TEXT PSYCHOMETRICS: ASSESSING PSYCHOLOGICAL CONSTRUCTS IN TEXT USING NATURAL LANGUAGE PROCESSING

Daniel M. Low [1,2,*] **Patrick Mair** [2] **Matthew K. Nock** [2,3,4] **Satrajit S. Ghosh** [4,5]

1. DAIR Center, Child Mind Institute, New York, NY, USA
2. Department of Psychology, Harvard University, Cambridge, MA, USA
3. Department of Psychiatry, Harvard Medical School, Boston, Massachusetts, United States
4. Franciscan Children's Hospital, Boston, Massachusetts, United States
5. McGovern Institute for Brain Research, MIT, Cambridge, United States
6. Department of Otolaryngology–Head and Neck Surgery, Harvard Medical School, Boston, Massachusetts, United States

* Correspondance to Daniel M. Low, daniel.low@childmind.org

ABSTRACT

Large language models (LLMs) have revolutionized natural language processing (NLP). Yet when used to assess psychological constructs in text, they are generally not evaluated for the types of validity, reliability, and standardization typically expected from traditional questionnaires with rating scales. This study bridges that gap by demonstrating how to evaluate the psychometric properties of text-based models, which we call Text Psychometrics.

We first review different NLP methods, compare their ability to address key challenges in psychological research such as explainability, and outline methods for evaluating them on many desirable psychometric properties. We then demonstrate this through two empirical studies. Study 1 classifies thousands of crisis counseling conversations and Reddit posts into different types of mental health issues and introduces a novel method to evaluate text models for content validity —the extent to which a test captures the full range of expressions of a construct. Study 2 examines prospective criterion validity by estimating how 49 known suicide risk factors predict imminent risk in crisis counseling conversations.

In sum, NLP studies in psychology often rely on only a few validation metrics; here, we demonstrate the need for broader psychometric evaluation and provide a practical blueprint and future directions for achieving it.

***K*eywords** natural language processing · large language models · psychometrics · artificial intelligence · validity · reliability

1 Introduction

Text data is ubiquitous: it is present in social media, direct messages, clinical interview transcripts, survey responses, electronic health records, and scientific documents. There are thousands of distinct well-studied constructs psychologists and social scientists might wish to quantify in text including emotions, moods, symptoms, attitudes, states, traits, and abilities, often through different dimensions such as valence, arousal, and severity. The goals of these assessments could be to screen for a health outcome, summarize a person's experience, trigger an intervention, or use measurements as variables in subsequent models to further understand a psychological theory (Demszky et al., 2023; Kjell, Kjell, & Schwartz, 2023). This can be used to make individual-level predictions, such as a suicide risk screening (Zuromski et al., 2024), or to do large-scale public-health monitoring such as understanding how a pandemic might be affecting different psychiatric populations over time (Low et al., 2020). The traditional approach Psychometrics has taken to measure psychological constructs has been through the use of rating scales (Figure 1). Rating scales ask a rater to convert a subjective observation into a specific value, usually on an ordinal scale. Compared to rating scales, systematically measuring constructs from natural language itself can seem more complex or inexact. However, language has been

shown to describe nuances and extremes with more detail and more ecologically than rating scales (Kjell et al., 2023). Natural language processing (NLP) offers many tools to quantify these constructs, and the recent growth in abilities of large language models (LLMs) –a subset of NLP models based on deep neural networks that have billions of parameters– shows promise for identifying constructs more reliably (Demszky et al., 2023; Kjell et al., 2023). Historically, the different types of validity, reliability, and standardization typically expected from questionnaires using rating scales (Rust, Kosinski, & Stillwell, 2020) are not normally carried out when evaluating NLP models for psychological measurement. Researchers would never publish scales without assessing these psychometric properties, but they do so for NLP and other behavioral or biological markers, which are also susceptible to poor psychometrics, a double standard that has been called the “psychometric bias" (Shankman, Kaiser, & Shenberger, 2020).

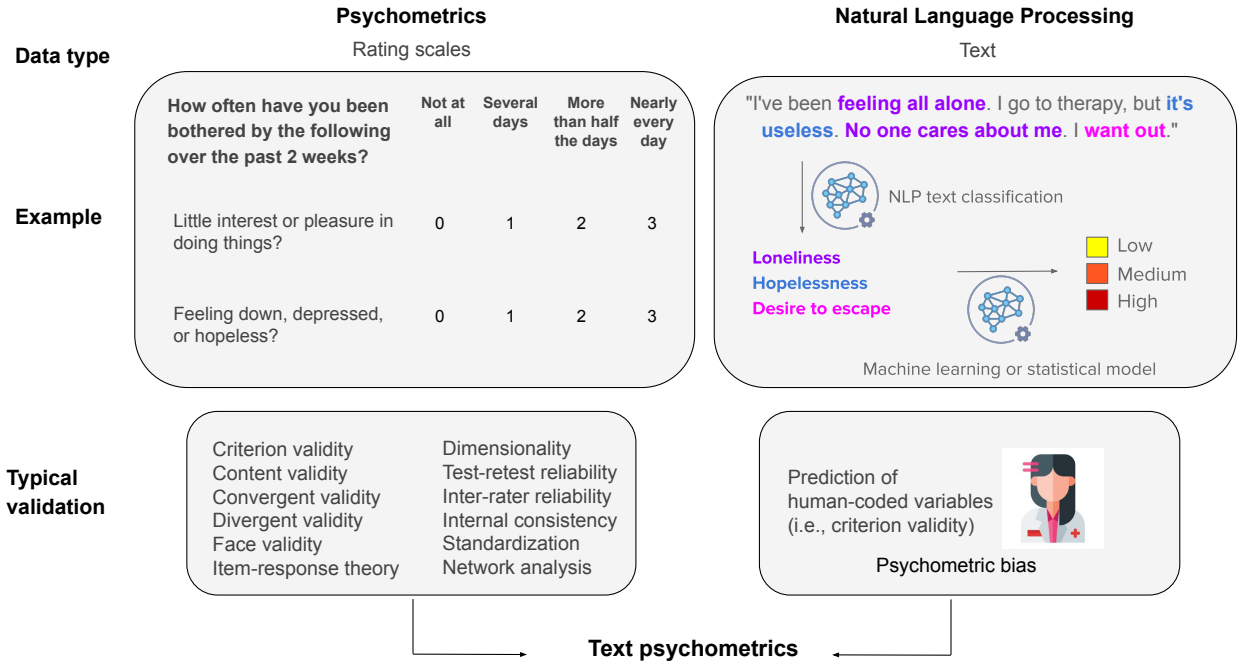

Figure 1: Psychometrics (left) focuses on using responses to rating scales to assess psychological constructs such as these from the PHQ-2. Natural language processing (NLP; right) can be used to create text classification models that output whether a text document expressed a given psychological construct. These outputs can be used as is or as independent variables in subsequent machine learning or statistical models (e.g., for risk assessments). While researchers tend to validate rating scale assessments on many psychometric properties using rich modeling choices, NLP studies in psychology mainly report the prediction of human-coded variables. This validation gap implies NLP suffers from a “psychometric bias," which text psychometrics could bridge. Definitions and examples of types of validity, reliability, and models are provided in section 4.

NLP studies and machine learning more broadly have largely focused model validation on a specific type of validation in psychology: prediction of human-coded labels (Cohen et al., 2022; Grimmer & Stewart, 2013; Riezler & Hagmann, 2022). [1] For instance, humans may code sentences from Reddit for different emotions, and then a machine learning model can learn to predict human-coded emotions from patterns in language using a variety of methods (Demszky et al., 2020). Predicting human-coded labels in NLP models is also the main type of validation for NLP benchmark leaderboards. The latter allow users to compare and rank models evaluated on many datasets and tasks. For instance, the Massive Text Embedding Benchmark covers 8 tasks (e.g., text classification, summarization, semantic similarity), 58 datasets, and 112 languages (Muennighoff, Tazi, Magne, & Reimers, 2022). Given that the goal of many applications is

[1] Prediction from a machine learning perspective is defined as predicting the value of the dependent variable for new samples the model did not use for any aspect of model fitting (i.e., out-of-sample prediction using cross-validation or other resampling methods) (Raschka, 2018); whereas association is a broader term that can refer to (a) out-of-sample prediction, (b) estimation of the value of the dependent variable of samples the model used during model fitting (i.e., within-sample association), or (c) estimation of group-level coefficients (e.g., correlations, within-sample regression). The latter two forms of association are often referred to as “prediction” in psychology, but these within-sample associations do not imply out-of-sample prediction capabilities (i.e., generalizability; see Appendix A) without validation. Many of the psychometric properties we will introduce in section 4 can be validated through association or prediction.

to try to automate human tasks, proving a model that generalizes and can predict new samples of human judgments it has not encountered is a good priority, but as we demonstrate here, should not be the only priority. A single metric is likely not capable of capturing many aspects of validity.

Our study aims to provide a framework for evaluating psychometric properties of NLP models and is structured as follows:

- In section 2, we argue why consolidating Text Psychometrics as a field is important and we review the non-obvious shared history between psychometrics and NLP.
- In section 3, we review different NLP methods and how they can be used to assign a score to a text document (i.e., text classification) for psychological assessments. We also propose key constraints psychology researchers face when using NLP models (e.g., small datasets, the need for explainable and private models) and how to overcome them with different NLP model characteristics. We also introduce an improved method of text classification with embeddings we call "Construct-Text Similarity."
- In section 4, we provide an overview of how psychometrics has evaluated the validity and reliability of psychological assessments and how these properties can be evaluated in text data.
- In section 5, study 1, we provide a use case and test these methods in detecting a diverse set of mental health issues including suicidal ideation, eating disorders, and physical abuse to compare methods (open-source and proprietary LLMs, lexicons, and embeddings methods) and observe their performance across constructs. We provide a novel way of quantitatively evaluating content validity in text, that is, whether a tool assesses most behaviors or expressions associated with the construct being measured.
- In section 6, study 2, we focus on a methodologically more challenging text classification task, that of predicting who out a set of suicidal individuals is at high imminent risk, which is a scenario with several common constraints in clinical NLP: an imbalanced dataset; and the need for models that are trustworthy and explainable. We predict this using initial parts of the conversation to evaluate prospective criterion validity. We also demonstrate how to inform theory by estimating how predictive each variable is of the dependent variable (i.e., feature importance).
- In section 7, the Discussion, we systematically compare NLP methods for text classification on a subset of these desirable properties so researchers can decide which NLP method to use, given each method may respond better to the different constraints researchers may have. And we discuss open areas of research for Text Psychometrics.

2 Text Psychometrics

2.1 Why text psychometrics?

We consider that Text Psychometrics is a concise term for the subfield of psychometrics that studies how to assess psychological constructs in text or natural language, especially when focusing on validity, reliability and standardization properties. Focusing on this subfield could help concentrate discussion on broadening the validation of NLP models for psychological assessment and showing limitations of the most common approaches. A related field uses NLP to improve the development and validation of rating-scale assessments, such as generating questionnaire items (e.g., Laverghetta Jr and Licato (2023)), which could also fall under Text Psychometrics, and under the more established 'computational psychometrics' (von Davier, Mislevy, & Hao, 2022), which is much more broad in scope. A second related area of research is assessing the artificial psychological profiles of artificial intelligence (AI) models (Shu et al., 2023a), which has been referred to as AI Psychometrics (Pellert, Lechner, Wagner, Rammstedt, & Strohmaier, 2023). We can distinguish this from Text Psychometrics in that the latter measures psychological properties in humans from their natural language or uses NLP to improve rating-scale assessments in humans. AI psychometrics would not be a fitting label for these applications given many NLP methods (e.g., lexicons, TF-IDF, sentence embeddings) should probably not be considered intelligent, at least orders of magnitude less so than current LLMs. Text Psychometrics is simply performing psychometrics from text and natural language instead of rating scales.

With this subfield as a bridge, we hope to consolidate efforts to taxonomize and standardize the evaluation of desirable properties of these NLP-based models in a broader sense than what has been done thus far.

There is a growing number of studies trying to improve the psychometric validation of text-based NLP models for psychological assessments (Ahmad et al., 2020; Fang, Nguyen, & Oberski, 2022; Kjell et al., 2023; Riezler & Hagmann, 2022; Van der Wal et al., 2022). Each study has focused on specific types of models and properties. Cohen et al. (2022) demonstrated that an NLP model detecting paranoia from text data had good test-retest reliability (see section 4

for definitions), it correlated positively with clinical ratings of paranoia and correlated negatively with measures of anxiety and depression. Fang et al. (2022) evaluated different types of validity (face, convergent, divergent, content, predictive) of several text embedding models using synthetic natural language data. Van der Wal et al. (2022) describe how to assess validity and reliability of NLP model bias such as gender bias. Hoyle et al. (2021) have demonstrated that automated evaluation of topic models using coherence metrics fail to capture human judgments. Kennedy, Bacon, Sahn, and von Vacano (2020) had annotators rate the degree of hate speech in text documents using an ordinal scale and then used item response theory and deep learning to build a model that outputs a debiased continuous measure of hate speech from text. We contribute to prior work by (1) describing how NLP and psychometrics have a shared history (section 2.2), (2) describing traditional and more recent NLP methods and proposing desirable NLP model characteristics (section 3.2), (3) proposing additional desirable psychometrics properties and methods for evaluating these properties in NLP models (section 4), (4) evaluating these properties across two studies (sections 5 and 6), and (5) discussing open areas of research and future directions (section 7).

2.2 Bridging natural language processing and psychometrics, two fields with a shared history

It is important to note that human language has been at the center of psychometrics' history. For a brief overview of psychometrics and related methods including item-response theory, network analysis, and causal inference, see Appendix B. In Figure 2 we highlight key developments in psychometrics, NLP, and AI throughout history. Psychometrics has discovered key attributes humans have (e.g., states, traits, attitudes) often by collecting large lists of words associated with psychological traits. Since early in psychometric history, it was clear that just because we have different words for traits does not mean these words refer to different attributes; therefore, there was a need to discover the underlying independent and general traits from all possible trait names (Allport & Odbert, 1936). Thurstone's (1934) "The Vectors of Mind" (Thurstone, 1934) suggested we can discover the key, non-redundant psychological traits by having individuals rate if they would describe someone with adjectives such as "friendly" or "cheerful." Using factor analysis, he grouped words with similar ratings and was surprised to discover there were only five factors underlying personality. Factor analysis would return how much each word is associated with each factor (i.e., factor loadings). He proposed plotting each adjective as a vector of points in space in five dimensions whereby similar traits would point in the same direction; if they are anticorrelated they would be pointing in opposite directions; and if they are independent, then they would be orthogonal (p. 15). One could measure this using cosine angles from the center of the space. Allport, Odbert, and Cattell between the 1930s and 1950s took this approach and gathered such ratings on thousands of words, an approach that ultimately led to the Big five factors of personality (Goldberg, 1981), which includes openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism.

Lexicons (i.e., lists of words and phrases associate with constructs) have also been used to measure psychological constructs in text. The General Inquirer gathered thousands of words related to different constructs, but used the first commercial computers from the 1960s to automatically count the appearance of different words in corpora (Stone, Dunphy, Smith, & Ogilvie, 1962) which later inspired the widely used Linguistic Inquiry and Word Count (LIWC) (Boyd, Ashokkumar, Seraj, & Pennebaker, 2022) and many other lexicons (Low et al., 2024).

Not only was language key in developing psychometrics, psychometrics has also influenced NLP, although this is relatively unknown. About three decades after Thurstone's (1934) address, vector space models by Salton and others in the 1960s and 70s took this factor analysis approach (Dubin, 2004), (without acknowledging Thurstone as far as we can tell) but vectorized words by counting how many times a word co-occurs with other words (dimensions of the vector) within small windows of text across different corpora (Dubin, 2004; Salton, Wong, & Yang, 1975). This was used to find documents that were similar to a query in information retrieval systems (e.g., finding articles about a topic in libraries).

This approach, along with advances in deep learning and hardware capabilities, contributed to the explosion of vector semantics that came with the creation of the word2vec word embedding algorithm in 2013, which was able to better quantify aspects of the meaning of words in vectors (i.e., ordered lists of numbers that represent positions in a multidimensional space). Improving upon word embeddings led to current state-of-the-art transformer-based embeddings, which can encode aspects of sentence and document meaning. These embeddings from transformer models have a wide variety of uses including information retrieval, paraphrase detection, categorizing text, summarization, and keyword extraction (Muennighoff et al., 2022). The origins of vector semantics is often credited to hypotheses in linguistics from the 1950s from Harris ("If A and B have almost identical environments we say that they are synonyms" Harris (2013)) and Firth ("You shall know a word by the company it keeps" Firth (1957)): vector semantics captured a word's meaning by counting which words it co-occurs with. However, we find factor analysis as previously described also had a large influence as well. We can add the following quote by Thurstone (1934) to the history of vector semantics: "each trait can be represented as a point on the surface of a ball. If two traits A and B tend to coexist, the two points will be close together on the surface of the ball." (Thurstone (1934) p. 14).

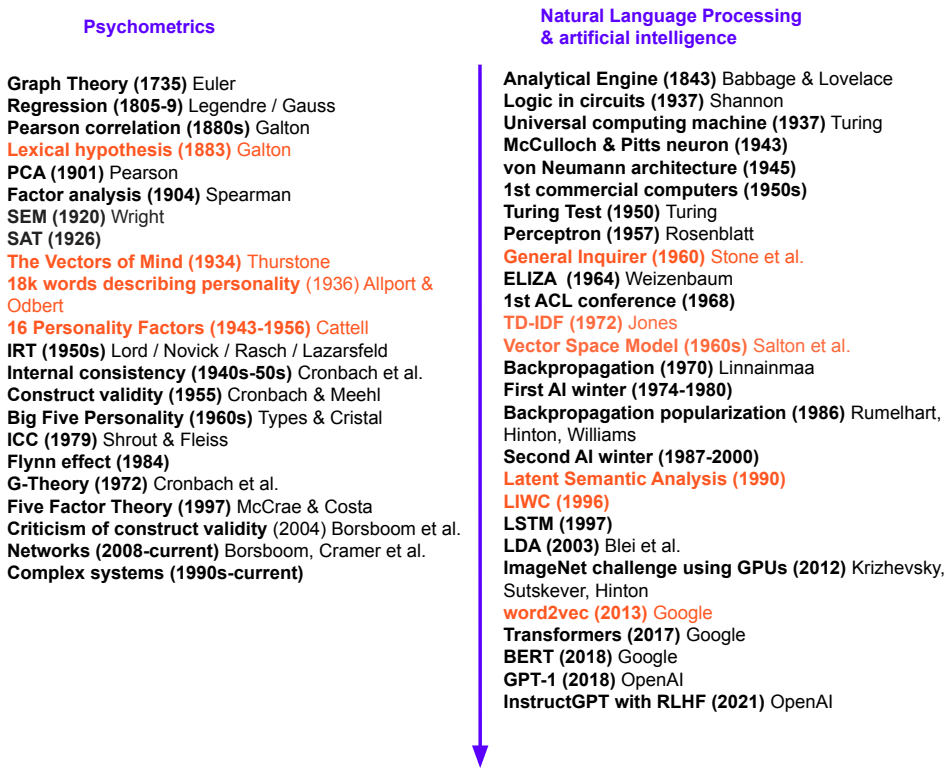

Figure 2: Psychometrics and natural language processing have a shared history. Several, non-exhaustive examples of key developments that influenced each other across fields are displayed in orange.

3 Using natural language processing for psychological assessments

3.1 Brief overview of text classification methods for psychological assessments

Here we provide different ways NLP tackles the task of assigning a score to a text document for a given construct (e.g., does the text mention loneliness?), which is called "text classification" in NLP. This is analogous to assessments with rating scales, which generally assign a number to a construct for a given individual [2]. We focus on describing prior work achieving text classification through multiple methods displayed in Figure 3 [3].

Lexicons or dictionaries. Using lexicons is a traditional way of performing text classification (Stone et al., 1962). Lexicons contain one or many constructs, each with an associated list of tokens (words and phrases) – which can be validated by domain experts – that are counted in a given document (Low et al., 2024). Higher counts indicate the construct is more likely to be present. Then these counts can be used as features in a statistical or predictive model, which would be highly interpretable and would avoid missing obvious, prototypical tokens (e.g., "anorexia" for eating concerns) that might be missed by a high-performing deep learning model (Low et al., 2024). Our *construct-tracker* packages allows for using generative AI to automatically create the initial lexicon, which can then be validated by human experts (Low et al., 2024).

Term Frequency–Inverse Document Frequency (TF-IDF) is a common technique for identifying words and phrases in documents that characterize a dependent variable based on how frequently they appear in a given text relative to a larger corpus (Jones, 1972). These scores can be used as features in predictive models (e.g., see Low et al. (2020) for words that characterize different mental health support groups on Reddit). Unlike lexicons, TF-IDF requires no expert input, making it data-driven, scalable, and domain-agnostic. However, as with lexicons, TF-IDF treats words independently and lacks semantic understanding. Furthermore, this data-driven approach that finds words that characterize a variable is likely to be dependent on the specific dataset used and might not generalize to other datasets.

[2] For a more formal description of text classification, see Appendix D.

[3] For more in depth introductions to machine learning, deep learning, and LLMs, see (Pargent, Schoedel, & Stachl, 2023; Prince, 2023; Raschka, 2024).

Other methods presented here (including lexicons and LLMs) have the advantage of detecting a broad coverage of words that are prototypical of a construct –if designed to do so– and not just associated in some way in a specific dataset (Low et al., 2024).

The remaining methods are generally carried out using deep learning methods. Rarely is training done from scratch on the target task (e.g., loneliness detection), given they generally require large datasets coded for the target task; rather, pretrained models like BERT (Devlin, Chang, Lee, & Toutanova, 2018) are trained on a very large corpus (the entire English Wikipedia and other sources) in a self-supervised manner (e.g., predicting masked words within the input) to generate rich, internal representations that quantify aspects of meaning of words and phrases. Then this model can be fine-tuned (i.e., adapted) for a specific task (e.g., detecting loneliness in text) on a smaller, annotated dataset (e.g., documents coded for loneliness or other constructs) by attaching a feedforward neural network (Sun, Qiu, Xu, & Huang, 2019). Pretraining can also be done in a supervised manner by training to predict labels in large, publicly-available datasets (e.g., for named-entity recognition). Then these pretrained models can be fine-tuned for the target task, which should improve performance in comparison to using the target dataset to train the model from scratch.

Named-entity recognition classifies each token as a possible entity belonging to a category or ontology (Ashok & Lipton, 2023). The presence or sum of entity types in a document would return similar results as text classification. The general approach is to train a model to output a probability for each entity type.

Natural language inference and textual entailment. This method can be used for text classification as well (Gretz et al., 2023; Schopf, Braun, & Matthes, 2022; Yin, Hay, & Roth, 2019). Entailment is inferring whether two sentences are a contradiction or an entailment (e.g., “Met my first girlfriend that way." is a contradiction of “I didn’t meet my first girlfriend until later."; “At 8:34, the Boston Center controller received a third transmission from American 11" is an entailment of “The Boston Center controller got a third transmission from American 11."; Williams, Nangia, and Bowman (2018)). In a classification setup, the document to be classified can be treated as the premise for which one can build a positive hypothesis using the template “this text is about <construct>" as well as a negative hypothesis using other labels; (Yin et al., 2019). Gretz et al (2023) Gretz et al. (2023) fine-tune natural language inference models which output entailment and contradiction scores that serve for text classification.

Fine-tuning deep learning models. Whereas traditional methods involve some preprocessing (e.g., lemmatization, data cleaning, and feature extraction), pretrained deep learning models like BERT allows for end-to-end prediction, taking raw data, skipping most preprocessing, and outputting a prediction (Devlin et al., 2018). However, this generally requires a substantial amount of labeled training data, at least in the thousands of samples (although new advances such as LoRA are improving this (Hu et al., 2021)).

Construct-text similarity. Since lexicons can miss important tokens that they do not have (e.g., a lexicon may have suicide attempt but miss “attempted suicide" if not in the lexicon), there is a method to find semantically similar tokens to the ones in lexicon using deep learning using relatively interpretable method called similarity-based zero-shot text classification (Schopf et al., 2022) or sentecon (Lin & Morency, 2023). We introduce a more generalized version that has more parameters to chose from for each step, which we call construct-text similarity (which we expect will be a more transparent term for psychologists), as shown in Appendix Figure 8. This method works as follows: it first computes the similarity between a target construct representation (e.g., text embedding for the phrase "eating disorders" or for many tokens related to eating disorders or the average of all of these) and each token in a given document (e.g., text embeddings for phrases). It then takes a summary statistic as the final output (e.g., maximum between all construct and document embeddings). This approach produces a zero-shot output, meaning the method requires no training data. The construct can be the average of the text embeddings for a lexicon’s category, which means that it can capture much of what the lexicon captures and also similar words while a lexicon can only capture exact matches. The process of being able to capture similar words has been demonstrated to outperform lexicons (Lin & Morency, 2023). It is a middle ground between interpretable lexicons which will miss similar words if not in the lexicon and standard supervised deep learning which is a black box and hard to explain. It can also combine counts (0 or N with semantic similarity if there is not an exact match). However, it does not learn the contextual cues or reasoning abilities found in the previously described deep learning models. The resulting score can be used as the output for zero-shot text classification. It was found to outperform the textual entailment approach in a zero-shot setting (Schopf et al., 2022), but not when textual entailment models were fine-tuned (Gretz et al., 2023). It is also more lightweight than deep learning methods because it can run more easily on standard CPUs, unlike deep learning, which often requires specialized GPUs. This is especially beneficial for psychology researchers who may not have the expertise to work with GPUs.[4]

Large language models that generate text can be used for text classification (i.e., generating a score) by designing prompts to score whether certain constructs are present in a given document (Gretz et al., 2023; Rathje et al., 2024)

[4]For a tutorial on construct-text similarity, see `https://github.com/danielmlow/construct-tracker/blob/main/tutorials/suicide_risk_lexicon_construct_text_similarity.ipynb`

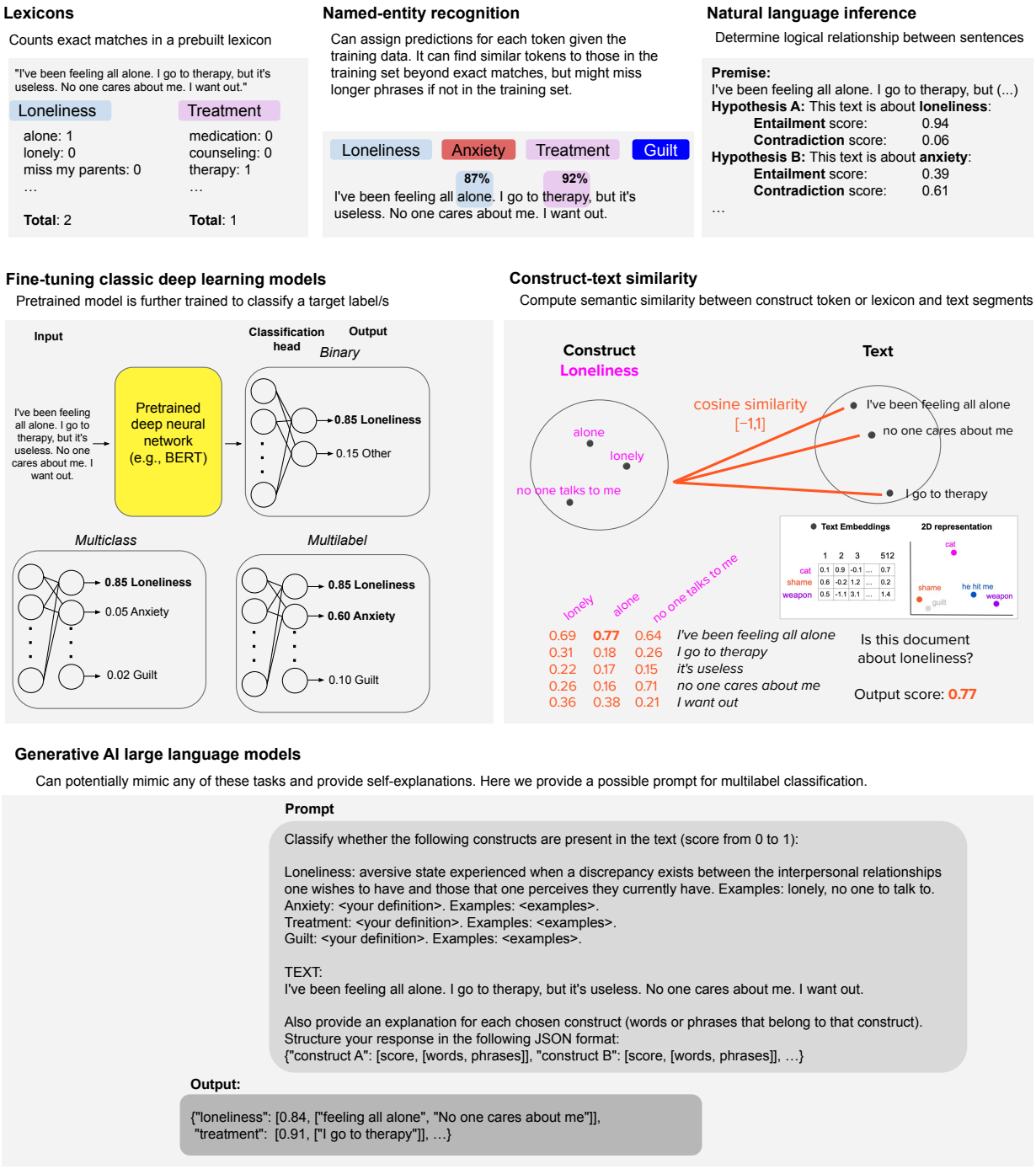

Figure 3: Natural language processing tasks and methods that can be used for text classification. Lexicon counts can be used as is or as variables in a machine learning model to further predict a target label. The rest of the methods tend to use deep neural networks. Text classification varies between binary classification (one out of two possible answers is correct), multi-class (one out of several answers is correct), and multi-label (more than one answer can be correct). Finetuning a model on a specific dataset (e.g., psychotherapy transcripts) that has previously been trained for the type of classification on more general data (non-mental health data) can help performance. Construct-text similarity turns words and phrases into embeddings, which are vectors in a semantic space where more similar phrases in meaning should be closer together. Here we take the maximum similarity between construct tokens from a lexicon and phrases from the target document as the final zero-shot classification score. Conceptually, this is similar to asking where does the document mention something most related to loneliness? Large language models (LLMs) tend to show state-of-the-art performance but may not be preferred due to difficulty of use or privacy limitations. Submitting many texts to LLMs can be automated through the use of LLM application programming interfaces (APIs).

–which can be considered LLM-assisted qualitative coding (Hämäläinen, Oksanen, Tavast, & Bhatnagar, 2024; Hämäläinen, Tavast, & Kunnari, 2023). However, the best-performing models (e.g., GPT-4.1, Gemini 2.5, Claude Opus 4) tend to be proprietary (Chiang et al., 2024) and therefore require users to submit the prompt to a company's server through a browser or API (Rathje et al., 2024), which may be infeasible if the data is highly sensitive. Another reason for submitting data to APIs are because generally for a fee, users can run inference or training quickly on cloud-based GPUs (e.g., OpenRouterAI, Replicate, Huggingface API, Microsoft Azure, Amazon Bedrock). Alternatively, users can download open-source LLMs (e.g., Gemma, Llama, Qwen, Mistral series). These tend to be smaller (e.g., 2B-70B parameters) and underperform their larger proprietary counterparts (with the exception of DeepSeek R1 which has 671B parameters Guo et al. (2025) and Llama-4 Maverick with 400B parameters (Meta AI, 2024)) and require considerable computational resources such as GPUs. Models can be quantized by reducing the precision of the weights for faster inference and fine-tuning while achieving similar performance (Dettmers, Pagnoni, Holtzman, & Zettlemoyer, 2024) which can lead lighter models to run on a personal computer's GPU or CPU (e.g., using packages such as transformers, Ollama, GPT4All, llama cpp), which is likely to become very prevalent as models become smaller and more efficient and hardware capabilities improve. Leaderboards include Huggingface Open LLM leaderboard[5] and the Chatbot Arena [6] (Chiang et al., 2024).

3.2 Desirable natural language processing model characteristics that can address key constraints in psychological research

Now that we have introduced several NLP methods, we highlight desirable NLP model characteristics as shown in Table 1. These tackle pragmatic constraints that psychology researchers often face. One is the limited availability of big training datasets, especially in clinical settings. This poses a challenge for fine-tuning standard supervised learning models such as BERT that require large datasets. A second challenge is the need for interpretable models to understand how a model is making predictions to increase trust that it is not biased (Kaur, Uslu, Rittichier, & Durresi, 2022). Explainability methods also allow researchers to rank which variables are most predictive of a target variable, which can be used to advance scientific understanding of the variables or discover novel risk factors (Molnar, 2020). A third challenge is the need to keep sensitive data private, which currently limits the use of submitting data to cloud services to use existing proprietary generative AI models, although cloud-based data security may improve in the near future.

Table 1: NLP and AI-specific desirable properties of psychological assessments that are specific to NLP and AI models, not traditional rating scales.

Desirable NLP model characteristic	Description
Zero-shot	Traditional supervised learning requires hundreds-to-thousands of examples of text-label pairs to train (e.g., documents coded for loneliness). Self-supervised learning enables training on even larger amounts of data because it does not need to be coded and often leads to general domain models (e.g., GPT-4) that can solve different types of downstream tasks (e.g., detect loneliness in text) zero-shot, that is without any further training or fine-tuning on the target dataset. This is extremely useful to avoid the resources involved in training especially when little annotated training data is available. This is expected in the social sciences: since there are thousands of constructs that could be measured, having annotated datasets for each construct is infeasible. **NLP example:** NLP has distinguished this type of text classification problem where categories are not pre-defined as topical text classification versus the more standard categories such as sentiment, genre, or spam, for which it is easier to find labeled datasets for fine-tuning; zero-shot methods are particularly useful to classify new categories (Gretz et al., 2023; Lang, 1995).
Explainable	One useful definition of explainability refers to the ease with which humans can simulate how a trained model takes an input and produces an output (i.e., simulatibility (Lipton, 2018)). This is straightforward in small decision trees and linear models and less feasible in tree ensembles and neural networks. **NLP example:** Counts from lexicons tokens trained on a linear model are interpretable; asking GPT-4o for an explanation of its predictions can seem plausible but might not be truthful with regard to the underlying prediction mechanism, which remains a black-box (Agarwal, Tanneru, & Lakkaraju, 2024).

[5] `https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard#/`
[6] `https://lmarena.ai/`

Reasonable errors	An error analysis can be performed by annotating a random subset of a model's false positives and false negatives (or of the largest errors in a regression task) to determine whether mistakes are reasonable or not. **NLP example:** A model detecting social media users who are actively suicidal (i.e., where the user posts about a desire and/or intent to die by suicide) may incorrectly classify text about someone else's suicide or prior suicidal ideation as a high-risk case, a more reasonable error than detecting non-mental health language as suicidal (Zuromski et al., 2024).
Semantically similar matching	Traditional keyword search methods using lexicons depend on an exact match and can miss related tokens. **NLP example:** A lexicon would not match the word "unworthy" even if it contains the word "worthless". Deep learning methods and Construct-Text Similarity are more likely to associate them.
Sensitive to context	It is important that even though a model detects a certain token (e.g., "depressed") as related to the dependent variable (depression), it should modify its prediction based on the token's context which could include negation, hedges, degree modifiers, modality markers (e.g, "not", "might be", "a bit", "very", "I must be."). **NLP example:** Rule-based methods such as counting tokens from a lexicon generally do not capture context as is better achieved with deep learning.
Ease of use & low computational resources	Deep learning models generally require using GPUs for efficient training and for inference (when they are large) and knowledge of running models, which is infeasible for most psychology researchers. However, the use of LLMs via graphical user interfaces (e.g., ChatGPT) and cloud-based APIs (e.g., OpenRouterAI) can reduce these barriers. **NLP example:** Using lexicons in statistical models can be run on a single CPU while fine-tuning or running inference for encoder BERT-style models and decoder generative AI models requires substantial technical and GPU capabilities (although light-weight models are becoming increasingly available; Dettmers et al. (2024)).
Data privacy	LLMs can be used through cloud APIs when they cannot be downloaded because they are either proprietary or when a user prefers to avoid setting up their own GPUs and pipelines. Sending sensitive data to APIs may not be secure. **NLP example:** Sharing names, professions, addresses, numbers, birthdates, medical history, and payment information. Open-source models can be downloaded to use in a local, secure way (demonstrated in the current study).
Robust to dataset shift	The ability of a model trained on a given dataset to generalize to data that differs from the original dataset in time, demographic distributions, language, feature extraction methods, etc. (Wiles et al., 2021) **NLP example:** Train model on data from 2021 and prove it performs well on data from 2024.

Given these constraints and NLP-model characteristics, researchers can choose different types of approaches to perform text classification, which we guide how to do in section 7.1. Once an NLP model is trained using machine learning, different psychometric properties can be evaluated. There is not a one-size-fits-all method, and choosing a method will depend on these constraints and on evaluating multiple psychometric properties.

4 Desirable psychometric properties of psychological assessments and how to evaluate these properties in text

There are many desirable psychometric properties that we may want a psychological assessment tool to fulfill. Here we provide a concise outline of these properties with examples of how to evaluate them with text-based models. This list of properties is helpful for several reasons: first, to adapt the definitions of classical and modern psychometric concepts to fit the needs of current NLP and AI goals. And second, to inspire new types of analyses that might be less frequent among NLP researchers so they can validate their models more broadly and help link their intuitive analyses to concepts that already have a vast literature in psychometrics.

Criterion validity. An instrument score is valid if it is associated with a criterion assumed or demonstrated to be valid Rust et al. (2020). The criterion is often a human or expert judgment but can also be a certain behavior or biological variable. Psychometrics has generally focused on within-sample associations, while NLP and machine learning focus on out-of-sample prediction. There are three types of criterion validity, depending on *when* the criterion and the instrument were assessed:

(a) **Concurrent validity**: The instrument's score is associated with a more standard or previously validated questionnaire total score or questionnaire item that at face value is valid, assessed at the same time point, either through correlations or out-of-sample prediction. *NLP example:* Correctly classifying crisis counseling conversations as being about physical abuse given the counselor's annotation (i.e., the criterion; demonstrated in section 5).
(b) **Prospective validity**: Rating scales should be associated with future scores or future behaviors within the same person, which may require repeated measures. This will also depend on whether the construct is an episodically-changing trait or a persistent trait. Higher scores on a suicide risk assessment questionnaire should have a positive correlation with future suicide attempts. This is related to the study of dataset shift in machine learning (see Table 1). This has been traditionally referred to as "predictive validity" (Rust et al., 2020); however, given correlation is not prediction (see Prediction entry below), we use "prospective validity" instead. *NLP examples:* (1) Demonstrate baseline language patterns are associated with future PTSD symptoms (Son et al., 2023). (2) Predict whether a conversation will be annotated as an imminent risk by the counselor (better assessed towards the end of the session) from the initial 25% of the conversation (demonstrated in section 6).
(c) **Retrospective validity**: Whether an instrument reflects past conditions or behaviors. For example, the score of a suicide risk assessment should correlate with lifetime suicidal thoughts and behaviors. *NLP example:* Train a model to detect language about depression and correctly identify which individuals had been diagnosed with a depressive disorder in the past, independent of current diagnosis.

Ecological validity. The assessment resembles real-world conditions vs. tightly controlled laboratory conditions. *NLP example:* Have participants journal by randomly prompting them during the day (i.e., ecological momentary assessment); social media data (Low et al., 2020).

Face validity. Whether a test appears to be valid for members of the target population based on a subjective judgment. For instance, individuals may not support the claim that the way they draw a person –a common projective assessment– clearly reflects their personality or mental health. If a math exam is clearly too hard for test-takers, then it will have low face validity. A scale assessing honesty may purposely try to have low face validity to avoid a social desirability bias. *NLP example:* The individuals being assessed could compare the construct being assessed (e.g., anxiety) with the question they were asked (e.g., "Describe what has been worrying you the most today", which would have high face validity versus "Describe this picture of a forest" which may have low face validity). If the model is explainable (see Table 1), they could assess the appropriateness of the explanation. For instance, if a model uses words from text like "bulimia" and "eating" to indicate eating and body image concerns but ignores "anorexia", it may indicate low face validity (Low et al., 2024)).

Construct validity. Some constructs such as pain, mood, attitudes, disability, and quality of life are not directly measurable and therefore do not have a clear gold-standard method of measurement (Davidson, 2014). Construct validation is thought to provide support that the measurement is measuring a given construct. A measurement score for a construct should not be highly associated with scores for unrelated constructs (discriminant validity) and associated scores for similar constructs (convergent validity). It can be demonstrated though a pairwise correlation matrix of related and unrelated scores or through prediction. Discriminant validity has a variety of methods (Rönkkö & Cho, 2022). If one of the two variables being associated is considered true (e.g., a gold standard), then the association would be a demonstration of criterion validity; when neither is, it is a demonstration of construct validity. A test can be similar to another, but they can both be invalid; therefore construct validity can be seen as an inferior form of validity than criterion validity with a gold standard, and there is good reason to suggest it does not assess validity properly (Borsboom, Cramer, Kievit, Scholten, & Franić, 2009). *NLP example:* Among individuals at clinical high-risk for psychosis, the NLP metric of semantic coherence was correlated with positive thought disorder severity (e.g. tangentiality and derailment) and the NLP metrics of complexity (sentence length and complementizer use) were correlated with negative thought disorder severity (poverty of speech and content) (Bilgrami et al., 2022). A type of construct validity is **known-groups validity**: When a test score discriminates between two groups known to differ (Davidson, 2014). *NLP example:* A valid measure of quality of life from natural language should likely differ between groups with and without chronic pain.

Content validity. The systematic examination of the test content by an expert to determine whether it covers a representative sample of the behavior to be measured (Anastasi & Urbina, 1997). A scale assessing depression should cover most common symptoms; however, depression scales perhaps should be expected to have low content validity given the dozens of diverse symptoms that are considered part of depression (i.e., heterogeneity, (Fried, Flake, & Robinaugh, 2022)). *NLP example:* Even though a classifier has relatively high performance on a test set, this does not guarantee it covers a representative sample of the behaviors of the construct being measured. A classifier can achieve higher performance detecting eating or body image concerns while missing "anorexia" if this token is relatively infrequent in the test set (Low et al., 2024). LIWC has low content validity for most mental health symptoms such as hopelessness or suicidal ideation. Content validity test sets compute the sensitivity of a given trained model on its ability to predict as positive a list of tokens that cover most types of instances of the construct (proposed in the current study).

Epistemic validity and clinical utility. Validity is generally seen as a demonstration that a test measures what it intends to measure (Borsboom, Mellenbergh, & Van Heerden, 2004; Kelley, 1927), independent of how it is being used (Borsboom, 2006). However, a test may be used to inform a diagnosis or treatment plan or to determine the efficacy of a therapeutic or educational intervention. Aspects of how a score is interpreted and used in practice ethically or unethically are related to its epistemic validity (Truijens, Cornelis, Desmet, De Smet, & Meganck, 2019) (see also, Jacobs and Wallach (2021)). Relatedly, determining whether a certain sensitivity and precision of a test is appropriate depends not on only their values but on how decisions will be made as a consequence of their values, their clinical utility. This can be quantified through a net benefit analysis which weighs potential benefits versus harms of a test, considering its true positive and false positive rates (Kennedy et al., 2021; Kessler, Bossarte, Luedtke, Zaslavsky, & Zubizarreta, 2020; Vickers, Van Calster, & Steyerberg, 2016). *NLP example:* Consider an NLP model that detects suicide ideation in the last clinical visit to determine whether their should be a follow-up; if the model has low precision, it might still lead to improved outcomes compared to not deploying it, a net benefit.

Fairness. Related to the concept of equivalence in psychometrics, a test should not unintentionally function differently for different groups (e.g., age, disability, gender, race, ethnicity, education level, sexual orientation) Rust et al. (2020), particularly if it is unknown to the test recipient. More broadly, fairness relates to evaluating whether discoveries (e.g., five dimensions of personality) generalize as universals beyond the socioculturally-narrow populations where they were tested (Gurven, Von Rueden, Massenkoff, Kaplan, & Lero Vie, 2013). *NLP example:* A model underpredicts the need for an intervention for a protected group. Metrics and algorithms exist to identify and mitigate bias (Hort, Chen, Zhang, Harman, & Sarro, 2023). Discovering social bias in LLMs is another active field of research given the data used to train models, while large in scale, may also be narrow socioculturally (Kotek, Dockum, & Sun, 2023).

Measures of uncertainty. Uncertainty can be considered as doubt about the validity of a measurement result and a dispersion of the values that could reasonably be attributed to the attribute being measured (Jcgm et al., 2008). While standard deviations and confidence intervals can be estimated as average effects, individual predictions should also convey level of confidence as well as a lack of confidence when a model has not been trained on similar samples. Uncertainty methods for individual predictions include prediction intervals, conformal sets, Monte Carlo dropout, ensembling, and Bayesian methods (Kompa, Snoek, & Beam, 2021). *NLP examples:* (1) If model predicts high risk with high uncertainty, do not deploy planned intervention and have expert inspect. (2) Generative AI models could provide confidence estimates and evidence that supports their opinions or inferences.

Standardization. Method by which a score's test is interpreted by comparing a score to a representative sample (populational norms) on standardized scores (e.g., stanines, T scores; Rust et al. (2020)). Lack of replication across studies can occur when different measurements of the same construct are done differently due to either different unstandardized research design or operationalization. *NLP examples:* (1) Leaderboards are ways of standardizing models' criterion validity, public tokenizers can help standardize preprocessing. (2) Standards for feature extraction could be developed to rule out results across studies are not different due to differences in pipeline implementations or choices in parameters.

Unidimensionality. Whether the target construct to be measured is homogenous, that is, indicators measure the same thing (e.g., a specific type of insomnia) or whether it is heterogeneous and thus measures multiple things (e.g., depression, Fried et al. (2022)). The amount of dimensions can be assessed through exploratory factor analysis (EFA), categorical principal component analysis loading plots, and item response theory (Mair et al., 2018). If items or more specific constructs have strong associations to certain latent variables and not to others, then each latent variable can be seen as a homogeneous construct. These could be grouped into a multidimensional scale (e.g., personality, job performance). A network approach does not need to assume latent variables and can assess dimensionality through community detection of observed variables. *NLP example:* The dimensionality of a lexicon can be assessed by encoding the tokens with embeddings and clustering, where the amount of clusters (i.e., construct dimensions) would have the highest average silhouette score. Embeddings could also be used in factor analysis. Features of a trained model could also be evaluated similarly.

Reliability. The consistency of a measure depends on the condition under which it is evaluated [7]:
(a) **Parallel-forms reliability and intra-method reliability**. Consistency of responses of a given individual to two versions of the same questionnaire that use different items (Van der Wal et al., 2022). *NLP example:* If we consider trained models as individuals, slightly changing prompts to a generative AI LLM can result in very different performance (Sanh et al., 2021; Shu et al., 2023b; Van der Wal et al., 2022). Testing sensitivity of a model to reinitializing random weights or parameters (Hoyle, Goel, Sarkar, & Resnik, 2022; Van der Wal et al., 2022).
(b) **Test-retest (across time)**. Measuring same individuals using same test across time points. When this is done by a same rater over time, it is called intra-rater reliability. As with prospective validity, this will depend on whether construct is expected to change (i.e., trait) or remain stable over time (i.e., state). *NLP example:* An NLP model

[7] For an extensive overview of reliability in NLP, see Riezler and Hagmann (2022).

detecting paranoia from text had good test-retest reliability over the course of several days using intra-class correlation (Cohen et al., 2022).

(c) **Internal consistency (across items)**. Degree to which items measure the same thing (Cronbach, 1951). The most widely-used Cronbach's alpha is not recommended as it tends to underestimate reliability due to many unrealistic assumptions, such as assuming each item on a scale contributes equally to the total score (tau equivalence) (McNeish, 2018). A measurement may have high internal consistency but not be unidimensional (a more stringent concept) if items are associated with multiple latent variables (see unidimensionality section above). *NLP example:* The Suicide Risk Lexicon (Low et al., 2024) as a whole covers a diverse set of risk factors, but each of its 49 constructs could be evaluated for internal consistency by computing the average cosine similarity of each token's embedding with a target label embedding representing the construct.

(d) **Inter-rater (across raters)**. Consistency across different raters' independent judgements as measured through absolute percentage agreement, Cohen's Kappa, intra-class correlation (ICC) and other methods (Mair et al., 2018). *NLP examples:* (1) Annotators judging whether to include tokens in a lexicon (Low et al., 2024). (2) Two distinct topic models could be both correct because of the subjective nature of categorization, topic models could be considered raters and reliability could be assessed through Rand Index (Hoyle et al., 2022).

Which of these properties should be evaluated will depend on the goals of the study. Ideally, a measurement is demonstrated to be valid when a measurement outcome (e.g., change in thermometer) is caused by an intervention on the target attribute (e.g., temperature) (Borsboom et al., 2004). This causal effect can be estimated through randomized controlled trials or causal inference methods for observational data (Hernán & Robins, 2020). This causal perspective on validity could discard many types of so-called validity approaches listed here as necessary to prove validity; in particular, construct validity (i.e., the degree to which a test correlates with what it should and does not correlate with what it should not) could be discarded Borsboom et al. (2009). Some of the properties we present may only demonstrate some expected property of a valid assessment, but not prove validity themselves (e.g., construct validity). An NLP example of this type of causal validity could be demonstrated by evaluating the validity of an NLP model that measures state anxiety from text data: participants could be randomly assigned to a relaxation- or anxiety-inducing task and asked immediately after to write about their affect –a valid score from this NLP model would be lower after the relaxation task than the anxiety task. [8] Having said this, we expect researchers would want to evaluate most of these different properties depending on their application, but might only evaluate some of them (e.g., standardization) after more basic evaluation (criterion validity) is assessed. Our goal is to describe most properties that be of interest, especially for NLP and psychology researchers who are less familiar with psychometrics or psychometricians who are less familiar with NLP examples.

The following two studies we performed demonstrate trade-offs of selecting a given NLP method as a function of the dataset, goal, and psychometric property that should be fulfilled. The studies try to assess variables related to psychological crises and mental health, tasks that highlight the need for models that are valid, explainable, and keep sensitive data private. In study 1, we compare predictive performance of a wide variety of NLP methods. Furthermore, we demonstrate why it is important to evaluate content validity, and we provide a novel way of doing so: evaluating whether a model trained to detect a construct can correctly classify highly prototypical and obvious words and phrases of said contruct, as defined by experts. In study 2, we showcase a common scenario of class imbalance where the outcome of interest is rare. We predict the target variable –which is in the future– using increasingly larger segments of the crisis counseling conversation (i.e., prospective validity). In this high-stakes scenario, we use an interpretable method to avoid unknown bias and to understand which symptoms most predict future risk.

5 Study 1: assessing multiple types of mental health issues to evaluate concurrent criterion validity and content validity

Here, we present a use case to demonstrate how to train and validate these different types of NLP methods to detect a diverse set of psychological constructs in two datasets: (a) issues discussed during de-identified Crisis Text Line conversations, as annotated by the crisis counselor immediately following the conversation ; and (b) similar types of issues from Reddit (described in section 5.1). The semantic variety of issues –from suicide to bullying to gender/sexual identity– should allow us to evaluate the sensitivity of the models to these different types of constructs. Demonstrating that an NLP model can detect what counselors coded as the conversation happened is a form of concurrent criterion validity. However, our main goal is to demonstrate that while being able to predict a human-annotated label is key, high-performing models may or may not be capturing a representative set of expressions of the target construct (i.e., low content validity). To assess this, we introduce a novel way of evaluating content validity (see section 5.4).

[8] For further discussion on validity and measurement, see Appendix C.

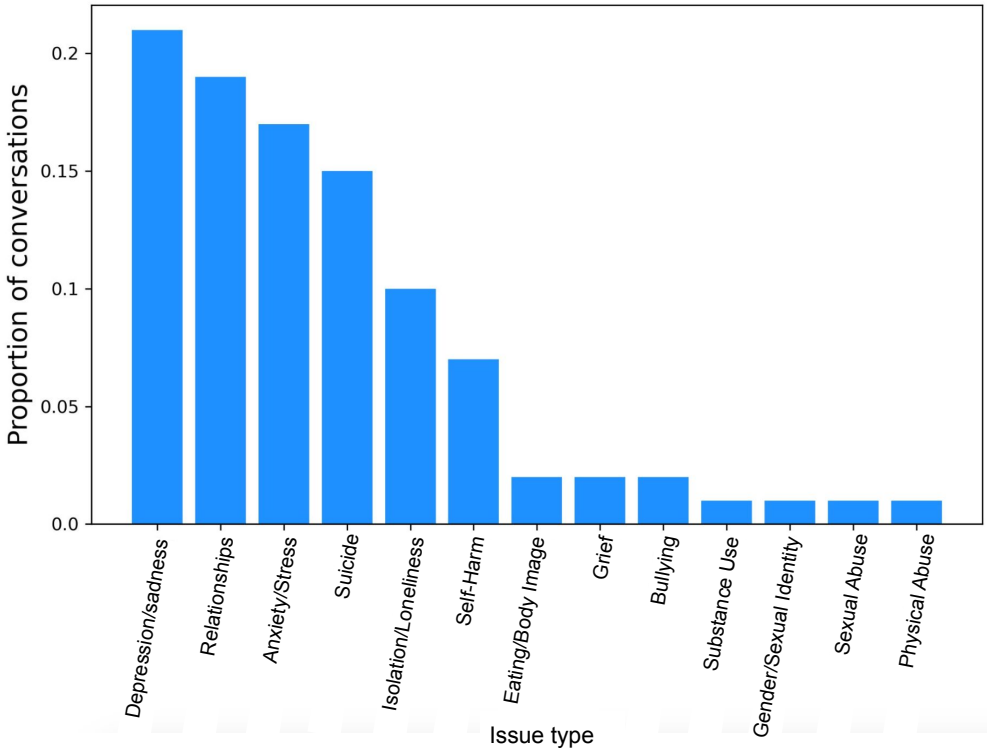

Figure 4: Distribution of types of issues

5.1 Datasets

Crisis Text Line. Crisis Text Line is a nonprofit organization that provides free, 24/7, confidential mental health support for individuals in crisis who can begin a conversation with a trained, volunteer counselor via text messaging, WhatsApp, or webchat. Crisis Text Line users have been reported to be predominantly young (76% under the age of 25), female (79%), white (53%) and heterosexual (52%) (Pisani et al., 2022). We built several different datasets using Crisis Text Line data, containing de-identified crisis counseling conversations between individuals in crisis and counselors from the years 2017 to 2022. Years 2020 and 2021 were omitted to avoid symptoms and concerns that were prevalent during the Covid-19 pandemic. In Figure 4 we show the distribution of the 13 most common types of issues as labelled by the counselors. We built balanced binary datasets (training set N=4000, test set N = 600) where only one of these issues was present vs. a random selection of the remaining types of issues from a total of 52,270 unique conversations. Binary models were used because multi-class classification could change radically depending on what constructs are chosen and we want to reflect a psychometric tests that results in a single score. The sample size was chosen to test how increases in training set affects performance and we consider 4000 samples would be among the larger samples in clinical studies.

Reddit. Reddit is a widely-used forum, used by 22%, of U.S. adults (Gottfried, 2024). It is organized into subreddits, many of which are lightly-moderated spaces for seeking and providing mental health support. Using the Python praw package (v7.8.1), we downloaded posts from 11 subreddits (4000 for training and 600 for testing for each subreddit), reflecting most of the Crisis Text Line issues (i.e., Depression, Anxiety, SuicideWatch, Lonely, SelfHarm, EatingDisorders, GriefSupport, Bullying, Addiction, AskLGBT, SexualAssault). Since these are public data, we can evaluate the performance of proprietary state-of-the-art LLMs in their ability to identify the subreddit of each post. Posts were from 2025, after each model's training data cutoff date, so these samples were not in the model's training data. This also allows us to evaluate how much our results change across two datasets, keeping constructs relatively constant. This dataset differs from Crisis Text Line in being a monologue (looking at just the post content), not a dialogue, and coming from a different population: while Reddit users also tend to be young, they are characterized by being less young (44% under 29 years old), mostly male, and with more Asian and Hispanic than White users (Gottfried, 2024).

5.2 Models

We include interpretable methods (lexicons, similarity-based text classification) given psychologists commonly use models for high-stakes decisions or to advance the understanding of what predicts a given construct. And we also include zero-shot methods (construct-text similarity and generative AI) given the previously discussed challenges of obtaining computational resources and the moving target of constructs that might need to be assessed.

LIWC-22 lexicon: LIWC is a lexicon that counts the proportion of tokens related to 117 categories including semantic constructs (family, anger) and nonsemantic linguistic features (1st person pronouns, unique words).

Suicide Risk Lexicon (SRL): this lexicon contains tokens (words and phrases) associated with 49 different risk factors for suicidal thoughts and behaviors (STBs) including direct-self injury, hopelessness, and discrimination (Low et al., 2024). It has been shown to surpass LIWC and offers higher content validity (i.e., more related constructs) than LIWC for language around psychological issues and imminent risk. Initial tokens were built using GPT-4 turbo and later validated by clinicians. Counts are normalized by document word count.

LIWC-22 and SRL were trained using logistic regression. Hyperparameter tuning was done using Bayesian optimization using the *scikit-optimize* package on a 5-fold cross-validation of the training set. Hyperparameter tuning of the Ridge alpha parameter included values [0.001, 0.01, 0.1, 1, 10, 100].

Construct-Text similarity (CTS): We apply this method, as described in section 3.1, twice: once by building the construct representation using only one prototypical token and again by using all prototypical tokens from the suicide risk lexicon as construct's tokens; prototypical tokens are those that scored 3 out of 3 by all clinicians in how much they represent a given construct (e.g., "anorexia" is very prototypical of eating disorders while "didn't eat much" is less so). We tokenize the document into independent clauses (with subject and predicate) to capture more minimal units of meaning than sentences separated by punctuation. We compute the maximum similarity to capture the document token with the highest similarity to any prototypical lexicon token. We then min-max scale the output of the cosine similarities between 0 and 1 to match the scale of other models.

Open-source LLM Gemma instruct models: Gemma 7B is an open-source generative LLM that was high-performing on multiple tasks involving reasoning, question-answering, math and science, and coding at the time of analysis (Team et al., 2024). Gemma 2B is a smaller version that is expected to underperform but can run efficiently on a personal CPU unlike its larger counterpart. These models were downloaded onto our private and secure computer cluster. We provided the prompt displayed in Appendix Figure 9. We did inference using Gemma 2B-it on a single NVIDIA GeForceGTX1080Ti GPU for 10 hours and Gemma 7B-it on a single QUADRORTX6000 for 16 hours.

Proprietary LLM Google Gemini 2.0 flash & OpenAI GPT-4o models. These models are among the highest performing models at the time of analysis [9]. Therefore, we used these for public Reddit data but not Crisis Text Line data. They also allow for a prompt that includes the entire codebook which reduces API requests considerably (see Figure 10). We randomized the order of the constructs in the prompt for every submission to avoid an ordering bias. To encourage the model to select the most highly predicted tokens, the temperature —a parameter controlling output randomness or creativity that typically ranges between 0 to 1— was set to a low value of 0.1. If the output of LLMs did not contain a parseable JSON (approximately 3% of cases), random output scores (i.e., between 0 and 1) were generated.

5.3 Concurrent criterion validity: predictive performance on binary classification task of 13 types of issues

In Table 2 we show the average and range of the performance of the different binary text classification models across the different types of issues in two datasets, summarized using ROC AUC, the probability that a randomly chosen positive instance will be ranked higher than a randomly chosen negative instance by the classifier (0.5 = random; 1 = perfect prediction). In both datasets, the SRL outperforms LIWC, which is expected given it actually captures more relevant constructs to mental illness (i.e., as higher content validity at face value). In the Crisis Text Line dataset, the CTS model —using just a single variable in a zero-shot fashion– surpasses LIWC and the Suicide Risk lexicon trained on 300 documents and matches LIWC trained on 4000 samples. In the Reddit dataset, the CTS models performed relatively well: they only marginally underperformed the lexicon models but used only one variable and did not require any training (i.e., zero-shot). The LLMs require no additional training data from the target dataset (e.g., Crisis Text Line, Reddit) and improve their performance likely as a function of parameter size and recent innovations in training which allow for larger training datasets (while model and training specifications are proprietary, GPT4o and Gemini 2.0 flash are likely larger than gemma 2–7b models; Minaee et al. (2024)).

[9] `https://huggingface.co/spaces/lmarena-ai/chatbot-arena-leaderboard`

Table 2: Performance of different types of text classification methods. Number of variables are shown in parentheses. Zero-shot models require zero additional training data and are therefore displayed separately. The models that performed best on the test set (highest ROC AUC) and on the content validity test sets (highest sensitivity of prototypical tokens) are highlighted in bold. Only public Reddit data was submitted to proprietary APIs (Google, OpenAI). SRL: Suicide Risk Lexicon; CTS: Construct-Text Similarity; single: single prototypical token; multi: multiple prototypical tokens.

		Crisis Text Line		Reddit	
Models (N variables)	**Train set**	**ROC AUC Mean [min-max]**	**Content validity Mean [min-max]**	**ROC AUC Mean [min-max]**	**Content validity Mean [min-max]**
LIWC-22 (117)	100	0.69 [0.58-0.77]	0.80 [0.38-1.00]	0.80 [0.73-0.89]	0.86 [0.56-1.00]
LIWC-22 (117)	300	0.75 [0.66-0.84]	0.74 [0.21-1.00]	0.84 [0.78-0.90]	0.82 [0.62-1.00]
LIWC-22 (117)	4000	0.82 [0.73-0.89]	0.77 [0.40-0.96]	0.90 [0.85-0.94]	0.88 [0.70-1.00]
SRL (50)	100	0.74 [0.66-0.79]	**1.00** [0.98-1.00]	0.84 [0.73-0.95]	**0.98** [0.86-1.00]
SRL (50)	300	0.80 [0.69-0.87]	**0.97** [0.67-1.00]	0.88 [0.81-0.94]	**0.96** [0.78-1.00]
SRL (50)	4000	**0.86** [0.73-0.95]	**0.99** [0.91-1.00]	**0.93** [0.86-0.98]	**0.98** [0.78-1.00]
CTS single (1)	0	**0.82** [0.59-0.95]	0.51 [0.16-0.79]	0.79 [0.65-0.89]	0.43 [0.13-0.69]
CTS multi (1)	0	**0.82** [0.71-0.92]	**1.00** [1.00-1.00]	0.83 [0.71-0.92]	0.84 [0.53-0.97]
gemma-2b-it	0	0.61 [0.41-0.79]	0.92 [0.40-1.00]	–	–
gemma-7b-it	0	**0.82** [0.62-0.90]	0.91 [0.50-1.00]	–	–
Gemini 2.0 flash	0	–	–	**0.91** [0.83-0.95]	**0.91** [0.78-1.00]
GPT-4o	0	–	–	**0.91** [0.83-0.96]	**0.94** [0.85-1.00]

5.4 Evaluating content validity with content validity test sets

To evaluate content validity quantitatively, we introduce content validity test sets: lists of tokens judged by experts to be highly prototypical of the construct of interest. Prototypical tokens are those that first come to mind when thinking of a construct ("I'm a burden" or "I'm a nuisance" for burdensomeness) –given a specific definition– and have fewest features in common with other constructs (Low et al., 2024). Prototypical tokens should not lead to many false positives where the the person is not expressing the construct (e.g., "I'm anxious" for burdensomeness). While these tokens are likely to appear in a large enough test set, their relative frequency will be different, as depicted in Figure 5. A random test set will evaluate a model's ability to generalize to real, unseen data; while a content validity test set will answer the question: how many of the obviously prototypical tokens can the model detect? These lists can be obtained from publicly available lists made by experts (e.g. symptoms in DSM-5) or more rigorously by creating a lexicon and validating it with clinicians, as done with the Suicide Risk Lexicon (SRL; Low et al. (2024)) [10]. We would expect high-performing models to capture all prototypical tokens, especially those scored 3 out of 3 in clinician evaluations of prototypicality as we test here. Therefore, the metric of choice is sensitivity: out of the prototypical tokens how many can the model correctly predict? While traditional test sets include non-obvious positive cases making a perfect prediction of 1 very difficult, we expect all models to have a perfect sensitivity close to 1 in detecting obvious tokens, especially those built using SRL (SRL and CTS). These content validity test sets provide a model-agnostic method of evaluating content validity, which can also serve as an explainability tool (i.e., what information does the model capture?). Fang and cols. (2022) Fang et al. (2022) evaluated content validity of text embeddings through probing experiments: text is encoded with a given embeddings model x (e.g., BERT) and with an interpretable feature f (e.g., sentence length); the embedding is assumed to contain the interpretable information if it can predict it, a common approach in explainable machine learning. However, this does not answer whether a model captures a representative set of prototypical expressions, which may be closer to the definition of content validity, and we try to demonstrate through our complementary approach, which unlike probing is dataset-independent and works for evaluating models and not just embeddings.

In Table 2 we show the average and range of the performance of the same binary text classification models as section 5.3 but now tested on lists of clinically-validated tokens that are highest on prototypicality (i.e., content validity test sets), computing the sensitivity for these tokens (i.e., content validity). Any misses are considered an obvious mistake. As expected, SRL-based models (SRL and CTS using prototypical SRL tokens) are able to capture the protoytipical tokens except CTS (1 prototype) which has low performance. Evaluating the model just on out-of-sample prediction using metrics such as ROC AUC or F1 score (i.e., criterion validity) miss this lack of content validity. Surprisingly, gemma models perform extremely well on average, although much worse for some constructs. Unsurprisingly, the more recent and larger Gemini 2.0 flash and GPT-4o models peform very well on average.

[10] We specifically used SRL tokens judged 3/3 by experts (srl_prototypes_v1-0) from construct-tracker package. For a tutorial, see: `https://github.com/danielmlow/construct-tracker/blob/main/tutorials/suicide_risk_lexicon.ipynb`

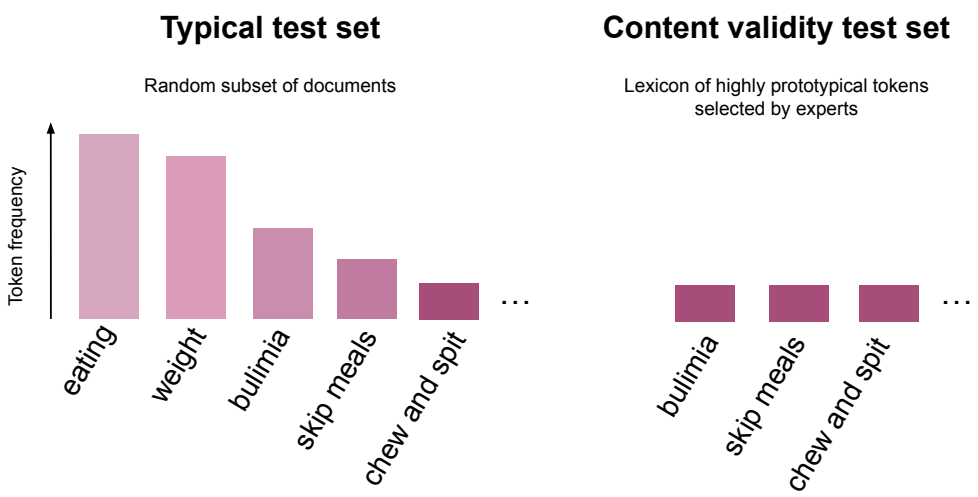

Figure 5: Typical test sets influence models to use tokens to improve prediction to the extent they are associate with the target construct. They test generalizability to unseen data. Certain infrequent tokens could be ignored and still achieve high performance. On the other hand, content validity test sets are lexicons of prototypical words and phrases related to the construct as validated by experts, where all tokens have equal weight to evaluate whether a model detects a good coverage of expressions of a construct. Prototypical tokens are expected to be both sensitive and specific by experts (i.e., "eating" would create many false positives). Individual tokens should be detected as positive by a model that claims to detect the target construct (i.e., high sensitivity of these tokens). This can also be evaluated by placing tokens in multiple sentences each, generated using LLMs. We use the most prototypical tokens (3/3) from relevant constructs of the Suicide Risk Lexicon (Low et al., 2024).

6 Study 2: Predicting whether to deploy an emergency service intervention for imminent suicide risk to evaluate prospective criterion validity under class imbalance

The goal of this second use case is to present a more challenging NLP task: predicting a low-base rate phenomenon: whether the outcome of the conversation resulted in requesting an emergency services intervention, resulting in an imbalanced dataset. Furthermore, we will predict this outcome using earlier parts of the conversation, which is a form of prospective criterion validity. We show how different classification metrics improve as more messages are included. We finish by providing an algorithmic explanation (i.e., feature importance), that is, revealing which risk factors are most predictive of the outcome which can help inform theory.

6.1 Dataset

During Crisis Text Line conversations analyzed, counselors assessed for suicidal thoughts, suicidal plans, access to means, and imminent 48-hour timeframe, in that order (see Low et al. (2024) for more details). If a texter is determined to be at a high risk by the crisis counselor, the counselor will involve a supervisor. The supervisor evaluates whether the texter is in immediate danger and unable to create a safety plan or deescalate, in which case the supervisor may contact local emergency services. To better dissociate the likely highest risk group in the dataset, we created two groups: (a) suicidal desire without plans, means, or timeframe, and (b) cases where an emergency service intervention was initiated, which has all symptoms present and cannot be fully deescalated. Such models could speed up risk assessments (e.g., alert the supervisor earlier), give a second opinion to the counselor and supervisor, or, after appropriate validation, be used in a different setting where there is no counselor (e.g., social media). We balanced the training set by downsampling suicidal desire cases to the imminent risk sample size to avoid the model over-predicting values closer to the majority value, resulting in a training set of 3500 (half in each group). The test set included a total of 2358 cases, with 2152 suicidal desire cases without plans, means, or timeframe cases (91% of test set) and 206 emergency services intervention cases (9% of test set).

6.2 Model

For this task, we wanted an explainable model to improve trustworthiness and so we could perform feature importance analysis to understand which risk factors are associated with imminent suicide risk needing an emergency service intervention. We chose CTS with all prototypical tokens given its high performance in the prior task in capturing constructs. We used all constructs from the Suicide Risk Lexicon as independent variables and trained a LGBMClassifier model. Hyperparameters [and values] tested for the LGBMRegressor were: num_leaves [30,45,60], colsample_bytree [0.1, 0.5, 1], max_depth: [0,5,15], min_child_weight: [0.01, 0.001, 0.0001], min_child_samples: [10, 20,40].

6.3 Prospective criterion validity: predicting whether to deploy an emergency service intervention for imminent suicide risk

In Figure 6 we report predicting the outcome of the conversation resulting in an emergency service intervention as the conversation advances (computed on the quartiles of the number of messages across all participants of the training set: 25% = 28, 50% = 43, 75% = 64, 100 % varies by conversation). Similar performance to using the entire conversation can be achieved with the first $3/4$ of the conversation, which could help speed up potential alert systems that could be designed in the future, if the number of messages or duration of conversation could also be reliably predicted, to help supervisors make a decision. Whether this model is clinically useful (one of the psychometric properties we described), that is, whether these classification metrics are sufficiently high for a model to be deployed –in particular, models with a low precision, which is common in low-base rate phenomena such as suicide risk assessments (Kessler et al., 2020)– could be evaluated quantitatively with a net benefit analysis, which should include weighing potential benefits versus risks of harm of emergency services interventions, which may vary across populations (Kennedy et al., 2021; Kessler et al., 2020; Vickers et al., 2016).

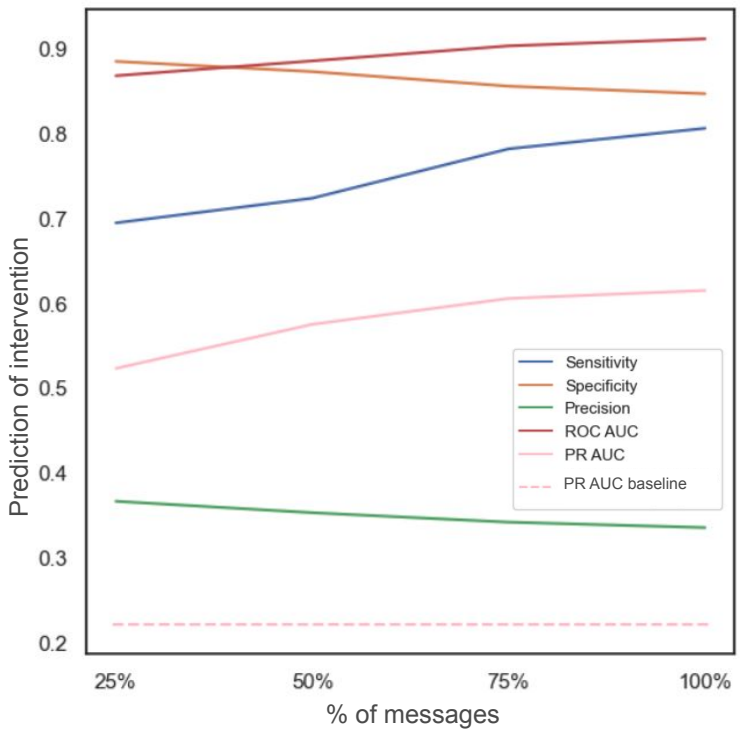

Figure 6: Predicting a request for an emergency service intervention (decided towards the end of the conversation) as the conversation advances

6.4 Feature importance

Understanding which variables are associated with suicide risk from a language perspective can help inform the theory of suicide risk: what risk factors should counselors pay particular attention to? To compute feature importance with the LGBMRegressor model, we calculated the improvement in the information gain (i.e., the splitting criterion) from using a given feature in a tree's split (i.e., how much a given feature helps to reduce the error). Other methods include SHAP

(Mosca, Szigeti, Tragianni, Gallagher, & Groh, 2022), leave-one-feature out (especially when there is colinearity)[11], and permutation feature importance.

In Table 3 we show the top 20 features for predicting imminent risk with emergency service intervention. Mentions of lethal means for suicide (e.g., cutting), Other substance use & overdose ("overdose", "pills"), and having the texter write more than the counselor were the most predictive features for detecting the need for an emergency service intervention.

Table 3: This interpretable method with high content validity allows us to report which known suicide risk factors from the literature are most predictive of an emergency service intervention scenario. Importance: feature importance, which is operationalized as the improvement in the information gain from using a given feature in a tree's split.

Feature	Importance
Lethal means for suicide	5106.62
Other substance use and overdose	2964.54
Texter-counselor word count ratio	1951.01
Police & incarceration	830.73
amount of messages counselor	688.64
Depressed mood	624.06
Anxiety	487.35
Anhedonia & uninterested	422.76
Panic	406.98
Borderline Personality Disorder	378.76
Impulsivity	341.36
amount of messages texter	341.35
Passive suicidal ideation	339.31
Rumination	296.11
Bipolar Disorder	295.19
Agitation	258.65
Entrapment & desire to escape	250.47
Active suicidal ideation & suicidal planning	235.09
Other suicidal language	190.90
Hopelessness	179.73

7 Discussion

We have introduced Text Psychometrics as an important area of research to improve the validity and reliability of NLP methods being developed for psychological assessments in text data. We reviewed how to validate models that use NLP to assess psychological constructs in text through examples and through two original studies. In particular, we have provided a novel method for evaluating content validity using lexicons and expanded a similarity-based text classification method which we call construct-text similarity which is interpretable and lightweight. Next we help guide researchers in choosing the optimal NLP method given their research and dataset constraints.

7.1 How to choose a natural language processing method

In Figure 7 we summarize the tradeoffs of different NLP methods as a function of different, key desirable properties. This is a qualitative analysis based on our results and prior studies. The first point to make is that using LIWC or an existing lexicon is rarely a good choice given it generally has lower predictive performance compared to other methods. We also argued it lacks content validity at face value since it does not generally capture the main target constructs of interest in any given study (Low et al., 2024); furthermore, when used to detect a given construct, the resulting model misses many prototypical tokens of the construct as shown with the content validity test sets (Table 2). An alternative is building one's own lexicon, and one can have an LLMs generate a draft lexicon as a starting point (Low et al., 2024). This is likely to achieve improved performance (criterion validity) and higher content validity. Custom lexicons can be used directly without training a model (zero-shot) in the sense that one can count whether their tokens appear in a document without training a model, although it is unlikely to perform very well. It is better to train a model to learn how much to weigh these counts to detect the dependent variable. The main limitaitons with lexicons is they cannot match tokens that are similar to the ones they have ("suicide attempt" will not match "attempted suicide") and they lack awareness of how a word's context might be associated with the target variable ('not sad', 'very sad', 'sort of sad'),

[11] `https://github.com/aerdem4/lofo-importance`

which explains their limited performance. Fine-tuning a deep learning model (e.g., BERT) overcomes this issue of context, which is one of the reasons why it tends to outperform lexicons, but requires substantially more training data and are black-boxes (Low et al., 2024). An intermediate approach between lexicons and fine-tuning a deep learning model is to use Construct-Text similarity, which is zero-shot, matches similar tokens, has high content validity, and is interpretable. However, it is also insensitive to context which will limit its predictive performance.

Recent LLMs perform well zero-shot, can be fine-tuned for improved performance, and can be requested to provide self-explanations of why they generated a certain result, which may or may not be truthful (Agarwal et al., 2024). They are sensitive to extremely large prompts (i.e., context windows) and can capture attempts at avoiding keyword detection (e.g., “s3lf harm3d yesterday" would be likely flagged as self harm). Open-source versions currently require substantial computational resources (GPUs, technical abilities) and tend to perform worse than proprietary versions Chiang et al. (2024); however, a proprietary version requires users to submit their data to APIs, which may not be secure or permitted for sensitive data until cloud-based systems become more secure or are trusted. It is also challenging to make generative LLMs output deterministic results, which may create unforeseen errors and is particularly problematic in high-stakes scenarios. Other approaches (BERT-type models) can be used in these settings for deterministic results as an alternative or used alongside generative LLMs.

Overall, lexicons or construct-text similarity might be prioritized over other deep learning approaches when (a) an interpretable is preferred for high-stakes decisions or scientific inquiry (perhaps used in tandem with LLMs to guarantee capturing tokens an LLM might miss); and (b) when models that might result in higher predictive performance are challenging to train or use due to lack of labeled data, computational resources, or data-privacy requirements that forbid sharing data with a LLM API or cost (e.g., running LLMs for millions of electronic health records). LLMs are already dominating other methods given their performance is extremely high. LLMs will gain even more traction once their size decrease in order to run more easily on personal computers and their self-explanations are deemed not only plausible, but also faithful (i.e., true) (Agarwal et al., 2024). More mechanistic interpretability methods and understanding for LLMs are also being developed (Lieberum et al., 2024; Lindsey et al., 2025).

Method \ **Desirable property**	Zero-shot (no training data)	Semantically similar matching	Sensitive to context	Performance (Criterion validity: predict human annotation)	Explainable	Content validity	Data Privacy (run locally)
Existing lexicon + model (LIWC)	—	✗	✗	—	✔	✗	✔
Your own lexicon + model	—	✗	✗	—	✔	✔	✔
Fine-tuning a deep learning model (BERT, RoBERTa)	✗	✔	—	—	—	✔	✔
Construct-Text Similarity	✔	✔	✗	—	✔	✔	✔
Small open-source LLMs (Gemma 7B, Llama 8B)	✔	✔	—	—	—	✔	✔
SOTA LLMs via APIs (GPT-4.1, Gemini 2.5)	✔	✔	✔	✔	—	✔	—

✔ Good
— Moderate
— Poor
✗ Bad

Figure 7: Desirable properties and how different types of NLP methods compare. LLM: large language model. SOTA: state-of-the-art.

7.2 Are text-based measurements objective if they predict a subjective criterion? Moving beyond the convergence with rating scales.

It is commonly thought that a biomarker (digital speech pattern, neural circuit, genetic marker) or behavioral marker (language sentiment) are objective because they can measure a variables automatically, that is, they are more objective than any one human’s subjective opinion. However, can these markers be considered objective if they are predicting subjective variables such as a self-report questionnaire? A first consideration is that the target labels may come not from self-report but from the opinion of experts who provided the labels, who may use more objective criteria than nonexperts. The label may be a gold-standard that is relatively objective, such as a diagnosis of vocal-fold paralysis through video endoscopy Low, Rao, Randolph, Song, and Ghosh (2023). A model that accurately predicts expert judgements could be more objective than most non-expert humans. Second, if a model is biased or unfair, such as one

that unjustly uses proxies for race to grant or deny parole to incarcerated individuals (Mayson, 2018), it could in theory be improved or corrected through bias mitigation methods (Hort et al., 2023). This model could end up being more objective than the average judge. Therefore, by increasing the amount of training datasets and correcting over time, a model might become more objective than the humans who provided annotations in any one of its training datasets. Self-supervised models such as GPT models achieve this even without an external criterion: it uses next-word prediction of the own training data on vast amounts of diverse corpora to allow models to gain rich representations useful in having better predictions than many humans in a variety of downstream tasks, especially when combined with reinforcement learning with human feedback (Bai et al., 2022).

Nevertheless, given text-based measures often predict subjective target variables, it might be important to evaluate the validity of a text-based measurement beyond assessing its convergence to a subjective rating scale (Kjell et al., 2023). Several studies have compared text models to rating scales without having to train the text model to optimize the prediction of a self-report rating scale variable. One study (Kjell, Kjell, Garcia, & Sikström, 2019) took the pictures of facial expressions of different emotions done by actors (happy, sad, or contemptuous, taken as the ground-truth) and randomly assigned participants to either describe each expression with words or rate each stimulus using rating scales for the corresponding expression (e.g., 1 = low to 5 = high happiness). Then both types of responses (i.e., text versus rating scale) were used to predict the ground-truth emotion category and the text-based model showed higher performance. Therefore, these types of analyses allow text-based models to independently be compared to rating scales using criterion validity of a ground truth. Another study (Sikström, Nicolai, Ahrendt, Nevanlinna, & Stille, 2024) first had one set of individuals write about a series of internal states (depression, anxiety); these narratives are safely assumed to be about those states (ground-truth). Then they had a second set of individuals use relevant rating scales (PHQ-9, GAD-7) to assign scores to those narratives. Then models were trained to predict the states from either the features extracted from the narratives or from the rating-scale scores and the text models outperformed the rating scales in detecting emotional states. However, going forward, comparing the utility of natural language and rating scales should be based on multiple desirable psychometric properties, consider the trade-offs in the time it takes to produce scores with rating scales and narratives, and how well a researcher can measure a specific construct of interest using both methods; for instance, perhaps heterogeneous constructs such as depression could be better assessed through natural language in order to capture nuanced and diverse symptoms, but a more specific construct like hopelessness is better assessed through a quick ordinal rating scales. Because if the person does not mention hopelessness in their response, this does not mean they are not experiencing it, which could quickly be assessed with a rating scale.

7.3 Ethical considerations of assessing psychological constructs in text

As NLP-based assessments improve, it might be possible for users to assess individuals without their consent. This could lead to spreading private information or rejecting individuals during jobs or insurance applications based on model predictions. European legislation has proposed individuals have a right to an explanation on high-stakes algorithmic decisions, which would make these malicious practices illegal (European Commission, 2023). As models become better, users may not doubt models' outputs, which is especially problematic if we trust AI to evaluate AI (section 7.5.3). If measurement is done via AI chatbots then all the known risks of generative AI hold: hallucinations; suicide encouragement; non-evidence-based medical recommendations; disinformation; illegal, toxic, and discriminatory content; leakage of personal information or copyrighted material from the training set; and plagiarism (Weidinger et al., 2021). Standardized AI safety benchmarks are being used to assess LLMS (Liang et al., 2022). Specific questions could be posed to a model to determine whether it will output harmful content such as suicide methods. Furthermore, training, fine-tuning, and particularly running inference with LLMs can produce a substantial impact on the environment, and reducing this impact is an active area of research (Chien et al., 2023).

7.4 Limitations

LLMs are known to suffer from small changes in the prompts, which can be evaluated more thoroughly (e.g., see Shu et al. (2023b), who refer to this as a lack of consistency), which is similar to parallel forms validity or intra-method validity in psychometrics (Van der Wal et al., 2022).

We evaluated different types of models in predicting 13 types of issues. A comparison of these text classification methods (and potentially other methods) on multiple datasets from different domains (e.g., clinical, nonclinical, political) would be useful. While we cannot assume the results will hold exactly in other datasets, we do hypothesize a similar trend in other datasets given the discussed advantages and limitations of each approach (e.g., amount of training data needed, how content validity reflects capturing relevant constructs, the lack of context-sensitivity in some methods, reasoning abilities and general-domain knowledge of newer generative AI models).

7.5 Future directions: exploring other models, other NLP tasks, and AI evaluating AI

In this section, we discuss possible extensions to the field more broadly, beyond the specific use case of crisis counseling discussed above. These do not reflect any plans or opinions of Crisis Text Line.

7.5.1 Other psychometric methods

In Appendix B we briefly reviewed factor analysis, item response theory, networks, and causal inference. Other multivariate modeling has been approached in psychometrics using MANOVA (independent regression models) and a series of path analyses such as structural equation models (which integrate confirmatory factor analysis into a path analytic framework), multidimensional scaling (Mair et al., 2018), and moderator-mediator models (Mair et al., 2018) that try to provide more insight into associations between variables. Future work could further explore how to combine these methods with text data; for instance, using embeddings instead of rating-scale responses for factor analysis.

7.5.2 Other large language model methods.

Multiple methods could be applied to improve performance including few-shot or exemplar prompting (i.e., including multiple examples in the prompt with their correct labels), chain-of-thought prompting (i.e., instructing the model to provide a step-by-step explanation in the prompt before providing the answer), ensemble refinement (e.g., include multiple chain-of-thought responses through temperature sampling in the prompt for the final answer), sampling different hyperparameters (temperature, top-k, and nucleus), and fine-tuning using efficient methods such as QLoRA (Dettmers et al., 2024; Singhal et al., 2023).

7.5.3 Other NLP tasks.

We focused on text classification, but there are many other NLP tasks that could also benefit from rigorous psychometrics. We briefly outline additional tasks and how they are currently validated. Many of these tasks and metrics can be evaluated before and after deployment.

NER: classification metrics can be used as well as more nuanced interpretability methods (Fu, Liu, & Neubig, 2020).

Summarization and question-answering. Summarization has many applications such as creating artificial memories in generative AI chatbots so it can access information about a user or provide summaries of electronic health records to clinicians before they see a patient. Question-answering also returns a text which needs to be validated to some ground truth. Validation of these tasks involves comparing the similarity to a ground-truth (criterion validity), although recent approach are reference-free (Scialom et al., 2021).

RAG (Retrieval-Augmented Generation) is an approach to improve generative AI giving them access to external databases (Gao et al., 2023). Even though LLMs are originally trained on large corpora, they will likely not have information that is proprietary or novel and, therefore, can benefit from searching through new data using semantic textual similarity and other techniques. There are many metrics and benchmarks that evaluate different properties of the retrieved information (Gao et al., 2023). Applying item-response theory (IRT) could help evaluate whether the system adapts appropriately to varying levels of question difficulty or ambiguity.

Chatbots. Evaluating the quality of a chatbot will require similar metrics as we have used, but also ethical considerations as well as metrics related to user engagement and conversation length. Fully automated chatbots are likely pose serious risks in high-stakes scenarios such as crisis counseling. A desirable property for a chatbot could be long-term memory of the interactions with the user, which may help the system retrieve details of the user over time and over longer documents. Another may be to have a reliable personality (Shu et al., 2023b).

Agentic AI. AI agents can take human input via prompts in order to interact with a device and take actions. Examples include virtual assitants that can set an alarm; AI research assistants that can search for articles, analyze data, and draft manuscripts; and code assistants that can generate and test scripts (Acharya, Kuppan, & Divya, 2025). They are typically evaluated using task completion metrics, benchmark suites (e.g., BabyAGI, AutoGPT), and qualitative human feedback on their goal achievement and adaptability. Incorporating psychometrics could enhance evaluation by modeling latent behavioral traits and tracking their evolution over time, task type, complexity, and novelty.

LLM-as-a-judge: AI evaluating AI for automated psychometrics. If AI is similar or better than humans at a given task, then AI could be the ground truth criterion we try to predict, perhaps to validate smaller, weaker, faster, or cheaper models. This is known as LLM-as-a-judge (Zheng et al., 2024). The gemma models we used incorporate this during training by having a larger, higher capability model express choose between responses it produces (Team et al., 2024). Models could choose the best hyperparameters of another model for training, suggest adjustments during training, suggest how to adapt trained models to new data, generate synthetic training data to improve measurement, continually

monitor predictions once deployed, and run tests to improve predictions. It is feasible that AI could one day help demonstrate all the desirable properties we outlined in section 4 on its own. However, generative AI has been shown to carry different types of racial and gender biases (Zack et al., 2024) as well as methodological biases (Zheng et al., 2024), which should be carefully evaluated and mitigated.

8 Conclusion

We have tried to promote Text Psychometrics as a field that studies how to assess psychological constructs in text. We first presented different types of validity and desirable properties that could be evaluated for text-based assessments. We then systematically compared a diverse set of models including lexicons, zero-shot construct-text similarity, and zero-shot generative AI on different desirable properties in their ability to predict differents types of mental health issues and imminent suicide risk. We argue there is a hyperfocus on criterion validity (e.g., out-of-sample prediction of a gold-truth label). We introduced content validity test sets as a way to demonstrate whether models miss any key prototypical expressions of the target variable. This helps us demonstrate that some models can have good criterion validity (e.g., ROC >0.75) but miss many prototypical tokens (e.g., content validity sensitivity < 0.80). In a second use case, we predicted the need for a future emergency service intervention (i.e., prospective criterion validity) and uncovered which risk factors most predicted the need for an emergency service intervention (e.g., lethal means for suicide more than depressed mood or hopelessness). We finish by providing a comparison of desirable properties to select models depending on the constraints different researchers may have (Figure 7.1). We hope this study will encourage broader validation of these promising NLP methods in psychology, medicine, and the social sciences.

Acknowledgments

We would like to thank the entire team at Crisis Text Line, especially Elizabeth Olson, Margaret Meagher, Shannon Green, and Devyani Singh; and to Philip Resnik and Sierra Bainter for helpful discussions. DML was supported by an NIMH training grant [5T32MH125815-03], an NIDCD training grant [5T32DC000038-28], a RallyPoint Fellowship, and an Amelia Peabody Professional Development Award. The work was supported by a gift to the McGovern Institute for Brain Research at MIT. DML has been a paid consultant for legal cases regarding a death by suicide. MKN receives publication royalties from Macmillan, Pearson, and UpToDate. MKN has been a paid consultant in the past three years for Apple, Microsoft, and COMPASS Pathways, and for legal cases regarding a death by suicide. He has stock options in Cerebral Inc. MKN is an unpaid scientific advisor for Empatica, Koko, and TalkLife. The other authors have no competing interests to disclose.

Author contributions

Conceptualization: DML; methodology: DML, SSG; data curation: DML; Software: DML; Formal Analysis: DML; Original Draft: DML; Review & Editing: all authors; supervision: MKN, SSG.

Reddit data and code are available at: `https://github.com/danielmlow/text_psychometrics`. Crisis Text Line data is sensitive and cannot be shared.

References

Acharya, D. B., Kuppan, K., & Divya, B. (2025). Agentic ai: Autonomous intelligence for complex goals–a comprehensive survey. *IEEE Access*.

Agarwal, C., Tanneru, S. H., & Lakkaraju, H. (2024, February). Faithfulness vs. plausibility: On the (Un)Reliability of explanations from large language models.

Ahmad, F., Abbasi, A., Li, J., Dobolyi, D. G., Netemeyer, R. G., Clifford, G. D., & Chen, H. (2020). A deep learning architecture for psychometric natural language processing. *ACM Transactions on Information Systems (TOIS)*, *38*(1), 1–29.

Allport, G. W., & Odbert, H. S. (1936). Trait-names: A psycho-lexical study. *Psychological monographs*, *47*(1), i.

Anastasi, A., & Urbina, S. (1997). *Psychological testing*. Prentice Hall.

Ashok, D., & Lipton, Z. C. (2023). Promptner: Prompting for named entity recognition. *arXiv preprint arXiv:2305.15444*.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., ... others (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Bilgrami, Z. R., Sarac, C., Srivastava, A., Herrera, S. N., Azis, M., Haas, S. S., ... others (2022). Construct validity for computational linguistic metrics in individuals at clinical risk for psychosis: associations with clinical ratings. *Schizophrenia research*, *245*, 90–96.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*(3), 425–440.
Borsboom, D., Cramer, A. O., Kievit, R. A., Scholten, A. Z., & Franić, S. (2009). The end of construct validity. In *The concept of validity: Revisions, new directions and applications, oct, 2008.*
Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological review*, *111*(4), 1061.
Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). The development and psychometric properties of liwc-22. *Austin, TX: University of Texas at Austin*, 1–47.
Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., ... Stoica, I. (2024). *Chatbot arena: An open platform for evaluating llms by human preference.*
Chien, A. A., Lin, L., Nguyen, H., Rao, V., Sharma, T., & Wijayawardana, R. (2023). Reducing the carbon impact of generative ai inference (today and in 2035). In *Proceedings of the 2nd workshop on sustainable computer systems* (pp. 1–7).
Cohen, A. S., Rodriguez, Z., Warren, K. K., Cowan, T., Masucci, M. D., Edvard Granrud, O., ... Strauss, G. P. (2022). Natural language processing and psychosis: on the need for comprehensive psychometric evaluation. *Schizophrenia bulletin*, *48*(5), 939–948.
Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *psychometrika*, *16*(3), 297–334.
Davidson, M. (2014). Known-groups validity. *Encyclopedia of quality of life and well-being research*, 3481–3482.
Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020). Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.
Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., ... others (2023). Using large language models in psychology. *Nature Reviews Psychology*, *2*(11), 688–701.
Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2024). Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, *36*.
Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
Dubin, D. (2004). The most influential paper gerard salton never wrote.
European Commission. (2023). *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonized Rules on Artificial Intelligence (Artificial Intelligence Act).* Retrieved from `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206` (COM/2021/206 final)
Fang, Q., Nguyen, D., & Oberski, D. L. (2022). Evaluating the construct validity of text embeddings with application to survey questions. *EPJ Data Science*, *11*(1), 39.
Firth, J. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*, 10–32.
Fried, E. I., Flake, J. K., & Robinaugh, D. J. (2022). Revisiting the theoretical and methodological foundations of depression measurement. *Nature Reviews Psychology*, *1*(6), 358–368.
Fu, J., Liu, P., & Neubig, G. (2020). Interpretable multi-dataset evaluation for named entity recognition. *arXiv preprint arXiv:2011.06854*.
Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., ... Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.
Goldberg, L. R. (1981). Language and individual differences: The search for universals in personality lexicons. *Review of personality and social psychology*, *2*(1), 141–165.
Gottfried, J. (2024). Americans' social media use. *Pew Research Center*, *31*.
Gretz, S., Halfon, A., Shnayderman, I., Toledo-Ronen, O., Spector, A., Dankin, L., ... others (2023). Zero-shot topical text classification with llms-an experimental study. In *The 2023 conference on empirical methods in natural language processing.*
Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, *21*(3), 267–297.
Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., ... others (2025). Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
Gurven, M., Von Rueden, C., Massenkoff, M., Kaplan, H., & Lero Vie, M. (2013). How universal is the big five? testing the five-factor model of personality variation among forager–farmers in the bolivian amazon. *Journal of personality and social psychology*, *104*(2), 354.
Hämäläinen, P., Oksanen, J., Tavast, M., & Bhatnagar, P. (2024). *LLMCode: A toolkit for AI-assisted qualitative data analysis.* Retrieved from `https://github.com/PerttuHamalainen/LLMCode`
Hämäläinen, P., Tavast, M., & Kunnari, A. (2023). Evaluating large language models in generating synthetic hci research data: a case study. In *Proceedings of the 2023 chi conference on human factors in computing systems* (pp. 1–19).
Harris, Z. S. (2013). *Papers in structural and transformational linguistics*. Springer.
Hernán, M. A., & Robins, J. M. (2020). *Causal inference: What if*. Boca Raton: Chapman Hall/CRC.

Hort, M., Chen, Z., Zhang, J. M., Harman, M., & Sarro, F. (2023). Bias mitigation for machine learning classifiers: A comprehensive survey. *ACM Journal on Responsible Computing*.

Hoyle, A., Goel, P., Hian-Cheong, A., Peskov, D., Boyd-Graber, J., & Resnik, P. (2021). Is automated topic model evaluation broken? the incoherence of coherence. *Advances in neural information processing systems*, *34*, 2018–2033.

Hoyle, A., Goel, P., Sarkar, R., & Resnik, P. (2022). Are neural topic models broken? *arXiv preprint arXiv:2210.16162*.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... Chen, W. (2021). Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Jacobs, A. Z., & Wallach, H. (2021). Measurement and fairness. In *Proceedings of the 2021 acm conference on fairness, accountability, and transparency* (pp. 375–385).

Jcgm, J., et al. (2008). Evaluation of measurement data—guide to the expression of uncertainty in measurement. *Int. Organ. Stand. Geneva ISBN*, *50*, 134.

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. In *Journal of documentation* (Vol. 28, pp. 11–21). Emerald Group Publishing Limited.

Kaur, D., Uslu, S., Rittichier, K. J., & Durresi, A. (2022). Trustworthy artificial intelligence: a review. *ACM computing surveys (CSUR)*, *55*(2), 1–38.

Kelley, T. L. (1927). *Interpretation of educational measurements*. World Book Company.

Kennedy, C. J., Bacon, G., Sahn, A., & von Vacano, C. (2020). Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application. *arXiv preprint arXiv:2009.10277*.

Kennedy, C. J., Mark, D. G., Huang, J., van der Laan, M. J., Hubbard, A. E., & Reed, M. E. (2021). Development of an ensemble machine learning prognostic model to predict 60-day risk of major adverse cardiac events in adults with chest pain. *MedRxiv*, 2021–03.

Kessler, R. C., Bossarte, R. M., Luedtke, A., Zaslavsky, A. M., & Zubizarreta, J. R. (2020). Suicide prediction models: a critical review of recent research with recommendations for the way forward. *Molecular psychiatry*, *25*(1), 168–179.

Kjell, O. N., Kjell, K., Garcia, D., & Sikström, S. (2019). Semantic measures: Using natural language processing to measure, differentiate, and describe psychological constructs. *Psychological Methods*, *24*(1), 92.

Kjell, O. N., Kjell, K., & Schwartz, H. A. (2023). Beyond rating scales: With targeted evaluation, language models are poised for psychological assessment. *Psychiatry Research*, 115667.

Kompa, B., Snoek, J., & Beam, A. L. (2021). Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digital Medicine*, *4*(1), 4.

Kotek, H., Dockum, R., & Sun, D. (2023). Gender bias and stereotypes in large language models. In *Proceedings of the acm collective intelligence conference* (pp. 12–24).

Lang, K. (1995). Newsweeder: Learning to filter netnews. In *Machine learning proceedings 1995* (pp. 331–339). Elsevier.

Laverghetta Jr, A., & Licato, J. (2023). Generating better items for cognitive assessments using large language models. In *Proceedings of the 18th workshop on innovative use of nlp for building educational applications (bea 2023)* (pp. 414–428).

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., ... others (2022). Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Lieberum, T., Rajamanoharan, S., Conmy, A., Smith, L., Sonnerat, N., Varma, V., ... Nanda, N. (2024). Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. *arXiv preprint arXiv:2408.05147*.

Lin, X. V., & Morency, L.-P. (2023). Sentecon: Leveraging lexicons to learn human-interpretable language representations. In *Findings of the association for computational linguistics: Acl 2023* (pp. 4312–4331).

Lindsey, J., Gurnee, W., Ameisen, E., Chen, B., Pearce, A., Turner, N. L., ... Batson, J. (2025, March). *On the biology of a large language model.* `https://transformer-circuits.pub/2025/attribution-graphs/biology.html`. (Published March 27, 2025)

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, *16*(3), 31–57.

Low, D. M., Rankin, O., Coppersmith, D., Bentley, K. H., Nock, M. K., & Ghosh, S. S. (2024). Using large language models to create lexicons for interpretable text models with high content validity: the suicide risk lexicon. *PsyArXiv*.

Low, D. M., Rao, V., Randolph, G., Song, P. C., & Ghosh, S. S. (2023). Identifying bias in models that detect vocal fold paralysis from audio recordings using explainable machine learning and clinician ratings. *medRxiv: the preprint server for health sciences*, 2020–11.

Low, D. M., Rumker, L., Talkar, T., Torous, J., Cecchi, G., & Ghosh, S. S. (2020, October). Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during

COVID-19: Observational study. *J. Med. Internet Res.*, *22*(10), e22635.
Mair, P., et al. (2018). *Modern psychometrics with r* (Vol. 10). Springer.
Mayson, S. G. (2018). Bias in, bias out. *YAle lJ*, *128*, 2218.
McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological methods*, *23*(3), 412.
Meta AI. (2024). *The llama 4 herd: The beginning of a new era of natively multimodal ai innovation.* `https://ai.meta.com/blog/llama-4-multimodal-intelligence/`. Retrieved from `https://ai.meta.com/blog/llama-4-multimodal-intelligence/` (Accessed: 2025-04-07)
Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024). Large language models: A survey. arxiv 2024. *arXiv preprint arXiv:2402.06196*.
Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.
Mosca, E., Szigeti, F., Tragianni, S., Gallagher, D., & Groh, G. (2022). Shap-based explanation methods: a review for nlp interpretability. In *Proceedings of the 29th international conference on computational linguistics* (pp. 4593–4603).
Muennighoff, N., Tazi, N., Magne, L., & Reimers, N. (2022). Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.
Pargent, F., Schoedel, R., & Stachl, C. (2023). Best practices in supervised machine learning: A tutorial for psychologists. *Advances in Methods and Practices in Psychological Science*, *6*(3), 25152459231162559.
Pellert, M., Lechner, C. M., Wagner, C., Rammstedt, B., & Strohmaier, M. (2023). Ai psychometrics: Using psychometric inventories to obtain psychological profiles of large language models. *OSF preprint*.
Pisani, A. R., Gould, M. S., Gallo, C., Ertefaie, A., Kelberman, C., Harrington, D., ... Green, S. (2022). Individuals who text crisis text line: Key characteristics and opportunities for suicide prevention. *Suicide and Life-Threatening Behavior*, *52*(3), 567–582.
Prince, S. J. (2023). *Understanding deep learning*. The MIT Press. Retrieved from `http://udlbook.com`
Raschka, S. (2018). Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808*.
Raschka, S. (2024). *Build a large language model (from scratch)*. Manning Publications.
Rathje, S., Mirea, D.-M., Sucholutsky, I., Marjieh, R., Robertson, C. E., & Van Bavel, J. J. (2024). Gpt is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences*, *121*(34), e2308950121.
Riezler, S., & Hagmann, M. (2022). *Validity, reliability, and significance: Empirical methods for nlp and data science*. Springer Nature.
Rönkkö, M., & Cho, E. (2022). An updated guideline for assessing discriminant validity. *Organizational Research Methods*, *25*(1), 6–14.
Rust, J., Kosinski, M., & Stillwell, D. (2020). *Modern psychometrics*. Routledge.
Salton, G., Wong, A., & Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613–620.
Sanh, V., Webson, A., Raffel, C., Bach, S. H., Sutawika, L., Alyafeai, Z., ... others (2021). Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.
Schopf, T., Braun, D., & Matthes, F. (2022). Evaluating unsupervised text classification: zero-shot and similarity-based approaches. In *Proceedings of the 2022 6th international conference on natural language processing and information retrieval* (pp. 6–15).
Scialom, T., Dray, P.-A., Gallinari, P., Lamprier, S., Piwowarski, B., Staiano, J., & Wang, A. (2021). Questeval: Summarization asks for fact-based evaluation. *arXiv preprint arXiv:2103.12693*.
Shankman, S. A., Kaiser, A. J., & Shenberger, E. R. (2020). The importance of examining psychometrics of biomarkers. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *5*(4), 379–380.
Shu, B., Zhang, L., Choi, M., Dunagan, L., Card, D., & Jurgens, D. (2023a). You don't need a personality test to know these models are unreliable: Assessing the reliability of large language models on psychometric instruments. *arXiv preprint arXiv:2311.09718*.
Shu, B., Zhang, L., Choi, M., Dunagan, L., Card, D., & Jurgens, D. (2023b). You don't need a personality test to know these models are unreliable: Assessing the reliability of large language models on psychometric instruments. *arXiv preprint arXiv:2311.09718*.
Sikström, S., Nicolai, M., Ahrendt, J., Nevanlinna, S., & Stille, L. (2024). Language or rating scales based classifications of emotions: computational analysis of language and alexithymia. *npj Mental Health Research*, *3*(1), 37.
Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., ... others (2023). Large language models encode clinical knowledge. *Nature*, *620*(7972), 172–180.
Son, Y., Clouston, S. A., Kotov, R., Eichstaedt, J. C., Bromet, E. J., Luft, B. J., & Schwartz, H. A. (2023). World trade center responders in their own words: predicting ptsd symptom trajectories with ai-based language analyses of interviews. *Psychological medicine*, *53*(3), 918–926.

Stone, P. J., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1962). The general inquirer: A computer approach to content analysis in the behavioral sciences. *Am. Sociol. Rev.*.

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune BERT for text classification? In *China national conference on chinese computational linguistics* (pp. 194–206).

Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., … others (2024). Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Thurstone, L. L. (1934). The vectors of mind. *Psychological review*, *41*(1), 1.

Truijens, F. L., Cornelis, S., Desmet, M., De Smet, M. M., & Meganck, R. (2019). Validity beyond measurement: Why psychometric validity is insufficient for valid psychotherapy research. *Frontiers in psychology*, *10*, 532.

Van der Wal, O., Bachmann, D., Leidinger, A., van Maanen, L., Zuidema, W., Schulz, K., et al. (2022). Undesirable biases in nlp: Averting a crisis of measurement. *arXiv preprint arXiv:2211.13709*.

Vickers, A. J., Van Calster, B., & Steyerberg, E. W. (2016). Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *bmj*, *352*.

von Davier, A. A., Mislevy, R. J., & Hao, J. (2022). *Computational psychometrics: New methodologies for a new generation of digital learning and assessment: With examples in r and python*. Springer Nature.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., … others (2021). Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Wiles, O., Gowal, S., Stimberg, F., Alvise-Rebuffi, S., Ktena, I., Dvijotham, K., & Cemgil, T. (2021). A fine-grained analysis on distribution shift. *arXiv preprint arXiv:2110.11328*.

Williams, A., Nangia, N., & Bowman, S. R. (2018). A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of naacl-hlt* (pp. 1112–1122).

Yin, W., Hay, J., & Roth, D. (2019). Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *2019 conference on empirical methods in natural language processing and 9th international joint conference on natural language processing, emnlp-ijcnlp 2019* (pp. 3914–3923).

Zack, T., Lehman, E., Suzgun, M., Rodriguez, J. A., Celi, L. A., Gichoya, J., … others (2024). Assessing the potential of gpt-4 to perpetuate racial and gender biases in health care: a model evaluation study. *The Lancet Digital Health*, *6*(1), e12–e22.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., … others (2024). Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, *36*.

Zuromski, K. L., Low, D. M., Jones, N., Kuzma, R., Kessler, D. T., Zhou, L., … K, N. M. (2024). Detecting suicide risk among US servicemembers and veterans: A deep learning approach using social media data. *Psychological Medicine*.

Appendix

A Correlation does not imply prediction

A key methodological distinction between psychology and NLP is in the definition of prediction and association. While it is common knowledge that correlation (or more generally, association) does not imply causation, it is less known in psychology that correlation does not imply prediction. In machine learning, prediction entails training a model from a set of samples and labels, and then testing the model on new samples the model did not observe. It is common in psychology to refer to within-sample associations (e.g., correlations, within-sample regressions) as predictions; however, a model fit using the entire dataset may not generalize well when predicting new individual samples (Yarkoni & Westfall, 2017). Therefore, it is useful to distinguish associations or explanations (i.e., within-sample testing) from predictions (i.e., out-of-sample testing) as defined in machine learning and natural language processing. Moreover, many have suggested to place more focus on prediction in psychology and related fields (Poldrack, Huckins, & Varoquaux, 2020; Shmueli, 2010; Yarkoni & Westfall, 2017). Prediction is a better test of generalizability, that is the ability for an instrument to be reliable across different demographics, measurement instruments, models, hardware, and time (related to the concepts of external validity and transferability).

B Brief history of psychometrics: from classical test theory to networks and causal inference

Here we review how psychometrics has dealt with modeling psychological assessments throughout its history. From a classical test theory perspective, $X = T + E$ (the true score model) where X is the observed score, T is the true unknown score and E is the unknown error between the two (Mair et al., 2018). A good quality measurement should show consistency among repeated attempts to measure the same thing, which implies observations X match true scores T (i.e., reliable scores, free from error across observations). Differences in repeated attempts could be due to measurement error; seasonal effects; the episodic, oscillating nature of the construct; carryover effects (i.e., people remember their answers if the time-lapse is short). Since true scores and errors are unknown, many different attempts at estimating reliability have been developed which involve either correlating multiple test administrations (i.e., parallel forms reliability, test-retest reliability) or calculating reliability across items of a single test administration (internal consistency) (McNeish, 2018). Instead of assuming there is one source of error as in the classical true score model, Cronbach (1972) (Cronbach, 1972) proposed generalizability theory to account for multiple sources of error (i.e., facets) including items, raters, and measurement instances (Mair et al., 2018). A test can reliably return the same score, but be measuring the wrong attribute, that is, not be valid; whereas, a test cannot logically be both valid and unreliable.

Item response theory (IRT, a.k.a., modern test theory) was later developed to account for the fact that different items have different difficulty and discrimination (i.e., how well an item can differentiate between individuals with different levels of the trait) (De Ayala, 2013). IRT achieves this by modeling the probability of a person's response to an item as a function of the person's ability and the item characteristics. This can provide more flexible, efficient, and informative testing than classical test theory approaches. When applied to biomarkers or behavioral makers (e.g., allostatic load from blood biomarkers, a psychological construct from natural language), IRT can help identify individual variables that are most informative and least informative for developing a risk assessment score, and determine if individual variables provide information along the entire risk score continuum, or if they provide information for only a portion of the continuum (Liu, Juster, Dams-O'Connor, & Spicer, 2021).

A limitation of IRT is that it assumes a specific causal structure where observed items (e.g., sadness, anhedonia, fatigue) are caused by an underlying latent variable (e.g., depression), that is, a reflective model (Edwards & Bagozzi, 2000). However, as Fried (2017) (Fried, 2017) points out, psychological constructs do not seem to be natural kinds (e.g., gold, water) that are discovered, unchanging, and exist independently of human measurement; rather they are constructed (human-defined), mutable, and have no sharp boundaries. Once constructed, certain psychological constructs seem to co-occur not necessarily because there is an underlying common cause (e.g., depression as a brain disorder causing changes in passive indicators), but because they cause each other (e.g., each symptom having unique brain circuits that affect each other without the need to consider a depression construct). Instead of a natural kind, this can be better thought of a homeostatic property cluster and statistically modeled through networks (i.e., a graph of variables and multivariate associations between them) (Borsboom, 2008; Borsboom et al., 2021; Cramer, Waldorp, Van Der Maas, & Borsboom, 2010; Fried, 2017) instead of through latent variable models like IRT. Unlike traditional models that view symptoms as passive outcomes of an underlying illness, the network approach sees symptoms as active contributors that causally trigger and reinforce each other: symptoms do not merely signal the presence of a disorder—they are the disorder (Roefs et al., 2022). For more on the history of psychometrics, see Jones and Thissen (2006).

A limitation of prior approaches is they do not prove causality as rigorously as structural causal models that use directed acyclic graphs (Pearl, 2010) or as the potential outcomes framework Hernán and Robins (2020). Having said

this, the use of directed acyclic graphs is challenging in psychology given acyclicality would rarely hold if constructs are causing each other, but recent advances that drop this assumption have been proposed (Park, Ryan, & Waldorp, 2023). Another approach is to use dynamical systems to obtain closed-formed equations or computational models of how constructs interact in a system Haslbeck, Ryan, Robinaugh, Waldorp, and Borsboom (2019); Robinaugh et al. (2024); Ryan, Haslbeck, and Robinaugh (2023). This approach is promising because, as with networks, it better reflects the fact that constructs (e.g., panic disorder, algebra proficiency) and exposures (cognitive behavioral therapy, racism) are not a single, unitary objects but a system of often nonlinear interacting objects; therefore, applying standard causal inference on the effect of CBT (which is practiced differently by different clinicians and has many active ingredients) on panic disorder may be an ill-posed problem, which complex systems may model better (Ryan et al., 2023). From this perspective, mental health disorders are systems that are stuck in negative stable states reinforced over time by harmful causal interactions (anxiety > insomnia > fatigue > anxiety).

C How to define measurement and validity?

A common view of measurement in psychology is that of assigning any type of numeral to a construct through some function or rule (Krantz, Luce, Suppes, & Tversky, 1971). This definition broadly applies to many types of relationships. For instance, the degree to which SAT scores is associated with future college success. So what constitutes a valid measurement? A classic definition for validity states that a test is valid if it measures what it purports to measure (Kelley, 1927). Borsboom et al. (2004) (Borsboom et al., 2004) argues that a test is valid for measuring an attribute if the attribute exists and if variations in the attribute causally produce variations in the outcomes of the measurement; if they do not, then perhaps one is measuring something different or nothing at all (Borsboom et al., 2004). Markus and Borsboom (2013, p.54) (Markus & Borsboom, 2013) suggest it might therefore be more accurate to think of measurement as assigning quantities related to a real attribute, which should be on at least an interval scale. If the difference between a score at the 10th percentile and a score 1 point above is not the same as the difference between a score at the 90th percentile and a score 1 point above (i.e., it is not interval), then changes in the real attribute assessed would not create precise changes in the score assigned as occurs in more natural attributes such as temperature and well-validated measurement tools such as thermometers. Furthermore, measurement is prototypically asymmetric: a thermometer measures temperature, but temperature does not measure a thermometer's value given the logical causal direction between the two. In this sense, finding that SAT scores correlate with GPA in college is statistically symmetric, so perhaps should not be considered a case of measurement but of prediction, association, or testing. Therefore, Markus and Borsboom (2013) (Markus & Borsboom, 2013) definition of measurement would not apply to most psychological assessments.

D Formal description of text classification

As described in (Riezler & Hagmann, 2022), first, a label y for a raw text input x is provided by a human annotator or algorithm. Then a machine learning model p_θ is trained (a function is learned) to map labels and inputs and predict output $\hat{y}$ where θ is the learned parameters. This can be formalized as $\hat{y} = p_\theta(x)$ for a regression setting. In classification settings, we use the maximum a posteriori prediction $\hat{y} = \text{argmax}_y\, p_\theta(y|x)$ where the category y_i with the maximum value is taken among possible categories y. Measuring psychological constructs in text can be done for binary or multi-class (i.e., multinomial) classification (i.e., categorical output variable) to answer questions such as "Is this text document about construct c?"; through multi-label classification to answer questions such as "Which of these constructs $c \in C$ are present?"; as well as through regression (which in machine learning refers to predicting ordinal or metric output variables) to answer questions such as "Can I detect the severity or degree that construct c is present in a person from a sample of their text?".

E Construct-Text Similarity

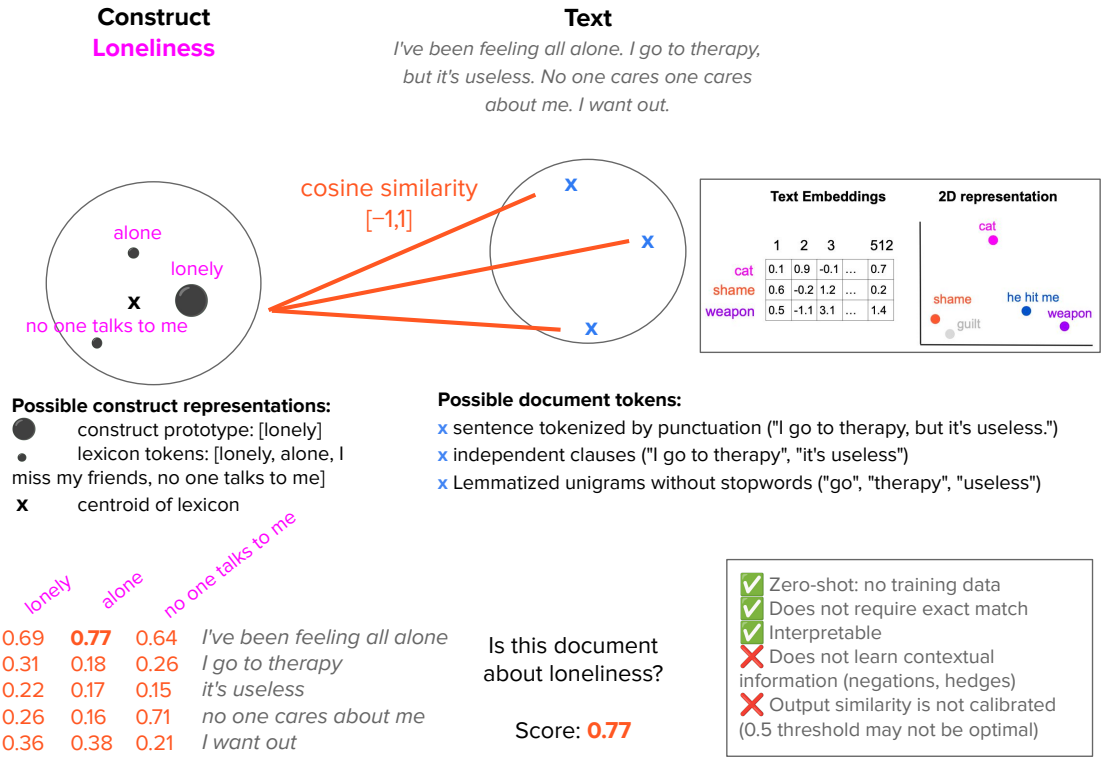

Figure 8: Construct-text similarity. Circles and crosses are text embeddings (quantitative representations of meaning in the form of vectors). There are multiple options for computing each step: construct representation, document representation, similarity function, and summary function. Here we build the construct through highly prototypical lexicon tokens, the document representation as independent clauses, compute the cosine similarity, and take the maximum similarity (in bold at the bottom). Conceptually, this is similar to asking where does the document mention something most related to loneliness? The document text is made-up.

F Prompts used for large language models

Small LLMs (e.g., gemma 2b and 7b) performed above chance only with a single classification task (i.e., one issue or label) per prompt (Figure 9). The more powerful models (GPT-4o and Gemini 2.0 flash) could perform well while still including all labels in the same prompt, so as to reduce the number of API submissions.

Prompt

You are a <setting> classification assistant. Classify the following <setting> (starts and ends with ```):

```

{document}

```

Assign probabilities for the following labels and return using this JSON format (do not provide additional notes, explanations, or warnings). Provide your best guess, only return JSON (both probabilities should sum to 1):

JSON:

{'texter mentions something related to anxiety': <your_probability>,

'texter does not mention anything related to anxiety': <your_probability>}

JSON:

Output

{'texter mentions something related to anxiety': 0.8,

'texter does not mention anything related to anxiety': 0.2}

Figure 9: Prompt for zero-shot text classification using small LLMs. <setting> could be "Crisis Text Line conversation" or "Reddit post". Prompt was changed for each issue and document. This elicits binary classification.

Prompt

In the following Reddit post, classify the following constructs, defined as follows:

Self harm or self injury: Mention of actual or desired direct and deliberate destruction by an individual of their own body tissue, including with suicidal intent (suicidal self-injury) our without suicidal intent (non-suicidal self-injury). Examples: self-injury; cut myself; cutting; burn myself; harm myself
Anxiety: Anxiety is a feeling of worry, nervousness, or unease, often about an imminent event or something with an uncertain outcome. We include anxiety disorders. Examples: anxiety; fear; afraid; worried; OCD; PTSD; panic attacks; social anxiety; phobia; nightmare; nervous
Suicidal thoughts or suicidal behaviors: Explicit suicidal thoughts of killing oneself; mention of methods, plan, preparations, time, and/or place of suicide attempt. Suicidal thoughts desiring one's own death or related states (e.g., disappearing, not waking up). Suicidal language that does not belong in the other suicidal categories (active and passive SI, self-injury, and lethal means for suicide). Examples: kill myself; jump off a bridge; jump in front of a train; overdose; hang myself; shoot myself wish I wasn't alive; wish I could disappear; hoping I wouldn't wake up; I want to die no reason to live; safety plan; I hate my life
Bullying: Bullying is defined as intentional, repeated, and harmful aggressive behavior (verbal, physical, social) often with an imbalance of power between the perpetrators and the victims. Examples: bullied; makes fun of me
Sexual abuse: Any non-consensual act or behavior of a sexual nature imposed on someone. It can range from unwanted sexual touching and coercion to rape and exploitation. We also include sexual harassment in this lexicon. Examples: rape; sexual assault; was abused; non-consensual; molested me; stalking; groped me
Bereavement or grief: Sadness from losing a loved one (grief) which may result in depressive symptoms (bereavement). This may also include losing a relationship or the health of a loved one. Examples: grief; grieving; bereavement; bereaving; divorce; breakup; died; passed away
Loneliness or social isolation: Loneliness: aversive state experienced when a discrepancy exists between the interpersonal relationships one wishes to have and those that one perceives they currently have. The perception that one's social relationships are not living up to some expectation (Heinrich & Gullone 2006). Isolation: the expression of solitary behavior that results from peer rejection (Rubin & burgess). Related to thwarted belongingness where the person perceives loneliness, isolation, and a failing to meet one's need to belong (Van Orden et al 2010). Examples: lonely; isolated
Depression: Low mood and sadness that does not go away after a few weeks. Examples: depression; sadness; low mood; melancholy
Gender identity: Gender identity refers to an individual's sense of their self as male, female, transgender, non-binary, genderfluid, agender, or something else. Gender identity is distinct from the cultural roles, behaviors, and attributes expected of women and men based on their sex assigned at birth. Sexual identity, also known as sexual orientation, refers to the pattern of romantic or sexual attraction to others. Common sexual identities include: heterosexual, homosexual, bisexual, pansexual, and asexual. Examples: gay; heterosexual; homosexual; lesbian; pansexual; bisexual; asexual; male; female; nonbinary; transgender; intersex; questioning
An eating disorder or body image issues: Eating disorders are mental health conditions characterized by unhealthy and obsessive behaviors related to food, eating, and body image. Common types include anorexia nervosa, bulimia nervosa, and binge-eating disorder. Examples: anorexia; bulimia; binge; purging; eating disorder
Substance use: Mentions of types of abused substances and related behaviors (snort), measures (8 ball), and states (high, buzz) Mentions of types of alcohol beverages and related behaviors (drinking), measures (shot), and states (drunk). Examples: cocaine; heroin; weed; cannabis; meth; amphetamine; snort; 8 ball; pain killer; opioid beer; cocktail; wine; hangover; drunk; buzz

---

Provide a probability score between 0 and 1 for each construct (0=construct is not mentioned, indirectly mentioned, or not clear, 1 = clear and prototypical mention of the construct) as to whether the text clearly mentions the construct and an explanation (words or phrases from the text that are very prototypical expressions of the construct). Only provide a score for these constructs.

Post: {document}

Structure your response in the following JSON format (no extra text):
{'construct_A': [[score], [words, phrases]], 'construct_B': [[score], [words, phrases]], ...}

JSON:

Output

{"Self harm or self injury": [1.0], ["harming a specific part of my body", "wanna harm"], "Anxiety": [0.3], ["don't want her to get mad at me"], ...}

Figure 10: Prompt for zero-shot text classification using proprietary LLMs via an API. Prompt was changed for each document. This elicits multi-label classification (scores do not add to 1).

References

Borsboom, D. (2008). Psychometric perspectives on diagnostic systems. *Journal of clinical psychology*, *64*(9), 1089–1108.

Borsboom, D., Deserno, M. K., Rhemtulla, M., Epskamp, S., Fried, E. I., McNally, R. J., ... others (2021). Network analysis of multivariate data in psychological science. *Nature Reviews Methods Primers*, *1*(1), 58.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological review*, *111*(4), 1061.

Cramer, A. O., Waldorp, L. J., Van Der Maas, H. L., & Borsboom, D. (2010). Comorbidity: A network perspective. *Behavioral and brain sciences*, *33*(2-3), 137–150.

Cronbach, L. J. (1972). The dependability of behavioral measurements. *Theory of generalizability for scores and profiles*, 1–33.

De Ayala, R. J. (2013). *The theory and practice of item response theory*. Guilford Publications.

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological methods*, *5*(2), 155.

Fried, E. I. (2017). What are psychological constructs? on the nature and statistical modelling of emotions, intelligence, personality traits and mental disorders. *Health psychology review*, *11*(2), 130–134.

Haslbeck, J., Ryan, O., Robinaugh, D., Waldorp, L., & Borsboom, D. (2019). *Modeling psychopathology: From data models to formal theories. psyarxiv.*

Hernán, M. A., & Robins, J. M. (2020). *Causal inference: What if*. Boca Raton: Chapman Hall/CRC.

Jones, L. V., & Thissen, D. (2006). 1 a history and overview of psychometrics. *Handbook of statistics*, *26*, 1–27.

Kelley, T. L. (1927). *Interpretation of educational measurements*. World Book Company.

Krantz, D., Luce, D., Suppes, P., & Tversky, A. (1971). Foundations of measurement, vol. i: Additive and polynomial representations.

Liu, S. H., Juster, R.-P., Dams-O'Connor, K., & Spicer, J. (2021). Allostatic load scoring using item response theory. *Comprehensive Psychoneuroendocrinology*, *5*, 100025.

Mair, P., et al. (2018). *Modern psychometrics with r* (Vol. 10). Springer.

Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. Routledge.

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological methods*, *23*(3), 412.

Park, K., Ryan, O., & Waldorp, L. (2023). Discovering cyclic causal models in psychological research.

Pearl, J. (2010). An introduction to causal inference. *The international journal of biostatistics*, *6*(2).

Poldrack, R. A., Huckins, G., & Varoquaux, G. (2020). Establishment of best practices for evidence for prediction: a review. *JAMA psychiatry*, *77*(5), 534–540.

Riezler, S., & Hagmann, M. (2022). *Validity, reliability, and significance: Empirical methods for nlp and data science*. Springer Nature.

Robinaugh, D. J., Haslbeck, J., Waldorp, L. J., Kossakowski, J. J., Fried, E. I., Millner, A. J., ... others (2024). Advancing the network theory of mental disorders: A computational model of panic disorder. *Psychological review*, *131*(6), 1482.

Roefs, A., Fried, E. I., Kindt, M., Martijn, C., Elzinga, B., Evers, A. W., ... Jansen, A. (2022). A new science of mental disorders: Using personalised, transdiagnostic, dynamical systems to understand, model, diagnose and treat psychopathology. *Behaviour Research and Therapy*, *153*, 104096.

Ryan, O., Haslbeck, J., & Robinaugh, D. (2023). Improving treatments for mental disorders using computational models.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122.