nature mental health

Article

https://doi.org/10.1038/s44220-025-00527-y

The empirical structure of psychopathology is represented in large language models

Received: 12 September 2023

Accepted: 24 September 2025

Published online: 18 November 2025

Check for updates

Joseph Kambeitz[1] ✉, Jason Schiffman[2], Lana Kambeitz-Ilankovic[1,3], Vijay A. Mittal[4], Ulrich Ettinger[5] & Kai Vogeley[1,6]

Clinical assessment and scientific research in psychiatry are largely based on questionnaires that are used to assess psychopathology. The development of large language models (LLMs) offers a new perspective for analysis of the language and terminology on which these questionnaires are based. We used state-of-the-art LLMs to derive numerical representations ('text embeddings') of the semantic and sentiment content of items from established questionnaires for the assessment of psychopathology. We compared the pairwise associations between empirical data from cross-sectional studies and text embeddings to test whether the empirical structure of psychopathology can be reconstructed by LLMs. Across four large-scale datasets ($n = 1{,}555$, $n = 1{,}099$, $n = 11{,}807$ and $n = 39{,}755$), we found a range of significant correlations between empirical item-pair associations and associations derived from text embeddings ($r = 0.18$ to $r = 0.57$, all $P < 0.05$). Random forest regression models based on semantic or sentiment embeddings predicted empirical item-pair associations with moderate to high accuracy ($r = 0.33$ to $r = 0.81$, all $P < 0.05$). Similarly, empirical clustering of items and grouping to established subdomain scores could be partly reconstructed by text embeddings. Our results demonstrate that LLMs are able to represent substantial components of the empirical structure of psychopathology. Consequently, the integration of LLMs into mental health research has the potential to unlock numerous promising avenues. These may encompass improving the process of developing questionnaires, optimizing generalizability and reducing the redundancy of existing questionnaires or facilitating the development of new conceptualizations of mental disorders.

The study of psychopathology holds great importance in psychiatric research because it serves as the foundation for establishing diagnoses, identifying therapeutic targets and assessing mental health outcomes. Although nonverbal and paraverbal behavior is also informative, language serves as the primary medium for conveying psychopathological descriptions of both inner experiences as expressed by the patients and observable behavior as described by clinicians. As a consequence, empirical studies typically assess psychopathology with the help of questionnaires that try to detect and quantify psychopathological phenomena related to extraordinary experiences such as perception, thought, emotion and behavior, including language and social interaction.

This primacy of language is common to all approaches to psychological assessments and implies that questionnaires include verbal descriptions of symptoms (for example, 'I felt sad and depressed'). Analysis of the possible co-occurrence of symptoms and the details of this low-dimensional structure in data acquired with the help of questionnaires is at the core of psychiatric research[1]. As an example, factor-analytic approaches suggest latent traits that give rise to positive correlations between symptoms. Similarly, unsupervised

A full list of affiliations appears at the end of the paper. ✉e-mail: joseph.kambeitz@uk-koeln.de

machine-learning techniques such as clustering can identify groups of patients[1] or symptoms that frequently co-occur[2]. This low-dimensional structure is the empirical basis for psychiatric syndromes, diagnoses or more complex taxonomies of psychopathology[3]. An alternative approach focuses on the mutual interactions between symptoms and conceptualizes psychopathology in network models[4]. In empirical research, some symptoms co-occur, whereas others do not, which gives rise to a specific low-dimensional structure.

Although the medium of language plays such a central role, surprisingly little is known about how empirical associations in psychopathology are affected by the way in which clinical questionnaires are constructed. As an example, the specific wording of questionnaire items is critical for the accuracy of an assessment and all subsequent interpretations. There is some evidence indicating low content overlap between questionnaires for the assessment of depression[5,6], bipolar disorder[7] or psychosis risk[8]. This is problematic because inconsistent research findings might result from differences between patients in the same diagnostic group or—alternatively—from differences between questionnaires. Moreover, it has been argued that two questionnaire items might correlate simply because of their semantic similarity (item wordings are formulated similarly)[9,10].

Notably, the relationship between language and psychological constructs has been debated for more than 150 years, since Galton proposed his so-called lexical hypothesis[11]. Galton's hypothesis holds that the efficient description of thought and behavior is highly relevant for successful communication and thus, specific descriptions of personality traits will become part of a group's vocabulary. As an example, consider a person who 'is pacing around the room with an intense and restless energy' and one who 'is fidgeting nervously and is unable to sit still'. A clinical psychiatrist might state that these behaviors are similar because they frequently co-occur. Thus, both people might be described as 'agitated'. However, this abstraction can be influenced by implicit assumptions of the psychiatrist (for example, the psychiatrist's belief that these behaviors co-occur). Thus, it has been argued that verbal descriptions of psychological constructs are not entirely neutral and objective descriptions, but are co-constructed by the judgment of assessing clinicians based on their knowledge and experiential background[12,13]. This motivates the identification of methodological approaches regarding how these language-mediated descriptions can be adequately studied. A very interesting approach is provided by the new development of so-called LLMs in the field of artificial intelligence.

LLMs, rooted in deep-learning techniques, represent advanced artificial intelligence systems capable of processing and generating text[14]. Typically, these models undergo pretraining using vast corpora comprising billions of words, followed by fine-tuning to improve performance in specific tasks. Most recent LLMs were built on the basis of so-called transformer architectures[15]. This allows the model to weigh the importance of specific words in a sentence to enable contextual understanding and the handling of long-range dependencies[15]. Current popular models such as GPT-3[16], GPT-4[17], Llama[18], BERT[19] or T5[20] have demonstrated impressive performance across a wide range of tasks encompassing language translation, text summarization and question answering, among others. Specifically in medicine, there is a range of promising applications of LLMs. As an example, LLMs allow the use of chat-bots for screening or diagnostic purposes, can provide support for administrative tasks in medicine and support medical education[21]. Moreover, recent LLMs have demonstrated impressive performance in answering questions from medical licensing exams[22]. Lastly, conversational systems based on LLMs have shown higher performance in diagnostic accuracy than medical doctors in the structured assessment of patient actors[23].

Interestingly, recent research indicates that LLMs can also provide information regarding the semantics of text and hence, could represent substantial aspects of our knowledge about the world[24]. Words can be represented by multidimensional vectors ('embeddings') and the semantic similarity of two words can be quantified by the distance between their vectors[24]. Interestingly, similarity judgments derived from word embeddings closely resemble the similarity judgments of human raters[25,26]. Moreover, similarities regarding the specific attributes of objects or animals (for example, size or color)[26] or similarities of emotions[27] can also be derived from LLMs. Finally, LLMs can detect emotional content ('sentiment') and provide not only simple coarse-grained emotions (for example, positive versus negative) but also detailed categories such as anger, fear, disgust or surprise[28]. In this line of research, previous results demonstrate how LLMs allow for the data-driven investigation of psychological constructs[29,30] by characterizing the semantics of various questionnaires[31]. In a study on personality questionnaires, pairwise similarities of adjectives were extracted from LLMs and found to correspond strongly with empirical ratings[31]. Earlier work has investigated the overlap between scales[32] or between constructs[33] using language models. Although initial studies thus exist in the area of personality assessment[34,35] or emotion[27], analysis of the semantic properties of clinical questionnaires using LLMs has not yet been applied to psychopathology.

LLMs alongside empirical data have the potential to identify the core structure of questionnaires in psychopathology and their language-based descriptions; for instance, by efficiently identifying redundant items. Moreover, the results of this process may improve our understanding of language-based descriptions and will potentially enable generalizability across diverse populations, intercultural comparisons and provide insights into the cultural influences on psychopathology.

Against this background, a rigorous and systematic investigation of the language used for the description of psychopathology and the structure of psychological constructs is urgently needed. In this work, we apply LLMs to the analysis of a range of psychopathology-related questionnaires. For this purpose, we made use of recently developed LLMs[16,18,19] to investigate both the structure and content of four clinical questionnaires on the basis of four large-scale datasets. We systematically explored the extent to which the empirical low-dimensional structure of psychopathology was represented in these models. To this aim, we extracted pairwise similarities of established questionnaire items from language models and predicted associations in empirical data.

By analyzing the language used to describe the symptoms and syndromes therein, this data-driven approach offers an innovative perspective on psychopathology. Using LLMs alongside empirical data has the potential to focus the process of questionnaire generation by efficiently identifying redundant items. Moreover, this integration may improve the generalizability across diverse populations, enable intercultural comparisons and provide insights into cultural influences on psychopathology.

In the current study, we use a language model to derive semantic and sentiment embeddings from items in four established questionnaires for the assessment of psychopathology. This allows us to systematically explore the extent to which the empirical low-dimensional structure of psychopathology is represented in these models.

Results

Representation of empirical item-pair associations in LLMs

Overall, embeddings-based associations correlated significantly with empirical associations in all questionnaires (Fig. 1), both in the semantic domain and the sentiment domain. The highest correlations were found for semantic embeddings of depression, anxiety and stress symptoms ($r = 0.57$, $P < 0.001$). For depression and anxiety symptoms (Depression–Anxiety–Stress Scale (DASS)) and schizotypal symptoms (Oxford–Liverpool Inventory of Feelings and Experiences (O-LIFE)), semantic embeddings show slightly stronger correspondence to empirical data compared to sentiment embeddings. Similar results were obtained using other language models (Supplementary Fig. 1 and

https://doi.org/10.1038/s44220-025-00527-y

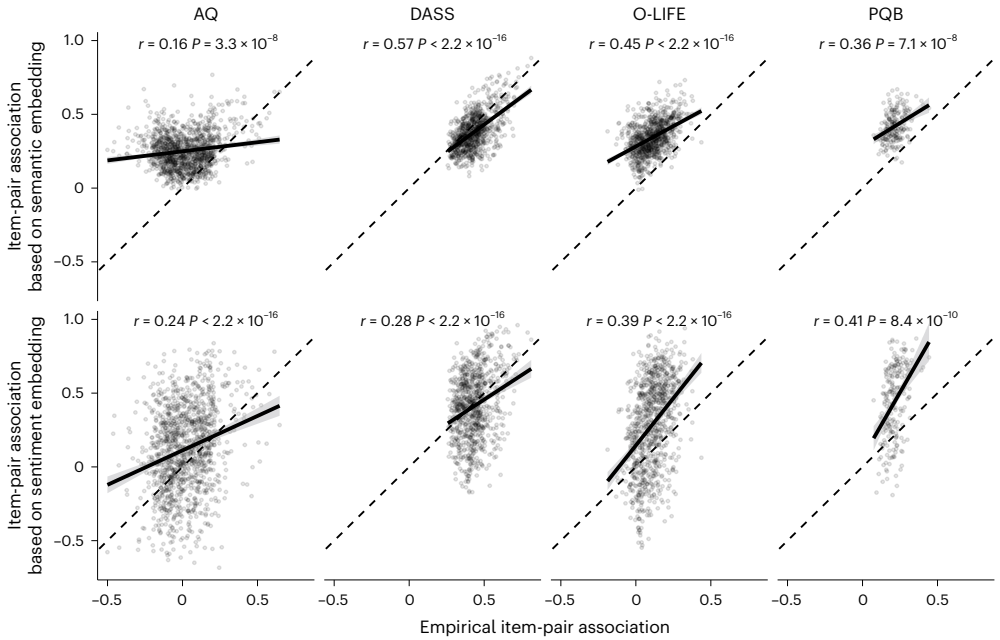

Fig. 1 | Regression analysis of item-pair correlation. Correlation coefficients between item-pair associations derived from empirical data and item-pair associations based on semantic embedding (upper row) or sentiment embedding (lower row). Each point represents a pair of two questionnaire items. There are $n = 1{,}225$ data points for the AQ, $n = 861$ for the DASS, $n = 903$ for the O-LIFE and $n = 210$ for the PQB. Dashed lines depict perfect correlations of $r = 1$, solid lines depict the line of best fit as identified by a linear regression model. Shaded gray areas indicate the 95% confidence intervals (CIs) for the predictions of the linear model. *P* values are reported for linear correlation analysis. No correction for multiple comparisons was made.

Supplementary Table 1). Randomization of semantic embeddings or word order in questionnaire items led to substantially smaller correlations between empirical data and semantic embeddings, indicating that our findings did not occur by chance (Supplementary Figs. 2, 3 and 4 and Supplementary Table 2). To ensure that empirical item-pair associations are estimated robustly in the investigated samples, we conducted multiple iterations of subsampling the data (Supplementary Fig. 5). This indicated that for all four questionnaires, even with 25% of the available data, empirical item-pair associations could be estimated with very high accuracy ($r > 0.99$). Moreover, the position of items in the questionnaire did not affect our findings (Supplementary Fig. 6 and Supplementary Table 3).

In subsequent analysis, we trained random forest regression models for the prediction of empirical item-pair associations based on either semantic or sentiment embeddings. For all investigated questionnaires we obtained high correspondences between predicted and empirical item-pair associations for both semantic and sentiment embeddings (all $P < 0.001$). In this analysis, depression, anxiety and stress symptoms could be predicted with the highest accuracy ($r = 0.79$ for semantic embeddings, $r = 0.75$ for sentiment embeddings). Overall, semantic and sentiment embeddings showed similar prediction accuracy for each questionnaire (Fig. 2).

Representation of item clustering in LLMs

For depression, anxiety and stress symptoms (DASS), autism-related symptoms (Autism Spectrum Quotient (AQ)) and psychotic-like experiences (O-LIFE), the Rand Index indicated strong correspondence between item clustering based on word embeddings and item clustering based on empirical data (Fig. 3a). A clear pattern emerged for these three scales: clustering based on semantic embeddings showed the highest Rand Index, whereas sentiment embeddings and random clustering showed Rand Index values around zero (as expected). This pattern was observed over the range of investigated models comprising between two and ten clusters. For attenuated psychosis symptoms (Prodromal Questionnaire (PQB)), no clear pattern emerged and overall comparably low values for the Rand Index were obtained (Fig. 4).

Replication of established subscales in LLMs

For each of the investigated questionnaires, there exist established subscales (or subdomains) that are indicative of more fine-grained psychopathological concepts. We tested the extent to which this previously established structure could be replicated using the empirical data available in the context of the current analysis, embedding vectors (semantic and sentiment embedding) and compared this to randomized data. *K*-means clustering was conducted for all items. Agreement of this clustering with the established structure was assessed by calculating the clustering error rate. The results indicated strong correspondence of empirical data to established subdomains (low clustering error rate) (Fig. 3b) and clustering based on randomized data showing the lowest correspondence (high clustering error rate). Clustering based on semantic embeddings showed error rates that were comparable to empirical data. Sentiment embeddings showed intermediate error rates for autistic symptoms (AQ), for schizotypal–psychosis-risk symptoms (O-LIFE) and for depression, stress and anxiety symptoms (DASS) (Fig. 3 and Table 1). Semantic embeddings showed lower error rates than sentiment embeddings for all questionnaires except for the PQB (Table 2).

Discussion

In this work, we investigated representation of the structure of psychopathology in LLMs across a range of established clinical questionnaires. Our results indicate that the empirical association between two questionnaire items can be predicted based on sentence embeddings in an LLM (Robustly Optimized BERT Pretraining Approach (ROBERTA)) with low to moderate accuracy using simple regression models ($r = 0.16$ to $r = 0.57$), and with moderate to high accuracy using random forest regression ($r = 0.30$ to $r = 0.79$). Exploratory analyses of different LLMs (Supplementary Fig. 1) indicated even greater performance for larger architectures such as the embeddings models of OpenAI (https://platform.openai.com/docs/guides/embeddings/embedding-models). Moreover, clustering of items based on sentence embeddings showed correspondence to not only clustering based on empirical data but also previously established subdomains of the questionnaires. In general, these results indicate that empirical correlations in clinical

https://doi.org/10.1038/s44220-025-00527-y

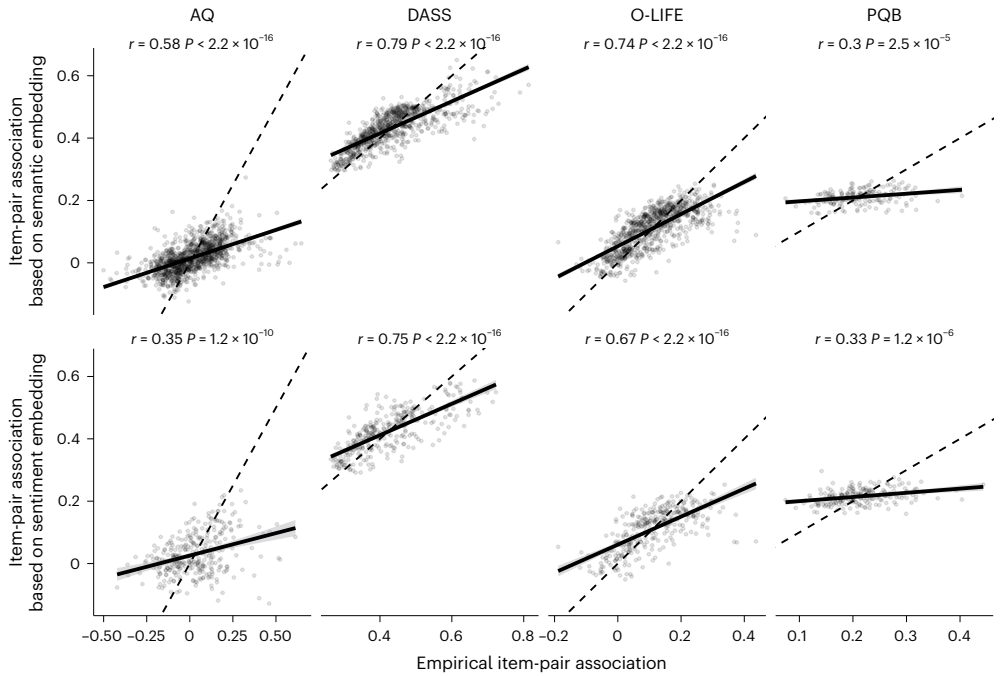

Fig. 2 | Regression analysis of item-pair correlation based on machine learning models. Correlation coefficients between item-pair associations derived from empirical data and item-pair associations as predicted by a random forest regression model based on semantic embedding (upper row) or sentiment embedding (lower row). Each point represents a pair of two questionnaire items. There are $n = 1{,}225$ data points for the AQ, $n = 861$ for the DASS, $n = 903$ for the O-LIFE and $n = 210$ for the PQB. Dashed lines depict perfect correlations of $r = 1$, solid lines depict the line of best fit as identified by a linear regression model. Shaded gray areas indicate the 95% CI for the predictions of the linear model. *P* values are reported for linear correlation analysis. No correction for multiple comparisons was made.

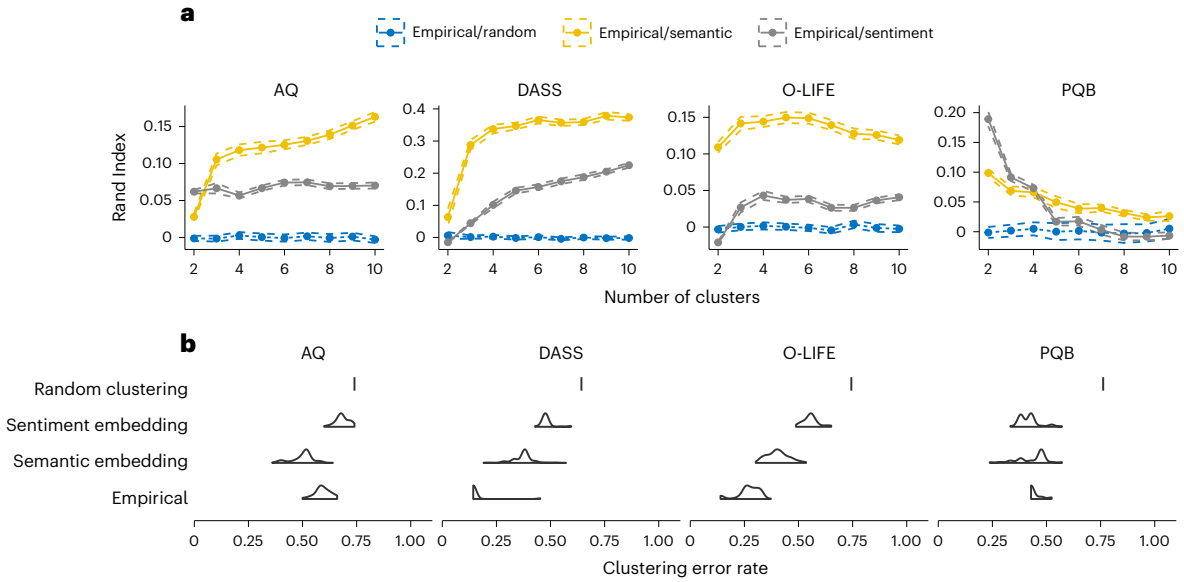

Fig. 3 | Clustering analysis of item-pair associations. a, Clustering consensus as measured by the adjusted Rand Index representing the correspondence of an empirical clustering solution of questionnaire items and a clustering solution based on semantic and sentiment embedding (*k*-means clustering with $n = 200$ repetitions). Dotted lines indicate the 95% CI. **b**, Correspondence of item clustering based on semantic embedding with clustering of items based on established subdomains.

questionnaires can be reconstructed to a large extent solely based on LLMs.

In particular, our findings show some variability when comparing different questionnaires and their related psychopathologies and embedding methods. Correlational analysis indicated that empirical correlations of item-pairs could be best predicted for anxiety, depression and schizotypal symptoms. Interestingly, the correlational analysis results depended on the type of embedding: we found a higher correspondence between correlations based on empirical data and correlations based on embeddings when using semantic embeddings compared to sentiment embeddings. Overall, greater accuracy was obtained when using random forests. This might

https://doi.org/10.1038/s44220-025-00527-y

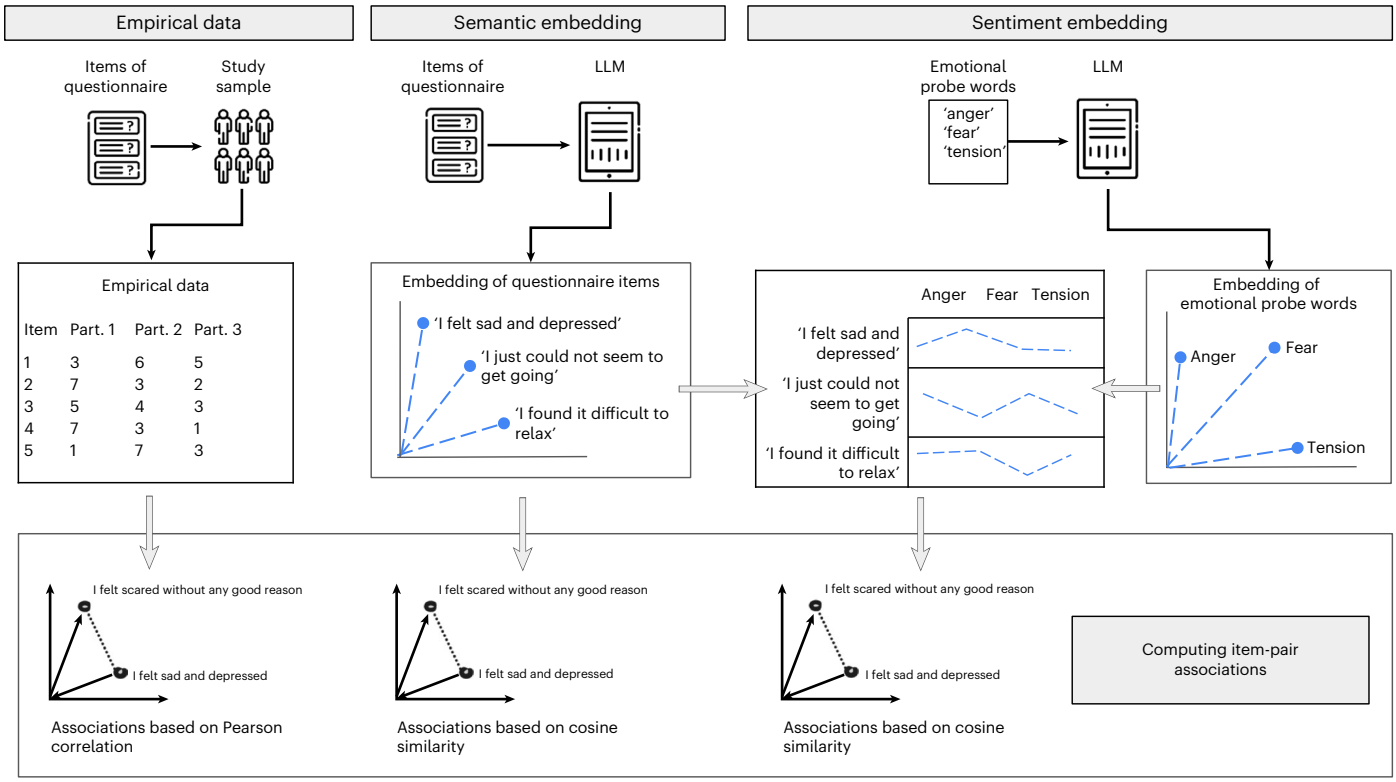

Fig. 4 | Overview of the analytic workflow. For each questionnaire empirical data from study samples were obtained. In parallel, semantic embeddings for each questionnaire item were derived from a state-of-the-art LLM (ROBERTA[14]). A sentiment embedding was obtained by comparing the embedding of each questionnaire item with the embeddings of a set of emotional words. Part., participant.

Table 1 | Clustering based on semantic and sentiment embedding corresponds to empirical clustering significantly better than expected by chance across four questionnaires and across solutions varying between two and ten clusters

		Number of clusters								
Scale	**Comparison**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
DASS	Empirical–semantic	$P<0.001$	$P<0.001$	$P<0.001$	$P<0.001$	$P<0.001$	$P<0.001$	$P<0.001$	$P<0.001$	$P<0.001$
DASS	Empirical–sentiment	$P<0.001$	$P<0.001$	$P<0.001$	$P<0.001$	$P<0.001$	$P<0.001$	$P<0.001$	$P<0.001$	$P<0.001$
O-LIFE	Empirical–semantic	$P<0.001$	$P<0.001$	$P<0.001$	$P<0.001$	$P<0.001$	$P<0.001$	$P<0.001$	$P<0.001$	$P<0.001$
O-LIFE	Empirical–sentiment	$P<0.001$	$P<0.001$	$P<0.001$	$P<0.001$	$P<0.001$	$P<0.001$	$P<0.001$	$P<0.001$	$P<0.001$
AQ	Empirical–semantic	$P<0.001$	$P<0.001$	$P<0.001$	$P<0.001$	$P<0.001$	$P<0.001$	$P<0.001$	$P<0.001$	$P<0.001$
AQ	Empirical–sentiment	$P<0.001$	$P<0.001$	$P<0.001$	$P<0.001$	$P<0.001$	$P<0.001$	$P<0.001$	$P<0.001$	$P<0.001$
PQB	Empirical–semantic	$P<0.001$	$P<0.001$	$P<0.001$	$P<0.001$	$P<0.001$	$P<0.001$	$P<0.001$	$P=0.001$	$P=0.025$
PQB	Empirical–sentiment	$P<0.001$	$P<0.001$	$P<0.001$	$P=0.033$	$P=0.051$	$P=0.539$	$P=0.557$	$P=0.441$	$P=0.184$

indicate that not all dimensions of the embeddings carry distinctive information and that the identification of relevant dimensions by the random forest algorithm might help to improve accuracy. Moreover, the superior performance of random forest regression models might indicate that embeddings relate to the empirical low-level structure in a nonlinear way. Interestingly, random forest analysis indicated that the empirical correlations between autistic symptoms, as measured with the AQ, could be partly reconstructed using semantic embeddings but was much worse using sentiment embeddings. By contrast, the questionnaires DASS, O-LIFE and PQB did not show this discrepancy between semantic and sentiment embedding. Because autism is associated with a high prevalence of alexithymia, it could be hypothesized that patients with autism struggle to identify and describe emotions and therefore the emotional connotation of questionnaire items does not correspond to empirical data as strongly as semantic embedding[36].

Overall, clustering analysis indicated results that were in accordance with the analysis of pairwise associations. Similar patterns were obtained when investigating symptoms related to anxiety and depression, autism or schizotypy. For all four questionnaires, the clustering

Table 2 | Analysis of the published subdomains of the DASS (three subdomains: depression, stress, anxiety), PQB (four subdomains: perceptual abnormalities, grandiose or unusual delusions, persecutory or thought delusions), O-LIFE (four subdomains: unusual experiences, cognitive disorganization, introvertive anhedonia, impulsive nonconformity) and AQ (five subdomains: social skill, attention switching, attention to detail, communication, imagination)

Questionnaire	Contrast	Estimate	Standard error	df	*P*
DASS	Empirical–semantic embedding	−0.2156	0.0043	796	<0.001
DASS	Empirical–sentiment embedding	−0.3243	0.0043	796	<0.001
DASS	Random clustering–semantic embedding	0.2705	0.0043	796	<0.001
DASS	Random clustering–sentiment embedding	0.1618	0.0043	796	<0.001
DASS	Semantic embedding–sentiment embedding	−0.1087	0.0043	796	<0.001
AQ	Empirical–semantic embedding	0.0942	0.0034	796	<0.001
AQ	Empirical–sentiment embedding	−0.0901	0.0034	796	<0.001
AQ	Random clustering–semantic embedding	0.2409	0.0034	796	<0.001
AQ	Random clustering–sentiment embedding	0.0566	0.0034	796	<0.001
AQ	Semantic embedding–sentiment embedding	−0.1843	0.0034	796	<0.001
O-LIFE	Empirical–semantic embedding	−0.1328	0.0039	796	<0.001
O-LIFE	Empirical–sentiment embedding	−0.2844	0.0039	796	<0.001
O-LIFE	Random clustering–semantic embedding	0.3409	0.0039	796	<0.001
O-LIFE	Random clustering–sentiment embedding	0.1893	0.0039	796	<0.001
O-LIFE	Semantic embedding–sentiment embedding	−0.1516	0.0039	796	<0.001
PQB	Empirical–semantic embedding	0.0100	0.0044	796	0.109
PQB	Empirical–sentiment embedding	0.0255	0.0044	796	<0.001
PQB	Random clustering–semantic embedding	0.3286	0.0044	796	<0.001
PQB	Random clustering–sentiment embedding	0.3440	0.0044	796	<0.001
PQB	Semantic embedding–sentiment embedding	0.0155	0.0044	796	0.003

Agreement of clustering based on embeddings and empirical data with published subdomains was quantified by the cluster error rate. Clustering based on semantic and sentiment embeddings corresponds to established subdomains of the questionnaires significantly better than expected by chance but significantly worse than empirical data.

solution based on semantic embeddings corresponded best to the empirical clustering.

It might be speculated that our results point to a fundamental difference between semantics and sentiment. The former is predominantly 'world-related', objective and deterministic, whereas the latter refers to more 'person-related' inner emotional states that are inherently subjective, probabilistic and potentially more fuzzy. This distinction parallels the dichotomy between 'things' and 'persons' in the fields of psychiatry and psychology[37,38] and has gained recent attention because of its association with underlying cerebral mechanisms[39]. The subjective and probabilistic nature of internal emotional states might render sentiment embeddings less well suited to represent the empirical data of psychopathological questionnaires.

Our findings corroborate with previous publications indicating that word embeddings from LLMs can encode the similarity of words[26], general medical knowledge[40], emotion[27] and personality ratings[31]. Here we extend this line of research to the field of altered states or deviances of inner experience covered by psychopathology. Even though the current analysis mainly focused on questionnaires that measure psychopathology, the theory and methodology are applicable to psychological constructs in general, including intelligence, cognitive capacities, emotions or personality traits[41].

In research relying on self-rating questionnaires, there is often little control regarding how participants conduct their ratings. Shallower processing of the questionnaire items during the rating might rely on superficial semantic aspects. By contrast, deeper processing might require more mental effort and include the evaluation of items in the context of one's own biography or past experiences above and beyond the mere semantic information contained in the items. Increasing fatigue due to lengthy questionnaires might facilitate shallower processing. A strong correspondence of empirical data to semantic similarity scores (for example, low manifest validity) might indicate superficial processing[34,42].

Questionnaire design guided by sentence embeddings

There are many potential applications for text embeddings to facilitate research in mental health[31]. Typically, the first step in the construction of a clinical instrument is the generation of questionnaire items that contain text-based descriptions of psychopathological phenomena or mental states experienced by patients (for example, thoughts, emotions, intentions). It is common practice to reduce such sets of questionnaire items based on preliminary data and statistical procedures by means of factor analysis or item-response analysis. Importantly, the initial generation of questionnaire items is often based on the idiosyncratic conceptualization of individual researchers or psychiatrists, but usually not on a hypothesis-generating approach using qualitative data (for example, content analysis of free patient interviews) in a systematic way leading to the abduction of hypotheses. This first step of the conceptualization of a questionnaire can potentially influence the low-level correlational structure that is later identified using the generated questionnaire. With the application of LLMs, this step can be implemented more efficiently and circumvent potential subjective biases. As an example, language-based models allow the investigation of hundreds or thousands of questionnaire items, if necessary. Hence, a large number of candidate items can be screened very efficiently and items with high similarity can be identified, considered redundant and might be excluded. Moreover, sensitivity analysis can be conducted by excluding individual items and assessing the effect on the correlational structure.

Another important aspect is the data collection procedure, which occurs during generation of a clinical questionnaire. Typically, this is a costly process conducted in larger populations to obtain

representative samples. However, this process might fail to include individuals from populations that are 'hard-to-reach' because of their geographical location or socioeconomic situation[43]. This might limit the generalizability of the developed questionnaires to specific subsets of the population[44]. Moreover, for some countries, generation of sufficient validation samples might not be feasible because of the high costs involved. In general, the data on which LLMs are based consist of text that is unstructured and does not take a specific corpus describing psychopathologically relevant experiences into account. However, data are substantially larger (for example, 160 GB, which corresponds approximately to 30 billion words in the case of ROBERTA[14], https://huggingface.co/roberta-base) than the empirical datasets of questionnaires and cover a larger proportion of the population. Thus, combining empirical data and information from LLMs during the process of generating questionnaires might improve the generalizability to all parts of the population. Moreover, the current approach can easily be adopted to investigate pairwise item correlations across questionnaires. As an example, the semantic similarity between existing and new questionnaires can be quantified to decrease redundancy in data gathering[32]. Lastly, the approach presented here offers new options to study intercultural differences in psychopathology. Notably, there is evidence that psychological concepts such as emotions are shaped by cultural context[45,46]. Analysis of psychopathology based on LLMs can provide a different angle on the analysis of such cultural influences.

Methodological considerations

It is important to keep in mind that all presented findings fully depend on the data and their degree of universality that was used to generate the used LLM. Thus, using a different database might affect the results and all potential biases present in the source data might be transferred to the sentence embeddings. As an example, previous studies indicated that the frequency of a word in a training corpus might affect the resulting embedding and measures such as word similarity[47,48]. Moreover, the model architecture will influence the subsequent analysis of sentence embeddings. Interestingly, recent advances allow for the systematic investigation of different corpora and Hugging Face (https://huggingface.com), for example, provides access to a large number of language-based models that can be queried for comparative analysis.

Another important future direction focuses on the procedures used to extract information from language models. As an example, the dimensionality of embeddings affects the performance in downstream language tasks[49]. In some cases, embeddings vectors have been shown to contain a correlational structure, and removing this might improve the expressiveness of the embeddings[50,51]. Lastly, it has recently been suggested that extracting information along dimensional axes of specific properties (for example, large–small, hard–soft) might provide embeddings that are highly relevant for the assessment of real data, but also provide highly interpretable values[26].

Limitations

It should be noted that the cohorts investigated in this study differ with respect to age, sex, access (general, help-seeking population) and language (English, German). Moreover, the investigated questionnaires differ with respect to multiple features such as the response format (binary versus Likert scale) or the way in which items are formulated, which might influence our results. As an example, DASS and AQ use a first-person perspective (for example, 'I felt that I was…'), whereas O-LIFE and PQB are formulated from a second-person view ('Do you think that you could…'). Despite this heterogeneity, our results indicate that LLMs can represent psychopathology across a range of different samples and questionnaires.

However, an important limitation of the current results is that we investigated only a small set of questionnaires in an explorative and hypothesis-generating approach. A large number of further questionnaires are not investigated here, which introduces the possibility of selection bias. Future studies will have to examine whether similar results can be obtained, for example by using other questionnaires for depression. Moreover, it remains to be demonstrated to what extent the current findings generalize to other aspects of psychopathology that were not investigated, such as the externalizing symptoms of inattention, impulsivity or aggression.

Another important aspect of concern for our results is the potential bias present in language models. Recent studies demonstrate that word embeddings are related to the cultural and social contexts of the text corpus on which they are based[52]. Thus, word embeddings closely track cultural and social contexts. Any potential bias present in the training data might subsequently result in biases of word embeddings, for example bias toward minorities. Recent empirical evidence supports the view that language use might differ between various communities[30] and depend on age[53]. In addition, the nuanced terminology developed in psychopathology to describe the inner experience of patients is presumably not a substantial part of the training data and hence at least underrepresented in LLMs. Finally, inner experiences frequently take the form of inner speech, but might also occur in experiences that are more difficult to express verbally such as unusual sensory experiences including hallucinations, feelings, inner images, thoughts including delusional experiences, delusional mood or unsymbolized thinking[54]. As a consequence, these aspects of inner experiences are not adequately represented in LLMs and are therefore difficult to investigate with approaches such as word embeddings.

A further important consideration in the context of our analysis is the aspect of model choice. Although we initially selected ROBERTA as an optimal model because of its bilingual training, accessibility and high performance in benchmark tests, our explorative results using more modern models indicated even greater performance. Thus, future studies should take into account the substantial influence that the choice of the model architecture could exert and should explore a wider range of modern model architectures to optimize performance.

Conclusion

In conclusion, we provide evidence of the representation of the structural aspects of psychopathology in LLMs. These initial results prompt further investigations into the potential of extracting information from such models to foster innovative research in mental health. Potential applications include the generation of questionnaires guided by LLMs, the generation of statistical priors in the investigation of psychological concepts and the assessment of psychopathology in a specific context (for example, hard-to-reach populations) including cross-cultural studies, for example comparison between collectivistic and individualistic cultures[31].

Methods

All research presented here was conducted in accordance with relevant ethical guidelines and regulations. Ethical approvals for the individual study samples included in this work are detailed separately below for each sample.

Investigated questionnaires for psychopathology

We investigated four questionnaires: the DASS[55], the short form of the O-LIFE, the brief version of the PQB[56] and the AQ[57]. These scales cover a wide range of symptoms including depression (DASS), anxiety (DASS), schizotypal (O-LIFE) and attenuated psychosis symptoms (PQB), and the symptoms associated with autism spectrum disorder (AQ). Moreover, the questionnaires range from scales that are broadly used for the assessment of manifest and sometimes severe clinical symptoms (DASS, AQ) to scales that assess mild or moderate symptoms and traits that might also occur in the healthy population or individuals at risk for psychiatric disorders (O-LIFE, PQB). For each questionnaire, we analyzed empirical data from large cohorts (Table 3).

Table 3 | Overview of the empirical samples investigated

	Depression, Anxiety and Stress Scale	Oxford–Liverpool Inventory of Feelings and Experiences (short form)	Autism Quotient	Prodromal Questionnaire–Brief
Abbreviation	DASS	O-LIFE	AQ	PQB
Symptom domains	Depression, anxiety, stress	Psychosis-proneness or schizotypy	Autism-related traits	Psychosis-proneness
Sample	*n*=39,775 Age 23.6 years Gender 30,367 female; 8,789 male; 552 other; 67 not disclosed	*n*=11,807 Age 30.4 years Gender 8,633 female; 3,174 male	*n*=1,555 Age 36.5 years Gender 535 female; 1,019 males; 1 missing	*n*=1,099 Age 18.7 years Gender 373 female; 305 male; 14 missing data
Sample	Population sample	Population sample	Help-seeking sample	Population sample
Number of items	42	43	50	21
Number of words in total	415	548	318	526
Words per questionnaire item (mean)	9.88	10.96	15.14	12.23
Response format	Four-point Likert scale, self-rating	Binary scale, self-rating	Four-point Likert scale, self-rating	Binary scale, self-rating
Number of subdomains	3	4	5	4

The DASS includes 42 self-reported items (for example, 'I found myself getting upset by quite trivial things') from 3 subdomains (depression, stress, anxiety) that are rated on a Likert scale with 4 levels (Did not apply to me at all; Applied to me to some degree, or some of the time; Applied to me to a considerable degree, or a good part of time; and Applied to me very much, or most of the time). We used a publicly available dataset consisting of $n = 39{,}755$ participants (https://openpsychometrics.org/_rawdata/, retrieved 1 June 2022). This represents a population sample without any restriction and no inclusion or exclusion criteria. There were no missing data in this sample.

The short form of the O-LIFE includes 43 self-reported items (for example, 'Does a passing thought ever seem so real it frightens you?') from 4 subdomains (unusual experiences, cognitive disorganization, introvertive anhedonia, impulsive nonconformity) that are rated on a binary scale (yes versus no). We used a publicly available dataset consisting of $n = 11{,}807$ participants[58]. The online questionnaire was advertised on mailing lists and online forums across Germany. There were no exclusion criteria except that only data from participants who were at least 18 years old were analyzed. There were no missing data in this sample.

The PQB includes 21 self-reported items (for example, 'Do familiar surroundings sometimes seem strange, confusing, threatening or unreal to you?') from 4 subdomains (perceptual abnormalities, grandiose or unusual delusions, persecutory or thought delusions) that are used to assess psychosis-risk symptoms[59]. Items are rated on a binary scale (true versus false) and for each item the resulting distress is rated (distress was not analyzed in the current analysis). We analyzed a sample of $n = 1{,}099$ healthy individuals from an undergraduate population[60]. One part of this sample was recruited via the University of Colorado Boulder's human subject recruitment pool and included students and community members from the general population. The second part of this sample consisted of undergraduate students recruited from introductory psychology courses at the University of Maryland, Baltimore County. For both groups there were no exclusion criteria except for a minimum age of 18 years. Twenty-one participants had between one and three missing values and were retained in our analysis. Four participants had only missing values and were excluded from any further analysis.

The AQ is based on 50 self-reported items (5 subdomains: social skill, attention switching, attention to detail, communication, imagination) to measure the expression of autistic traits in individuals (for example, 'When I'm reading a story, I find it difficult to work out the characters' intentions'). The items are rated on a Likert scale with four levels (Definitely agree; Slightly agree; Slightly disagree; and Definitely disagree). In the current analysis, we included scores from $n = 1{,}555$ individuals presenting at the Autism Outpatient Clinic at the Department of Psychiatry and Psychotherapy at the University Hospital in Cologne (Germany). There were no exclusion criteria except for a minimum age of 18 years. There were 260 participants who had between 1 and 10 missing values and were retained in our analysis. Two participants had more than ten missing values and were excluded from any further analysis. Correlations between questionnaire items were estimated using Pearson correlation with pairwise complete observations.

Before participation, written informed consent was obtained from all participants in the case of the DASS, O-LIFE and PQB. In the case of the DASS and O-LIFE, participants provided questionnaire answers online and also opted in the use of their data online. In the case of AQ, we collected data as part of our clinical routine and studied the data retrospectively and anonymously. In this case, data can be used for research without the formal consent of patients according to German regulations. The ethnicity of the investigated participants was not assessed in the four investigated samples. No distinction was made between 'sex' and 'gender' in the reported studies. We refer to the self-description of participants.

Embedding of questionnaire items using LLMs

We used the LLM ROBERTA[14] because of its fine-tuning on English and German data and its high performance on benchmark tests for both languages (https://huggingface.co/T-Systems-onsite/cross-en-de-roberta-sentence-transformer). The accessibility of the ROBERTA model aligns with principles of reproducibility and open science. In general, the ROBERTA model follows a transformer-based architecture, consisting of a stack of transformer layers. It is trained on a large text corpus (approximately 2.5 TB) using the objectives of predicting masked words or the next sentence. This training enables the model to incorporate contextual information and generate text embeddings that capture semantic meaning. One notable advantage of the ROBERTA model used in this study is that it has been fine-tuned through multilingual training, specifically on combined samples in English and German. As a result, the model can generate similar sentence embeddings for both languages[61].

For each item in each questionnaire, the ROBERTA model can provide an embedding vector with the dimension 1×768. The semantic similarity between two questionnaire items was calculated as the cosine distance between their embedding vectors (from now on referred to

as semantic embedding)[25] (Fig. 3). In a parallel analysis, we derived sentiment-based similarity for each pair of questionnaire items. To this aim, we used a set of probe words that captured distinct categories of self-reported emotional experience. As probe words we selected 27 terms that had been identified and empirically justified in a previous study (admiration, adoration, esthetic appreciation, amusement, anger, anxiety, awe, awkwardness, boredom, calmness, confusion, craving, disgust, empathetic pain, entrancement, excitement, fear, horror, interest, joy, nostalgia, relief, romance, sadness, satisfaction, sexual desire and surprise)[62]. For each of these probe terms, we derived a 1×768 dimensional embedding vector using ROBERTA. Then, for each questionnaire item we calculated its similarity to each probe word based on the cosine distance of both embedding vectors. This provided a 'sentiment profile' for each questionnaire item with 27 values representing the relevance of sentiment dimensions (from now on referred to as the sentiment embedding) (Fig. 3).

Item-pair associations and item clustering

In this work, we investigated two complementary aspects of the structure of psychopathology. First, we investigated pairwise associations of questionnaire items in empirical data and tested whether these associations could be predicted by embedding vectors. For a specific questionnaire, we first calculated the empirical correlations between each pair of items. In a parallel procedure, the associations between pairs of questionnaire items were estimated using (semantic and sentiment) embedding vectors and by using the cosine similarity as a distance metric. We tested the degree to which empirical item-pair associations related to associations in their (semantic and sentiment-based) embeddings by linear regression analysis. Subsequently, we used random forest regression models[63] to generate prediction of empirical pairwise item associations based on semantic and sentiment embeddings. Here we tuned the following hyperparameters: the maximum depth of each tree (between 2 and 10), the ratio of candidate variables considered at each split (between 0.5 and 1.0), the number of regression trees in the ensemble (between 500 and 2,000), and the proportion of randomly drawn observations (between 0.5 and 1.0)[64]. Hyperparameter tuning was embedded in a nested cross-validation scheme with 5 folds in the inner and outer loop, and the hyperparameter spaces were explored by random search with 20 iterations at each step with the coefficient of determination as the optimization target. Model performance in the outer loop was aggregated across folds to estimate the accuracy of predictions on unseen data.

In a complementary analysis, we investigated the clustering of questionnaire items and the degree to which this cluster structure could be replicated based on the embedding of questionnaire items. First, we used k-means clustering to assign each questionnaire item to a varying number of clusters ($k = 2$ to $k = 10$) based on empirical data. The same procedure was applied to (semantic and sentiment) embedding vectors to create clustering solutions. Subsequently, the correspondence of the empirical clustering solution and the clustering solution based on embeddings was tested by calculating the adjusted Rand Index[65].

$$\text{Rand Index} = \frac{a+b}{a+b+c+d}$$

$$\text{Adjusted Rand Index} = \frac{\text{Index} - \text{Expected Index}}{\text{Max Index} - \text{Expected Index}}$$

For two cluster solutions X and Y, a is the number of observations that are in the same cluster for X and Y, and b is the number of observations that are in different clusters in both clustering solutions. The variables c and d represent the number of observations that are in the same cluster for one cluster solution but in different clusters for the other cluster solution. The adjusted Rand Index corrects the Rand Index for the probability that two observations show correspondence of their cluster assignments by chance. In brief, this measure quantifies the degree to which two random items show similar cluster membership in two different clustering solutions while controlling for the potential that two items occur in the same cluster by chance. Thus, the adjusted Rand Index is high if two items are assigned to the same cluster in both clustering solutions or to two different clusters in both clustering solutions.

All analyses were performed in R (v.4.3.3) and Python (v.3.11.0).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data of the DASS (https://openpsychometrics.org/_rawdata/) and the O-LIFE (https://osf.io/epfvq/) are open and available. The data of the PQB is available in anonymized form upon reasonable request to the authors of the original publication (J.S., V.A.M.). AQ data are not publicly available, as the participants did not consent to data sharing with other researchers.

Code availability

Analysis code is available via https://github.com/kambeitzlab/LLM_Psychopathology.

References

1. Lefort-Besnard, J. et al. Patterns of schizophrenia symptoms: hidden structure in the PANSS questionnaire. *Transl. Psychiatry* **8**, 237 (2018).
2. Chekroud, A. M. et al. Reevaluating the efficacy and predictability of antidepressant treatments: a symptom clustering approach. *JAMA Psychiatry* **74**, 370–378 (2017).
3. Kotov, R. et al. The Hierarchical Taxonomy of Psychopathology (HiTOP): a dimensional alternative to traditional nosologies. *J. Abnorm. Psychol.* **126**, 454–477 (2017).
4. Borsboom, D. & Cramer, A. O. J. Network analysis: an integrative approach to the structure of psychopathology. *Annu. Rev. Clin. Psychol.* **9**, 91–121 (2013).
5. Fried, E. I. The 52 symptoms of major depression: lack of content overlap among seven common depression scales. *J. Affect. Disord.* **208**, 191–197 (2017).
6. Vilar, A. et al. Content agreement of depressive symptomatology in children and adolescents: a review of eighteen self-report questionnaires. *Eur. Child Adolesc. Psychiatry* **33**, 2019–2033 (2024).
7. Chrobak, A. A., Siwek, M., Dudek, D. & Rybakowski, J. K. Content overlap analysis of 64 (hypo)mania symptoms among seven common rating scales. *Int. J. Methods Psychiatr. Res.* **27**, e1737 (2018).
8. Bernardin, F., Gauld, C., Martin, V. P., Laprévote, V. & Dondé, C. The 68 symptoms of the clinical high risk for psychosis: low similarity among fourteen screening questionnaires. *Psychiatry Res.* **330**, 115592 (2023).
9. Arnulf, J. K., Larsen, K. R., Martinsen, Ø. L. & Bong, C. H. Predicting survey responses: how and why semantics shape survey statistics on organizational behaviour. *PLoS ONE* **9**, e106361 (2014).
10. D'Andrade, R. G. Trait psychology and componential analysis. *Am. Anthropol.* **67**, 215–228 (1965).
11. Galton, F. Measurement of character. *The Fortnightly Review* **36**, 179–185 (1884).
12. Block, J., Weiss, D. S. & Thorne, A. How relevant is a semantic similarity interpretation of personality ratings? *J. Pers. Soc. Psychol.* **37**, 1055–1074 (1979).

13. Shweder, R. A. & D'Andrade, R. G. Accurate reflection or systematic distortion? A reply to Block, Weiss, and Thorne. *J. Pers. Soc. Psychol.* **37**, 1075–1084 (1979).
14. Liu, Y. et al. RoBERTa: a robustly optimized bert pretraining approach. Preprint at http://arxiv.org/abs/1907.11692 (2019).
15. Vaswani, A. et al. Attention is all you need. Preprint at http://arxiv.org/abs/1706.03762 (2017).
16. Brown, T. B. et al. Language models are few-shot learners. Preprint at http://arxiv.org/abs/2005.14165 (2020).
17. OpenAI. GPT-4 Technical report. Preprint at http://arxiv.org/abs/2303.08774 (2023).
18. Touvron, H. et al. LLaMA: open and efficient foundation language models. Preprint at http://arxiv.org/abs/2302.13971 (2023).
19. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. Preprint at http://arxiv.org/abs/1810.04805 (2018).
20. Raffel, C. et al. Exploring the limits of transfer learning with a unified text-to-text transformer. Preprint at http://arxiv.org/abs/1910.10683 (2019).
21. Thirunavukarasu, A. J. et al. Large language models in medicine. *Nat. Med.* **29**, 1930–1940 (2023).
22. Singhal, K. et al. Towards expert-level medical question answering with large language models. *Nat. Med.* **31**, 943–950 (2025).
23. Tu, T. et al. Towards conversational diagnostic AI. *Nature* **642**, 442–450 (2025).
24. Turney, P. D. & Pantel, P. From frequency to meaning: vector space models of semantics. *J. Artif. Intell. Res.* **37**, 141–188 (2010).
25. Pereira, F., Gershman, S., Ritter, S. & Botvinick, M. A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. *Cogn. Neuropsychol.* **33**, 175–190 (2016).
26. Grand, G., Blank, I. A., Pereira, F. & Fedorenko, E. Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nat. Hum. Behav.* **6**, 975–987 (2022).
27. Li, M. et al. Language-specific representation of emotion-concept knowledge causally supports emotion inference. *iScience* **27**, 111401 (2024).
28. Sailunaz, K., Dhaliwal, M., Rokne, J. & Alhajj, R. Emotion detection from text and speech: a survey. *Soc. Netw. Anal. Min.* **8**, 28 (2018).
29. Jackson, J. C. et al. From text to thought: how analyzing language can advance psychological science. *Perspect. Psychol. Sci.* **17**, 805–826 (2022).
30. Schwartz, H. A. et al. Personality, gender, and age in the language of social media: the open-vocabulary approach. *PLoS ONE* **8**, e73791 (2013).
31. Cutler, A. & Condon, D. M. Deep lexical hypothesis: Identifying personality structure in natural language. Preprint at http://arxiv.org/abs/2203.02092 (2022).
32. Rosenbusch, H., Wanders, F. & Pit, I. L. The Semantic Scale Network: an online tool to detect semantic overlap of psychological scales and prevent scale redundancies. *Psychol. Methods* **25**, 380–392 (2020).
33. Nimon, K., Shuck, B. & Zigarmi, D. Construct overlap between employee engagement and job satisfaction: A function of semantic equivalence? *J. Happiness Stud.* **17**, 1149–1171 (2016).
34. Arnulf, J. K., Larsen, K. R., Martinsen, Ø. L. & Nimon, K. F. Editorial: Semantic algorithms in the assessment of attitudes and personality. *Front. Psychol.* **12**, 720559 (2021).
35. Neuman, Y. & Cohen, Y. A vectorial semantics approach to personality assessment. *Sci. Rep.* **4**, 4761 (2014).
36. Kinnaird, E., Stewart, C. & Tchanturia, K. Investigating alexithymia in autism: a systematic review and meta-analysis. *Eur. Psychiatry* **55**, 80–89 (2019).
37. Stern, W. *Person und Sache. System der philosophischen Weltanschauung (3. d.: System des kritischen Personalismus.)* (Barth, 1906).
38. Heider, F. *The Psychology of Interpersonal Relations* (Wiley, 1958).
39. Vogeley, K. Two social brains: neural mechanisms of intersubjectivity. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **372**, 20160245 (2017).
40. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
41. Fried, E. I. What are psychological constructs? On the nature and statistical modelling of emotions, intelligence, personality traits and mental disorders. *Health Psychol. Rev.* **11**, 130–134 (2017).
42. Larsen, K. R., Nevo, D. & Rich, E. Exploring the semantic validity of questionnaire scales. In *Proc. 41st Annual Hawaii International Conference on System Sciences (HICSS 2008)* 440 (IEEE Computer Society, 2008).
43. Dowrick, C. et al. Researching the mental health needs of hard-to-reach groups: managing multiple sources of evidence. *BMC Health Serv. Res.* **9**, 226 (2009).
44. Braslow, J. T. et al. Generalizability of studies on mental health treatment and outcomes, 1981 to 1996. *Psychiatr. Serv.* **56**, 1261–1268 (2005).
45. Jack, R. E., Garrod, O. G. B., Yu, H., Caldara, R. & Schyns, P. G. Facial expressions of emotion are not culturally universal. *Proc. Natl Acad. Sci. USA* **109**, 7241–7244 (2012).
46. Jackson, J. C. et al. Emotion semantics show both cultural variation and universal structure. *Science* **366**, 1517–1522 (2019).
47. Zhou, K., Ethayarajh, K. & Jurafsky, D. Frequency-based distortions in contextualized word embeddings. Preprint at https://arxiv.org/abs/2104.08465 (2021).
48. Liang, Y., Cao, R., Zheng, J., Ren, J. & Gao, L. Learning to remove: towards isotropic pre-trained BERT embedding. Preprint at http://arxiv.org/abs/2104.05274 (2021).
49. Wang, Y. Single training dimension selection for word embedding with PCA. Preprint at http://arxiv.org/abs/1909.01761 (2019).
50. Raunak, V., Gupta, V. & Metze, F. Effective dimensionality reduction for word embeddings. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)* 235–243 (2019); https://aclanthology.org/W19-4328/?ref=https://githubhelp.com
51. Mu, J., Bhat, S. & Viswanath, P. All-but-the-top: simple and effective postprocessing for word representations. Preprint at http://arxiv.org/abs/1702.01417 (2017).
52. Garg, N., Schiebinger, L., Jurafsky, D. & Zou, J. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proc. Natl Acad. Sci. USA* **115**, E3635–E3644 (2018).
53. Dubossarsky, H., De Deyne, S. & Hills, T. T. Quantifying the structure of free association networks across the life span. *Dev. Psychol.* **53**, 1560–1570 (2017).
54. Heavey, C. L. & Hurlburt, R. T. The phenomena of inner experience. *Conscious Cogn.* **17**, 798–810 (2008).
55. Lovibond, S. H. & Lovibond, P. F. *Depression Anxiety Stress Scales (DASS-21, DASS-42)* (APA PsycTests, 1995).
56. Loewy, R. L., Pearson, R., Vinogradov, S., Bearden, C. E. & Cannon, T. D. Psychosis risk screening with the Prodromal Questionnaire–brief version (PQ-B). *Schizophr. Res.* **129**, 42–46 (2011).
57. Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J. & Clubley, E. The Autism-Spectrum Quotient (AQ): evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *J. Autism Dev. Disord.* **31**, 5–17 (2001).
58. Polner, B. et al. The network structure of schizotypy in the general population. *Eur. Arch. Psychiatry Clin. Neurosci.* **271**, 635–645 (2021).
59. Azis, M. et al. Structure of positive psychotic symptoms in individuals at clinical high risk for psychosis. *Early Interv. Psychiatry* **15**, 505–512 (2021).

60. Lång, U., Mittal, V. A., Schiffman, J. & Therman, S. Measurement invariance of psychotic-like symptoms as measured with the Prodromal Questionnaire, Brief Version (PQ-B) in adolescent and adult population samples. *Front. Psychiatry* **11**, 593355 (2020).
61. Reimers, N. & Gurevych, I. Sentence-BERT: sentence embeddings using siamese BERT-Networks. Preprint at http://arxiv.org/abs/1908.10084 (2019).
62. Cowen, A. S. & Keltner, D. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proc. Natl Acad. Sci. USA* **114**, E7900–E7909 (2017).
63. Wright, M. N. & Ziegler, A. ranger: a fast implementation of random forests for high dimensional data in C++ and R. Preprint at http://arxiv.org/abs/1508.04409 (2015).
64. Bischl, B. et al. Hyperparameter optimization: foundations, algorithms, best practices and open challenges. *WIREs Data Min. Knowl. Discov.* **13**, e1484 (2023).
65. Hubert, L. & Arabie, P. Comparing partitions. *J. Classification* **2**, 193–218 (1985).

Acknowledgements

The original studies analyzed in this work were supported by the National Institute of Mental Health (Grant R01MH112612) to J.S. and the Deutsche Forschungsgemeinschaft (DFG) ET 31/7-1 to U.E. K.V. was supported within the project SIMSUB (Grant 01GP2215) of the German Ministry of Research, Technology and Space (BMFTR). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

J.K. and K.V. generated the analysis strategy. J.S., L.K.-I., V.A.M. and U.E. provided expertise in the refinement of the analysis. All authors contributed to the interpretation of the findings and to the writing of the manuscript. All authors approved the final version of the manuscript. All authors have agreed both to be personally accountable for the author's own contributions and to ensure that questions related to the accuracy or integrity of any part of the work, even ones in which the author was not personally involved, are appropriately investigated, resolved and the resolution documented in the literature.

Funding

Open access funding provided by Universität zu Köln.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44220-025-00527-y.

Correspondence and requests for materials should be addressed to Joseph Kambeitz.

Peer review information *Nature Mental Health* thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

[1]Department of Psychiatry and Psychotherapy, Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany. [2]Department of Psychology, University of California, Irvine, CA, USA. [3]Department of Psychiatry and Psychotherapy, Ludwig-Maximilian University, Munich, Germany. [4]Department of Psychology, Northwestern University, Evanston, IL, USA. [5]Department of Psychology, University of Bonn, Bonn, Germany. [6]Institute of Neuroscience and Medicine - Cognitive Neuroscience (INM3), Research Center Jülich, Jülich, Germany. ✉e-mail: joseph.kambeitz@uk-koeln.de

nature portfolio

Corresponding author(s): Joseph Kambeitz

Last updated by author(s): Joseph Kambeitz (Sept 10, 2025)

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Please do not complete any field with "not applicable" or n/a. Refer to the help text for what text to use if an item is not relevant to your study.
For final submission: please carefully check your responses for accuracy; you will not be able to make changes later.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection: - No software was used for data collection

Data analysis: R and Python

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The data of the DASS (https://openpsychometrics.org/_rawdata/) and the O-LIFE (https://osf.io/epfvq/) are open and available. The data of the PQB is available in anonymized form upon reasonable request to the authors of the original publication (JS, VAM). The data of the AQ is not publicly available.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	No distinction was made between “sex” and “gender” in the reported studies.
Reporting on race, ethnicity, or other socially relevant groupings	Was not aquired in the samples
Population characteristics	Is reported on the manuscript
Recruitment	does not apply
Ethics oversight	We provide IDs of the ethical approval for the relevant data sets.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	n=39775 for the DASS, n=11807 for the OLIFE, n=1555 for the AQ, n=1099 for the PQB
Data exclusions	none
Replication	none
Randomization	does not apply
Blinding	does not apply

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Studies compares correlation structure of psychopathology in empirical data and LLMs All data analysed are quantitative data.
Research sample	Samples of patients and population samples
Sampling strategy	Secondary analysis of available studies
Data collection	Does not apply
Timing	Does not apply
Data exclusions	none
Non-participation	Does not apply
Randomization	Does not apply

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	
Research sample	
Sampling strategy	
Data collection	
Timing and spatial scale	
Data exclusions	
Reproducibility	
Randomization	
Blinding	

Did the study involve field work? ☐ Yes ☐ No

Field work, collection and transport

Field conditions	
Location	
Access & import/export	
Disturbance	

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
[X]	[ ] Antibodies
[X]	[ ] Eukaryotic cell lines
[X]	[ ] Palaeontology and archaeology
[X]	[ ] Animals and other organisms
[X]	[ ] Clinical data
[X]	[ ] Dual use research of concern
[X]	[ ] Plants

Methods

n/a	Involved in the study
[X]	[ ] ChIP-seq
[X]	[ ] Flow cytometry
[X]	[ ] MRI-based neuroimaging

Antibodies

Antibodies used	
Validation	

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	
Authentication	
Mycoplasma contamination	
Commonly misidentified lines (See ICLAC register)	

Palaeontology and Archaeology

Specimen provenance	
Specimen deposition	
Dating methods	

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Ethics oversight	

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals	
Wild animals	
Reporting on sex	
Field-collected samples	
Ethics oversight	

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	does not apply
Study protocol	does not apply
Data collection	does not apply
Outcomes	does not apply

Dual use research of concern

Policy information about dual use research of concern

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes	
X	☐	Public health
X	☐	National security
X	☐	Crops and/or livestock
X	☐	Ecosystems
X	☐	Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes	
X	☐	Demonstrate how to render a vaccine ineffective
X	☐	Confer resistance to therapeutically useful antibiotics or antiviral agents
X	☐	Enhance the virulence of a pathogen or render a nonpathogen virulent
X	☐	Increase transmissibility of a pathogen
X	☐	Alter the host range of a pathogen
X	☐	Enable evasion of diagnostic/detection modalities
X	☐	Enable the weaponization of a biological agent or toxin
X	☐	Any other potentially harmful combination of experiments and agents

Plants

Seed stocks

Novel plant genotypes

Authentication

ChIP-seq

Data deposition

☐ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links
May remain private before publication.

Files in database submission

Genome browser session
(e.g. UCSC)

Methodology

Replicates

Sequencing depth

Antibodies

Peak calling parameters

Data quality

Software

Flow Cytometry

Plots

Confirm that:

☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation

Instrument

Software

Cell population abundance

Gating strategy

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type

Design specifications

Behavioral performance measures

Imaging type(s)

Field strength

Sequence & imaging parameters

Area of acquisition

Diffusion MRI ☐ Used ☐ Not used

Preprocessing

Preprocessing software

Normalization

Normalization template

Noise and artifact removal

Volume censoring

Statistical modeling & inference

Model type and settings

Effect(s) tested

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference

(See Eklund et al. 2016)

Correction

Models & analysis

n/a	Involved in the study
X	☐ Functional and/or effective connectivity
X	☐ Graph analysis
☐	X Multivariate modeling or predictive analysis

Functional and/or effective connectivity

Graph analysis

Multivariate modeling and predictive analysis: Random Forest with hyperparameter tuning in repeated nested CV