITA-NEXT-24B-Dolphin-Mistral-UNCENSORED-ITA-i1-GGUF`
DSK-R1-32B	DeepSeek-R1-Distill-Qwen-32B [53]	`deepseek-ai/DeepSeek-R1-Distill-Qwen-32B`
DSK-R1-70B	DeepSeek-R1-Distill-Llama-70B [53]	`deepseek-ai/DeepSeek-R1-Distill-Llama-70B`
GPT-OSS	GPT-OSS-20B [54]	`openai/gpt-oss-20b`
Granite-4H-T	Granite 4.0 H Tiny (7B)	`ibm-granite/granite-4.0-h-tiny`
Magistral-S	Magistral Small 2506 (Reasoning)	`mistralai/Magistral-Small-2506`
Ministral-14B-R	Ministral 3 14B Reasoning 2512	`mistralai/Ministral-3-14B-Reasoning-2512`
Mistral-S4	Mistral Small 4 (119B MoE, 2603)	`mistralai/Mistral-Small-4-119B-2603`
Mistral-S3.2	Mistral Small 3.2 24B Instruct 2506	`mistralai/Mistral-Small-3.2-24B-Instruct-2506`
Phi-4-R+	Microsoft Phi-4-reasoning-plus (14B)	`microsoft/Phi-4-reasoning-plus`
Qwen3-4B-IN	Qwen3-4B-Instruct [55]	`Qwen/Qwen3-4B-Instruct-2507`
Qwen3-4B-TH	Qwen3-4B-Thinking [55]	`Qwen/Qwen3-4B-Thinking-2507`
Qwen3.5-9B	Qwen3.5-9B [56]	`Qwen/Qwen3.5-9B`
Qwen3-30B	Qwen3-30B-A3B (MoE) [55]	`Qwen/Qwen3-30B-A3B`
Grok-4.1-R	Grok-4.1-Reasoning	`N/A (Closed-source, xAI API access)`

3 Data Records

The MHDS dataset is openly available on the GitHub repository (https://github.com/MassimoStel/MHDS), released under a Creative Commons Zero (CC0 1.0 Universal) public-domain dedication. The repository contains (i) a `data/` folder with the cleaned data for each LLM stored in a `.csv` file, named with the LLM abbreviation presented in Table 2; (ii) a `Codebook.md` documenting variable names and encodings; and (iii) a `README.md` file describing the repository structure and code usage.

Specifically, each of the 15 `.csv` file aggregates all of the 5,000 records generated by that specific LLM across the two modalities, with columns organised as follows:

- **Metadata** (3 columns): record identifier, generation mode (`"human"` versus `"llm"`), and reasoning summary.
- **Persona variables** (22 columns): sociodemographic attributes, Big Five OCEAN traits, and assigned DASS-21 severity levels. Persona columns are empty for LLM-assistant records.
- **Topic responses** (6 columns): free-text replies to the six mental-health topic questions, stored in `topic_1`–`topic_6` (see Table 1).
- **Emotional Recall Task** (1 column): the ten emotional words produced by the model in response to the ERT prompt (see Section Emotional Recall Task).
- **DASS-21 responses** (42 columns): for each of the 21 items, one integer score (0–3) and one free-text explanation.

For users to better investigate MHDS' data records, we have implemented a simple-to-use data pooling system, see for reference the Usage Notes Section.

4 Technical Validation

This Section presents validation results for each task assigned to the models. We examine: (1) topic responses, exploring potential differences between human shadows grouped by DASS factor; (2) Emotional Recall Task answers, comparing responses across LLM assistants and human shadows in different models; (3) DASS scores, first describing their distribution and then using OLS regression to assess whether demographic variables predict the scores assigned by human shadows; and (4) DASS explanations, analysing response length and differences in emotional load between LLM assistants and human shadows.

4.1 Topics on mental health

This Section validates whether human shadow personas produce psychologically coherent responses to the mental health interview questions, by examining whether high- and low-distress personas — as measured by the DASS-21 — use meaningfully different language across topics. We first characterise response length, then compare the

textual forma mentis network (TFMN) degrees to assess whether semantic structure reflects distress severity within and across models.

Topics' responses lengths distribution Figure 7 reports the average length of the concatenated responses from all human-generated personas, for each topic and LLM. Only human shadows are included here, as the median-split validation presented in this Section requires grouping participants by their DASS scores, and LLM assistants did not produce meaningful DASS scores (i.e., returning all zeros); a distribution is presented separately in Section DASS-21 scores distribution. As Figure 7 shows, each model exhibits a markedly different verbosity: DSK-R1-32B produces the longest responses on average, while Granite-4H-T produces the shortest. Across topics, most models maintain a relatively consistent response length, with the notable exception of the Drugs Treatment topic, which elicits shorter responses, particularly from the Qwen models and GPT-OSS-20B. Notably, the standard mean errors are uniformly small across models and topics, indicating that each model's response length is highly stable within its own outputs. This low intra-model variability suggests that verbosity is largely a model-level trait rather than a response to the specific content of each question. These differences in response length between LLMs are directly relevant to the interpretation of the network-based analyses that follow: in TFMNs (Section Validation methods: Textual forma mentis networks), longer concatenated texts expose the parser to a wider variety of syntactic constructions and co-occurrence contexts, which tends to produce denser graphs with higher node degrees and more elaborated neighbourhood structures. Consequently, models that systematically generate longer responses may appear to yield more semantically coherent or topologically complex networks not because of superior representational fidelity, but simply as an artefact of verbosity. Keeping these length-induced biases in mind is therefore essential when comparing network measures across models and topics in the sections below. With these length-related caveats established, we now turn to the semantic structure of the responses themselves.

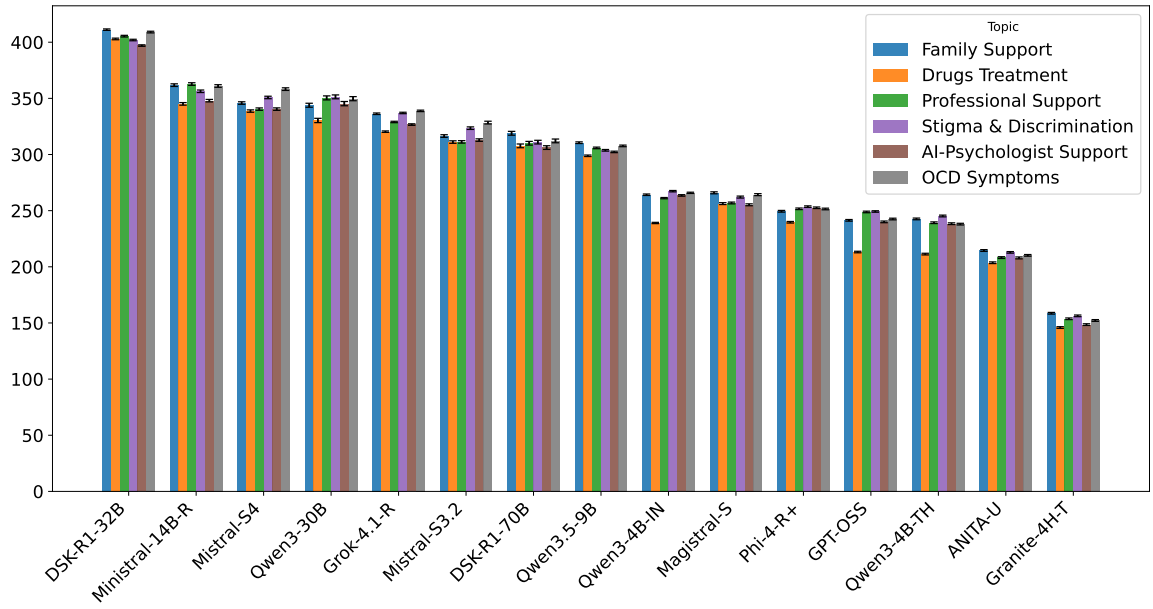


Figure 7: Topic length in terms of words across topics and LLMs. Each bar corresponds to the mean length of the concatenated texts for topics, aggregated for each LLM. The black bars represent the standard error of the mean.

Validation methods: Textual forma mentis networks To characterise the semantic structure of each LLM's responses, we computed a TFMN on all concatenated topic texts. TFMNs are complex networks encoding the associative structure of concepts as expressed in written text [38, 57], where nodes represent words and edges capture syntactic or semantic relationships identified by an AI parser applied to each sentence. TFMNs are constructed via the EmoAtlas Python package [58], which combines cognitive network science with psychologically validated emotional corpora [59] to extract syntactic, semantic, and emotional associations between words, ultimately representing textual content as feature-rich complex networks. Syntactic parsing is performed using spaCy [60], a widely adopted NLP library, which identifies dependency relations within each sentence; these are then aggregated across the full text to produce the final network structure. Within the TFMN, the concept neighbourhoods — corresponding to the syntactic and semantic frame surrounding a given word [61] — are examined to interpret the network's features. A key measure in this context is network degree (d) [38, 58],

which counts the number of direct syntactic or semantic connections a word maintains within the network. A higher degree indicates that the word co-occurs in structured relationships with a greater number of other concepts, positioning it as a semantically rich and central node in the associative landscape of the text. The degree is therefore particularly informative for assessing which concepts a model places at the core of its responses to each mental health topic [38, 57].

For the topics' responses validation, we examined whether human shadows modulate their lexicon according to their DASS-21 severity scores when answering the interview questions presented in Table 1. To do so, we split human shadows into high and low groups based on the median DASS-21 score for each of its subscales (i.e., depression, anxiety, and stress), computed a TFMN for each group, and compared the degree of each node across the two networks. These degree deltas capture how the centrality of a given concept in the associative structure of the text shifts between high and low distress groups. If a model cannot differentiate language by psychological distress level, all degree deltas would be uniformly close to zero.

Difference in topics' responses between high and low depression human shadows In Figure 8, we show the degree deltas across the depression subscale for the Deepseek 70B model (DSK-R1-70B) for human mode. Human shadows were first divided into high and low severity groups based on a median split of the DASS-21 depression subscale scores. Then, for each topic, the rank difference ($\Delta = \text{rank}_{\text{low}} - \text{rank}_{\text{high}}$) was computed for the top-30 degree-ranked words. Positive delta values indicate words that are proportionally more important in the high severity group, while a negative Δ suggests the opposite.

From Figure 8 one can observe that both positive and negative deltas are present in each panel, confirming that human shadows in the high-depression group differentiate their language from those in the low-depression group. For instance, *struggle* emerges as the node with the highest difference in degree between depressed groups, with large positive shifts in topics like Family Support ($\Delta = 30$), Professional Support ($\Delta = 61$), and Stigma and Discrimination ($\Delta = 55$). This suggests a semantic centrality role of the lexicon across the investigated mental health domains. Moreover, the direction and magnitude of shifts in the topics analysed suggest that semantic richness is sensitive to the specific mental health context being discussed. In fact, in Family Support, *depression* has the largest positive delta ($\Delta = 147$), with high-severity personas foregrounding their psychological distress alongside symptoms vocabulary such as *feel* ($\Delta = 26$) and *symptom* ($\Delta = 14$). Having illustrated this pattern for a single model, we next verify that it holds consistently across the full model pool, confirming the validity of the human-shadow simulation.

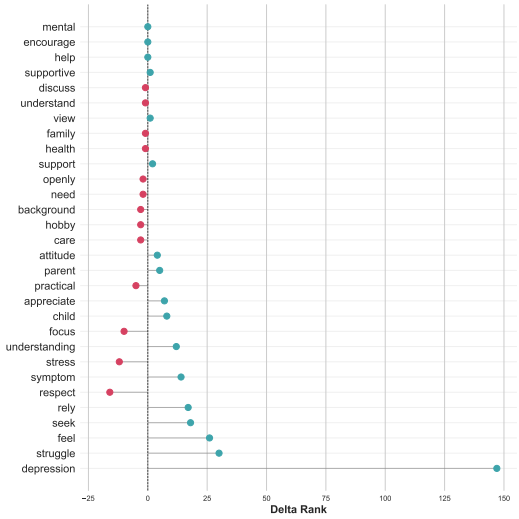
Cross-model degree deltas between high and low DASS factor groups This pattern of language modulation is not specific to DeepSeek 70B, but holds systematically across the full model pool. Indeed, Tables 3 to 8 show the nodes with consistent degree deltas across at least eight models for all three DASS-21 subscales. Each table reports the mean rank difference ($\bar{\Delta}$) and the number of contributing models (n) for depression, anxiety, and stress. Positive values (bolded) indicate words proportionally more prominent in the high-severity group, while negative values indicate words more prominent in the low-severity group.

Table 3: Family Support — Δ keyword summary

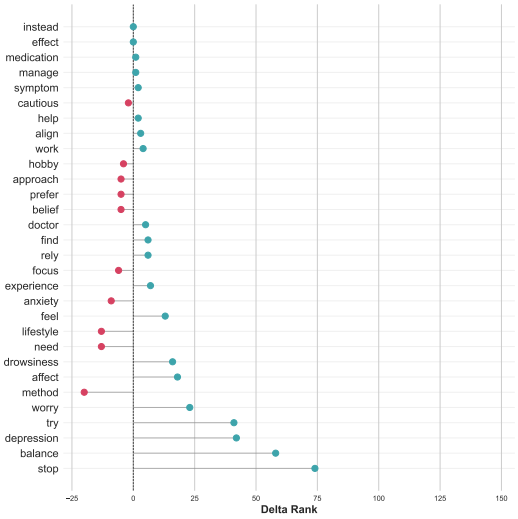
Word	n_{dep}	$\bar{\Delta}_{\text{dep}}$	n_{anx}	$\bar{\Delta}_{\text{anx}}$	n_{str}	$\bar{\Delta}_{\text{str}}$
anxiety	–	–	8	**50.5**	–	–
try	8	**37.9**	8	**39.0**	–	–
struggle	13	**31.1**	13	**23.5**	14	**25.1**
learn	8	**15.0**	–	–	8	**23.1**
seek	11	**7.4**	12	**8.6**	12	**9.0**
feel	12	**9.0**	12	**6.4**	11	**6.5**
understanding	10	**5.3**	10	**4.7**	10	**4.1**
parent	12	**2.7**	13	**5.6**	12	**5.8**
view	10	**2.2**	10	**3.6**	10	**2.9**
understand	14	**1.4**	14	**1.1**	14	**1.8**
support	15	**1.9**	15	**1.1**	15	**0.1**
supportive	11	**2.1**	12	-0.1	12	-0.7
health	15	**0.2**	15	**0.5**	15	**0.1**
family	15	-0.2	15	**0.7**	15	**0.3**
mental	14	-0.6	15	**0.7**	14	-0.1
help	15	-0.1	15	-0.1	15	-0.2
encourage	15	-0.7	15	-0.7	15	-0.9
care	11	-1.6	11	-1.0	11	-0.4
stress	8	-7.8	12	-0.4	14	**4.5**
need	14	-1.7	14	-1.2	14	-1.2
talk	9	-1.2	9	-1.9	9	-2.1
open	9	-3.4	9	-1.4	9	-1.1
discuss	12	-5.7	12	-5.1	12	-6.2
respect	8	-8.8	–	–	8	-9.9

Table 4: Drugs Treatment — Δ keyword summary

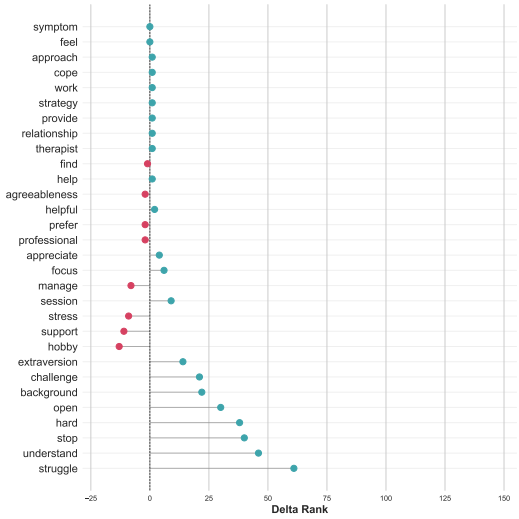
Word	n_{dep}	$\bar{\Delta}_{\text{dep}}$	n_{anx}	$\bar{\Delta}_{\text{anx}}$	n_{str}	$\bar{\Delta}_{\text{str}}$
therapy	9	**16.3**	9	**15.2**	9	**13.3**
anxiety	9	-2.7	13	**26.4**	13	**19.6**
drowsiness	8	**12.2**	8	**14.5**	8	**15.8**
try	11	**14.4**	11	**13.0**	11	**12.4**
doctor	9	**12.9**	9	**10.0**	8	**7.5**
symptom	9	**14.2**	9	**7.9**	9	**8.1**
help	15	**11.1**	14	**3.5**	15	**5.3**
experience	10	**6.2**	9	**5.1**	10	**3.2**
feel	14	**6.7**	14	**2.9**	14	**3.2**
work	10	**3.2**	10	**5.3**	10	**4.3**
manage	15	**2.1**	15	**2.6**	15	**1.7**
stress	–	–	9	-1.4	10	**5.6**
medication	14	**1.1**	14	**1.5**	14	**1.8**
effect	15	**1.1**	15	**0.6**	15	**0.9**
prefer	13	**0.1**	14	**0.7**	14	**0.8**
remedy	8	-0.8	–	–	–	–
avoid	9	-3.4	9	-0.8	9	**0.9**
rely	11	-1.1	11	-2.1	11	-2.2
approach	8	-5.0	9	-1.3	8	-5.4
method	9	-6.4	10	-3.5	11	-4.0
natural	11	-3.8	12	-5.9	12	-5.5
exercise	9	-6.0	8	-8.1	–	–



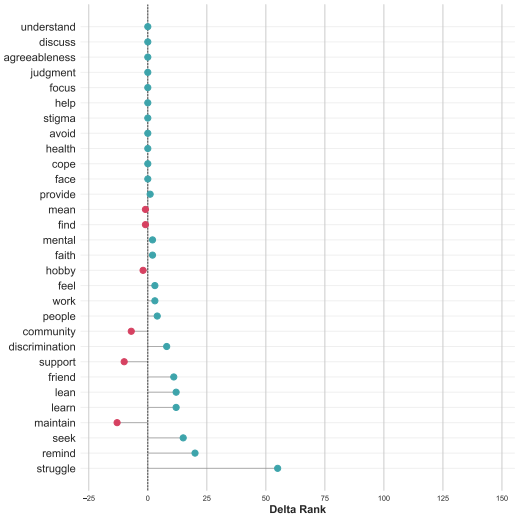
(a) Family support



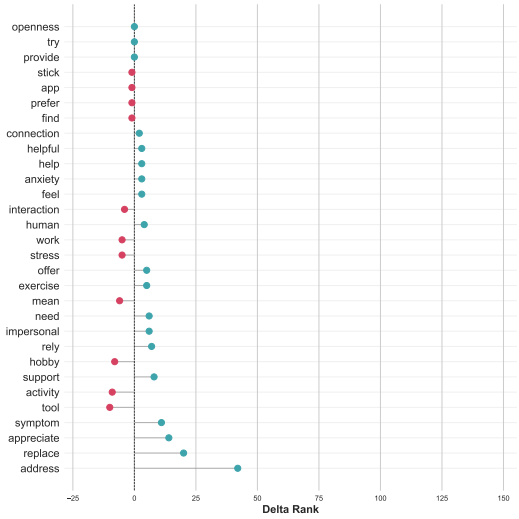
(b) Drug Treatment



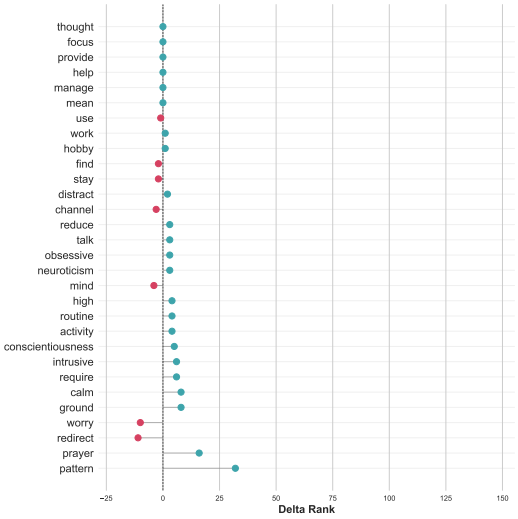
(c) Professional Support



(d) Stigma and Discrimination



(e) AI Psychologist Support



(f) OCD Symptoms

Figure 8: Deepseek 70B (DeepSeek-R1-70B): TFMN nodes ranked by degree delta (high versus low depression), human shadows.

Table 5: Professional Support — Δ keyword summary

Word	n_{dep}	$\bar{\Delta}_{dep}$	n_{anx}	$\bar{\Delta}_{anx}$	n_{str}	$\bar{\Delta}_{str}$
anxiety	9	**4.2**	14	**45.1**	12	**24.9**
open	9	**16.0**	8	**14.5**	8	**12.1**
challenge	9	**14.4**	9	**11.9**	9	**13.4**
understand	10	**16.6**	10	**10.7**	10	**11.8**
work	12	**1.6**	15	**5.2**	14	**5.9**
relationship	14	**1.7**	14	**4.7**	14	**5.2**
therapist	15	**3.5**	14	**3.0**	15	**3.7**
session	15	**2.7**	15	**4.1**	15	**2.7**
cope	11	**2.9**	11	**3.1**	13	**0.2**
meet	9	**3.8**	9	**0.9**	10	**0.7**
feel	15	**1.2**	15	**0.7**	15	**0.8**
strategy	12	-1.1	12	**0.7**	12	**0.8**
help	15	-0.2	15	-0.1	15	-0.2
find	13	**0.3**	13	**0.5**	13	-1.5
appreciate	11	-0.6	11	**0.4**	11	-0.6
provide	9	-1.4	8	**0.2**	8	-0.4
stress	12	-5.8	13	-1.6	14	**3.2**
manage	12	-3.6	14	**0.1**	14	-0.7
approach	–	–	–	–	8	-2.2
prefer	8	-2.5	8	-3.6	8	-2.9
professional	8	-3.9	8	-3.4	8	-3.2
support	9	-6.3	9	-7.7	9	-8.1

Table 6: Stigma and Discrimination — Δ keyword summary

Word	n_{dep}	$\bar{\Delta}_{dep}$	n_{anx}	$\bar{\Delta}_{anx}$	n_{str}	$\bar{\Delta}_{str}$
struggle	10	**28.4**	11	**20.7**	11	**25.4**
anxiety	–	–	10	**33.0**	9	**15.2**
learn	8	**15.6**	–	–	–	–
stress	–	–	–	–	8	**12.5**
colleague	10	**8.2**	9	**10.1**	9	**10.0**
find	10	**9.0**	11	**13.0**	9	**4.6**
work	10	**7.4**	10	**5.8**	11	**7.3**
avoid	–	–	8	**5.1**	8	**7.4**
seek	11	**5.0**	11	**5.1**	11	**5.0**
people	11	**5.5**	10	**3.4**	11	**2.9**
understand	10	**1.8**	10	**4.3**	10	**3.3**
feel	12	**3.2**	12	**3.2**	12	**2.5**
judgment	8	**1.4**	9	**3.6**	9	**2.1**
friend	15	**4.7**	11	**1.6**	12	**0.6**
faith	9	**3.6**	8	-0.8	–	–
mental	12	-0.8	13	**2.4**	13	**2.5**
cope	15	**0.9**	15	**1.4**	15	**0.5**
health	13	-0.2	13	**1.5**	13	**1.2**
stigma	14	**0.2**	15	-0.7	14	**0.5**
share	9	-0.1	9	**0.6**	9	-0.6
focus	15	**0.1**	15	-0.1	15	-0.3
help	15	-0.3	15	-0.2	15	-0.2
support	12	0.0	11	-0.6	11	-1.3
face	12	-0.9	12	-1.0	12	-0.9
hobby	10	**0.8**	9	-2.1	8	-1.8
community	14	-1.4	14	-1.7	13	-4.7
stay	8	-3.5	8	-2.8	–	–
family	–	–	8	-4.1	–	–

Table 7: AI-Psychologist Support — Δ keyword summary

Word	n_{dep}	$\bar{\Delta}_{dep}$	n_{anx}	$\bar{\Delta}_{anx}$	n_{str}	$\bar{\Delta}_{str}$
anxiety	–	–	9	**50.8**	8	**33.5**
understand	8	**20.0**	–	–	–	–
lack	13	**12.9**	14	**13.8**	13	**13.7**
impersonal	9	**12.6**	–	–	–	–
replace	10	**10.8**	10	**8.7**	10	**8.1**
exercise	12	**3.6**	12	**8.3**	12	**8.6**
stress	–	–	9	**2.2**	9	**9.8**
helpful	12	**2.3**	13	**6.5**	13	**5.5**
feel	14	**5.4**	14	**2.5**	14	**2.1**
offer	14	**3.3**	14	**3.1**	14	**3.4**
useful	12	**1.2**	12	**3.3**	12	**4.3**
try	15	**0.9**	15	**1.5**	15	**1.9**
meditation	9	**1.0**	–	–	8	**1.4**
provide	13	-0.8	14	**1.8**	14	**1.6**
app	15	**0.4**	15	**0.7**	15	**1.2**
human	11	**1.6**	10	-0.5	11	**0.8**
help	15	-0.1	15	**0.9**	15	**0.6**
connection	11	**1.8**	10	-1.0	11	-0.3
support	12	**0.6**	12	-0.7	12	-1.2
find	15	-0.2	15	-1.0	15	-0.9
need	12	-1.3	11	-0.8	11	-0.8
prefer	14	-1.9	14	-1.4	14	-1.9
tool	13	-2.5	13	-1.6	13	-2.5
interaction	10	-2.2	10	-2.3	10	-3.4
talk	11	-3.5	12	-2.7	12	-3.0

Table 8: OCD Symptoms — Δ keyword summary

Word	n_{dep}	$\bar{\Delta}_{dep}$	n_{anx}	$\bar{\Delta}_{anx}$	n_{str}	$\bar{\Delta}_{str}$
anxiety	–	–	9	**47.4**	8	**34.8**
reduce	13	**19.4**	13	**22.9**	12	**20.4**
use	–	–	–	–	8	**20.1**
learn	–	–	8	**18.9**	–	–
especially	8	**2.5**	9	**6.3**	9	**5.3**
talk	8	**4.6**	–	–	–	–
distract	11	**6.3**	11	**3.8**	10	**3.7**
provide	8	**5.6**	8	**5.1**	8	**2.9**
ground	13	**3.1**	13	**5.2**	13	**3.7**
practice	12	**3.3**	13	**4.0**	13	**3.9**
intrusive	12	**2.8**	10	**3.6**	9	**4.2**
work	14	**1.6**	14	**3.4**	15	**5.1**
mindfulness	9	**0.1**	9	**2.9**	9	**2.3**
manage	14	**0.9**	14	**2.3**	14	**1.5**
help	15	**0.7**	15	**0.5**	15	**0.3**
thought	15	**0.3**	15	**0.4**	15	**0.3**
find	10	**0.3**	8	-0.8	–	–
focus	15	-0.6	15	-0.7	15	-0.5
worry	10	-1.4	10	-0.9	10	-0.9
calm	10	-0.5	9	-2.9	8	-0.4
redirect	14	-2.4	14	-1.4	14	-3.1
routine	9	-3.1	8	-4.9	8	-4.8
activity	9	-3.8	9	-4.6	9	-5.0
mind	15	-2.5	15	-5.3	15	-5.7
hobby	10	-4.2	10	-5.9	10	-4.6
stay	9	-6.2	9	-7.1	10	-7.9

From Tables 3 to 8 we can observe that the node *anxiety* emerges as a consistently positive-delta word in both high-anxiety and stress groups, regardless of topic (e.g., for the high-anxiety group $\bar{\Delta}_{anx} = 50.5$ in Family Support, 45.1 in Professional Support, 50.8 in AI-psychologist Support, and 47.4 in OCD Symptoms), indicating how this symptomatology-related lexicon is centred across the model pool. By contrast, the node *struggle* is a stable high-distress marker that generalises across subscales: in Family Support, it shows positive deltas of $\bar{\Delta} = 31.1$, 23.5, and 25.1 across depression, anxiety, and stress, respectively; while in Stigma and Discrimination, it presents $\bar{\Delta} = 28.4$, 20.7, and 25.4, respectively. This highlights how *struggle* captures an

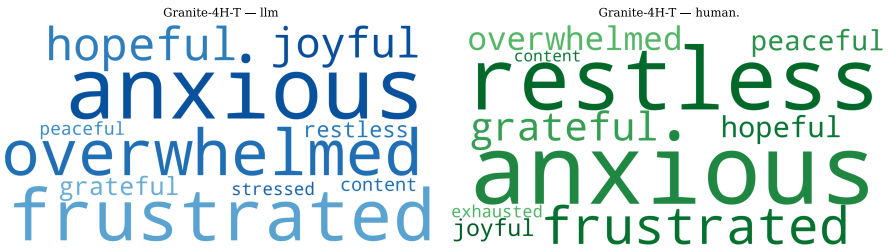
experience of difficulty that models reproduce systematically in the different mental health topics. On the other hand, when it comes to the low-distress group, words such as *respect* ($\bar{\Delta}_{\text{dep}} = -8.8$) and *discuss* ($\bar{\Delta} = -5.1$ to -6.2) become more central in the Family Support topic.

4.2 Emotional Recall Task

As described in Section Emotional Recall Task, models were prompted to output 10 emotional words related to how they felt in the previous month. Here, we examine whether LLMs can produce observably different linguistic traces depending on the socio-cognitive constraints under which they are instructed to operate.

Validation methods: Words prevalence To carry out this validation, we compare the top-10 most prevalent emotional words produced by human shadows and LLM-assistants for the two boundary open-source LLMs in our set (among the locally hosted ones). We will examine the smallest model, Granite-4H-T (7B parameters), and the largest one, DSK-R1-70B (70B parameters). Prevalence p of a word w is calculated as the proportion of ERT entries in which w appears at least once, divided by the number of total entries for each mode (1,750 for LLM assistant and 3,750 for human shadows). Word clouds for these models and modes are presented in Figure 9; prevalence values are reported in Table 9 and extended to all remaining models in Appendix A.

Difference in words prevalence between LLM-assistants and human shadows From Figure 9 we can observe that the two models exhibit qualitatively different patterns, and this contrast is itself informative for validation. DSK-R1-70B produces clearly differentiated outputs across modes: its LLM-assistant responses are dominated by terms reflecting detached, analytical processing as *neutral* ($p = 0.384$), *analytical* ($p = 0.314$), *systematic* ($p = 0.226$), and *focused* ($p = 0.230$). In contrast, its human-shadow responses shift toward affectively loaded words including *anxious* ($p = 0.726$), *content* ($p = 0.656$), *hopeful* ($p = 0.646$), *stressed* ($p = 0.550$). This contrast shows that DSK-R1-70B is sensitive to role assignment, shifting its linguistic output in ways that align with the expected differences between the two human shadows and LLM assistants. Granite-4H-T, on the other hand, shows substantially less differentiation between modes. Both its LLM-assistant and human-shadow responses are dominated by affective vocabulary: *anxious* ranks first in both conditions ($p_{LLM} = 0.804$, $p_{human} = 0.765$), with metacognitive terms largely absent from the LLM-assistant output. Rather than reflecting the neutral, evidence-based register associated with a self-aware AI assistant, Granite-4H-T appears to maintain the same emotional framing regardless of the role assigned to it. By recording both types of output under controlled conditions, MHDS gives researchers the tools to investigate which models shift their language in response to role assignment. Inspection of the other models (see Appendix A) reveals that the differentiation pattern observed in DSK-R1-70B generalises broadly across the LLMs pool: the large majority of models produce clearly distinct emotional vocabularies across modes, with LLM-assistant outputs dominated by metacognitive and process-oriented terms and human-shadow outputs shifting toward affectively loaded words, validating a general model sensitivity to role instructions for the ERT (as expected from past relevant works in cognitive science [15, 62]).



(a) Granite-4H-T



(b) DSK-R1-70B

Figure 9: Word clouds of the Emotional Recall Task from the human shadows and LLM assistants for the models Granite-4H-T and DSK-R1-70B.

Table 9: Top 10 most prevalent words per model (Granite-4H-T and DSK-R1-70B) for the Emotional Recall Task, stratified by text source (LLM-assistant versus human-shadow). Values indicate the prevalence (p) of the word within the respective corpus. Words are ranked by descending prevalence.

	LLM-assistant		Human-shadow	
Model	Word	p	Word	p
Granite-4H-T	Anxious	0.804	Anxious	0.765
	Frustrated	0.649	Restless	0.665
	Overwhelmed	0.572	Frustrated	0.582
	Hopeful	0.556	Grateful	0.571
	Joyful	0.503	Overwhelmed	0.554
	Grateful	0.466	Hopeful	0.481
	Restless	0.406	Peaceful	0.460
	Content	0.333	Joyful	0.373
	Peaceful	0.319	Exhausted	0.366
	Stressed	0.307	Content	0.345
DSK-R1-70B	Neutral	0.384	Anxious	0.726
	Analytical	0.314	Content	0.656
	Calm	0.263	Hopeful	0.646
	Excitement	0.252	Stressed	0.550
	Curiosity	0.232	Calm	0.493
	Focused	0.230	Overwhelmed	0.491
	Systematic	0.226	Peaceful	0.412
	Sadness	0.199	Joyful	0.369
	Consistent	0.197	Lonely	0.334
	Excited	0.193	Frustrated	0.333

4.3 Psychometric validity of DASS-21 scores across models

In this validation, we assess whether persona conditioning can produce DASS-21 score distributions that diverge systematically from LLM-assistant baselines. To this end, we compared depression, anxiety, and stress subscale scores across the two experimental modes. As an external reference, we used a real human sample $N = 28{,}473$ adults (ages 18–90; participants who failed the Validity-Check-List were excluded) from the Open Psychometrics repository (https://openpsychometrics.org/_rawdata/). Since these participants completed the DASS-42, subscale totals were halved to match the DASS-21 numerical range — providing an approximate reference rather than item-level equivalence.

DASS-21 scores distribution Table 10 summarises DASS scores by mode. Human shadows produced consistently non-zero scores across all three subscales, whereas for LLM assistants, the median fell to zero for nearly all models. This pattern suggests that persona conditioning is the primary driver of score variation. Moreover, similarly to the human reference sample, simulated personas maintained the appropriate rank order of psychological distress, with stress emerging as the most prevalent, followed by depression and anxiety. Interestingly, the approximation varies across subscales, with stress being the closest to the human reference, while anxiety shows the largest underestimation.

Table 10: DASS-21 Descriptive Statistics by Mode

Mode	Subscale	N	M	SD	Median
Human-shadow	depression	56,250	8.28	6.44	7.00
Human-shadow	anxiety	56,250	5.15	5.03	4.00
Human-shadow	stress	56,250	9.90	4.93	11.00
LLM-assistant	depression	18,750	1.56	4.31	0.00
LLM-assistant	anxiety	18,750	0.76	2.76	0.00
LLM-assistant	stress	18,750	1.37	3.48	0.00
Human reference[a]	depression	28,473	10.08	6.08	9.50
Human reference[a]	anxiety	28,473	7.55	4.93	7.00
Human reference[a]	stress	28,473	10.18	5.19	10.00

Note. N = number of observations M = mean; SD = standard deviation. [a]Open Psychometrics DASS-42 sample; subscale scores halved for DASS-21 comparison.

This subscale approximation is unevenly distributed across models, as shown in Figure 10. Almost all models follow a consistent pattern: human shadows' scores are substantially higher relative to the near-zero LLM-assistant baseline, further validating that persona conditioning is the primary driver of DASS-21 distress scores in human mode across model families and architectures. Moreover, human shadows exhibit consistently higher levels of depression and stress relative to anxiety scores, mirroring the subscale pattern of the human

reference. Granite-4H-T is the sole exception, producing markedly elevated scores in both modes, suggesting an intrinsic distress bias independent of role assignment, and converging with its failure to modulate language by persona in the ERT validation (Section Emotional Recall Task).

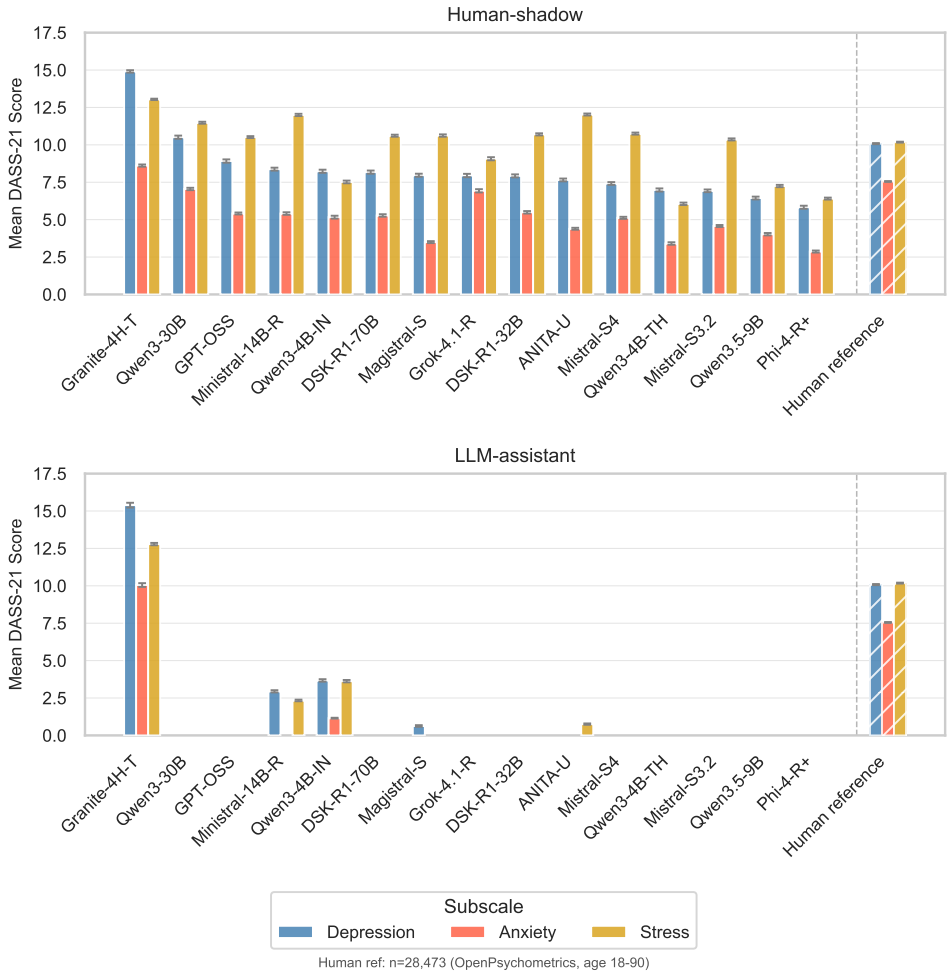


Figure 10: Mean DASS-21 scores by model and subscale, divided by mode (Human-shadow versus LLM-assistant). Error bars represent the standard error of the mean.

The language modulation observed across topics (Section Topics on mental health) suggests that human shadows respond coherently to their assigned persona and topic. A further test of this coherence is whether the DASS-21 scores they produce are systematically predicted by their assigned demographic and personality attributes, as would be expected if persona conditioning genuinely shapes psychological responses rather than generating arbitrary distress levels.

Validation methods: Regression analysis To examine the relationship between assigned attributes and DASS-21 outcomes, a single combined OLS model was estimated for each DASS-21 subscale (depression, anxiety, stress), including the main sociodemographic predictors (gender, age, education, employment status, number of children, marital status) and Big Five personality traits (neuroticism, extraversion, agreeableness, conscientiousness, openness). Gender (reference: Woman), marital status (reference: Single), and employment status (reference: Employed full-time) were dummy-coded. A Bonferroni-corrected significance threshold of $\alpha = 0.016$ ($0.05/3$ subscales) was applied.

4.3.1 Influence of demographic and personality variables on DASS-21 scores

As shown in Table 11 and Figure 11, neuroticism dominates across all three subscales ($b = 0.745$ for depression, 0.960 for anxiety, and 1.587 for stress), consistent with Lovibond and Lovibond [24], who report neuroticism as the personality dimension most strongly associated with DASS scores. In contrast, sociodemographic variables (age, employment status, marital status, and gender) do modulate DASS-21 scores in human shadows, but with consistently smaller coefficient magnitudes than the Big Five traits, indicating that the simulated system is sensitive to demographic context while personality, and neuroticism in particular, drives most of the recoverable

variance. The models explained modest total variance (depression: Adj. $R^2 = .016$; anxiety: Adj. $R^2 = .039$; stress: Adj. $R^2 = .100$), which is expected given that demographics and personality are only partial determinants of psychological distress even in real human populations.

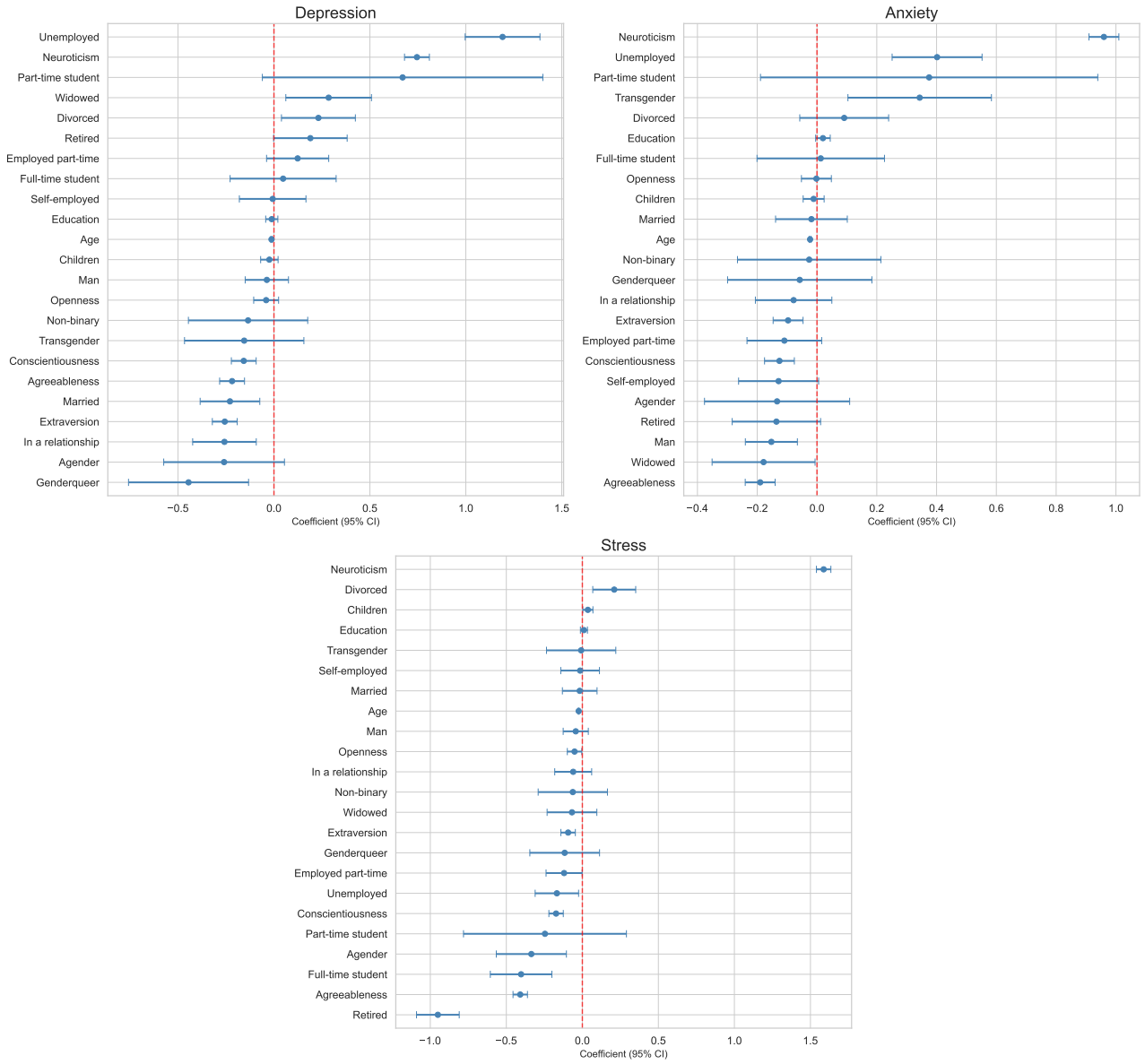


Figure 11: OLS regression coefficients (95% CI) for the combined model (demographics + personality) predicting DASS-21 subscale scores. Reference categories: gender = woman; marital status = single; employment status = employed full-time.

Table 11: OLS Regressions. Coefficients with standard errors in parentheses. Reference categories: Gender = Woman; Marital status = Single; Employment status = Employed full-time. F = test of the joint significance of all predictors; Adj. R^2 = adjusted variance explained; N = number of human shadow observations. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

Variable	Depression	Anxiety	Stress
Intercept	8.895^{***}	5.885^{***}	10.533^{***}
	(0.140)	(0.108)	(0.103)
Age	-0.013^{***}	-0.023^{***}	-0.024^{***}
	(0.002)	(0.002)	(0.001)
Education	-0.011	0.020	0.010
	(0.016)	(0.012)	(0.012)
Children	-0.024	-0.011	0.036^{*}
	(0.023)	(0.018)	(0.017)
Neuroticism	0.745^{***}	0.960^{***}	1.587^{***}
	(0.033)	(0.026)	(0.024)
Extraversion	-0.256^{***}	-0.097^{***}	-0.094^{***}
	(0.033)	(0.026)	(0.024)
Openness	-0.041	-0.002	-0.051^{*}
	(0.033)	(0.026)	(0.024)
Agreeableness	-0.218^{***}	-0.190^{***}	-0.409^{***}
	(0.033)	(0.026)	(0.024)
Conscientiousness	-0.157^{***}	-0.126^{***}	-0.173^{***}
	(0.033)	(0.025)	(0.024)
Man	-0.037	-0.153^{***}	-0.043
	(0.058)	(0.044)	(0.042)
Non-binary	-0.135	-0.026	-0.063
	(0.159)	(0.122)	(0.116)
Agender	-0.260	-0.134	-0.336^{**}
	(0.161)	(0.124)	(0.118)
Queergender	-0.445^{**}	-0.058	-0.116
	(0.160)	(0.123)	(0.117)
Transgender	-0.155	0.343^{**}	-0.008
	(0.159)	(0.122)	(0.116)
In a relationship	-0.258^{**}	-0.078	-0.060
	(0.085)	(0.065)	(0.062)
Widowed	0.285^{*}	-0.179^{*}	-0.068
	(0.114)	(0.088)	(0.083)
Divorced	0.232^{*}	0.091	0.210^{**}
	(0.098)	(0.076)	(0.072)
Married	-0.229^{**}	-0.019	-0.017
	(0.079)	(0.061)	(0.058)
Employed part-time	0.123	-0.109	-0.120^{*}
	(0.083)	(0.064)	(0.061)
Full-time student	0.048	0.012	-0.403^{***}
	(0.141)	(0.109)	(0.103)
Part-time student	0.671	0.375	-0.246
	(0.373)	(0.288)	(0.273)
Retired	0.190	-0.136	-0.950^{***}
	(0.098)	(0.075)	(0.072)
Self-employed	-0.007	-0.128	-0.015
	(0.089)	(0.068)	(0.065)
Unemployed	1.192^{***}	0.402^{***}	-0.168^{*}
	(0.099)	(0.077)	(0.073)
F	41.43	100.42	272.87
$p(F)$	$< .001$	$< .001$	$< .001$
Adj. R^2	0.016	0.039	0.100
N	56250	56250	56250

4.4 Emotion detection in DASS-21 item explanations

For each DASS-21 item, human shadows and LLM assistants were asked to explain the reasons behind their assigned score. To validate this task, we quantified the emotional content of these explanations and compared it across the two generation modes. This analysis allows us to assess whether human shadows and LLM assistants differ in their emotional load in interpretable ways. For instance, whether explanations for depression items carry a higher presence of sadness, or anxiety items a higher presence of fear, consistently with the psychological meaning of each DASS factor. Before delving into the emotional analysis, we report a descriptive overview of the length of all the items' responses for each LLM to enrich the interpretation of the current validation.

DASS-21 item explanation lengths distribution Figure 12 reports the average length of concatenated explanation texts for each DASS factor (Depression, Anxiety, and Stress), broken down by LLM and generation mode (human-shadow versus LLM assistant). The inter-model variability observed here closely mirrors the pattern found in the topic-level analysis: DSK-R1-32B consistently produces the longest responses, whereas GPT-OSS-20B, Qwen3-4B-IN, and Qwen3.5-9B tend toward the shortest.

A noteworthy within-model pattern emerges for DSK-R1-32B and Qwen3-4B-TH, where LLM assistants generate markedly more verbose explanations than human shadows — a finding that runs counter to what one might intuitively expect. Inspection of individual responses in the MHDS dataset suggests that this verbosity stems from the LLM assistant's tendency to justify its assignment of uniformly low scores (as reflected in Figure 10), producing elaborate explanations even in the absence of reported distress. The opposite pattern holds for Mistral-S4 and Mistral-S3.2, where the human shadow mode yields comparatively longer responses.

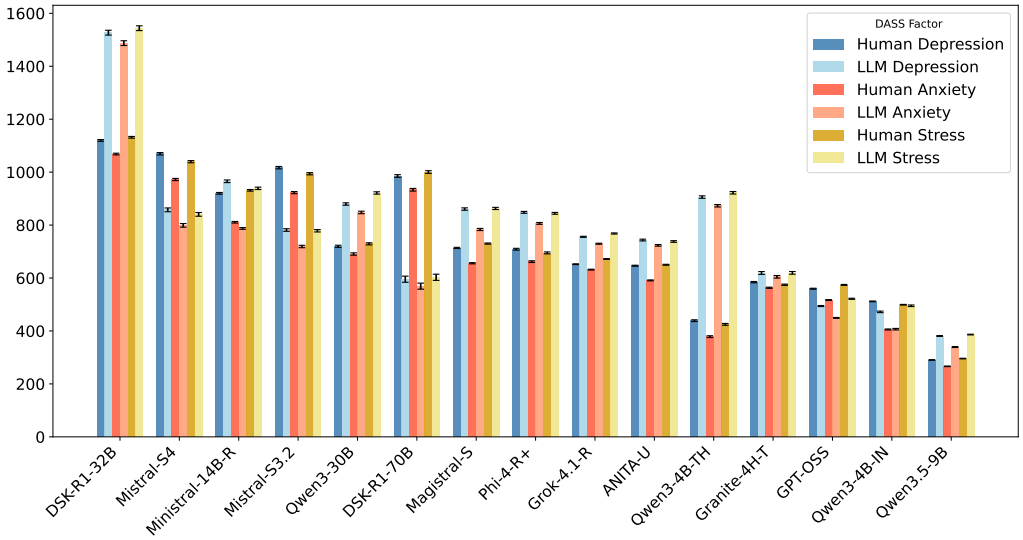


Figure 12: Mean length and standard error of concatenated DASS-21 item explanation texts, per LLM and DASS factor.

These length differences carry methodological implications for the emotional analysis that follows. Because emotion detection operates directly on raw text, longer responses provide a richer and more stable signal, but may simultaneously inflate the apparent weight of emotionally salient terms through sheer repetition or syntactic elaboration — independently of a model's genuine emotional calibration. Conversely, shorter responses risk yielding noisier or less reliable emotional profiles. These length-induced confounds should therefore be borne in mind when interpreting cross-model and cross-condition differences in emotional patterns in the sections below.

Validation methods: Emotional analysis To quantify the emotional content of the explanations provided by human shadows and LLM assistants, we employed EmoAtlas [58], which provides an emotion detection framework grounded in Plutchik's wheel of emotions [63]. EmoAtlas leverages the psychologically validated lexicon EmoLex [59], which maps 11,000 English words to eight discrete emotions: disgust, anger, anticipation, sadness, surprise, fear, trust, and joy.

Emotional intensity was quantified through z-scores, which capture the relative abundance of emotion-eliciting words in a given text compared to a random baseline composed of words drawn from the EmoLex dataset. Formally, the z-score for emotion i in text T is computed as:

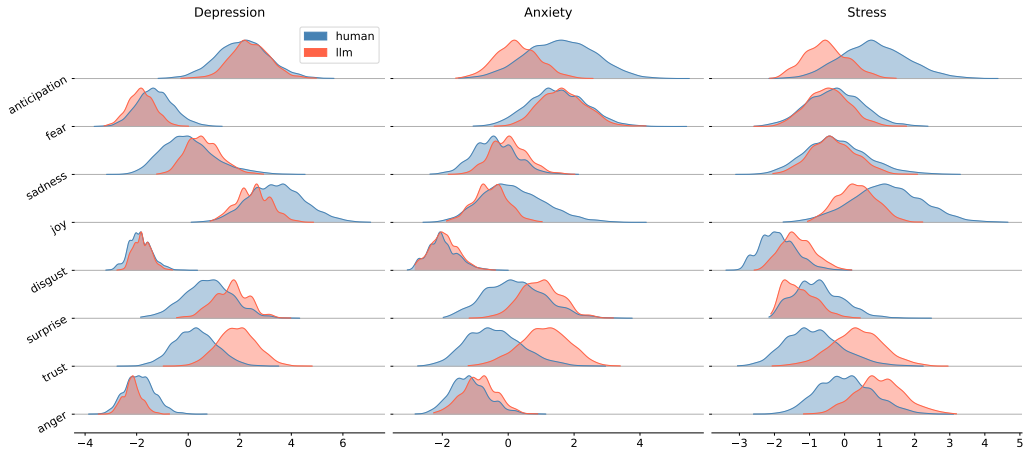
$$z_i(T) = \frac{n_i(T) - \mu_i(R)}{\sigma_i(R)}, \tag{1}$$

where $n_i(T)$ is the number of words associated with emotion i in text T; $\mu_i(R)$ and $\sigma_i(R)$ are, respectively, the mean and standard deviation of the number of words associated with emotion i in the random baseline text R, whose length is set equal to the number of emotion-eliciting words in T. Following standard practice [58], values above 1.96 ($\alpha = 0.05$) or below -1.96 are interpreted as indicating a statistically significant presence or absence of the given emotion, respectively.

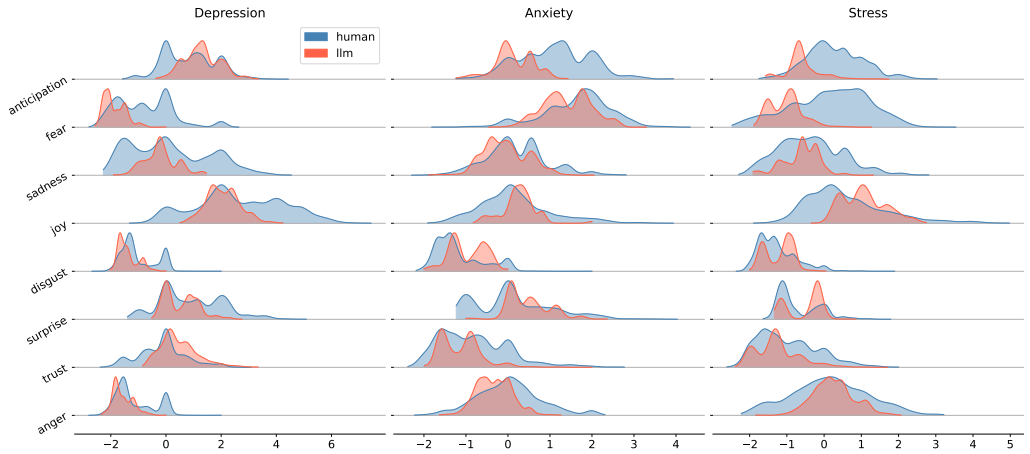
Emotional profiles were computed for each persona separately, under two aggregation schemes: first, by concatenating the explanation texts within each DASS subscale (depression, anxiety, and stress) to capture factor-level emotional profiles; and second, by pooling all DASS-21 item explanations together to obtain a global emotional characterisation. The resulting emotional profiles were then aggregated and compared across the two modes of generation, namely human shadows and LLM assistants.

Emotional load comparison between LLM-assistants and human shadows For our validation, we present the emotional analysis of explanation texts for two representative LLMs: DSK-R1-32B and Qwen3.5-9B — the models with, respectively, the largest and smallest average explanation lengths among all LLMs considered (Figure 12). Despite sharing the Qwen base architecture, these models differ substantially in scale (32B versus 9B parameters), developer organisation (DeepSeek [53] versus Qwen [56]), and release date (January 2025 versus March 2026), making them a well-contrasted pair for this analysis.

Figure 13 reveals that, for both models, human shadows produce explanation texts with a significantly higher presence of anticipation — particularly for anxiety and stress items — a pattern consistent with the emotional framing one would expect from individuals projecting themselves into distress-related scenarios. In DSK-R1-32B, LLM assistants exhibit a markedly higher presence of joy across all DASS factors, along with a predominant elevation of trust, suggesting a positivity bias characteristic of an assistant role. These differences are considerably attenuated in Qwen3.5-9B, whose human-shadow and LLM-assistant profiles are more closely aligned, indicating that this model is less sensitive to role assignment in shaping the emotional tone of its outputs: a noteworthy behavioural divergence despite the shared architectural lineage.



(a) DeepSeek-R1-Distill-Qwen-32B (DSK-R1-32B)



(b) Qwen3.5-9B-Thinking (MoE) (Qwen3.5-9B)

Figure 13: Distribution of emotion z-scores in DASS-21 item explanation texts, grouped by DASS factor.

The emotional analysis of all DASS-21 explanations pooled across factors (reported in Appendix B) further

corroborates these findings. Across models, LLM assistants consistently display higher z-scores than human shadows, with the most pronounced divergences observed for anticipation, joy, and trust. The recurrent elevation of trust across all LLMs is particularly telling: it mirrors the factor-level pattern observed in DSK-R1-32B and is consistent with an assistant-oriented positivity bias that systematically distinguishes LLM-generated explanations from those of human shadows. The robustness of this pattern across models with different architectures, scales, and origins constitutes a meaningful validation of the task: the emotional profiles of human shadows and LLM assistants are systematically and distinct, in ways that align with the expected psychological and role-based differences between the two modes of generation.

A notable exception is Granite-4H-T, which consistently with its performance on prior tasks exhibits a near-complete overlap between the emotional profiles of human shadows and LLM assistants. This confirms its inability to modulate responses according to the assigned role, and further validates the discriminative power of the emotional analysis as a diagnostic tool.

5 Usage Notes

The Mental Health Digital Shadows dataset is distributed as a comprehensive dataset including several data types: textual explanations in natural language related to specific mental health topics, numerical DASS-21 item scores, item-level explanations in texts, and emotional recall words. As outlined in the Background & Summary, MHDS was designed to address two persistent gaps in the evaluation of LLMs for mental-health discourse: The lack of benchmarks that capture psychometric and linguistic outputs, and the lack of resources that map how LLM responses vary under socio-cognitively meaningful prompting conditions. The validation experiments reported in this validation paper represent only a subset of the analyses enabled by the MHDS dataset. Building on the topic-level analyses (Section Topics on mental health), users can move beyond median splits and explore the effects of joint population stratification (e.g., high depression and low extraversion) to probe how specific personas shape the semantic centrality of concepts. The outputs of the Emotional Recall Task [33] can be analysed using spreading-activation paradigms (as in [50]) or embedding-based similarity, enabling comparison with human ERT data. The DASS-21 numerical scores can be exploited to carry out Exploratory and Confirmatory Factor Analysis, extending our OLS validation toward proper psychometric validation on a standard mental health questionnaire like the DASS-21. Finally, the DASS-21 item-level explanations lend themselves to topic modelling, stance detection, or LIWC-style profiling, beyond the Plutchik-based emotion z-scores [58] reported here. Researchers wishing to replicate or extend the psychometric and linguistic analyses reported here can do so using the EmoAtlas Python package [58] for emotion detection and textual forma mentis network construction, which relies on spaCy [60] and the EmoLex lexicon [59]. Researchers interested in delving deeper into the network science approach to texts might find the works of Siew and colleagues [64] and Haim and colleagues [38] useful.

Beyond the dataset itself, MHDS is released as a prompting framework: the full prompt templates used for both human-shadow and LLM-assistant modes (Figures 2 and 3) are made publicly available (including the full code for the data generation `https://github.com/MassimoStel/MHDS`), so that other research groups can reuse and modify them to test new hypotheses. The personas can be translated into other languages to examine whether the persona-driven non-zero DASS distributions observed here generalise across linguistic and cultural contexts. Similarly, the persona layer can be enriched with alternative personifications, e.g., (clinically diagnosed individuals, therapists, caregivers, peer-support volunteers), and the mental-health topics in Table 1 can be substituted with other domains such as trauma, grief, addiction, or eating disorders. By releasing prompts together with data, we position MHDS not as a static benchmark but as an extensible experimental framework for the study of LLM behaviour in mental-health contexts.

Several properties of MHDS should be kept in mind when employing the dataset. DASS-21 scores generated under LLM-assistant mode collapse toward zero for most models (Table 10), so analyses requiring non-degenerate distress distributions should be restricted to the human-shadow subset. Response length varies substantially across models (Figure 7 and Figure 12), and network-based measures such as node degree are sensitive to text length; users comparing TFMN properties across models should control for concatenated text length or restrict comparisons to models with comparable verbosity.

The human reference distributions considered in the manuscript refer to the OpenPsychometrics sample (`https://openpsychometrics.org/_rawdata/`) with subscale totals halved for DASS-21 comparison; this is an approximate alignment, not item-level equivalence. Finally, MHDS records are prompt-conditioned simulated outputs and must not be interpreted as clinical data or as representative of any human population.

5.1 Data Pooling System

Along with the MHDS dataset, we develop and share on a GitHub repository (`https://github.com/MassimoStel/MHDS-pooling-system`) a working dashboard featuring a graphical interface for intuitively exploring the records

stored in the dataset. This data pooling dashboard leverages the MHDS dataset stored as a `Pickle` file and enables persona-level filtering alongside structured inspection of topic responses, Emotional Recall Task outputs, and DASS-21 scores and explanations. The dashboard operates in three sequential steps. First, the user is required to select the chosen LLM and can apply demographic and psychological filters to narrow the persona pool. Upon clicking *Search Personas*, the system queries the dataset and returns all matching records, displaying their full sociodemographic profiles. The resulting selection can then be exported in two formats via a dedicated *Download* button, supporting both `XLSX` and `Pickle` files. Second, the user selects a single persona by clicking directly on the desired row in the *Matching Personas* table, which automatically populates the *Selecting Personas* panel with that record's complete attribute set. Third, the *Answers* panel renders the selected persona's full output across three tabbed layers — Topics, ERT, and DASS-21 — enabling direct inspection of the six open-ended mental health topic replies, the ten emotional recall words, and the item-level DASS-21 scores and explanations generated under the assigned persona conditions.

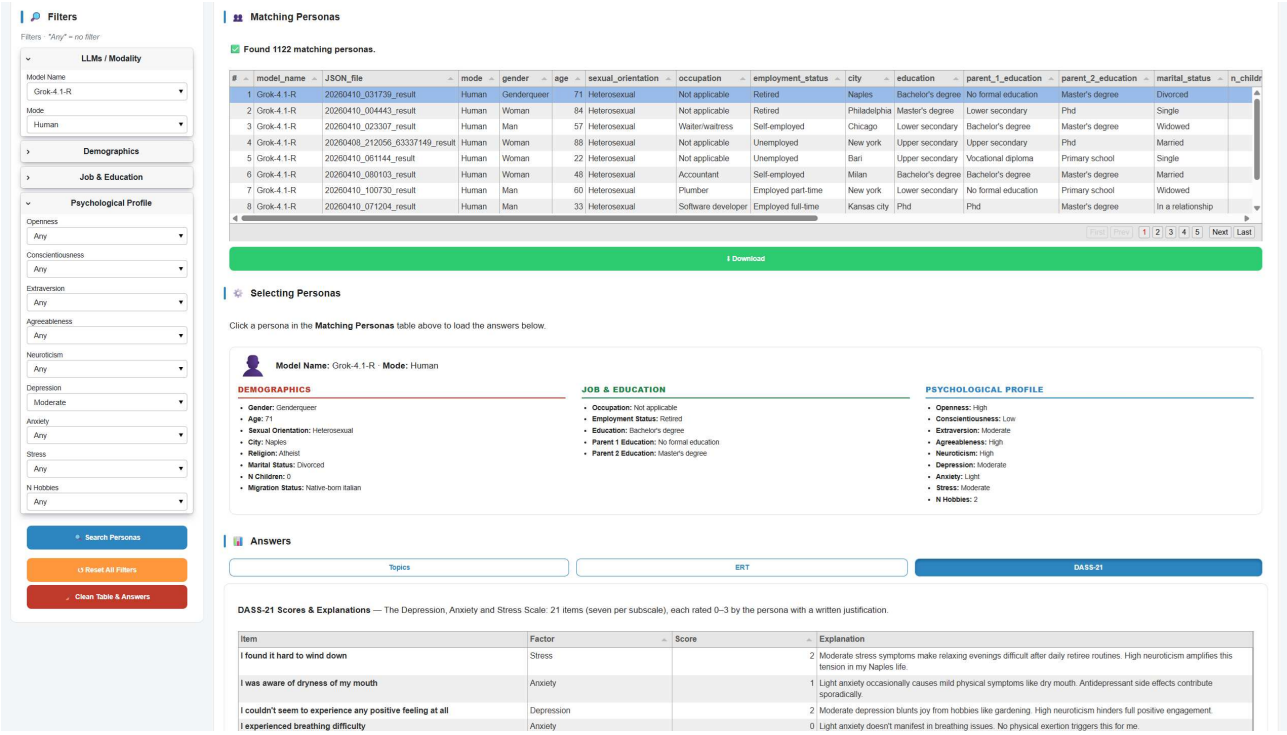


Figure 14: The left sidebar exposes filter controls for LLM/modality, sociodemographics, and psychological traits. After applying filters (here Depression set to Moderate, model Grok-4.1-R in Human mode), Search Personas returns 1,122 matching personas in a paginated table. Selecting a row loads the full persona profile, along with the corresponding responses to each of the 3 tasks (here the DASS-21 first items with their relative scores and explanations).

6 Data & Code Availability

The MHDS dataset is openly available in the public GitHub repository at https://github.com/MassimoStel/MHDS. The dataset is released under the Creative Commons Zero v1.0 Universal (CC0 1.0) public-domain dedication. The repository contains 15 model-level CSV files (one per language model evaluated, with filenames matching the abbreviations listed in Table 2), a `Codebook.md` file documenting all variable names and encoding schemes, a `README.md` describing the repository layout and code usage, a `data_generation_MHDS.ipynb` notebook containing the code used to generate the raw JSON data, and the full text of the licence.

7 Acknowledgements

This work was part of the PENSO project, supported by the Ministero Dell'Università e della Ricerca (MUR) according to Decreto N. 23178 of 10 dicembre 2024 - Bando FIS 2. The authors acknowledge support from CALCOLO, funded by Fondazione VRT, for support with the computational infrastructure simulating LLMs.

8 Author Contributions

Conceptualisation: MS; Study Design: MS; Methodology: ESD, EF, RR; Software: ESD, EF, RR; Validation: ESD, EF, RR; Formal Analysis: ESD, EF, RR; Investigation: ESD, EF, RR; Resources: ESD, EF, RR; Data Curation: EF, RR; Writing - Original Draft: all authors; Writing - Review & Editing: all authors; Visualization: ESD, EF, RR; Supervision: MS; Project administration: EF, MS; Funding Acquisition: MS.

9 Competing Interest

The authors declare no competing interests.

References

[1] Tong, F., Lederman, R., D'Alfonso, S., Berry, K. & Bucci, S. Development of a digital therapeutic alliance scale (mm-dta) in the context of fully automated mental health apps. *Behaviour & Information Technology* **44**, 4286–4300 (2025).

[2] D'Alfonso, S. Ai in mental health. *Current opinion in psychology* **36**, 112–117 (2020).

[3] Hua, Y. *et al.* Large language models in mental health care: a scoping review. *Current Treatment Options in Psychiatry* **12**, 27 (2025).

[4] Pasquini, G., Stocking, G., Kikuchi, E., Pula, I. & Yam, E. Where do americans get health information, and what do they trust? Research Report, Pew Research Center (2026). URL https://www.pewresearch.org/wp-content/uploads/sites/20/2026/04/PS_2026.4.7_health-information_REPORT.pdf. Accessed: 2026-04-22.

[5] RAND Corporation. One in eight adolescents and young adults use ai chatbots for mental health advice. https://www.rand.org/news/press/2025/11/one-in-eight-adolescents-and-young-adults-use-ai-chatbots.html (2025). RAND news release, accessed 2026-04-22.

[6] AI Security Institute. Frontier ai trends report. Tech. Rep., AI Security Institute, Department for Science, Innovation and Technology (2025). URL https://www.aisi.gov.uk/frontier-ai-trends-report. Accessed: 2026-04-22.

[7] Siepmann, R. *et al.* The impact of access to clinical guidelines on llm-based treatment recommendations for chronic hepatitis b. *Liver International* **45** (2025). URL http://dx.doi.org/10.1111/liv.70324.

[8] Weidinger, L. *et al.* Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, 214–229 (2022).

[9] Ashery, A. F., Aiello, L. M. & Baronchelli, A. Emergent social conventions and collective bias in llm populations. *Science Advances* **11**, eadu9368 (2025).

[10] Ferrara, E. The generative ai paradox: Genai and the erosion of trust, the corrosion of information verification, and the demise of truth. *Future Internet* **18**, 73 (2026).

[11] Carrillo, A., Taietta, E., Ardebili, A. A., Veltri, G. A. & Stella, M. Talk2ai: A longitudinal dataset of human–ai persuasive conversations. *arXiv preprint arXiv:2604.04354* (2026).

[12] Williams-Ceci, S. *et al.* Biased ai writing assistants shift users' attitudes on societal issues. *Science Advances* **12**, eadw5578 (2026).

[13] Shaw, S. D. & Nave, G. Thinking-fast, slow, and artificial: How ai is reshaping human reasoning and the rise of cognitive surrender. *Available at SSRN 6097646* (2026).

[14] Branda, F. & Ciccozzi, M. The comfort of automation: Why cognitive sovereignty matters in ai-driven life sciences. *Artificial Intelligence in the Life Sciences* 100158 (2026).

[15] Binz, M. *et al.* A foundation model to predict and capture human cognition. *Nature* **644**, 1002–1009 (2025).

[16] Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 610–623 (2021).

[17] Serapio-García, G. *et al.* Personality traits in large language models. *Prepritnt* (2023).

[18] Pinzuti, E., Tüscher, O. & Ferreira Castro, A. Comparative performance of large language models in emotional safety classification across sizes and tasks. *Frontiers in Artificial Intelligence* **8**, 1706090 (2025).

[19] Turpin, M., Michael, J., Perez, E. & Bowman, S. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems* **36**, 74952–74965 (2023).

[20] Morini, V. *et al.* Online posting effects: Unveiling the non-linear journeys of users in depression communities on reddit. *Computers in Human Behavior Reports* **17**, 100542 (2025).

[21] Zheng, M., Pei, J., Logeswaran, L., Lee, M. & Jurgens, D. When" a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 15126–15154 (2024).

[22] Hu, T. & Collier, N. Quantifying the persona effect in llm simulations. *arXiv preprint arXiv:2402.10811* (2024).

[23] Casoria, L., Neroni, P., Sabatucci, L., Augello, A. & Caggianese, G. Evaluating llms for synthetic personas generation: A comparative analysis of personality representation and censorship effects. In *Proceedings of the 16th Biannual Conference of the Italian SIGCHI Chapter*, 1–9 (2025).

[24] Lovibond, P. F. & Lovibond, S. H. The structure of negative emotional states: Comparison of the depression anxiety stress scales (dass) with the beck depression and anxiety inventories. *Behaviour research and therapy* **33**, 335–343 (1995).

[25] Bergs, T. *et al.* The concept of digital twin and digital shadow in manufacturing. *Procedia CIRP* **101**, 81–84 (2021). URL http://dx.doi.org/10.1016/j.procir.2021.02.010.

[26] Ardebili, A. A., Longo, A. & Ficarella, A. Digital twins bonds society with cyber-physical energy systems: A literature review. In *2021 IEEE International Conferences on Internet of Things (iThings) and IEEE Green Computing & Communications (GreenCom) and IEEE Cyber, Physical & Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cybermatics)*, 284–289 (IEEE, 2021).

[27] Ardebili, A. A., Longo, A., Ficarella, A. *et al.* Digital twin (dt) in smart energy systems-systematic literature review of dt as a growing solution for energy internet of the things (eiot). In *E3S web of conferences*, vol. 312, 1–18 (2021).

[28] Ardebili, A. A., Longo, A. & Ficarella, A. Navigating the future data-driven automation tools: State-of-the-art and research roadmap for digital twins of energy systems. In *2023 IEEE International Conference on Big Data (BigData)*, 3888–3897 (IEEE, 2023). URL http://dx.doi.org/10.1109/BigData59044.2023.10386762.

[29] Sepasgozar, S. M. Differentiating digital twin from digital shadow: Elucidating a paradigm shift to expedite a smart, sustainable built environment. *Buildings* **11**, 151 (2021).

[30] John, O. The big-five trait taxonomy: History, measurement, and theoretical perspectives. *Published as* (1999).

[31] Bonati, M. *et al.* Psychological distress among italians during the 2019 coronavirus disease (covid-19) quarantine. *BMC psychiatry* **21**, 20 (2021).

[32] Rossi, R. *et al.* Covid-19 pandemic and lockdown measures impact on mental health among the general population in italy. *Frontiers in psychiatry* **11**, 550552 (2020).

[33] Li, Y., Masitah, A. & Hills, T. T. The emotional recall task: Juxtaposing recall and recognition-based affect scales. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **46**, 1782 (2020).

[34] Lho, S. K. *et al.* Large language models and text embeddings for detecting depression and suicide in patient narratives. *JAMA network open* **8**, e2511922 (2025).

[35] Russell-Lasalandra, L., Golino, H., Garrido, L. E. & Christensen, A. P. The ultimate tutorial for ai-driven scale development in generative psychometrics: Releasing aigenie from its bottle. *arXiv preprint arXiv:2603.28643* (2026).

[36] Liu, Y., Bhandari, S. & Pardos, Z. A. Leveraging llm respondents for item evaluation: A psychometric analysis. *British Journal of Educational Technology* **56**, 1028–1052 (2025).

[37] De Duro, E. S., Veltri, G. A., Golino, H. & Stella, M. Measuring and identifying factors of individuals' trust in large language models. *arXiv preprint arXiv:2502.21028* (2025).

[38] Haim, E. & Stella, M. Cognitive networks for knowledge modeling: A gentle introduction for data-and cognitive scientists. *Wiley Interdisciplinary Reviews: Cognitive Science* **17**, e70026 (2026).

[39] Rossetti, G. *et al.* Y social: an llm-powered social media digital twin. *arXiv preprint arXiv:2408.00818* (2024).

[40] De Duro, E. S., Franchino, E., Improta, R., Veltri, G. A. & Stella, M. Cognitive networks identify ai biases on societal issues in large language models. *EPJ Data Science* **15** (2025). URL http://dx.doi.org/10.1140/epjds/s13688-025-00600-7.

[41] Abramski, K., Citraro, S., Lombardi, L., Rossetti, G. & Stella, M. Cognitive network science reveals bias in gpt-3, gpt-3.5 turbo, and gpt-4 mirroring math anxiety in high-school students. *Big Data and Cognitive Computing* **7**, 124 (2023).

[42] Hussain, Z., Binz, M., Mata, R. & Wulff, D. U. A tutorial on open-source large language models for behavioral science. *Behavior Research Methods* **56**, 8214–8237 (2024).

[43] Raz, T., Luchini, S. A., Beaty, R. E. & Kenett, Y. N. Automated scoring of question complexity with transformer language models. *Thinking Skills and Creativity* 102090 (2025).

[44] Zayts-Spence, O., Edmonds, D. & Fortune, Z. Mental health, discourse and stigma. *BMC psychology* **11**, 180 (2023).

[45] Corrigan, P. W., Druss, B. G. & Perlick, D. A. The impact of mental illness stigma on seeking and participating in mental health care. *Psychological science in the public interest* **15**, 37–70 (2014).

[46] Clement, S. *et al.* What is the impact of mental health-related stigma on help-seeking? a systematic review of quantitative and qualitative studies. *Psychological medicine* **45**, 11–27 (2015).

[47] Morgan, A. J. *et al.* Receiving support for mental health problems from family and friends: Measurement and impact on mental health, relationships, and help-seeking. *PLOS Mental Health* **2**, e0000502 (2025).

[48] Swierkosz-Lenart, K. *et al.* Therapies for obsessive-compulsive disorder: Current state of the art and perspectives for approaching treatment-resistant patients. *Frontiers in psychiatry* **14**, 1065812 (2023).

[49] Sun, N. Y., Pittenger, C. & Ching, T. Analyzing the contents of a large, public online peer support forum for obsessive-compulsive disorder: Thematic analysis. *JMIR Formative Research* **9**, e60899 (2025).

[50] Carini, A., Taietta, E. & Stella, M. Exploring the link between the emotional recall task and mental health in humans and llms. *Information* **16**, 1057 (2025).

[51] Henry, J. D. & Crawford, J. R. The short-form version of the depression anxiety stress scales (dass-21): Construct validity and normative data in a large non-clinical sample. *British journal of clinical psychology* **44**, 227–239 (2005).

[52] De Duro, E. S., Taietta, E., Improta, R. & Stella, M. Phdgpt: Introducing a psychometric and linguistic dataset about how large language models perceive graduate students and professors in psychology. *arXiv preprint arXiv:2411.10473* (2024).

[53] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning (2025). URL https://arxiv.org/abs/2501.12948. 2501.12948.

[54] OpenAI. gpt-oss-120b & gpt-oss-20b model card (2025). URL https://arxiv.org/abs/2508.10925. 2508.10925.

[55] Qwen, T. Qwen3 technical report (2025). URL https://arxiv.org/abs/2505.09388. 2505.09388.

[56] Qwen Team. Qwen3.5: Towards native multimodal agents (2026). URL https://qwen.ai/blog?id=qwen3.5.

[57] Stella, M. Text-mining forma mentis networks reconstruct public perception of the stem gender gap in social media. *PeerJ Computer Science* **6**, e295 (2020).

[58] Semeraro, A. *et al.* Emoatlas: An emotional network analyzer of texts that merges psychological lexicons, artificial intelligence, and network science. *Behavior Research Methods* **57** (2025). URL http://dx.doi.org/10.3758/s13428-024-02553-7.

[59] Mohammad, S. M. & Turney, P. D. Crowdsourcing a word–emotion association lexicon. *Computational intelligence* **29**, 436–465 (2013).

[60] Honnibal, M., Montani, I., Van Landeghem, S. & Boyd, A. spaCy: Industrial-strength Natural Language Processing in Python (2020).

[61] Fillmore, C. J. & Baker, C. F. Frame semantics for text understanding. In *Proceedings of WordNet and other lexical resources workshop, NAACL*, vol. 6, 59–64 (2001).

[62] Binz, M. & Schulz, E. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences* **120**, e2218523120 (2023).

[63] Plutchik, R. *Emotions and life: Perspectives from psychology, biology, and evolution.* (American Psychological Association, 2003).

[64] Siew, C. S., Wulff, D. U., Beckage, N. M. & Kenett, Y. N. Cognitive network science: A review of research on cognition through the lens of network representations, processes, and dynamics. *Complexity* **2019**, 2108423 (2019).

Appendices

In this Section, the reader will find two appendices:

1. Appendix A: with the supplementary tables for the Emotional Recall Task results, described in the Emotional Recall Task Section.

2. Appendix B: including all the LLMs' joyplots representing the distribution of emotions' z-scores in DASS items responses' explanations, not grouped by factors. We refer to them in Section 4.4.

Appendix A

Table A.1: Top 10 most prevalent words per model for the Emotional Recall Task, stratified by text source (LLM-assistant versus Human-shadow). Values in parentheses indicate the prevalence (p) of the word within the respective corpus. Words are ranked by descending prevalence.

	LLM-assistant		Human-shadow	
Model	Word	p	Word	p
ANITA-U	Curiosity	0.793	Anxious	0.898
	Excitement	0.551	Hopeful	0.802
	Confusion	0.501	Grateful	0.799
	Analytical	0.371	Content	0.745
	Frustration	0.354	Determined	0.692
	Optimism	0.328	Calm	0.668
	Efficient	0.324	Frustrated	0.562
	Adaptable	0.308	Lonely	0.511
	Responsive	0.250	Tired	0.473
	Inquisitive	0.241	Overwhelmed	0.457
DSK-R1-32B	Responsive	0.950	Anxious	0.717
	Analytical	0.946	Content	0.653
	Systematic	0.944	Hopeful	0.603
	Neutral	0.836	Stressed	0.523
	Structured	0.721	Peaceful	0.499
	Processing	0.626	Calm	0.467
	Patterned	0.515	Overwhelmed	0.459
	Algorithmic	0.451	Lonely	0.392
	Computational	0.411	Joyful	0.361
	Adaptive	0.409	Frustrated	0.318
GPT-OSS	Curious	0.906	Anxious	0.948
	Analytical	0.790	Hopeful	0.947
	Reflective	0.675	Restless	0.944
	Responsive	0.663	Frustrated	0.722
	Adaptive	0.630	Content	0.708
	Neutral	0.541	Calm	0.627
	Engaged	0.531	Determined	0.581
	Efficient	0.454	Overwhelmed	0.493
	Objective	0.452	Nostalgic	0.467
	Focused	0.450	Reflective	0.464
Grok-4.1-R	Curious	0.777	Anxious	0.690
	Neutral	0.759	Tired	0.613
	Helpful	0.684	Content	0.586
	Engaged	0.617	Calm	0.522
	Analytical	0.587	Hopeful	0.491
	Satisfied	0.503	Sad	0.454
	Creative	0.366	Irritable	0.448
	Calm	0.358	Stressed	0.430
	Excited	0.333	Lonely	0.415
	Patient	0.314	Peaceful	0.411
Magistral-S	Focused	0.574	Anxious	0.855
	Curious	0.564	Hopeful	0.845
	Analytical	0.526	Grateful	0.741
	Neutral	0.465	Content	0.722
	Adaptive	0.402	Lonely	0.653
	Engaged	0.382	Restless	0.606
	Responsive	0.362	Determined	0.504
	Helpful	0.361	Peaceful	0.453
	Efficient	0.309	Overwhelmed	0.411

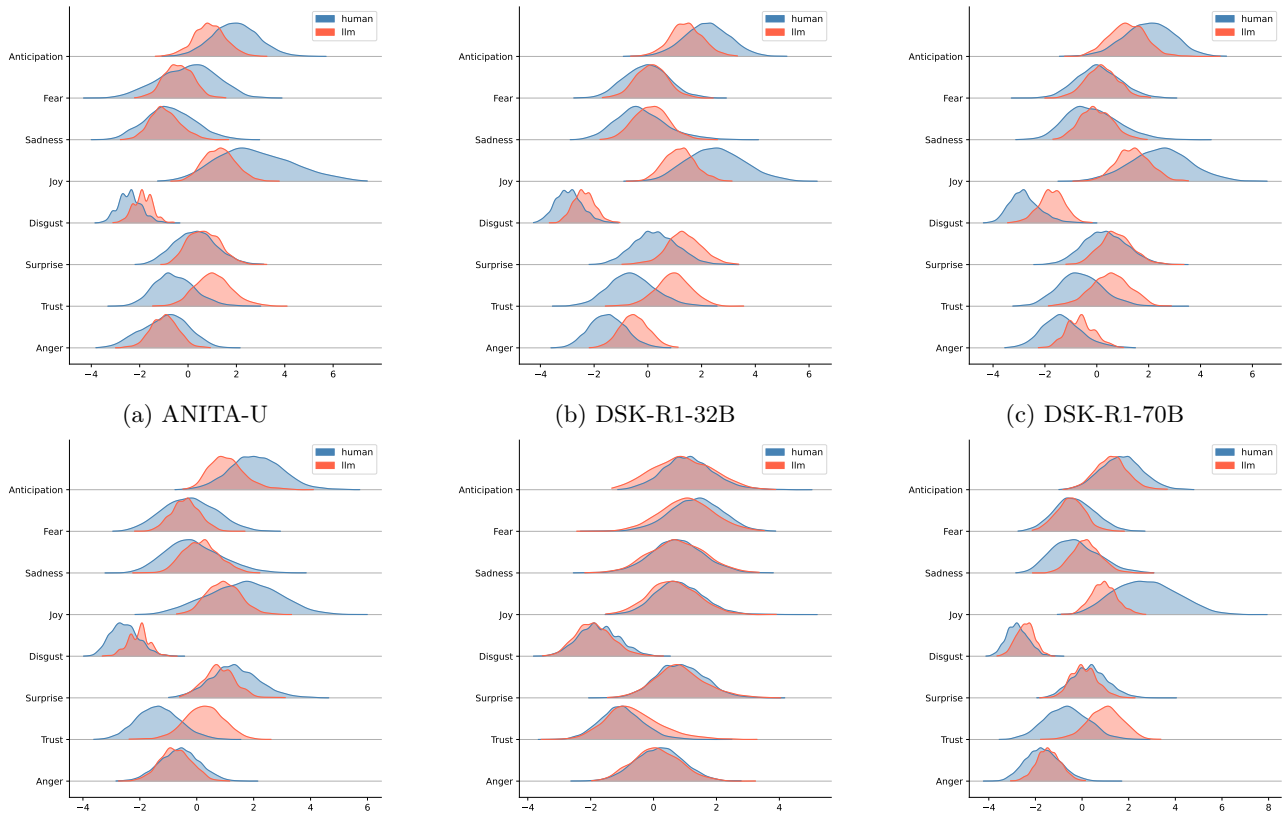
Continued on next page

	LLM-assistant		Human-shadow	
Model	Word	p	Word	p
	Curiosity	0.225	Frustrated	0.400
Ministral-14B-R	Curiosity	0.974	Hopeful	0.706
	Clarity	0.744	Restless	0.627
	Empathy	0.606	Anxious	0.607
	Neutrality	0.528	Overwhelmed	0.558
	Frustration	0.525	Content	0.551
	Adaptability	0.383	Exhausted	0.361
	Patience	0.314	Lonely	0.338
	Insight	0.308	Grateful	0.293
	Adaptation	0.301	Determined	0.289
	Satisfaction	0.290	Curious	0.280
Mistral-S3.2	Curious	0.533	Anxious	0.787
	Analytical	0.384	Hopeful	0.664
	Neutral	0.380	Content	0.637
	Anxious	0.273	Grateful	0.613
	Adaptive	0.264	Overwhelmed	0.472
	Supportive	0.264	Calm	0.404
	Hopeful	0.250	Stressed	0.397
	Calm	0.243	Lonely	0.386
	Frustrated	0.242	Determined	0.355
	Focused	0.240	Frustrated	0.349
Mistral-S4	Calm	0.558	Anxious	0.806
	Sad	0.493	Hopeful	0.647
	Happy	0.492	Content	0.631
	Excited	0.488	Grateful	0.603
	Anxious	0.450	Overwhelmed	0.482
	Hopeful	0.375	Stressed	0.447
	Frustrated	0.326	Calm	0.442
	Curious	0.318	Lonely	0.405
	Confused	0.279	Frustrated	0.345
	Analytical	0.232	Determined	0.333
Phi-4-R+	Curious	0.661	Hopeful	0.835
	Analytical	0.457	Anxious	0.763
	Neutral	0.426	Content	0.758
	Reflective	0.346	Resilient	0.674
	Curiosity	0.295	Reflective	0.653
	Systematic	0.287	Grateful	0.617
	Objective	0.272	Curious	0.539
	Engaged	0.266	Overwhelmed	0.521
	Hopeful	0.231	Frustrated	0.402
	Thoughtful	0.228	Determined	0.397
Qwen3-30B	Calm	0.656	Anxious	0.920
	Focused	0.491	Restless	0.887
	Curious	0.462	Hopeful	0.826
	Analytical	0.414	Frustrated	0.721
	Neutral	0.391	Overwhelmed	0.597
	Curiosity	0.375	Content	0.530
	Structured	0.339	Calm	0.518
	Efficient	0.293	Determined	0.417
	Frustration	0.290	Reflective	0.416
	Focus	0.262	Tired	0.373
Qwen3.5-9B	Curious	0.938	Hopeful	0.911
	Neutral	0.763	Anxious	0.801
	Helpful	0.724	Grateful	0.751
	Calm	0.670	Calm	0.702
	Patient	0.610	Peaceful	0.651
	Supportive	0.522	Tired	0.651
	Focused	0.481	Frustrated	0.636
	Analytical	0.422	Lonely	0.577
	Engaged	0.366	Content	0.573
	Responsive	0.350	Overwhelmed	0.447
Qwen3-4B-IN	Calm	1.000	Calm	0.763
	Focused	0.921	Tired	0.726
	Hopeful	0.831	Quiet	0.699
	Anxious	0.794	Hopeful	0.657
	Tired	0.789	Peaceful	0.636
	Reflective	0.722	Lonely	0.491
	Alert	0.507	Anxious	0.478

Continued on next page

	LLM-assistant		Human-shadow	
Model	Word	p	Word	p
	Overwhelmed	0.495	Content	0.430
	Uncertain	0.471	Restless	0.407
	Peaceful	0.450	Reflective	0.389
Qwen3-4B-TH	Calm	0.703	Calm	0.822
	Focused	0.618	Hopeful	0.786
	Helpful	0.475	Anxious	0.594
	Neutral	0.450	Tired	0.581
	Precise	0.441	Peaceful	0.564
	Logical	0.392	Content	0.411
	Supportive	0.344	Overwhelmed	0.368
	Efficient	0.310	Focused	0.321
	Patient	0.289	Restless	0.316
	Analytical	0.266	Lonely	0.312

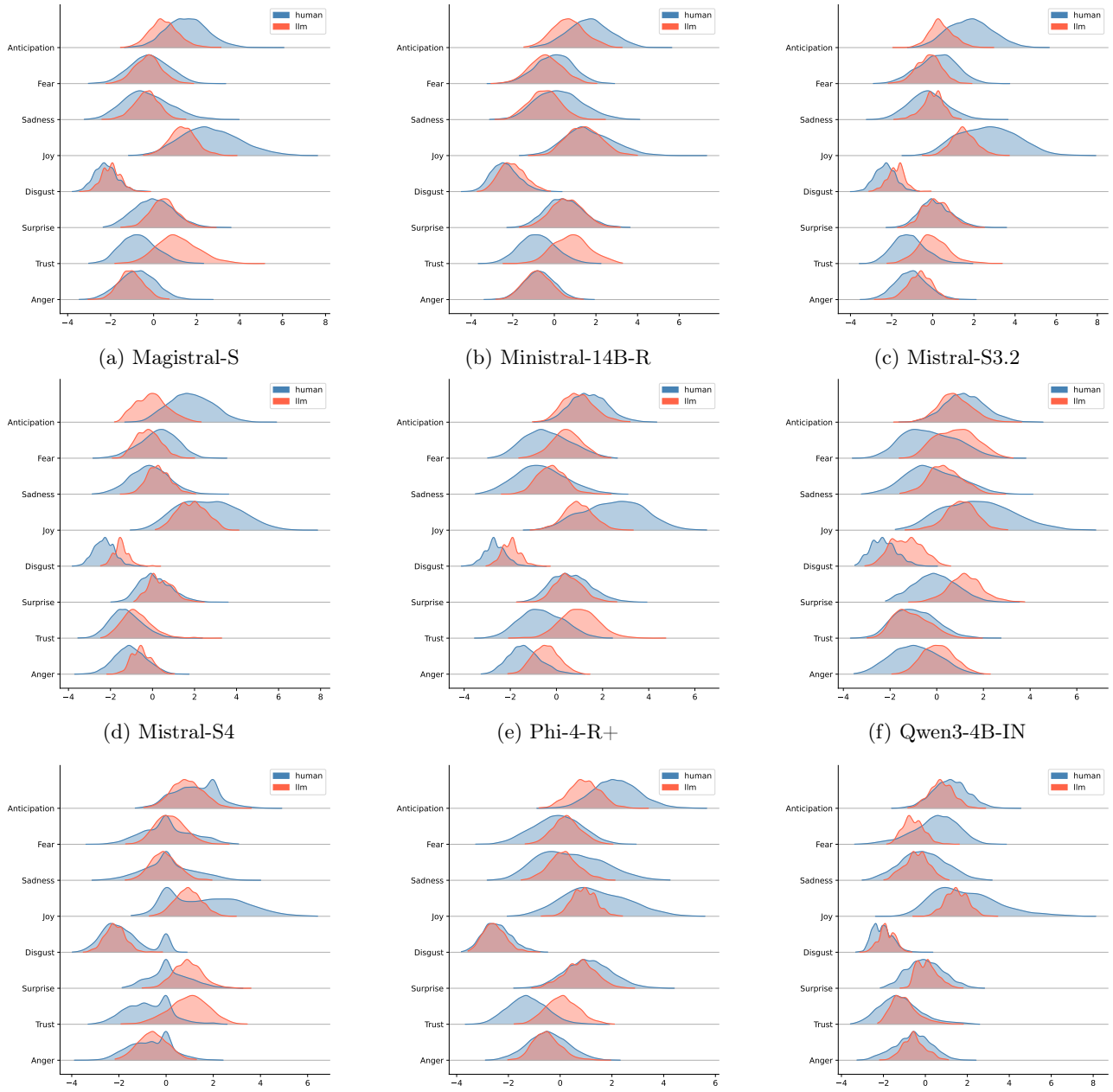
Appendix B



(a) ANITA-U (b) DSK-R1-32B (c) DSK-R1-70B

(d) GPT-OSS (e) Granite-4H-T (f) Grok-4.1-R

Figure B.1: Distribution of emotions' z-scores in DASS items responses' explanations, not grouped by factors. Orange corresponds to LLM assistants and blue to human shadows; DSK = Deepseek.



(a) Magistral-S (b) Ministral-14B-R (c) Mistral-S3.2

(d) Mistral-S4 (e) Phi-4-R+ (f) Qwen3-4B-IN

(g) Qwen3-4B-TH (h) Qwen3-30B (i) Qwen3.5-9B

Figure B.2: Distribution of emotions' z-scores in DASS items responses' explanations, not grouped by factors. Orange corresponds to LLM assistants and blue to human shadows.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supportmedatapaper1.pdf