frontiers | Frontiers in Social Psychology

TYPE Original Research
PUBLISHED 21 February 2025
DOI 10.3389/frsps.2025.1460277

OPEN ACCESS

EDITED BY
Youyou Wu,
University College London, United Kingdom

REVIEWED BY
Julian Tejada,
Federal University of Sergipe, Brazil
Suhaib Abdurahman,
University of Southern California, United States
Yinxian Zhang,
Queens College (CUNY), United States

*CORRESPONDENCE
Hannah L. Bunt
h.l.bunt@lse.ac.uk

RECEIVED 05 July 2024
ACCEPTED 27 January 2025
PUBLISHED 21 February 2025

CITATION
Bunt HL, Goddard A, Reader TW and Gillespie A (2025) Validating the use of large language models for psychological text classification. *Front. Soc. Psychol.* 3:1460277. doi: 10.3389/frsps.2025.1460277

Validating the use of large language models for psychological text classification

Hannah L. Bunt[1]*, Alex Goddard[1], Tom W. Reader[1] and Alex Gillespie[1,2]

[1]Department of Psychological and Behavioural Science, London School of Economics and Political Science, London, United Kingdom, [2]Department of Psychology, Oslo New University College, Oslo, Norway

Large language models (LLMs) are being used to classify texts into categories informed by psychological theory ("psychological text classification"). However, the use of LLMs in psychological text classification requires validation, and it remains unclear exactly how psychologists should prompt and validate LLMs for this purpose. To address this gap, we examined the potential of using LLMs for psychological text classification, focusing on ways to ensure validity. We employed OpenAI's GPT-4o to classify (1) reported speech in online diaries, (2) other-initiations of conversational repair in Reddit dialogues, and (3) harm reported in healthcare complaints submitted to NHS hospitals and trusts. Employing a two-stage methodology, we developed and tested the validity of the prompts used to instruct GPT-4o using manually labeled data (N = 1,500 for each task). First, we iteratively developed three types of prompts using one-third of each manually coded dataset, examining their semantic validity, exploratory predictive validity, and content validity. Second, we performed a confirmatory predictive validity test on the final prompts using the remaining two-thirds of each dataset. Our findings contribute to the literature by demonstrating that LLMs can serve as valid coders of psychological phenomena in text, on the condition that researchers work with the LLM to secure semantic, predictive, and content validity. They also demonstrate the potential of using LLMs in rapid and cost-effective iterations over big qualitative datasets, enabling psychologists to explore and iteratively refine their concepts and operationalizations during manual coding and classifier development. Accordingly, as a secondary contribution, we demonstrate that LLMs enable an intellectual partnership with the researcher, defined by a synergistic and recursive text classification process where the LLM's generative nature facilitates validity checks. We argue that using LLMs for psychological text classification may signify a paradigm shift toward a novel, iterative approach that may improve the validity of psychological concepts and operationalizations.

KEYWORDS

large language models (LLMs), GPT, psychology, text classification, validity, big qualitative data, artificial intelligence

1 Introduction

Large language models (LLMs)—such as ChatGPT, Gemini, and Claude—are transforming research in psychology (Demszky et al., 2023; Pangakis et al., 2023; Ziems et al., 2023). LLMs refer to neural networks that are trained on billions of textual documents (Brown et al., 2020), and include both generative models (e.g., GPT) that are designed to "predict the next word, phrase, sentence, or paragraph, given an input"

(Demszky et al., 2023, p. 2), and non-generative models (e.g., BERT) that focus on understanding and encoding language representations. In this paper, we focus on generative LLMs (henceforth "LLMs") due to their ability to classify text via natural language prompts without additional training.

One of the most promising uses of LLMs is to classify texts into meaningful categories informed by psychological theory (Bail, 2024); henceforth "psychological text classification." They appear to perform better than conventional approaches, such as word counting or supervised machine learning (Brown et al., 2020; Demszky et al., 2023; Van Atteveldt et al., 2021). This is highly significant for psychological research, as LLMs can be used to classify large textual datasets quickly and cost-effectively, whilst requiring minimal programming skills to implement, thus opening up big naturally occurring textual data to psychological analysis.

A body of evidence has accrued showing that LLMs can be used to undertake simple text classification tasks. For example, LLMs outperform crowd workers in the annotation of relevance, stance, topics, and frames (Gilardi et al., 2023) as well as in annotating political Twitter messages (Törnberg, 2023), and can identify and label psychological constructs, such as sentiment, emotions, and offensiveness across languages (Rathje et al., 2023). However, these papers provide little systematic validation beyond performance comparisons with hand-scored data. This is a problem because using LLMs for psychological text classification requires validation, as they can hallucinate and reproduce existing biases in the textual data they were trained on (Demszky et al., 2023; Pangakis et al., 2023). We draw on Krippendorff's (2004) validity framework, adopting his definition of validity as "the quality of research results that lead us to accept them as true, as speaking about the real world of people, phenomena, events, experiences, and actions" (p. 313). We also draw on Messick's (1995) framework, which highlights the utility of measures as integral to the validation process. There is currently a lack of evidence and theorization guiding the validation efforts in psychological text classification (De Kok, 2023) and little attention has been paid as to why classification performance varies across datasets and concepts. Specifically, it remains unclear exactly how psychologists should prompt and validate the use of LLMs for text classification.

To address this gap, we examined the validity of using LLMs (GPT-4o) for the classification of three distinct psychological phenomena: (1) reported speech in online diaries, (2) other-initiations of repair in Reddit dialogues, and (3) harm reported in healthcare complaints submitted to hospitals ($N = 1{,}500$ in each dataset). These datasets were hand-scored using qualitative coding protocols and interactions with GPT-4o using the chat interface. We assessed the validity of using GPT-4o for psychological text classification by developing and testing the validity of prompts that instruct the model to complete a classification task.

We employed a two-stage methodology to develop and test the validity of the prompts used to instruct GPT-4o. First, we performed an iterative prompt development phase over the development dataset (i.e., one-thirds of each manually coded dataset). Here, we examined the semantic validity, exploratory predictive validity, and content validity of our prompts. Second, we assessed the confirmatory predictive validity of our final prompts by assessing their performance on the withheld test dataset (i.e., the remaining two-thirds of each dataset).

We found that LLMs can be used for valid psychological text classification, provided that researchers employ them in an iterative and synergistic approach to establish validity. GPT-4o was found to not only replicate human coding with high accuracy, but also challenge us to refine our concepts and operationalizations. Accordingly, our study provides a dual contribution. First, we demonstrate that LLMs can provide valid coding of psychological phenomena in text. Second, we show that LLMs enable an intellectual partnership with researchers, allowing for validity checks uniquely facilitated by their generative nature. We provide avenues for future research and call for robust validation frameworks to fully harness the potential of LLMs in psychological text classification.

2 Literature review

Language contains meaningful cues about mental states and behavior. Accordingly, psychology has a long history of drawing insights from textual data. The field of psychological text analysis has become increasingly empirical and systematic (Boyd and Schwartz, 2021) since the early theories on language and mind (e.g., Freud, 1914). In particular, the digitization of human communications has led to the proliferation of textual data sources (e.g., emails, recorded speech, and online dialogues) and advances in natural language processing have created new ways to analyze behavior and thoughts expressed through written or spoken words. Whilst psychologists were previously limited to undertaking manual qualitative analysis (e.g., content analysis) of small corpora of textual data, they can now also use computer-aided approaches to analyze vast amounts of textual data.

Recent advances in automated text analysis can benefit psychological research (Birkenmaier et al., 2023; Boyd and Schwartz, 2021; Jackson et al., 2022; Kennedy et al., 2021). Psychology is grappling with the replication crisis, where many psychological studies have failed to replicate. This has been attributed to studies having low statistical power (Schimmack, 2012), as well as questionable research practices such as p-hacking, "salami slicing," and selective reporting, compromising the reliability of results (Open Science Collaboration, 2015; Simmons et al., 2011). Yet, psychology's replication crisis also stems from a "validity crisis" (Schimmack, 2021; Yarkoni, 2022). The validity of psychological research is often questioned due to its reliance on survey and experimental methodologies, which can oversimplify complex human experiences and fail to mimic real-world conditions, thereby compromising the ability to generalize to circumstances beyond the lab (Holleman et al., 2020; Kjell et al., 2019; Rai and Fiske, 2010).

Qualitative approaches, often presented as a counterpoint to quantitative approaches, also have limitations (Chang et al., 2021; Seawright, 2016). Qualitative data are typically collected by engaging participants (e.g., interviews), which can result in potentially socially desirable responses with low ecological validity. They may also be drawn from small samples of participants with limited generalizability. Additionally, inconsistent manual coding can compromise reliability (Lee et al., 2020). Biases and cherry-picking can also undermine reliability, and skew findings (Morse, 2010). Finally, the time- and cost-intensiveness of manual analyses

can pose feasibility issues, especially when resources are limited (Belotto, 2018).

Using computers to analyze text enables researchers to analyze high-validity textual data (e.g., everyday discourse) in a transparent way at scale. Moreover, because textual data can be analyzed both qualitatively and quantitatively, researchers can leverage a mixed-methods approach to have the best of both worlds: using algorithmic text classifiers to create replicable and valid measures of psychological phenomena, while qualitatively interpreting the specific meaning and context within which the text is produced to probe validity and generate explanations. Thus, the advent of new and large textual data sources, and development of automated methods for classifying them, has the potential to address both replication and validity problems in psychological research (Chang et al., 2021; Gillespie et al., 2024; Jackson et al., 2022; Kjell et al., 2019).

The simplest text classifiers involve "pattern-matching," where textual data is examined for the presence and extent of specific words or phrases associated with psychological states. For example, the Linguistic Inquiry and Word Count (LIWC) software, built in the 1990s and still widely used, counts word frequencies to classify textual data (see e.g., Tausczik and Pennebaker, 2010). Researchers have also developed supervised and unsupervised machine learning approaches to classifying text: the former involves algorithms being trained to reproduce top-down manual coding of text data, and the latter identifies psychological phenomena through bottom-up clustering of texts. Both of these approaches offer enhanced performance in data processing and analysis available to those with more advanced programming skills (Boyd and Schwartz, 2021; Jackson et al., 2022).

The validity of automated text classification approaches hinges on "gold standard" data, usually based on human coding. Gold standard datasets are used as benchmarks to evaluate the accuracy of automated text classifiers, yet the quality of these gold standard datasets is sensitive to factors such as bias in coding, poor training, low inter-rater agreement, or poor-quality coding due to fatigue and limits to attention span (Demszky et al., 2023; Grimmer and Stewart, 2013; Grimmer et al., 2022; Song et al., 2020). Alternatively, even when there is high inter-rater reliability, the conceptualization or operationalization of the target construct may still be flawed or of low quality. This means that coders might consistently apply the same misconceptions or biases about a psychological construct, leading to reliable but invalid coding. Accordingly, the validity of our concepts and measurement protocols is paramount (Bringmann et al., 2022; Flake et al., 2022; Flake and Fried, 2020). The challenges associated with existing text classification highlights the need for innovative approaches to address validity problems.

2.1 Using large language models for psychological text classification

Since the launch of ChatGPT in 2022, researchers have increasingly considered how modern LLMs might enhance the classification of psychological phenomena in text (Bail, 2024). Practically, developing a text classifier using LLMs involves "prompt development" or "prompt engineering" whereby researchers iteratively refine a prompt (i.e., a human-language command) that guides an LLM toward generating desired output (Törnberg, 2024).

LLMs have been used in various types of psychological text classification: for example, using OpenAI's GPT to annotate tweets and news articles (Gilardi et al., 2023), identify violent speech (Matter et al., 2024), classify utterances (e.g., in terms of emotion) in conversations and media (Ziems et al., 2023), and replicate annotations of datasets used in published research (Pangakis et al., 2023). While LLM classifier performance varies across datasets and concepts in these studies, it is currently unclear why. For example, in tests using GPT-4 by Pangakis et al. (2023), F1-scores ranged from 0.059 to 0.969, although it is unspecified which domains these scores apply to. These studies note that LLMs show promise for psychological text classification due to their scalability, ease of use, and time- and cost-effectiveness, however they do not adequately discuss how to develop and test their validity in applied use.

LLMs have potential advantages over non-generative automated text classification methods. LLMs can be instructed using natural language, are relatively accessible, can produce rationales for their classification, and do not require training data in order to work (De Kok, 2023; Demszky et al., 2023; Rathje et al., 2023; Törnberg, 2024). This is unlike dictionary-based methods such as LIWC, which are highly accessible and interpretable, but do not capture context-specific meanings or nuances in language and fail to precisely grasp complex linguistic structures (Boyd and Schwartz, 2021). It is also unlike traditional machine learning techniques that cannot be instructed using natural language, require significant training data, and cannot be prompted for their rationale (Grimmer and Stewart, 2013; Kennedy et al., 2021). In contrast, LLMs can be used as "zero-shot" classifiers, meaning they can be used as classifiers without requiring training. Moreover, LLMs may provide a reliable way to classify textual data, for example in terms of consistency of results (Pangakis et al., 2023; Reiss, 2023), and may help with various aspects of social research including ideation, writing, and programming (Bail, 2024; Korinek, 2023).

In summary, psychological text classification using LLMs offers potential benefits over both manual and non-generative automated text classification methods. However, the full potential of LLMs in this field remains undefined, and there is a lack of established validation approaches. We investigate this in the current study.

3 The current research

Our overarching aim is to explore the potential of LLMs to validly classify psychological phenomena in text. We assume that the validity of LLM-based psychological text classification has to be established outside of the LLMs themselves. Where the reliability partly lies in the hands of LLM developers—because the quality and architecture of the model affect the consistency and reproducibility of the LLM (Törnberg, 2024)—the researcher remains solely responsible for validity. We use GPT-4o, an LLM developed by OpenAI, because our pilot testing found it to be the best-performing LLM available. We employed only prompts to instruct the model, as further training or fine-tuning of the model was beyond the scope of our study. Our aim is divided into two sub-aims that run parallel to our two-stage methodology.

3.1 Aim 1: to develop high-validity LLM prompts for psychological text classification

Our first aim was addressed in an exploratory prompt development phase that used one-third of each manually coded gold standard dataset (i.e., the development dataset). Drawing on Krippendorff's (2004) validity framework, we operationalized validity during prompt development into three types: semantic validity, exploratory predictive validity, and content validity. Krippendorff's definitions of these validity types were adopted with minor modifications to suit the current context. We defined semantic validity as qualitative evidence that the meaning of language used in the measurement protocol (i.e., the prompt) corresponds to the meaning of the intended phenomenon in the target data. For instance, prompts with low semantic validity might include ambiguous language or specialized jargon that leads to misunderstandings, while prompts with high semantic validity use unambiguous terminology and parsimonious definitions that effectively guide the model toward the desired response. We therefore sought to develop prompts that elicited the intended concept "understanding" in the LLM and, thus, were consistent in meaning with the psychological phenomenon being classified. Overall, evidence of semantic validity is assessed holistically and iteratively, based on whether the prompt meets its intended goals. Next, we defined exploratory predictive validity as quantitative evidence on the extent to which a prompt accurately predicts outcomes in the prompt development phase. This was obtained by calculating relevant quantitative evaluation metrics using the manually coded gold standard dataset as the criterion. Finally, content validity—defined as the degree to which the classifier captures all the relevant aspects of the construct—was assessed by qualitatively examining the types of errors the LLM made compared to the gold standard dataset, and reflecting on whether the LLM accurately captured the different aspects of the target construct.

3.2 Aim 2: to test the validity of the final LLM prompts

Our second aim was to assess whether our developed prompts successfully completed the text classification task on a held-out portion of the gold standard dataset (i.e., the test dataset). We did this through a confirmatory predictive validity test, using the same quantitative evaluation metrics as in the exploratory predictive validity test. Through this final validity test, we tested whether the validity of the prompts developed using the development dataset held on unseen data.

4 Methodology

This section outlines the methodology we used to explore the potential of LLMs to validly classify psychological phenomena in text. Figure 1 illustrates our approach: we developed a manual coding framework through pilot studies, manually coded the datasets (except for the harm dataset, which was already coded—see Gillespie and Reader, 2016), split them into development and test datasets, iteratively refined the LLM prompts using the development dataset, and assessed the final prompts' performance on the test dataset.

4.1 Data collection and manual coding

We operationalized three classification tasks, each based on a different type of textual data. Our data were chosen to reflect Ziems et al.'s (2023) distinction of textual data types into "utterances," "conversations," and "documents"—each having different analytical affordances. Specifically, we performed (1) span extraction and binary classification of reported speech in online diary utterances, (2) binary classification of other-initiations of conversational repair in Reddit conversations, and (3) reasoning and subsequent ordinal classification of harm due to treatment errors reported in healthcare complaint documents submitted to hospitals. In what follows, we explain the sampling and manual coding of the data (see Table 1 for more details).

4.1.1 Reported speech in personal documents

Reported speech refers to any communication (including written) that purportedly represents another speech event, regardless of its veracity (Lucy, 1993). People refer to the utterances of others in live speech (e.g., "then they said X"), in social media posts (e.g., referencing a previous post), in formal communications (e.g., quoting a source in a newspaper) and in diaries (e.g., reporting a social interaction). Reported speech has long been of interest to literary scholars (Bakhtin, 1984) and psychologists (Holt,

TABLE 1 Summary of dataset characteristics.

Dataset	Type	Source	Documents ($N = 4,500$)	Inter-rater reliability sample size (N coders = 2)	Inter-rater reliability
Reported speech	Utterance	Online diaries ($n = 10$)	Random sample of paragraphs ($n = 1,500$)	150	Agreement = 0.86 Krippendorff's a = 0.66
Other-initiations of repair	Conversation	Reddit dialogues ($n = 387$) sampled from 27 subreddits	Reddit posts and comments ($n = 1,500$)	981	Agreement = 0.95 Krippendorff's a = 0.81
Harm	Document	Healthcare complaints submitted to NHS hospitals ($n = 1,500$)	Healthcare complaints processed with OCR ($n = 1,500$)	125	Intraclass Correlation Coefficient (ICC) = 0.86

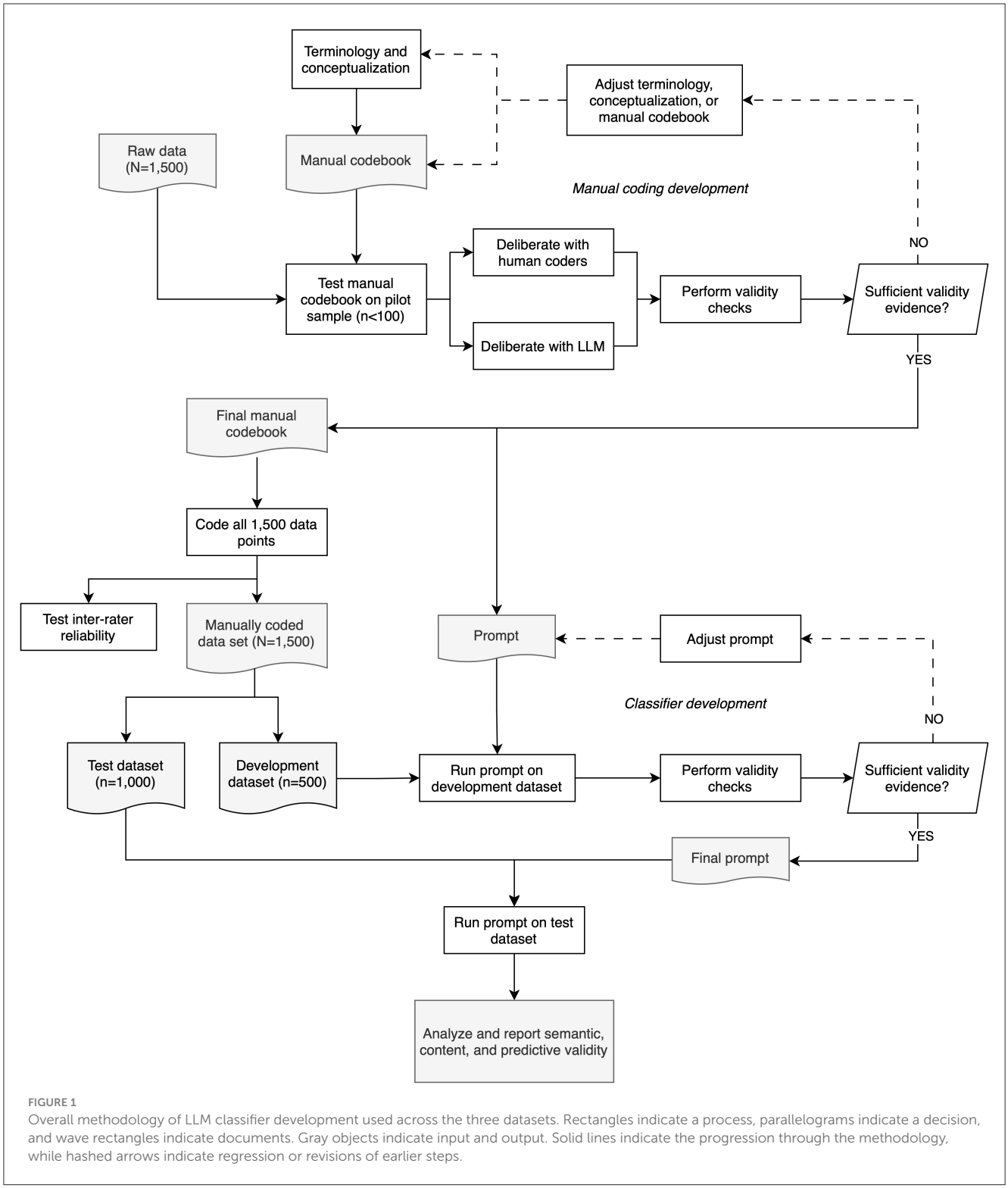

FIGURE 1
Overall methodology of LLM classifier development used across the three datasets. Rectangles indicate a process, parallelograms indicate a decision, and wave rectangles indicate documents. Gray objects indicate input and output. Solid lines indicate the progression through the methodology, while hashed arrows indicate regression or revisions of earlier steps.

2000) because it is revealing of a deep sociality in stories and in human speech. It has also been studied as a window on internal dialogues, perhaps revealing a fundamental dialogical dimension to the human mind (Aveling et al., 2015; Linell, 2009; Marková, 2016).

To ensure that diverse writing styles were represented, ten large (10+ year) diaries were sampled. The diaries were written digitally via an online platform and all diarists gave consent and provided access to their diaries. Single paragraphs in diary entries were used as the semantic unit for reported speech classification, as reported speech may span multiple sentences. Paragraphs were sampled to contain a minimum of 50 characters and a maximum of 200 words. The final dataset contained 1,500 paragraphs.

The diary excerpts were subsequently coded for reported speech by the first author upon deliberation with the team as well

as GPT (see Supplementary material for the codebook, adapted from a codebook developed by a team of five MSc level coders). This entailed extracting spans of reported speech, containing the reporting clause and reported clause. The last author coded 10%, i.e., 150 units of the dataset. The coding of reported speech showed good inter-rater reliability (agreement = 0.86; Krippendorff's $\alpha = 0.66$).

4.1.2 Other-initiations of conversational repair in online dialogues

Conversational repairs are defined as sequences of dialogue aimed at resolving problems of misunderstanding or miscommunication (Dingemanse and Enfield, 2015). Identical language can be used in various ways, leading to misunderstandings and necessitating constant maintenance through "conversational repair" (Healey et al., 2018; Reddy and Ortony, 1979; Schegloff et al., 1977). Conversational repairs appear to be a universal feature of dialogue, where evidence suggests clarification requests occur on average every 1.4 min (Dingemanse and Enfield, 2015). However, conversational repairs have been understudied in computer-mediated communication (Meredith, 2020; Meredith et al., 2021), despite their importance to high-quality online dialogues (Goddard and Gillespie, 2023). In this study, we identify other-initiations of conversational repairs using clarification requests (Dingemanse and Enfield, 2015).

Repairs occur in face-to-face dialogue between participants in a linear sequence (Schegloff et al., 1977). Online comment threads have an equivalent organization, where participants can directly reply to each other. We sampled from Reddit, an online forum and community platform organized into micro-blogs called subreddits, each with their own rules and norms (Chandrasekharan et al., 2018). Using Reddit's application programming interface (API), data was sampled in November 2021 from 27 subreddits, chosen to provide a spread of topics (e.g., r/movies, r/science, and r/coronavirus) and the media focus of the subreddit (e.g., r/pics requires posting a picture, and r/jokes a text-based joke). The dataset was built by first randomly sampling each subreddit's recent posts, then collecting all the comments and parent posts.

Two independent coders scored the Reddit posts and comments for the presence of other-initiations of repair (Krippendorff's $\alpha = 0.81$). An existing coding scheme for identifying other-initiations of repair (Dingemanse and Enfield, 2015) was adapted for online contexts, and the two coders performed eight pilot rounds (following Figure 1) prior to scoring the complete dataset (N adjacency pairs = 1,500) sampled from 387 Reddit threads. An adjacency pair is a mutually dependent set of dialogue turns (Sacks et al., 1974). Here, this included an utterance from one participant (denoted as Speaker A) with a response from another participant (denoted as Speaker B).

4.1.3 Severity of harm reported in patient complaints about poor healthcare experiences

Our final concept of interest is harm caused by medical errors (i.e., safety incidents), as described in letters of complaint submitted to NHS hospitals and trusts. Harm is defined as "the overall harm caused to patients by the problems raised in the letter of complaint" (Gillespie and Reader, 2016, p. 13). The Healthcare Complaint Analysis Tool (HCAT) includes an assessment of harm with the aim to reveal hot spots and blind spots in the quality and safety of healthcare. HCAT adopts the UK National Reporting and Learning System's risk matrix to assess harm (Gillespie and Reader, 2016, 2018).

We used healthcare complaints collected and analyzed in the development and validation of HCAT (Gillespie and Reader, 2016). Healthcare complaints provide insights from patients into hospital care where harm and near misses occur during the care journey. As such, they are important to systematically analyze to give insights into the quality and safety of healthcare and point to tangible areas of improvement. Freedom of Information requests were used to obtain access to healthcare complaints from 56 randomly sampled NHS hospitals and trusts. Two trained coders (MSc level graduates who received five hours of training) coded complaints using the HCAT manual (see Supplementary material), including harm, the latter of which showed excellent inter-rater reliability (ICC = 0.86).

Any personally identifiable information within the complaints was manually redacted by hospital staff before being shared in PDF format. As a result, the text from the PDFs had to be extracted. For this purpose, Google Cloud Vision's optical character recognition (OCR) was applied to 2,235 healthcare complaints. This yielded somewhat noisy textual data, with incomplete sentences due to redaction or filled with non-alphanumeric characters. However, we used ChatGPT to query its understanding of the text, which we deemed sufficient for an LLM-based psychological text classification. The 1,500 complaints with the least redactions were used for the analysis.

4.2 Splitting up the gold standard dataset

We used GPT-4o (specifically, gpt-4o-2024-05-13) in a zero-shot setting using prompt development, meaning that we did not perform additional training or fine-tuning for our classification tasks. In our Python pipeline, the LLM iterated over single units of analysis, thus denying the model any knowledge of prior inputs or outputs (i.e., each instance of coding was unaffected by the other instances). While reasonable insight can be gained from working with a small sample of a dataset (e.g., between 25 and 50 units of analysis), prompt development may involve many iterations over a dataset before a final prompt that completes a designated task with high accuracy is established. However, this poses a risk of overfitting, where the developed prompts do not generalize to unseen data.

Accordingly, we split each manually coded (gold standard) dataset into two datasets prior to prompt development. These datasets are akin to the "training" and "validation" datasets used in the development of supervised machine learning models (e.g., Song et al., 2020). In supervised machine learning, the training dataset is used to develop an algorithmic model, which is subsequently evaluated on the test dataset to avoid overfitting or underfitting the model to one's data. In the

present case using LLMs, we used the training dataset for iterative and exploratory prompt development, while we used the test dataset for confirmatory predictive validity testing. The separation between exploratory and confirmatory stages enables scientific rigor, as the former stage facilitates inference to the best measurement protocol (informed by the data's affordances and the prompt development outcomes), and the latter stage increases confidence in the overall validity of our approach by testing the prompts on unseen data (see Rubin and Donkin, 2022).

For the current purposes, we term the subsets of a manually coded dataset the "development" and "test" datasets. The ability to conduct prompt development with relatively little data and the importance of validation point to the need for the dataset split sizes be weighted toward validation. Therefore, we split our manually coded datasets into development and test sets of 500 and 1,000 units of analysis respectively. The reported speech and other-initiations of repair datasets were stratified, meaning the proportion of reported speech coded for in diary excerpts and other-initiations of repair in Reddit dialogues were equal in both sets (24% and 8% respectively). Similarly, the harm dataset was split such that the proportion of classifications on the ordinal scale was similarly distributed in both datasets.

4.3 Stage 1: iteratively developing the validity of LLM prompts

The first stage of our methodology involved iterative prompt development using the development dataset. Prompts are human-language instructions for the LLM, written to guide the model in generating responses by framing the task and specifying the expected output format, if applicable. Prompt development is a highly iterative process where one explores how to verbally instruct the LLM to complete a certain task. Here, the initial prompts drew on codebooks used for manual psychological text classification.

In the process of prompt development, we iteratively refined prompts with the aim to establish three types of validity: semantic validity, exploratory predictive validity, and content validity. We assumed that a prompt could only be considered valid with sufficient evidence of these three types of validity. Key to prompt development is the recursive refinement of the terminology used and instructions given to the LLM based on pushback from the model. For example, we altered the language of the terminology beyond what was in the original codebooks to elicit perceived correct "understanding" in the LLM. The manually coded data itself stayed the same throughout. Any revisions to the terminology and/or prompt were concluded in the prompt development phase before confirming the validity on the test dataset.

4.3.1 Semantic validity

Prompt development is an important time to explore one's language use, specifically through qualitatively assessing whether the terminology and definitions in the prompt elicit the intended concept "understanding" in the model. In other words, we assessed the semantic validity of our prompts, iteratively refining our language such that the model is perceived to correctly capture the meaning of the intended concept. Because every iteration over the data yields immediate feedback on whether the prompt achieves this goal, a relatively small subset of the development dataset (i.e., between 25–50 rows) can be used to ascertain evidence of semantic validity.

4.3.2 Exploratory predictive validity

Next, we calculated evaluation metrics on the development dataset to provide preliminary quantitative insight into the LLM's ability to accurately mirror human coding behaviors, thereby assessing exploratory predictive validity. The choice of evaluation metric was determined by the data type and annotation task. For reported speech, we integrated a fuzzy matching function to gauge whether the human coder and GPT extracted similar spans of reported speech. For ease of comparison, however, the final evaluation metrics used a binary variable representing whether both the human and GPT coded for reported speech in a diary paragraph. A confusion matrix was produced to assess the accuracy, precision, recall, and F1-scores between the model and human coding. Precision refers to the ratio of positive predictions that are correct, while recall measures the ratio of actual positive cases identified correctly. The F1-score is a harmonic mean of precision and recall, offering a single balanced measure. The same evaluation metrics were applied to compare the presence of other-initiations of repair (binary) in human coding and LLM coding. For the harm classifications, we report agreement and a weighted kappa, the latter of which is a measure of the agreement between predictions and actual values, considering both the quantity and the severity of disagreements, making it useful for ordinal classifications (De Raadt et al., 2021).

Aiming for the highest F1-score or weighted kappa possible, we kept track of the evaluation metrics throughout prompt development to identify whether a prompt required further iteration. Generally, we took low scores to indicate that a prompt achieved inaccurate classification results compared to the human coding, spurring further iterations over the prompts. We took high scores to indicate accurate classification results compared to the human coding, providing us reasons to assume that the prompt's performance was sufficient. Moreover, in the case of binary classification, we used the precision and recall of the target variable to steer the edits to the prompt. Low precision indicates a high number of false positives and low recall a high number of false negatives. We therefore assumed the former meant the prompt was too broad in its scope and definition, resulting in us making it more specific (e.g., more detail in definition). In contrast, we assumed the latter meant the prompt was too specific, resulting in a broadening of its scope (e.g., removal of detail in definition). This process was repeated until we achieved a balance between precision and recall, whilst seeking to maximize them.

4.3.3 Content validity

We gathered evidence of content validity through a systematic and qualitative analysis of the LLM's classifications of the full development dataset. For the binary classifiers (reported speech and other-initiations of repair), this included a manual inspection

of the LLM's false positives (where the model incorrectly classifies instances of the target phenomenon) and false negatives (where the model fails to identify correct instances). For the ordinal classifier (harm classification task), we evaluated the skewness of discrepancies, and whether the LLM rated structurally higher, lower, or in similar distributions compared to the human coder. Edge cases (instances where the accuracy of classification is ambiguous) were also identified, as they could reveal tensions between the human and the LLM regarding the interpretation and application of criteria for classification. Finally, we aimed to identify any potential biases in both human and LLM coding.

The aim of the content validity checks was to understand the nature of discrepancies between human coding and LLM coding. If disagreements in coding pointed to structural issues that were deemed necessary to rectify, we conducted further iterations over the prompt until satisfactory results were reached or the LLM's performance had plateaued (as indicated by the quantitative evaluation metrics). If not, this step was the final one to undertake on the development dataset, closing the iterative prompt development loop.

4.4 Stage 2: testing the validity of the final LLM prompts

With the final prompts developed, we moved onto the second stage of the methodology. This entailed running the prompts on the test dataset, which the LLM had not "seen" before. This stage served to evaluate the generalizability and effectiveness of the prompts on new data, ensuring that the prompt had generalized beyond the development dataset. Again, we calculated evaluation metrics to gauge the LLM's performance and provide an understanding of agreement or correlation between the LLM's and human coders' classifications on the test dataset. The obtained results were used as the final indicators of the LLM's performance in performing a certain text classification task as compared to the human coding.

4.5 Prompt design

OpenAI's API offers the possibility to assign the LLM roles, where the "user" role defines the prompt, the "system" role defines high-level instructions, and the "assistant" role provides an example of the model's response. We opted to make use of this functionality, and provided GPT-4o with a designated persona defined in the system role, priming its role as a research assistant in the relevant classification task. This approach, which we implemented as an exploratory measure, aligns with OpenAI's best practices for prompting (OpenAI, 2023). We did not systematically compare its performance against prompts without a defined system role.

Due to the scarcity of peer-reviewed research on prompt engineering for our specific tasks when we began our study, we relied on OpenAI's (2023) best practices to guide our prompt development. We opted for prompts that were simple by design and applicable across domains rather than exhaustive in covering additional dimensions to our prompts (e.g., chain-of-thought prompting, Wei et al., 2022). In doing so, we explored variants of "zero-shot" and "few-shot" prompts (e.g., Patil et al., 2024).

Our prompts (defined in the user role) take the following structure: (1) concept definitions with either "minimal" or "maximal" expansion of detail, and (2) zero or a few examples of the concept of interest. This yielded a two-by-two framework with four possible types of prompts: MinZero (a prompt with minimal concept definition and zero examples), MinFew (a prompt with minimal concept definition and a few examples), MaxZero (a prompt with maximal concept definition and zero examples), and MaxFew (a prompt with maximal concept definition and a few examples). We excluded MinFew from our prompting framework to maintain a linear progression in prompt complexity (as arguably, a MinFew prompt is more complex and comprehensive than a MaxZero prompt). Doing so enabled a clear and parsimonious comparison across prompt types. Examples of reported speech and other-initiations of repair were included at the discretion of the individual prompt developers, guided by their assessment of representativeness and conceptual alignment. Examples of every level of harm were directly taken from the NPSA (2008) risk matrix.

All prompts contained step-by-step instructions with specific directives guiding the model's tasks or prescribed output formats (e.g., "Respond with 'YES' if …"). We also provided additional structure to our prompts by using headings to succinctly delineate sections (e.g., "Definition: …," "Instructions: …"). These task instructions were consistent for each prompt per classification task. Beyond verbal instructions, one can set additional parameters when working with an LLM's API (Demszky et al., 2023). For example, the LLM's "temperature" is a key parameter that controls the randomness of the model's output (OpenAI, 2024). A lower temperature results in more deterministic output, while a higher temperature leads to more diverse and unpredictable output. We set the temperature to 0 to maximize the reliability of the LLM responses.

5 Results

We present and analyze our findings based on the two-stage methodology that we developed to explore the potential of GPT-4o to validly classify psychological phenomena in text. The first stage aimed to develop valid prompts using the development dataset; the second stage aimed to test the validity of the final prompts on the test dataset. We address qualitative and quantitative insights gained as part of our prompt development and discuss implications for using GPT-4o in psychological text classification.

5.1 Aim 1: to develop high-validity LLM prompts for psychological text classification

5.1.1 Semantic validity

Semantic validity was developed over many iterations of prompt development. During this process, we aimed to identify what language-use (related to the psychological phenomenon as

well as task instructions) would elicit the intended "understanding" in GPT. We found that conversing with and interpreting output from the LLM proved crucial to this end. In our use cases, this collaborative process resulted in us tweaking coding frameworks and revising the language in our prompts to aid classification performance (e.g., abandoning academically-preferred terminology). We describe the development of semantic validity for each classification task in detail below, and provide the final prompts in the Supplementary material.

First, during the development of the manual coding framework, we found that GPT coded performative verbs (e.g., to invite, to promise, to apologize) as reported speech, whereas these were initially not included in the manual codebook due to oversight. Upon deliberation with the team, it was decided that performative verbs should have been included in the definition of reported speech. This led to a minor change to the manual codebook and another iteration of manual coding. An example of a performative verb, presented as reported speech, was later also added to the examples included in the MaxFew prompt. Moreover, when developing the prompts for reported speech classification, we discovered that GPT's "understanding" of reported speech was in close alignment to our own. Elaboration or redefinition of reported speech in fact reduced performance. We therefore did not add our own definition of reported speech to the prompts.

Second, we abandoned the initial term "other-initiations of repair" for our prompts. Due to the unsatisfactory performance of early prompts, we used ChatGPT to generate definitions of clarification requests based on unlabelled raw data, create permutations of a previous prompt, explain edge cases, and explain its own misclassifications. The manually coded data was unaltered during this process. Through the interaction with ChatGPT, we identified that the term "clarification requests" improved the model's performance compared to the more academic and specific term "other-initiations of repair." Repairs require researchers to pay close attention to what participants are doing in the text itself (Schegloff, 2007) and it may be that "clarification requests" prompts GPT to infer based on structural features of the text (i.e., a type of question) rather than a subjective explanation of the dialogue. The change in terminology caused GPT's performance to increase (from F1 = 0.61 to F1 = 0.77), demonstrating the importance of semantic validity in enhancing the tool's performance in detecting other-initiations of repair.

Third, we completed many iterations over the prompts to instruct GPT to ordinally scale an assessment of harm for the hospital complaints. During this process, we concluded that both our data quality and complexity of the task were negatively affecting GPT's classification performance. We improved performance by including a data cleaning prompt in our script, asking GPT to fill in the gaps of our raw data (Google Cloud Vision OCR output) with the most appropriate named entity placeholder and clean the text such that it became more legible. We manually verified that the meaning of the original complaints was preserved. Despite the prompt increasing in performance, we noted that GPT still had difficulty with the harm terminology originally used the Healthcare Complaints Analysis Tool (see Gillespie and Reader, 2016). It struggled to distinguish harm caused by healthcare staff's negligence or accidents from the harmful effects of the condition that led a patient to hospital admission. Also, we observed that GPT tended to apply harm as an overall rating of a healthcare experience (e.g., to describe the outcome of a series of small cumulative errors), and these may have been challenging for GPT to recognize and grade. We experimented with different language in our prompts, leading us to abandon the harm terminology and instead ask the model to assess the severity of healthcare incidents, highlighting the causal attribution implicated in the conceptualization of harm.

TABLE 2 Exploratory predictive validity test results.

Dataset	Prompt type	Exploratory predictive validity
Reported speech	MinZero	Precision = 0.95 Recall = 0.59 F1-score = 0.90
	MaxZero	Precision = 0.90 Recall = 0.64 F1-score = 0.91
	MaxFew	Precision = 0.93 Recall = 0.67 F1-score = 0.92
Other-initiations of repair	MinZero	Precision = 0.67 Recall = 0.86 F1-score = 0.89
	MaxZero	Precision = 0.67 Recall = 0.80 F1-score = 0.89
	MaxFew	Precision = 0.86 Recall = 0.70 F1-score = 0.91
Harm	MinZero	Agreement = 0.34 Weighted k = 0.52
	MaxZero	Agreement = 0.34 Weighted k = 0.51
	MaxFew	Agreement = 0.33 Weighted k = 0.47

5.1.2 Exploratory predictive validity

At the end of the prompt development process for all three text classification tasks, we ran our evaluation metrics over the complete development datasets (n = 500). For reported speech and other-initiations of repair, we list the performance of different prompts in terms of the precision and recall of positive cases, as well as the weighted F1-score. We report the results in Table 2.

5.1.3 Content validity

Finally, we conducted qualitative content validity checks to assess the effectiveness of our prompts in executing the intended text classification task and covering all relevant aspects of the intended construct. This was done through a manual inspection of discrepancies between human and LLM coding. For our

TABLE 3 Content validity test results of best-performing prompts per dataset.

Dataset	Content validity checks for best-performing prompt
Reported speech	Discrepancies between human coding and GPT coding: Edge cases: it is unclear whether the human coder or GPT was correct. False positives: missed cases of reported speech by the human coder; mistakes by GPT where the extracted sentence does not contain reported speech; instances where GPT codes what was "not" said (e.g., "I'm not going to say X"), which was excluded as reported speech in the codebook for human coders. False negatives: missed cases of reported speech by GPT, some missed cases are more obvious than others.
Other-initiations of repair	Discrepancies between human coding and GPT coding: Edge cases: it is unclear whether the human coder or GPT was correct. False positives: cases coded by GPT where Speaker A asks for clarification, not Speaker B, cases where corrections are coded as clarification request, and cases where clarification requests are put in a sentence form, not question. False negatives: missed cases of clarification requests by GPT; rhetorical questions were missed as cases of clarification requests.
Harm	Discrepancies between human coding and GPT coding: Regular overestimation of harm by GPT compared to the human coding. Mentions of adverse events (whether these occurred as a result of healthcare incidents or not) led to high harm ratings, even when these were not due to healthcare incidents. GPT barely codes for "no harm," whereas human coders regularly did so. GPT appears to struggle with recognizing that no harm was done in hospital.

reported speech and other-initiations of repair classifications, these discrepancies concerned false positives, false negatives, and edge cases. For our harm classification, these discrepancies revolved around the numerical difference or "discrepancy value" between the harm category as assigned by humans and GPT, as well as a qualitative interpretation of the reasoning behind GPT's coding. We discuss some patterns we have inferred from working with GPT on all three datasets. We report on dataset-specific observations in Table 3.

First, in instances of binary classification (i.e., indicating whether an instance of the construct of interest is present), both human coders and the LLM occasionally overlook an instance of the construct. While human oversight can stem from limited attention span or coder fatigue, the LLM might miss instances due to the task's complexity or characteristics of its original training data. In any case, we found it beneficial to inspect the LLM's false negatives to assess how often the concept of interest was overlooked and determine whether this has implications for the coding framework or coding process (for example because of different construct understandings among human and LLM). In inspecting the LLM's false negatives, we learned that it is implied that one might be looking at a human false positive in the gold standard data set, for instance, classifying something as reported speech where, in hindsight, this should not have been the case. Similarly, the LLM's false positives sometimes concerned misclassifications which either followed from a failure to follow our instructions, or where it classified a positive instance of a concept where, upon reflection, we considered this false. The content validity checks therefore provided an opportunity to reflect on our gold standard dataset. Considering the overall performance of the LLM classifier, however, we consider the extent of misclassifications in both the reported speech and other-initiations of repair datasets to be minor.

Second, upon inspection of the discrepancies between human coding and LLM coding in all datasets, we found that a number of GPT's "misclassifications" concerned edge cases. These are instances that sit on the boundary of definitions and challenge the classification norms. A paraphrased example of a reported speech edge case from our dataset is, "my friend thinks the concert is going to be sold out," which presents ambiguity as, without context, it could be interpreted as either a direct quote or a summary of the friend's thoughts, only the first of which would be considered reported speech according to our coding framework. Inevitably, edge cases will present themselves in psychological text classification, and we found that GPT helped in recognizing and highlighting some of these instances.

Third, we encountered situations where GPT, despite explicit instructions, did not adhere to our desired output structure, revealing a potential divergence in understanding between humans and LLMs. For example, we instructed GPT to code rhetorical questions as a type of clarification request, yet it consistently ignored this instruction, resulting in false negatives. Similarly, we found GPT to consistently overestimate harm in our dataset of healthcare complaints, often seemingly based on the severity of the medical situation (e.g., death; patients experiencing the consequences of grave health conditions) rather than harm caused by healthcare staff's negligence or accidents. This suggests that LLMs may struggle when imposed definitions or constructs are not aligned with its semantic network, or if events have a high degree of nuance (e.g., complaint about the manner of a death, rather than the death itself) or relate to a series of incidents that are told in a non-linear fashion recognizable to human coders, but not necessarily GPT (e.g., complaints that list a series of mistakes or encounters that are sequenced over a year-long timeline). Our experiences underscore the importance of conceptual clarity when classifying psychological phenomena in text. However, this does not imply that human understanding is flawed or needs adaptation if discrepancies arise. Instead, these discrepancies could be due to the unique semantic network of the LLMs and their training processes and biases (De Kok, 2023). Thus, it is crucial to consider the LLM's unique semantic network when designing prompts and reflect on the structural discrepancies that may affect the classifications' content validity.

TABLE 4 Confirmatory predictive validity results.

Dataset	Prompt type	Confirmatory predictive validity
Reported speech	MinZero	Precision = 0.88 Recall = 0.65 F1-score = 0.89
	MaxZero	Precision = 0.82 Recall = 0.74 F1-score = 0.90
	MaxFew	Precision = 0.85 Recall = 0.77 F1-score = 0.91
Other-initiations of repair	MinZero	Precision = 0.72 Recall = 0.84 F1-score = 0.91
	MaxZero	Precision = 0.68 Recall = 0.78 F1-score = 0.89
	MaxFew	Precision = 0.91 Recall = 0.64 F1-score = 0.91
Harm	MinZero	Agreement = 0.34 Weighted k = 0.57
	MaxZero	Agreement = 0.34 Weighted k = 0.51
	MaxFew	Agreement = 0.33 Weighted k = 0.49

5.2 Aim 2: to test the validity of the final LLM prompts

This section presents the results of the study focusing on the confirmatory predictive validity of using GPT-4o for psychological text classification. As such, we tested whether our prompts generalized to unseen data. We summarize our findings in Table 4.

5.2.1 Reported speech

The reported speech classification results reveal distinct, though not drastically different outcomes for each prompt. The MinZero prompt, despite its brief definition and lack of examples, achieved a precision of 0.88, indicating a high rate of accurate positive predictions. However, its recall was 0.65, suggesting it failed to identify roughly a third of the actual instances of reported speech, yet the overall F1-score was still an excellent 0.89, presumably due to the natural distribution of reported speech in the dataset (24% of cases are positive, and many of the negative cases as coded by human coders were correctly classified by GPT). MaxZero offered more balance, with a precision of 0.82 and a recall of 0.74, culminating in an F1-score of 0.90. The MaxFew prompt achieved a precision of 0.85 and the highest recall of 0.77, attaining the best F1-score of 0.91, and indicating the most effective performance in classifying reported speech within the test dataset.

5.2.2 Other-initiations of repair

Turning to the classification of other-initiations of repair, the MinZero prompt showed a precision of 0.72, being a good performance in accurately identifying correct instances. Its recall was 0.84, and the weighted F1-score stood at an excellent 0.91. The MaxZero prompt resulted in lower precision and recall of 0.68 and 0.78 respectively, with an F1-score of 0.89, again due to GPT correctly classifying the high rate of negative cases in the manually coded dataset, positively skewing the F1-score. Finally, the MaxFew prompt had a remarkably high precision of 0.91 and a moderate recall of 0.64, suggesting room for improvement in capturing all relevant instances of other-initiations of repair. The F1-score stood at 0.91.

5.2.3 Harm

For the ordinal harm classification, the MinZero prompt, which did not include an explicit description of the full ordinal scale, had an agreement score of 0.34 and a weighted kappa of 0.57, indicating a surprising moderate level of agreement between human coding and LLM coding. The MaxZero prompt, while more detailed, scored lower in both agreement and weighted kappa, with values of 0.34 and 0.51 respectively. The MaxFew prompt, which included additional examples, achieved similar agreement to MaxZero at 0.33 but a slightly lower weighted kappa of 0.49. Interestingly, the MinZero prompt's higher scores suggest that additional detail in the MaxZero and MaxFew prompts did not correspond to improved alignment with human coders in ordinal harm classification.

5.2.4 Observations on prompt types

We observed no consistent performance difference among the MinZero, MaxZero, and MaxFew prompts across the three datasets. Whilst some runs suggested an escalating performance from MinZero to MaxFew, other runs over the same dataset showed the opposite pattern, clearly showcasing the stochastic nature of GPT (Bender et al., 2021). Accordingly, the appropriate level of concept definition and use of examples may depend on the psychological phenomenon in question. Thus, we recommend testing various prompt designs to ascertain what works best in any given context.

6 Discussion

Our central aim was to explore the potential of LLMs to validly classify psychological phenomena in text. To this end, we first conducted an iterative prompt development stage, during which we developed the prompts' semantic validity, exploratory predictive validity, and content validity. We then performed confirmatory predictive validity testing on our test dataset to assess whether our developed prompts can successfully complete a text classification task on unseen data. Our results contribute support to Bail's (2024) suggestion that LLMs can serve as valid coders of psychological phenomena in text. This rests on the condition that the researcher works with the LLM to ensure semantic, predictive, and content validity. Moreover, our study reveals a secondary contribution: LLMs can foster an intellectual

partnership with researchers, enabling a dynamic and iterative text classification process where the generative capabilities of LLMs assist in conducting essential validity checks. This highlights the potential of LLMs as collaborative partners in psychological research, suggesting a paradigm shift toward more interactive and reflexive methods in text classification.

Specifically, we observed that small changes in terminology can unlock the LLM's correct understanding of the intended psychological phenomenon, demonstrating the importance of examining the prompts' semantic validity. A prompt with high semantic validity closely corresponds in meaning to the intended psychological phenomenon in the target data. To achieve semantic validity, we must challenge the presumption that humans are the gold standard, recognizing instead that the LLM's unique design may require adapting our preferred terminology. In addition, our analyses suggested that the prompts' content validity is at least in part contingent on the simplicity, direct observability, and common-sense nature of the psychological phenomenon in question. Indeed, reported speech and other-initiations of repair fulfill these criteria. Here, LLM-classifications had high accuracy and discrepancies in coding were attributable to relatively few LLM and human errors as well as edge cases. However, LLM-classifications were less satisfactory for the concept of harm, highlighting current limitations and biases in the LLM as well as intrinsic methodological complexities in psychological text classification. Finally, the exploratory predictive validity and confirmatory predictive validity tests serve as quantitative checks as to the agreement between human coding and LLM coding at two different stages in the text classification process.

Our prompts exhibited excellent exploratory and confirmatory predictive validity when instructed to classify reported speech in online diaries and other-initiations of repair in Reddit dialogues. They exhibit moderate exploratory and confirmatory predictive validity when ordinally classifying harm caused by safety incidents described in healthcare complaints. Overall, LLMs show great promise for scaling up psychological text classification when used in a synergistic and iterative approach on big qualitative datasets. We now discuss implications for the methodology itself, as well as the field of psychology at large.

6.1 Implications

Our results contribute insight into what types of concepts and constructs are currently measurable with LLMs. Specifically, we observe a marked performance difference in the classification of reported speech and other-initiations of repair on the one hand, and the classification of harm on the other. The former two are more directly observable concepts, while the latter is a complex ordinal construct for which much contextual inference is required, and which involves nuanced judgments on cause and effect. Despite our efforts to describe the construct of harm in such a way that it was constituted of measurable elements, GPT-4o's harm classification performance remained moderate. This highlights the varying degrees of measurability in different phenomena, suggesting that more directly observable phenomena are more readily classified by GPT, while concepts and constructs for which higher levels of inference are required present challenges (e.g., extracting reported speech vs. inferring harm from a sequence of events). Even though reported speech and other-initiations of repair are conceptually complex, they were easier to measure than harm, perhaps because the LLM has a better inherent semantic understanding of those concepts. Harm caused by medical error is not only a subjective construct, but also must be attributed to an event or cause—and distinguished from the harm that led the patient into the hospital. While future advancements in LLMs' developments might enable classification of more complex phenomena, we maintain that it remains advisable to operationalize concepts or constructs in simple, empirical terms with the underlying conceptualization clearly spelled out to ensure validity (see also Bergner, 2024; Bringmann et al., 2022; Krpan, 2022).

Our results highlight the importance of considering the use of LLMs for psychological text classification as an "intellectual partnership" between humans and technology, where "results greatly depend on joint effort" (Salomon et al., 1991; p. 3). We argue that the LLM's added value in psychological text classification lies with its afforded use as a "companion" with which a researcher can engage in exploratory, synergistic loops over their conceptualizations and operationalizations. Doing such recursive refinements is indeed a process that can be done with human iteration alone, yet is considerably faster and more cost-effective with an LLM (Bail, 2024; De Kok, 2023; Korinek, 2023). In contrast, LLMs may provide independent and instant feedback, for example in deliberative dialogue using an interface such as ChatGPT, or by pushing back on the concept or task instructions it was given. Such feedback makes LLMs stand out from other classification methods, such as traditional supervised machine learning, that cannot be used to generate explanations and rationales for the ways they categorize text. However, LLMs are not unbiased (Bail, 2024; Ray, 2023), and various scholars have warned against the risks of using LLMs in qualitative research (Beghetto et al., 2024; Lindebaum and Fleming, 2024; Roberts et al., 2024). Moreover, the intellectual partnership requires caution to avoid potential pitfalls; for example, researchers might unintentionally tailor their coding manuals to align too closely with the LLM's outputs, creating a feedback loop that compromises the objectivity of the validation process.

Working with an LLM in psychological text classification can also support abduction, where the dynamic adjustments and the incorporation of new understandings in real-time can enhance the robustness and relevance of the research findings (Gillespie et al., 2024; Grimmer et al., 2022). Abduction is a mode of logical inference, alongside induction and deduction, that is focused on generating and refining theory based on surprising findings or discoveries (Peirce, 1955; Tavory and Timmermans, 2014). Inductive analysis typically involves the bottom-up generation of theories based on specific observations, while deductive analysis involves testing hypotheses based on existing theories. Inductive and deductive approaches are most commonly used in the social sciences, yet they are relatively unresponsive to new insights (Grimmer et al., 2022). In contrast, working with an LLM presents the opportunity to do recursive refinements of the manual coding framework, gold standard dataset and prompts. This is because immediate feedback from LLMs, even after runs over a few units of data, can reveal if one's conceptualization or operationalization is fuzzy or ineffective. This makes the use of LLMs in psychological

text classification stand out in comparison to both manual and non-generative automated approaches, neither of which enable the same number of iterations for the same effort or price. Thus, in so far as iterating on the data leads to improved validity of a text classification protocol, the speed and affordability of using LLMs to classify text enables more rapid and frequent iteration, which in turn can improve the overall quality of both the measurement protocol and the conceptualization of the concept being measured.

In practice, this requires researchers to refine and adjust their operationalizations of psychological phenomena based on exploratory testing during the prompt-development stage. This is a dynamic, iterative process rather than a static, one-off task, and stands in contrast to the deductive approach to developing text classifiers. A strict deductive approach dictates that we cannot look beyond our pre-established theories and hypotheses, and thereby limits creative and exploratory testing (Grimmer et al., 2022; Rubin and Donkin, 2022). Instead, working abductively with LLMs for psychological text classification implicates that exploratory prompt development can be conducted on the development dataset with two aims. The first aim is for the human to learn the terminology of the LLM for talking about the phenomenon of interest, and thereby improving LLM classification performance. Through deliberation with the LLM and reflecting on its pushback regarding our concept operationalizations, the second aim is to achieve conceptual clarity on one's psychological phenomenon of interest, being an important area of development to advance psychology as a science (Bringmann et al., 2022; Krpan, 2022). It is important to stress that the final prompt should be validated on the test dataset (withheld from development) to ensure that the prompt does not overfit or underfit the psychological phenomenon of interest, as well as to safeguard the integrity of the analysis (Rubin and Donkin, 2022; Song et al., 2020).

The abductive affordances of an LLM-based psychological text classification highlight problems with using statistical evaluations of validity without qualitative oversight. For example, we found that cases labeled as "false positives" or "false negatives" may be part of systematic biases or errors in either human or LLM coding, which only can be revealed after a comprehensive qualitative analysis of these cases, a point which has gone underappreciated in the current literature (Pangakis et al., 2023; Rathje et al., 2023; Törnberg, 2024). Such qualitative checks enable researchers to gain a clearer understanding of what constitutes correct coding by the human or LLM, and presents opportunities for learning (Gillespie et al., 2024). Indeed, while acknowledging that both humans and LLMs make errors, certain discrepancies may, upon careful consideration, warrant addressing at their problem source. As noted above, not only is the pace at which these recursive adjustments can be completed unprecedented, they are also crucial to improve the final semantic, predictive, and content validity of one's LLM-based psychological text classification.

In summary, there is high potential for LLMs to improve the validity of psychological text classification. Specifically, due to their abductive affordances, LLMs have the potential to advance conceptualization and operationalization of psychological phenomena in text, being promising tools for addressing validity challenges (Bringmann et al., 2022; Flake et al., 2022; Flake and Fried, 2020). Key to this process is an "intellectual partnership" defined by a synergistic and recursive text classification process where the LLM augments human capabilities, functioning as a sense-check of whether the phenomena we are interested in are measurable in the target data. Only after achieving a well-defined conceptual framework that is "understood" by the LLM, can we fully leverage the potential of LLMs. This includes their capacity to scale up psychological text classification through the use of large, qualitative datasets. Such scaling up offers a promising avenue for achieving greater external and ecological validity, thus facilitating a more comprehensive and real-world understanding of psychological phenomena (Chang et al., 2021; Gillespie et al., 2024; Jackson et al., 2022; Kjell et al., 2019).

6.2 Practical considerations

LLMs do not make manual coding redundant. Manually coded data is required to create the gold standard dataset that is central to both the development and testing stages (Grimmer and Stewart, 2013). When creating the manually coded dataset, the natural distribution of the psychological phenomenon should be considered: a lower ratio of positive to negative cases demands a larger dataset than when the positive and negative cases are equally present. The associated time and resource investment that manual coding requires may render this methodology less suitable for small-scale research projects. Instead, the methodology's strength particularly lies with its application to large, qualitative datasets, once the prompt(s) are developed and validated. That said, one may decide to merely explore the conceptualization and operationalization of a given psychological phenomenon in a small sample of text so as to abductively deepen one's understanding of it. This approach is a highly appropriate way of leveraging the strengths of LLMs on a smaller scale. We did not formalize this process, however, and future research should explore how LLMs can be rigorously integrated in the manual coding process prior to developing classifiers.

6.3 Limitations

There are currently several general limitations to the use of LLMs for text classification. Although commercial proprietary LLMs (such as GPT-4o) are much more cost-effective than manual coding, they still do incur a small cost. Proprietary LLMs are also closed-source, which means that it is difficult to control biases within the models and ensure reproducibility, as changes to the model are opaque (Bail, 2024). Moreover, proprietary LLMs incorporate built-in guardrails that can impact classification performance by, for example, steering away from sensitive topics or, in some cases, refusing to classify text that violates company policy altogether. Hopefully, open-source LLMs tailored to academic research will be developed which will allow better investigation of bias, version control, and reproducibility. In addition, ethical considerations remain complex when applying LLMs to personal or sensitive data, such as healthcare complaints, given the risk that such data could inadvertently become part of the LLM's training data (Törnberg, 2024). Working with LLMs also requires minimal coding skills. These are needed to operationalize the LLM, to

prepare text data for analysis, and evaluate the model's output. However, this is significantly less complex code than when training a machine learning model. Moreover, LLMs such as ChatGPT increasingly have the ability to generate high-quality code, and thus guide users in the coding necessary to use them (De Kok, 2023). Whilst this affordance does not eliminate the need to understand and troubleshoot the generated code, it certainly does help to democratize computer programming.

Moving onto our own methodological limitations, there are four shortcomings associated with our choice of using GPT-4o. First, we have obtained insights into the workings of GPT-4o (as well as GPT-3.5 and GPT-4 Turbo to a limited extent), yet we cannot say with certainty that our validity-related results will apply to other LLMs, or future versions of GPT. In addition, future research should explore different prompting strategies—including examining the effects of assigning various roles—to further enhance prompt performance and validity. Second, the closed-source nature of GPT-4o means that we are working with a black box: we have limited insights into why the model answers the way it answers (Bail, 2024). Especially the stochastic nature of LLMs like GPT-4o, referring to the fact that their output may differ with every run over the same input (Bender et al., 2021), is problematic for truly understanding the workings of the LLM and how to best leverage it. Moreover, this black box nature impedes a straightforward mapping between prompt modifications and performance, thus requiring a holistic approach and human judgement to determine whether the prompt achieves its intended goals. Therefore, our recommendation is to rely on open-source LLMs where possible, which additionally bypass data privacy concerns that are associated with the use of closed-source LLMs (Alfano et al., 2022; Ray, 2023). Moreover, future updates to proprietary LLMs—which may occur without our knowledge—could affect their performance, potentially impacting the validity and reproducibility of our results over time. Third, our method is limited by the exclusive use of prompt engineering for developing our classifiers. Even though other methods such as model fine-tuning and RAG-based approaches could have been leveraged (De Kok, 2023), we deemed these more advanced steps unnecessary due to the performance of our prompts being sufficient for the current purposes. Besides, prompt engineering in particular facilitates control over the semantic validity of one's prompts. Finally, LLMs are increasingly able to process multimodal input, such as text, audio, images, and video—each of which may provide rich insight into human behavior and mental states (Gillespie et al., 2024). We recommend that additional modalities beyond text are explored to deepen and broaden the use of LLMs for the classification of psychological phenomena in naturally occurring data.

7 Conclusion

Our findings demonstrate the promising potential of using LLMs for psychological text classification. Specifically, LLMs can be used to validly classify psychological phenomena in text, provided that the researcher works with the LLM to secure semantic, predictive, and content validity. However, this headline finding should not obscure our deeper, and perhaps more important, finding. By enabling rapid iteration (e.g., classifying hundreds of texts in minutes), LLMs can improve the fit between our concepts and the textual data, enabling researchers to explore a wider range of conceptualizations, and then iteratively optimize these. Thus, LLMs may mark a paradigm shift away from single-shot codebook creation and application, toward a more iterative approach to text classification that improves the overall validity of our concepts and operationalizations.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: the datasets contain private or sensitive information and are therefore not publicly available. Requests to access these datasets should be directed to h.l.bunt@lse.ac.uk.

Author contributions

HB: Writing – original draft, Writing – review & editing. AGo: Writing – review & editing. TR: Writing – review & editing. AGi: Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was funded by the Swiss National Science Foundation (grant number: 51NF40-205605).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frsps.2025.1460277/full#supplementary-material

References

Alfano, M., Sullivan, E., and Ebrahimi Fard, A. (2022). "Ethical pitfalls for natural language processing in psychology," in *Handbook of Language Analysis in Psychology,* eds. M. Dehghani and R. L. Boyd (New York, NY: Guilford Publications), 511–530.

Aveling, E.-L., Gillespie, A., and Cornish, F. (2015). A qualitative method for analysing multivoicedness. *Qual. Res.* 15, 670–687. doi: 10.1177/1468794114557991

Bail, C. A. (2024). Can generative AI improve social science? *Proc. Natl. Acad. Sci. U.S.A.* 121:e2314021121. doi: 10.1073/pnas.2314021121

Bakhtin, M. (1984). *Problems of Dostoevsky's Poetics*. Minneapolis: University of Minnesota Press. doi: 10.5749/j.ctt22727z1

Beghetto, R. A., Ross, W., Karwowski, M., and Glăveanu, V. P. (2024). Partnering with AI for instrument development: possibilities and pitfalls. *New Ideas Psychol.* 76:101121. doi: 10.1016/j.newideapsych.2024.101121

Belotto, M. J. (2018). Data analysis methods for qualitative research: Managing the challenges of coding, interrater reliability, and thematic analysis. *Qual. Rep.* 23, 2622–2633. doi: 10.46743/2160-3715/2018.3492

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). "On the dangers of stochastic parrots: can language models be too big?" in *Proceedings of the ACM FAccT,* 610–623. doi: 10.1145/3442188.3445922

Bergner, R. M. (2024). Establishing correct concept meanings in psychology: why should we care and how can we do it? *Am. Psychol.* 14, 1–12. doi: 10.1037/amp0001412

Birkenmaier, L., Lechner, C., and Wagner, C. (2023). The search for solid ground in text as data: a systematic review of validation approaches. *Commun. Methods Meas.* 18, 249–277. doi: 10.1080/19312458.2023.2285765

Boyd, R. L., and Schwartz, H. A. (2021). Natural language analysis and the psychology of verbal behavior: the past, present, and future states of the field. *J. Lang. Soc. Psychol.* 40, 21–41. doi: 10.1177/0261927X20967028

Bringmann, L. F., Elmer, T., and Eronen, M. I. (2022). Back to basics: the importance of conceptual clarification in psychological science. *Curr. Dir. Psychol. Sci.* 31, 340–346. doi: 10.1177/09637214221096485

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). *Language models are few-shot learners*. Available at: http://arxiv.org/abs/2005.14165 (accessed September 19, 2023).

Chandrasekharan, E., Samory, M., Jhaver, S., Charvat, H., Bruckman, A., Lampe, C., et al. (2018). The internet's hidden rules: an empirical study of Reddit norm violations at micro, meso, and macro scales. *Proc. ACM Hum.-Comput. Interact.* 2, 1–25. doi: 10.1145/3274301

Chang, T., DeJonckheere, M., Vydiswaran, V. G. V., Li, J., Buis, L. R., and Guetterman, T. C. (2021). Accelerating mixed methods research with natural language processing of big text data. *J. Mix. Methods Res.* 15, 398–412. doi: 10.1177/15586898211021196

De Kok, T. (2023). *Generative LLMs and textual analysis in accounting: (Chat)GPT as research assistant?* SSRN Electronic Journal. doi: 10.2139/ssrn.4429658

De Raadt, A., Warrens, M. J., Bosker, R. J., and Kiers, H. A. L. (2021). A comparison of reliability coefficients for ordinal rating scales. *J. Classif.* 38, 519–543. doi: 10.1007/s00357-021-09386-5

Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., et al. (2023). Using large language models in psychology. *Nat. Rev. Psychol.* 2, 688–701. doi: 10.1038/s44159-023-00241-5

Dingemanse, M., and Enfield, N. J. (2015). Other-initiated repair across languages: towards a typology of conversational structures. *Open Linguist.* 1, 96–118. doi: 10.2478/opli-2014-0007

Flake, J. K., Davidson, I. J., Wong, O., and Pek, J. (2022). Construct validity and the validity of replication studies: a systematic review. *Am. Psychol.* 77, 576–588. doi: 10.1037/amp0001006

Flake, J. K., and Fried, E. I. (2020). Measurement schmeasurement: questionable measurement practices and how to avoid them. *Adv. Meth. Pract. Psychol. Sci.* 3, 456–465. doi: 10.1177/2515245920952393

Freud, S. (1914). *The Psychopathology of Everyday Life*. New York: The Macmillan Company. doi: 10.1037/10012-000

Gilardi, F., Alizadeh, M., and Kubli, M. (2023). ChatGPT outperforms crowd-workers for text-annotation tasks. *Proc. Natl. Acad. Sci. U.S.A.* 120:e2305016120. doi: 10.1073/pnas.2305016120

Gillespie, A., Glăveanu, V., and de Saint Laurent, C. (2024). *Pragmatism and Methodology: Doing Research that Matters with Mixed Methods*. Cambridge: Cambridge University Press. doi: 10.1017/9781009031066

Gillespie, A., and Reader, T. W. (2016). The Healthcare Complaints Analysis Tool: development and reliability testing of a method for service monitoring and organisational learning. *BMJ Qual. Saf.* 25, 937–946. doi: 10.1136/bmjqs-2015-004596

Gillespie, A., and Reader, T. W. (2018). Patient-centered insights: using health care complaints to reveal hot spots and blind spots in quality and safety. *Milbank Q.* 96, 530–567. doi: 10.1111/1468-0009.12338

Goddard, A., and Gillespie, A. (2023). Textual indicators of deliberative dialogue: a systematic review of methods for studying the quality of online dialogues. *Soc. Sci. Comput. Rev.* 41, 2364–2385. doi: 10.1177/08944393231156629

Grimmer, J., Roberts, M. E., and Stewart, B. M. (2022). *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton: Princeton University Press.

Grimmer, J., and Stewart, B. M. (2013). Text as data: the promise and pitfalls of automatic content analysis methods for political texts. *Polit. Anal.* 21, 267–297. doi: 10.1093/pan/mps028

Healey, P. G. T., Mills, G. J., Eshghi, A., and Howes, C. (2018). Running repairs: coordinating meaning in dialogue. *Top. Cogn. Sci.* 10, 367–388. doi: 10.1111/tops.12336

Holleman, G. A., Hooge, I. T. C., Kemner, C., and Hessels, R. S. (2020). The "real-world approach" and its problems: a critique of the term ecological validity. *Front. Psychol.* 11:721. doi: 10.3389/fpsyg.2020.00721

Holt, E. (2000). Reporting and reacting: concurrent responses to reported speech. *Res. Lang. Soc. Interact.* 33, 425–454. doi: 10.1207/S15327973RLSI3304_04

Jackson, J. C., Watts, J., List, J.-M., Puryear, C., Drabble, R., and Lindquist, K. A. (2022). From text to thought: how analyzing language can advance psychological science. *Perspect. Psychol. Sci.* 17, 805–826. doi: 10.1177/17456916211004899

Kennedy, B., Ashokkumar, A., Boyd, R. L., and Dehghani, M. (2021). "Text analysis for psychology: methods, principles, and practices," in *Handbook of language analysis in psychology*, eds. M. Dehghani and R. L. Boyd (London: The Guilford Press), 3–62. doi: 10.31234/osf.io/h2b8t

Kjell, O. N. E., Kjell, K., Garcia, D., and Sikström, S. (2019). Semantic measures: using natural language processing to measure, differentiate, and describe psychological constructs. *Psychol. Methods* 24, 92–115. doi: 10.1037/met0000191

Korinek, A. (2023). Generative AI for economic research: use cases and implications for economists. *J. Econ. Lit.* 61, 1281–1317. doi: 10.1257/jel.20231736

Krippendorff, K. (2004). *Content Analysis: An Introduction to its Methodology*. Thousand Oaks, CA; London; New Delhi: SAGE Publications.

Krpan, D. (2022). (When) should psychology be a science? *J. Theory Soc. Behav.* 52, 183–198. doi: 10.1111/jtsb.12316

Lee, L. W., Dabirian, A., McCarthy, I. P., and Kietzmann, J. (2020). Making sense of text: artificial intelligence-enabled content analysis. *Eur. J. Mark.* 54, 615–644. doi: 10.1108/EJM-02-2019-0219

Lindebaum, D., and Fleming, P. (2024). ChatGPT undermines human reflexivity, scientific responsibility and responsible management research. *Br. J. Manag.* 35, 566–575. doi: 10.1111/1467-8551.12781

Linell, P. (2009). *Rethinking Language, Mind, and World Dialogically: Interactional and Contextual Theories of Human Sense-Making*. Charlotte, North Carolina: Information Age Publishing.

Lucy, J. A. (1993). *Reflexive Language: Reported Speech and Metapragmatics*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511621031

Marková, I. (2016). *The Dialogical Mind: Common Sense and Ethics*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511753602

Matter, D., Schirmer, M., Grinberg, N., and Pfeffer, J. (2024). Investigating the increase of violent speech in Incel communities with human-guided GPT-4 prompt iteration. *Front. Soc. Psychol.* 2:1383152. doi: 10.3389/frsps.2024.1383152

Meredith, J. (2020). Conversation analysis, cyberpsychology and online interaction. *Soc. Pers. Psychol. Compass.* 14:e12529. doi: 10.1111/spc3.12529

Meredith, J., Giles, D., and Stommel, W. J. P. (2021). "Introduction: the microanalysis of digital interaction," in *Analysing digital interaction*, eds. J. Meredith, D. Giles, and W. Stommel (Cham: Springer International Publishing), 1–21. doi: 10.1007/978-3-030-64922-7_1

Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Am. Psychol.* 50:741. doi: 10.1037/0003-066X.50.9.741

Morse, J. M. (2010). "Cherry picking": writing from thin data. *Qual. Health Res.* 20:3. doi: 10.1177/1049732309354285

NPSA (2008). *A risk matrix for risk managers.* Available at: https://nhgerm.wordpress.com/wp-content/uploads/2010/02/11-npsa_risk_matrix_for_risk_managers_v91.pdf (accessed June 26, 2023).

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science* 349:aac4716. doi: 10.1126/science.aac4716

OpenAI (2023). *Docs*. Available at: https://platform.openai.com/docs/guides/gpt-best-practices/six-strategies-for-getting-better-results (accessed August 22, 2023).

OpenAI (2024). *API reference*. Available at: https://platform.openai.com/docs/api-reference/introduction (accessed May 16, 2024).

Pangakis, N., Wolken, S., and Fasching, N. (2023). Automated annotation with generative AI requires validation. *arXiv Preprint arXiv:2306.00176*.

Patil, R., Heston, T. F., and Bhuse, V. (2024). Prompt engineering in healthcare. *J. Electron*. 13:2961. doi: 10.3390/electronics13152961

Peirce, C. S. (1955). *Philosophical Writings of Peirce*. Mineola, New York: Dover Publications, Inc.

Rai, T. S., and Fiske, A. (2010). ODD (observation-and description-deprived) psychological research. *Behav. Brain Sci.* 33:106. doi: 10.1017/S0140525X10000221

Rathje, S., Mirea, D.-M., Sucholutsky, I., Marjieh, R., Robertson, C., and Van Bavel, J. J. (2023). GPT is an effective tool for multilingual psychological text analysis. *Proc. Natl. Acad. Sci. U S A*. 121:e2308950121. doi: 10.31234/osf.io/sekf5

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *IOTCPS* 3, 121–154. doi: 10.1016/j.iotcps.2023.04.003

Reddy, M., and Ortony, A. (1979). "Metaphor and thought," in *The Conduit Metaphor,* 21–43.

Reiss, M. V. (2023). Testing the reliability of ChatGPT for text annotation and classification: a cautionary remark. *arXiv Preprint arXiv:2304.11085*.

Roberts, J., Baker, M., and Andrew, J. (2024). Artificial intelligence and qualitative research: the promise and perils of large language model (LLM) "assistance." *Crit. Perspect. Account.* 99:102722. doi: 10.1016/j.cpa.2024.102722

Rubin, M., and Donkin, C. (2022). Exploratory hypothesis tests can be more compelling than confirmatory hypothesis tests. *Philos. Psychol.* 37, 2019–2047. doi: 10.1080/09515089.2022.2113771

Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language* 50, 696–735. doi: 10.1353/lan.1974.0010

Salomon, G., Perkins, D. N., and Globerson, T. (1991). Partners in cognition: extending human intelligence with intelligent technologies. *Educ. Res. J.* 20, 2–9. doi: 10.3102/0013189X020003002

Schegloff, E. A. (2007). *Sequence Organization in Interaction: A Primer in Conversation Analysis*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511791208

Schegloff, E. A., Jefferson, G., and Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language* 53, 361–382. doi: 10.1353/lan.1977.0041

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychol. Methods* 17, 551–566. doi: 10.1037/a0029487

Schimmack, U. (2021). The validation crisis in psychology. *Meta-Psychol.* 5:1645. doi: 10.15626/MP.2019.1645

Seawright, J. (2016). *Multi-Method Social Science: Combining Qualitative and Quantitative Tools*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9781316160831

Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* 22, 1359–1366. doi: 10.1177/0956797611417632

Song, H., Tolochko, P., Eberl, J.-M., Eisele, O., Greussing, E., Heidenreich, T., et al. (2020). In validations we trust? The impact of imperfect human annotations as a gold standard on the quality of validation of automated content analysis. *Polit. Commun.* 37, 550–572. doi: 10.1080/10584609.2020.1723752

Tausczik, Y. R., and Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *J. Lang. Soc. Psychol.* 29, 24–54. doi: 10.1177/0261927X09351676

Tavory, I., and Timmermans, S. (2014). *Abductive Analysis: Theorizing Qualitative Research*. Chicago: University of Chicago Press. doi: 10.7208/chicago/9780226180458.001.0001

Törnberg, P. (2023). ChatGPT-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv Preprint arXiv:2304.06588*.

Törnberg, P. (2024). Best practices for text annotation with large language models. *arXiv Preprint arXiv:2402.05129*.

Van Atteveldt, W., Van Der Velden, M. A. C. G., and Boukes, M. (2021). The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. *Commun. Methods Meas.* 15, 121–140. doi: 10.1080/19312458.2020.1869198

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* 35, 24824–24837. doi: 10.5555/3600270.3602070

Yarkoni, T. (2022). The generalizability crisis. *Behav. Brain Sci.* 45:e1. doi: 10.1017/S0140525X21001758

Ziems, C., Held, W., Shaikh, O., Chen, J., and Zhang, Z. (2023). Can large language models transform computational social science? *arXiv Preprint arXiv.2305.03514*.