Cognitive and Behavioral Science Digest
WISDOM ACADEMIC https://doi.org/10.63808/cbsd.v2i2.486 Volume 2 , Issue 2

Research Article

Personality Language Markers from Lexicons to Large Language Models: A Two-Axis Review of Inference Mechanisms and Validity Evidence

Maojun Zhao[1,*]
[1]Warwick Business School, The University of Warwick, Coventry, CV4 7AL, United Kingdom
[*]**Correspondence:** Maojun Zhao, imogen_maojun_zhao@163.com

© The Author(s) 2026.
This article is published by Wisdom Academic Press Ltd. under a Creative Commons Attribution 4.0 International License. The license permits use, sharing, adaptation, distribution, and reproduction in any medium or format, provided that appropriate credit is given to the original author(s) and the source, a link to the license is supplied, and any modifications are indicated. Unless otherwise stated, all images and third-party materials included in this article are also covered by the same license. For any material not covered by this license, if the intended use is neither permitted by applicable law nor allowed under the license terms, direct permission must be obtained from the copyright holder. To view a copy of this license, please visit http://creativecommons.org/licenses/by/4.0/.

Abstract

Personality assessment is shifting from self-report questionnaires to automated inference from natural language, yet each new method is still certified against the instrument it was designed to replace. This review organises the literature through a two-axis framework that locates studies jointly by inference mechanism (from lexicons and embeddings to large language models) and by validity evidence type. Mapping a core corpus of twelve publications onto the framework shows that validation effort concentrates in a single cell where prompted models are benchmarked against questionnaire scores, while behavioural, temporal, and cross-domain evidence is sparse within the corpus reviewed; where questionnaire convergence serves as the sole criterion, the arrangement becomes circular, since agreement with the earlier instrument is reported as accuracy. A case analysis of low agreeableness asks whether language preserves the distinction between hostile and narcissistic expressions of antagonism that domain-level scales collapse, and finds the available linguistic evidence limited in magnitude. The review proposes behavioural prediction, temporal stability, and cross-corpus transfer as complementary forms of evidence that do not depend on the replaced instrument, and considers how they might inform admission standards for deployment and an agenda for research.

Keywords

Personality assessment, Language markers, Large language models, Validity, Questionnaire benchmarking

1. Introduction

Personality assessment is increasingly being supplemented, and in some applied settings displaced,

by other methods. For most of the last hundred years, standing individual differences were measured by having people rate themselves on questionnaire items; now, machines are increasingly used to analyse the language people produce in posts, essays, messages and answers to estimate these differences without any self-reporting at all. Transformer-based analysis of natural language has shown that it can predict traditional self-report measures of psychological constructs and subjective well-being among them with high accuracy, close to the theoretical upper limit set by the reliability of these instruments [1], which has been read as technical support for substitution. Based on the above evidence, text is no longer a fallback option but can serve as a measurement medium in its own right. Yet the replacement is not symmetrical. The mode of inference has changed; it is no longer a rating scale but a lexicon, then learned semantic representations, and now large language models that operate via prompting, fine-tuning and representation probing; however, the standard by which all new methods are validated remains the questionnaire they were originally designed to replace. Throughout this trajectory, the scale has served not as one stage among many but as the fixed basis for accepting all of them.

This asymmetry is, in the estimation of this review, the main unsolved problem of the field, and existing syntheses have not addressed it. Previous syntheses of computational personality assessment have characteristically organised the literature by method (e.g., which architecture, which features, which benchmark accuracy), and thus the progress in representing language has been implicitly equated with progress in measuring people. What such method-ordered accounts cannot show is whether the evidence for these systems has developed at all, or where trait distinctions that language may register and questionnaires collapse, the divergent expressions of shared antagonistic dispositions among them, are left without an arbiter. A synthesis is required to treat the means of inference and the grounds for validity as separate problems.

This review provides a synthesis by introducing a two-axis framework to locate any study jointly by its inference mechanism and by the type of validity evidence it invokes, and then critically reads the resulting map rather than describing it. It reviews the development of inference from lexicons to large language models, maps the surveyed literature onto the framework, studies a case where the dominant certification standard is structurally unable to adjudicate between competing readings of the same text, and considers whether the concentration of the reviewed literature's validation effort in a single cell is better read as circularity than as an achievement. Based on the above, it proposes three yardsticks that do not depend on the replaced instrument (behavioural prediction, temporal stability and cross-corpus transfer) and considers how they could inform admission standards for applied deployment and an agenda for research. The significance of the exercise is both measurement-theoretic and practical: it redefines what accuracy claims in this literature actually certify, and it specifies what text-based personality inference would need to demonstrate before its estimates could stand on evidence of their own.

2. Review Framework and Approach

2.1. Review design, search strategy, and selection

This review is a narrative review resting on a documented but non-exhaustive search, and it makes no claim of systematic coverage. It draws on empirical and methodological studies published between 2021 and 2026, identified through searches of Web of Science, Scopus, the ACL Anthology, and Google Scholar using combinations of personality-related terms (Big Five, dark traits, narcissism) with method-related terms (lexicon, LIWC, embedding, language model, prompt, fine-tuning, probing); a representative string is ("personality" OR "Big Five" OR "narcissism") AND ("language model" OR "embedding" OR "LIWC" OR "prompt") AND ("prediction" OR "assessment" OR "validity"). Eligible studies had to infer or assess human personality from natural language, or to bear directly on the evaluation of such inference; work concerned solely with simulating personality in artificial agents was excluded, except where it supplies the methodological basis of an inference route. From the eligible pool, a compact core of twelve publications was retained, selected to represent each inference mechanism, and each type of validity evidence for which the search returned a candidate, with its most load-bearing example rather than to enumerate the field.

Because selection was purposive rather than exhaustive, records were not screened in stages and stage-level counts are not reported. The consequence is observed throughout the sections that follow. The density and sparsity of the cells described below characterise the twelve-study core corpus rather than the field, and cannot support claims of field-wide absence. Four further publications ([13] to [16]) are cited as contextual references, bearing on the horizontal axis, on the linguistic evidence examined in the case probe, on cross-lingual transfer, and on alternative metric standards; they are discussed in the text but form no part of the core corpus, are assigned no cells, and in one instance fall outside the 2021 to 2026 window.

2.2. Coding procedure

Each core publication was assigned to a single cell on the basis of its primary evidence, defined as the evidence its authors themselves advance in support of the claim that their estimates index personality. Where a study reports more than one kind of evidence, priority follows the validation the authors treat as principal, and the remaining evidence types are not duplicated across cells. Studies whose contribution is methodological or argumentative rather than evidential are left unassigned. Two assignments are borderline and are flagged in both tables. [1] validates against rating-scale measures of an adjacent construct, and is retained because its result on approaching the reliability ceiling is load-bearing for the substitution claim; the criterion in [9] consists of report-style ratings that include observer as well as self reports, and is coded as questionnaire benchmarking because that criterion remains a

rating-scale score. Coding was carried out by the author alone, without independent duplicate coding, which limits the auditability of the assignments. Both borderline cases were reassigned to their alternatives and the mapping re-read; the concentration of evidence in the questionnaire column and the sparsity of the remaining columns persist under each change.

Table 1 records the data source, construct, inference mechanism, validation criterion, outcome metric, and principal finding for each core publication, so that the assignments underlying the map can be inspected individually.

Table 1

Study-level evidence for the twelve core publications

Study	Sample / data source	Construct	Inference mechanism	Validation criterion	Outcome metric	Principal finding
[1]	Open-ended written responses, several English-language samples	Subjective well-being (adjacent construct)	Embedding-based	Questionnaire benchmarking	Correlation with rating-scale scores	Transformer-based text measures approach the reliability ceiling of the rating scales
[2]	Dictionary development and validation corpora	Psychological process categories	Lexicon-based	Not mapped (methodological basis)	Category-level psychometric properties	Supplies the validated dictionary on which the lexicon route depends
[3]	Posts from several social media platforms	Big Five	Embedding-based	Questionnaire benchmarking	Accuracy and F1	Averaging several pre-trained encoders improves trait prediction over single encoders
[4]	Benchmark personality corpora under few-shot conditions	Big Five and type indicators	LLM: fine-tuning (prompt tuning)	Questionnaire benchmarking	Accuracy and F1	Encoding trait-descriptive semantics into trainable prompts improves few-shot recognition
[5]	Model-internal editing benchmark	Trait expression within the model	LLM: representation probing	Not mapped (methodological basis)	Editing success rate	Trait-relevant information can be located and modified inside model parameters

[6]	Social media status updates	Big Five	LLM: prompt-based (zero-shot)	Questionnaire benchmarking	Correlation with self-report scores	Zero-shot estimates correlate with self-reports at levels comparable to supervised baselines
[7]	Established personality text corpora	Big Five	LLM: prompt-based	Questionnaire benchmarking	Correlation and accuracy by trait	Performance is uneven; trait knowledge in the prompt aids generalisation while fine-grained estimation stays near baseline
[8]	Twitter accounts	Narcissism within the dark triad	Lexicon-based features with machine learning	Questionnaire benchmarking against trait labels	Classification accuracy	Lexical features separate users scoring high on narcissism from other users
[9]	Asynchronous video interview responses	Big Five	LLM: prompt-based	Report-style ratings, including observer reports	Convergence, reliability, fairness and rating-pattern indices	Model ratings converge with report-based scores while reliability and rating patterns vary across traits
[10]	Conceptual and programmatic argument	Not applicable	Not mapped	Not mapped (evaluative argumentation)	Not applicable	Argues for targeted evaluation against intended states and outcomes rather than imitation of rating scales
[11]	Two distinct text domains	Big Five	Embedding-based	Cross-corpus transfer	Predictive accuracy retained after domain change	High-dimensional models lose most predictive power across domains while lower-dimensional theory-informed

						models are less domain-dependent
[12]	Curricula vitae and short free-text responses from applicants	Big Five	Embedding-based	Questionnaire benchmarking	Correlation with self-report scores	Applicant personality is predicted from application materials at modest accuracy

2.3. The evolution axis: from lexicons to large language models

The path of text-based personality inference can be organised according to the development of representation for linguistic meaning. At the beginning of this was the lexicon paradigm, which converted texts into frequency profiles of psychologically curated word lists; its most widely used tool has since been iteratively improved into a dictionary system with explicitly validated categories for psychometric applications [2]. The paradigm also has limitations: it provides transparent, theory-driven categories but cannot capture meanings outside its predefined vocabulary.

Embedding-based representation relaxes this constraint by placing words and utterances in continuous semantic space, and then learns personality-related signals from the context rather than enumerating them in advance. Work in this stage showed that aggregating several pre-trained transformer encoders through model averaging improved trait prediction across social media platforms [3]. Thus, the representative unit is no longer a word category but a distributed vector, and feature engineering has been replaced by representation learning.

Large language models have extended this path and divided it into three operational modes. Prompt-based inference elicits trait judgments directly from a general-purpose model without task-specific training, and the model serves as the assessor. Fine-tuning adapts the model's parameters or a lightweight surrogate of them to the assessment task, and personality-descriptive prompt tuning is a typical case within pre-trained language models that encodes trait semantics into trainable prompts under few-shot conditions [4]. Representation probing is further away from the output and investigates or alters how trait-relevant information is arranged within the model's parameters, as shown by research on locating and modifying personality inside the model [5]. The three routes are different in terms of where inference occurs (at the interface, in the adapted weights, or in the internal representation), but they all share the same basic idea that personality can be inferred from linguistic features. A modularised variant of the prompt-based route separates content generation from embedding-based classification within a single pipeline [13], which indicates that the three modes describe where inference is located rather than

a fixed unit of analysis, and that one pipeline may present several distinct points at which validity could be assessed.

Two features of this axis deserve emphasis. Almost all of the movement along it has been driven by advances in representing language, but the criterion for judging the resulting estimates has not changed accordingly. And the axis is not a third dimension of the framework introduced below; its stages, once unfolded, simply constitute the categories of the inference-mechanism axis, so the evolution axis provides the generative logic for the horizontal dimension rather than adding another one.

2.4. The two-axis framework: inference mechanisms × types of validity evidence

Based on the axis of development, this review has been arranged in a two-dimensional analytical system that can be employed to find all empirical studies according to how their inferences are derived and how their validity is verified. The horizontal axis contains the categories generated at the evolution level: lexicon-based word counting, embedding-based representation, and the three paths for large language model deployment, namely prompt-based inference, fine-tuning, and representation probing. These categories share the same question: by what technical means is a personality assessment derived from text, and thus they function as mutually comparable positions on a single dimension rather than as independent families of methods.

The vertical axis divides studies based on the kind of evidence used to support the claim that the results estimate personality. These four kinds are reviewed here. Questionnaire benchmarking takes the degree of agreement with self-report scale scores as the standard for accuracy. Behavioural criteria determine whether the inferred traits predict behaviour outside the text theoretically. Temporal stability refers to the consistency of estimates for the same person over time, as trait constructs need this. Cross-corpus transfer determines if a model trained in one text domain can still perform well on another, thus indicating whether the learned signal is related to personality or merely domain-specific surface features. These four categories are types of evidence rather than degrees of validity, and they correspond to the standard vocabulary of validation in the following way. Questionnaire benchmarking supplies convergent evidence, and yields discriminant evidence only where patterns of non-convergence are examined alongside it. Behavioural criteria supply criterion-related evidence. Temporal stability belongs to reliability yet bears on construct validity, since trait constructs are defined partly by relative endurance. Cross-corpus transfer addresses generalizability across settings. Construct validity is not a fifth column but the claim that the four are marshalled to support, which is why no single column establishes it alone.

The two axes intersect at a grid; the density of cells indicates where verification has been carried out among the studies mapped, and empty cells indicate where the corpus assembled here contains no primary evidence of that type. The framework also ensures the separation that the literature tends to blur: progress along the mechanism axis does not automatically constitute progress in terms of evidence; a

more refined inference procedure may still be based on the same narrow criterion as that for previous research. The above division sets up the mapping presented in the following section.

3. Results: Mapping the Literature onto the Two-Axis Framework

3.1. Distribution of validity evidence across inference mechanisms

Table 2 locates the reviewed studies within the framework, with inference mechanisms as rows and types of validity evidence as columns. Reading the table by column shows questionnaire benchmarking to be the default certification across every mechanism generation: lexicon-based and embedding-based studies alike report their accuracy as agreement with scale scores, and the pattern intensifies rather than weakens as the mechanisms grow more capable.

Table 2

Distribution of reviewed studies by inference mechanism and validity evidence

Inference mechanism	Questionnaire benchmarking	Behavioural criteria	Temporal stability	Cross-corpus transfer
Lexicon-based	[8]	—	—	—
Embedding-based	[1][a], [3], [12]	—	—	[11]
LLM: prompt-based	[6], [7], [9][b]	—	—	—
LLM: fine-tuning	[4]	—	—	—
LLM: representation probing	—	—	—	—

Notes. a. Adjacent construct (subjective well-being). b. Benchmarked against report-style ratings, including observer reports. [2], [5], and [10] are not assigned cells: the former two supply the methodological bases of the lexicon and probing routes; [10] contributes evaluative argumentation. With the nine assigned publications these make up the twelve-study core corpus, and contextual references [13] to [16] are not mapped. Dashes denote the absence of primary evidence of that type within this corpus.

Prompt-based rows are typical examples of this continuity. Zero-shot inference from social media language has been shown to correlate with self-reported Big Five scores at the same level as supervised models trained specifically for the task [6], and this result has been widely regarded as evidence that general-purpose models can assess personality without dedicated training. A system-wide assessment of the same route shows that its performance is inconsistent; while adding trait knowledge to the prompt

enhances generalisation, fine-grained estimation is close to that of the baseline [7]. Both results, regardless of whether they are optimistic or not, are expressed in the same unit of proximity to questionnaire scores and are thus in the same column.

On the other hand, the remaining columns are sparsely populated. Within the twelve-study core corpus no publication was coded as supplying primary behavioural or temporal-stability evidence, and cross-corpus transfer is represented by a single study; this holds for the corpus reviewed and is not a demonstration that such work is absent from the field. Therefore, the resulting grid is relatively dense in one column and sparse in others. The following case examines a distinction that the main column is not equipped to adjudicate, and the paradox this entails for the field is then addressed directly.

3.2. A case probe: linguistic differentiation of low agreeableness: hostile versus narcissistic expression

Low agreeableness is a stress test for the framework because it is an area of overlap for trait taxonomies. Within the Big Five, low agreeableness occupies the antagonistic pole of a broad dimension, while dark personality traits (narcissism is a representative example) share this antagonistic core but are expressed through different motivational paths. In language, the shared core has been described as diverging into recognisably different signatures. Hostile expressions are characterised as direct antagonism, with anger vocabulary, profanity, and confrontation directed at others; narcissistic expressions are characterised as antagonism presented as self-exaltation, with status display and derogation of the audience rather than open conflict. The empirical support for this contrast is thinner than the description suggests. A pooled analysis of fifteen samples using a validated dictionary found grandiose narcissism to be unrelated to first-person singular pronouns, identified swear words among its strongest positive correlates, and reported effects below $|r| = .10$ throughout [14], so profanity does not cleanly separate the two styles and self-reference does not mark narcissism in the manner the description assumes. Machine learning classifiers of social media language have separated users scoring high on narcissism from other users using lexical features [8], which establishes that a classification pattern exists but not that the pattern tracks the underlying construct rather than correlated surface features. The case is therefore presented as a conceptual probe of what the dominant certification standard can adjudicate, rather than as evidence that language preserves a distinction the questionnaire loses.

The case matters to the framework because the main certification standard cannot rule on it. The domain-level agreeableness score assigns the same low value to both expressive styles, and thus the very construct used to benchmark text-based estimates fails to show a visible difference in language. When the two inference systems disagree, one interprets the text as hostile antagonism and the other as narcissistic antagonism; however, both have the same agreement with the questionnaire, and the

benchmark returns no verdict on which reading is closer to the person. Any information advantage that language holds over the scale would therefore lie in the part of the signal that questionnaire benchmarking is not equipped to assess.

Where a text-based inference is finer grained than its criterion, the criterion cannot certify the additional distinction. Whatever the finer reading contributes must either be discarded to fit the scale or retained without a recognised standard of adequacy. A field that certifies its methods only against instruments of coarser resolution has, in this region, no established basis for judging when those methods are correct. The mapping presented above bears on this point directly.

3.3. The concentration cell of prompt-based inference × questionnaire benchmarking, and the validity paradox

As shown in the map, the validation efforts are concentrated in a single column of questionnaire benchmarking, and within the current generation of studies, only one cell verifies the estimates obtained by prompting large language models against questionnaire scores. Focus is not limited to the above explorations. Even one of the most methodologically comprehensive evaluations available that investigated the validity, reliability, fairness and rating patterns of model-based personality assessment in employment interview settings still based its accuracy criteria on report-style ratings as the gold standard [9]. Thus, the scope of psychometric scrutiny has expanded, but the standard of judgment for all such scrutiny remains the same.

Thus, what this paper calls a validity paradox arises. Text-based inference is proposed because questionnaires are inconvenient to distribute, often reactive in nature, and thus limited in their ecological scope; it is then asserted that this replacement has succeeded by reproducing the scores obtained from the questionnaires. Questionnaire convergence is not for that reason an illegitimate criterion. Where the target construct is itself defined in terms of self-perception, agreement with a well-validated scale is the appropriate convergent evidence, and where convergence is one criterion among several it contributes to a construct-validity argument in the ordinary way. It becomes insufficient in three situations that recur in this literature: when it is the sole criterion, so that nothing bears on the estimate other than agreement with the instrument being replaced; when the inference is finer grained than the scale, as the preceding case illustrates; and when the estimate is applied in settings for which the scale itself has not been validated. In those situations whatever systematic error the questionnaire carries (such as socially desirable responding, reference-group effect, momentary self-perception) may be inherited by the new method and reported as accuracy, and correlations with self-reports primarily index convergence with the earlier instrument rather than proximity to the trait itself.

The deeper issue is what the concentration cell conceals. As automated psychological profiling has rarely been evaluated against a standard independent of the questionnaire, there is little basis for

determining whether a text-based estimate is superior to the questionnaire it emulates. Therefore, progress along the mechanism axis accumulates without corresponding movement in the direction of evidence, and the remaining cells of the framework (behavioural, temporal, and cross-domain evidence) remain sparsely populated in the corpus reviewed. These vacancies serve as the alternative standards for the discussion.

4. Discussion

4.1. Three alternative yardsticks: behavioural criteria, temporal stability, and cross-corpus transfer

If questionnaire convergence cannot serve as the sole basis, then the sparse columns in the framework indicate where independent standards would have to be obtained. Three such standards are set out below, each assessing validity outside the instruments these new methods are designed to replace.

Behavioural criteria anchor an estimate in what it predicts rather than in what it agrees with. A text-based trait score that predicts relevant behaviour (cooperation, persistence, interpersonal choices) is justified by the causal role of the trait, rather than by convergence with another measure of the same self-perception. This position is in line with the arguments that have recently emerged in psychological assessment: programmatic work has suggested evaluating language-based methods according to the psychological state and outcome they aim to capture, and rating scales should be viewed as one of several fallible indicators rather than a goal for imitation [10]. The yardstick proposed here extends that contention toward a certification requirement.

Temporal stability is also an internal standard that does not need an external instrument; rather, trait constructs are defined by relative endurance, so estimates for the same person at different times and in different situations should be consistent. No study in the corpus reviewed here reports such a test, although the repeated-measurement design required to supply one has long been available.

Cross-corpus transfer has been examined more directly than the other two, and the results qualify what within-domain accuracy demonstrates. Direct examination of model generalisation across text domains showed that high-dimensional models with excellent within-domain accuracy lost their predictive power almost entirely when transferred to a new domain, whereas lower-dimensional, theory-informed models were less dependent on the training domain [11]. The finding revises what the correlation between a model and questionnaires actually verifies: the model may be learning the language use of a platform rather than the dispositions of its users. Cross-corpus transfer should therefore be viewed not as a robustness addition but as a validity condition that separates personality signals from domain noise. Transfer is not only a matter of platform. Recent multilingual work trains personality recognition across languages by distilling cross-lingual attention and augmenting low-resource languages with generated personality-informed text [15], which indicates that claims of generalisation

should also be examined across languages and cultural settings, where the availability of training material is uneven.

The three yardsticks are complementary rather than interchangeable; prediction, endurance and portability certify different aspects of the same claim, and together they define what an independent standard of adequacy for psychological profiling should be. Their relative weight depends on the use to which an estimate is put rather than being fixed in advance.

4.2. Application implications and a research agenda

The yardsticks carry direct consequences for practice, because deployment has been running ahead of certification. Text-based personality inference already operates in personnel selection: machine learning systems have been shown to predict applicant personality from curricula vitae and short free-text responses [12]. Read against the framework, such results occupy the concentration cell (agreement with report-based scores in a single context) precisely where the stakes of the validity paradox are highest. The practical implication is not that such applications should halt, but that their admission standard should change: an estimate that predicts relevant workplace behaviour, remains stable across an applicant's writing occasions, and survives transfer from the training corpus to the hiring context has earned a warrant that questionnaire convergence alone cannot confer. What such a change requires can be stated at the level of evidence rather than of fixed numerical cut-offs, which the present literature is not in a position to specify. For research use, an estimate should carry at least one form of evidence other than questionnaire convergence. For low-stakes screening, it should in addition remain stable across an individual's writing occasions and retain its accuracy on a corpus other than the one on which it was developed. For high-stakes deployment, evidence that the estimate predicts the relevant outcome behaviour should be obtained in the deployment setting itself under a pre-specified analysis, together with the reliability and fairness checks already customary in selection contexts [9]. Work on aligning metrics across competing trait theories and on handling ambiguous cases indicates that such alternative standards can be formalised rather than left to discretion [16].

The same logic yields a research agenda, which is best stated as instructions for populating the grid. Studies pairing each inference mechanism with behavioural outcomes would fill the column that certification currently ignores; repeated-measurement designs would supply the stability evidence that this corpus does not contain; systematic cross-domain evaluation, extended beyond the isolated demonstrations now available, would establish whether any mechanism generation captures signal that travels. Comparative work along the mechanism axis becomes informative only once such columns exist, since mechanisms that tie on questionnaire agreement may separate sharply on prediction, endurance, or portability.

This review's own limits mark the boundaries of these claims. It proceeds narratively rather than

systematically, over a deliberately compact corpus weighted toward English-language sources, and its cell assignments were made by a single coder without independent duplicate coding. Statements about sparse or empty cells accordingly describe the corpus assembled here rather than the field. The framework is offered accordingly: not as a census of the field, but as a coordinate system that any subsequent study, systematic or empirical, can use to declare where it stands, and which empty cell it intends to fill.

5. Conclusion

This study found that although the means of personality assessment have changed, the standards for acceptance have not. Tracing inference from lexicon-based word counting to embedding representations, then to prompting, fine-tuning and representation probing in large language models, a review found that each generation was certified by the same instrument: the self-report questionnaire; that is to say, text-based measurement was intended to replace it. A two-axis framework has been introduced here to display this asymmetry and analyse it. Separating the mechanism of inference from the type of validity evidence transforms the literature into a grid whose geography is informative: verification efforts cluster in a single column of questionnaire benchmarking, densest where prompted models are scored against scale totals, and the columns that could independently ground the enterprise are sparsely populated in the corpus reviewed. Based on the above grid, the field's headline correlations suggest convergence with the old ruler rather than proximity to the person, and the case of low agreeableness indicates where the cost falls: where language may preserve distinctions within a shared antagonistic core, the domain-level benchmark collapses them and provides no basis for adjudicating the finer reading. The constructive response is not to cease benchmarking but to demote it. Behavioural prediction, temporal stability and cross-corpus transfer provide standards that are independent of the instrument of replacement, and together they define what an independent warrant for text-based profiling would require: graded admission requirements for already operating applications and a map of missing studies. Lexicons and large language models have settled the question of whether language gives estimates that align with questionnaire results. Whether those estimates reflect what they claim remains open, and the answer will come from the empty cells of the grid, not the crowded ones.

Acknowledgments

Not applicable.

Author contributions

Maojun Zhao individually undertook all roles including conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, original draft writing, review &

editing, visualization, supervision and project administration for this study, and has read and agreed to the published version of the manuscript.

Funding

Not applicable.

Data availability

Not applicable.

Declarations

Conflict of interest

Not applicable.

Ethics approval

Not applicable.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Additional information

Received: 29 July 2026

Accepted: 3 September 2026

Published online: 25 November 2026

Publisher's Note

Wisdom Academic Press Ltd. remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

[1] Kjell, O. N. E., Sikström, S., Kjell, K., & Schwartz, H. A. (2022). Natural language analyzed with AI-based transformers predict traditional subjective well-being measures approaching the theoretical upper limits in accuracy. *Scientific Reports, 12*(1), 3918. https://doi.org/10.1038/s41598-022-07520-w

[2] Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). *The development and psychometric properties of LIWC-22*. Austin, TX: University of Texas at Austin. https://www.liwc.app/static/documents/LIWC-22%20Manual%20-%20Development%20and%20Psychometrics.pdf

[3] Christian, H., Suhartono, D., Chowanda, A., & Zamli, K. Z. (2021). Text based personality prediction from multiple social media data sources using pre-trained language model and model averaging. *Journal of Big Data, 8*, 68. https://doi.org/10.1186/s40537-021-00459-1

[4] Wen, Z., Cao, J., Yang, Y., Wang, H., Yang, R., & Liu, S. (2023). DesPrompt: Personality-descriptive prompt tuning for few-shot personality recognition. *Information Processing & Management, 60*(5), 103422. https://doi.org/10.1016/j.ipm.2023.103422

[5] Mao, S., Wang, X., Wang, M., Jiang, Y., Xie, P., Huang, F., & Zhang, N. (2025). Editing personality for large language models. *Natural Language Processing and Chinese Computing, NLPCC 2024, Lecture Notes in Computer Science, 15360*, 241–254. https://doi.org/10.1007/978-981-97-9434-8_19

[6] Peters, H., & Matz, S. C. (2024). Large language models can infer psychological dispositions of social media users. *PNAS Nexus, 3*(6), 231. https://doi.org/10.1093/pnasnexus/pgae231

[7] Ganesan, A. V., Lal, Y. K., Nilsson, A. H., & Schwartz, H. A. (2023). Systematic evaluation of GPT-3 for zero-shot personality estimation. *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, 390–400. https://doi.org/10.18653/v1/2023.wassa-1.34

[8] Haz, L., Rodríguez-García, M. Á., & Fernández, A. (2022). Detecting narcissist dark triad psychological traits from Twitter. *Proceedings of the 14th International Conference on Agents and Artificial Intelligence ICAART 2022, 2*, 313–322. https://doi.org/10.5220/0010839100003116

[9] Zhang, T., Koutsoumpis, A., Oostrom, J. K., Holtrop, D., Ghassemi, S., & de Vries, R. E. (2024). Can large language models assess personality from asynchronous video interviews? A comprehensive evaluation of validity, reliability, fairness, and rating patterns. *IEEE Transactions on Affective Computing, 15*(3), 1769–1785. https://doi.org/10.1109/TAFFC.2024.3374875

[10] Kjell, O. N. E., Kjell, K., & Schwartz, H. A. (2024). Beyond rating scales: With targeted evaluation, large language models are poised for psychological assessment. *Psychiatry Research, 333*, 115667. https://doi.org/10.1016/j.psychres.2023.115667

[11] Berggren, M., Kaati, L., Pelzer, B., Stiff, H., Lundmark, L., & Akrami, N. (2024). The generalizability of machine learning models of personality across two text domains. *Personality and Individual Differences, 217*, 112465. https://doi.org/10.1016/j.paid.2023.112465

[12] Grunenberg, E., Peters, H., Francis, M. J., Back, M. D., & Matz, S. C. (2024). Machine learning in recruiting: Predicting personality from CVs and short text responses. *Frontiers in Social Psychology, 1*, 1290295. https://doi.org/10.3389/frsps.2023.1290295

[13] Tan, J. J., Kwan, B.-H., Ng, D. W.-K., Hum, Y. C., Mokraoui, A., & Lo, S.-Y. (2026). Prompting-in-a-series: Psychology-informed contents and embeddings for personality recognition with decoder-only models. *IEEE Transactions on Computational Social Systems, 13*(1), 333–347. https://doi.org/10.1109/TCSS.2025.3593323

[14] Holtzman, N. S., Tackman, A. M., Carey, A. L., Brucks, M. S., Küfner, A. C. P., Große Deters, F., Back, M. D., Donnellan, M. B., Pennebaker, J. W., Sherman, R. A., & Mehl, M. R. (2019). Linguistic markers of grandiose narcissism: A LIWC analysis of 15 samples. *Journal of Language and Social Psychology, 38*(5–6), 773–786. https://doi.org/10.1177/0261927X19871084

[15] Tan, J. J., Kwan, B.-H., Ng, D. W.-K., Hum, Y. C., Kawarazaki, N., Takano, K., & Mokraoui, A. (2026). Cross-lingual attention distillation with personality-informed generative augmentation for multilingual personality recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 1–16. https://doi.org/10.1109/TCDS.2026.3682672

[16] Tan, J. J., Kwan, B.-H., Ng, D. W.-K., Hum, Y. C., Lo, S.-Y., Chen, P.-A., Kawarazaki, N., Takano, K., & Mokraoui, A. (2026). Large-language-models-as-a-judge in theory-agnostic adaptive metric-

alignment for prototypical networks in personality recognition. *IEEE Transactions on Affective Computing*, 1-16. https://doi.org/10.1109/TAFFC.2026.3712379