# THE AI INTERPRETER

## A Criterion-Referenced Study of the Validity and Reliability of Large Language Models in Interpreting Personality Profiles

Prepared by Dr. James Niblick

## Abstract

Large language models are increasingly being used to interpret personality-assessment results. Individuals routinely submit assessment scores to consumer AI systems and ask what those scores mean, while organizations are beginning to incorporate similar interpretive functions into AI-enabled coaching, communication, talent, and decision-support systems. Yet relatively little research has examined whether different language models accurately interpret the same structured psychometric information or whether they reach consistent conclusions when presented with identical profiles. The present study evaluates the criterion-referenced interpretive validity and cross-model consistency of five large language models interpreting ten standardized DISC Plus personality profiles. Each profile contained natural and adaptive DISC scores together with seven Values Index scores. Before model administration, each profile was associated with a predetermined behavioral criterion describing the interpretations that should follow from the score configuration. Models were therefore evaluated not on whether their descriptions sounded plausible or were favorably received, but on the degree to which their interpretations agreed with the predefined criterion. Each model completed three tasks for each profile: a natural-language interpretation, a structured behavioral-rating task including unsupported "trap" items, and an identification of strengths and weaknesses. Model performance was evaluated using exact criterion agreement, mean absolute deviation, directional error, discrimination between similar and dissimilar profiles, confidence on unsupported inferences, and qualitative analysis of negative-trait attenuation. In the sample demonstration data, model-level exact agreement ranged from 51.3% to 75.0%, with an unweighted mean of 63.7%. Two models exceeded the preregistered 70% accuracy ceiling, disconfirming that hypothesis. Error magnitude increased monotonically as accuracy declined. Models distinguished strongly dissimilar profiles but reproduced 117 of 120 judgments identically across deliberately similar profile pairs, indicating limited sensitivity to small criterion-relevant differences. Models generally preserved negatively valenced interpretations rather than reversing them into positive traits, although their narrative language frequently softened those interpretations. Confidence on items that could not legitimately be inferred from the supplied personality scores varied sharply by model. The findings demonstrate why apparent plausibility should not be treated as evidence of accurate psychometric interpretation. They also distinguish several separate questions that are often conflated: whether an LLM interpretation agrees with the assessment criterion, whether different models agree with one another, whether a model preserves meaningful differences between profiles, and whether it recognizes the limits of what the assessment data can support. A repeated independent-session design is required to extend the present work from cross-model agreement to within-model test-retest reliability.


## 1. Introduction

Personality-assessment data do not interpret themselves. An assessment produces scores; an interpreter converts those scores into statements about behavioral tendencies, communication patterns, motivational priorities, potential strengths, and possible limitations. Historically, that interpretive function has been performed through standardized reports, trained practitioners, statistical rules, or some combination of the three.

Large language models now provide another potential interpreter.

A user can enter a series of personality-assessment scores into a general-purpose AI system and ask, "What does this mean about this person?" Within seconds, the model will produce a fluent and often highly persuasive interpretation. Similar functionality can be incorporated into AI systems used for coaching, employee development, communication support, hiring, education, and personalized digital interaction.

The fluency of such interpretations creates a methodological problem. A convincing interpretation is not necessarily a correct interpretation.

Psychological and educational testing standards distinguish the quality of an instrument from the validity of the interpretations made from its scores. Evidence supporting one does not automatically establish the other. The relevant question is therefore not simply whether an LLM can generate personality-related language. It is whether the model can correctly translate a particular set of standardized scores into the behavioral meanings those scores are intended to represent.

The present study examines that question directly.

Ten fabricated individuals were created to represent deliberately varied DISC Plus score configurations. Their score profiles were presented independently to five large language models. The models received the same information and were asked to interpret it under a common protocol.

***Given identical standardized personality-assessment scores, how accurately and consistently do major large language models reproduce the criterion-supported behavioral interpretation of those scores?***

That broad question contains several related but distinct measurement problems.

First is criterion-referenced interpretive validity: whether a model's interpretation agrees with the predetermined behavioral meaning assigned to the supplied score configuration.

Second is cross-model consistency, or inter-model agreement: whether different language models presented with identical assessment data reach substantively similar conclusions.

Third is discriminant sensitivity: whether models preserve meaningful differences between profiles, particularly when those differences are relatively small.

Fourth is epistemic restraint: whether a model recognizes when a requested inference cannot reasonably be supported by the supplied assessment data.

A secondary question concerns negative-trait attenuation: whether models preserve unfavorable but criterion-supported interpretations or soften them to the point that their substantive meaning changes.

These constructs should not be confused with the external validity of DISC Plus itself. The present study does not attempt to establish whether DISC Plus predicts job performance, life outcomes, or behavior in the general population. The instrument is held constant. The object being evaluated is the interpreter.

***Assuming these scores are to be interpreted according to the DISC Plus framework, does the language model interpret them correctly?***

# 2. Literature Review

## 2.1 Psychometric Interpretation as an Evidentiary Claim

Professional testing standards treat interpretation as more than a narrative exercise. The Standards for Educational and Psychological Testing, jointly developed by the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, emphasize that interpretations and uses of test scores require supporting evidence appropriate to their intended purpose.

This distinction is important for AI-mediated assessment. A validated instrument does not automatically confer validity on every interpretation subsequently generated from its scores. An automated

interpreter can introduce additional error even when the underlying scores themselves are psychometrically sound.

The validity question in the present study therefore concerns the correspondence between supplied assessment scores and the interpretation generated from them. That is distinct from asking whether the underlying assessment adequately measures personality in the first place.

### 2.2 Human Interpretation Is Not an Infallible Benchmark

Human professional judgment is often treated implicitly as the comparison standard for AI interpretation, but decades of psychological research caution against assuming that expert interpretation is inherently accurate.

Meehl's (1954) classic analysis compared clinical judgment with mechanical or actuarial prediction and found substantial evidence favoring formal combination rules. Grove et al. (2000) subsequently examined 136 studies comparing clinical and mechanical prediction and found that mechanical approaches were generally as accurate as or more accurate than clinical judgment. Ægisdóttir et al. (2006), reviewing 56 years of research relevant to counseling and clinical prediction, similarly reported an overall advantage for statistical approaches.

These findings do not imply that human interpreters are ineffective. They demonstrate that interpretation itself is an empirical performance problem.

An AI interpreter therefore should not be judged merely by whether its language resembles what a human practitioner might say. Its interpretation should be evaluated against an explicit criterion whenever such a criterion is available.

### 2.3 Perceived Accuracy Versus Criterion Accuracy

Another relevant literature concerns the Barnum or Forer effect: the tendency for individuals to accept broad, favorable, and high-base-rate personality descriptions as uniquely descriptive of themselves.

Reviews by Dickson and Kelly (1985) and Furnham and Schofield (1987) demonstrate that subjective acceptance of personality feedback can remain high even when the feedback is not individually diagnostic.

This distinction has direct consequences for evaluating AI interpretation. A fluent explanation may feel highly accurate to a user while nevertheless failing to reproduce the behavioral meaning implied by the underlying scores. Consequently, user satisfaction or perceived accuracy cannot by itself establish interpretive validity.

The present study therefore uses a predetermined criterion rather than participant acceptance as its principal benchmark.

### 2.4 Prior Evidence That LLMs Can Process Personality Information

Research increasingly demonstrates that large language models can extract, express, and manipulate personality-related information.

Peters, Cerf, and Matz (2024) found that GPT-4-based systems could infer Big Five traits from free-form conversation with moderate accuracy, although performance depended strongly on conversational structure.

Jiang et al. (2024), in the PersonaLLM project, showed that LLMs instructed to express particular Big Five traits generated language from which human observers could detect some intended personality characteristics with accuracy reaching approximately 80% under some conditions.

Rosenfelder, Levitin, and Gilead (2025) compared zero-shot GPT-4 with human judges attempting to infer psychological characteristics from self-descriptive texts. Across 30 psychological measures, model

predictions correlated with participant self-report at approximately r = .41 after correction for attenuation, compared with approximately r = .23 for human judges. Performance was strongest for personality and coping constructs and weaker for values and sociopolitical attitudes.

Serapio-García and colleagues developed a psychometric framework for measuring personality-like characteristics expressed by language models themselves. Their later peer-reviewed work demonstrated that some sufficiently scaled and instruction-tuned models can produce personality measurements with evidence of reliability and construct validity.

Collectively, these studies demonstrate that contemporary LLMs process personality-relevant information in nontrivial ways. They do not answer the specific question addressed here.

The task in the present study is not to infer personality from conversational behavior or to cause an LLM to express a target personality. The model receives structured psychometric scores and must accurately interpret what those scores mean.

### 2.5 Social Desirability and Negative Interpretive Attenuation

Research has also identified systematic pressures toward socially desirable output.

Salecha et al. (2024) found that large language models completing personality questionnaires showed social-desirability bias, shifting responses toward socially preferred poles when the models appeared to recognize that personality was being evaluated.

A related but conceptually different phenomenon is sycophancy. Sharma et al. (2023) defined sycophancy in terms of a model changing or selecting responses to align with a user's expressed beliefs or preferences, sometimes at the expense of truthfulness.

The present study does not directly manipulate user beliefs or ask models to agree with a preferred interpretation. It therefore does not provide a clean experimental test of sycophancy.

Instead, the relevant secondary construct is negative-trait attenuation.

An AI model may preserve the basic direction of an unfavorable interpretation while softening its language. For example, a criterion describing an individual as controlling or highly conflict-avoidant might be translated into coaching-oriented language such as preferring structure or seeking harmony.

Such phrasing is not necessarily inaccurate. However, if the substantive meaning becomes materially weaker than the criterion interpretation, attenuation becomes relevant to interpretive validity.

For that reason, negative-trait attenuation is treated here as an exploratory secondary analysis rather than as evidence of sycophancy.

### 2.6 Confidence and Unsupported Inference

A valid interpreter must know not only what the data support, but what they do not support.

Recent research on LLM confidence suggests that stated certainty often does not correspond cleanly with actual accuracy. Chhikara (2025) documented calibration problems across multiple models and question-answering datasets. Michael, BenShushan, Bien, and Moore (2026), in a preregistered study across multiple reasoning tasks, similarly found that model confidence exceeded accuracy on average, particularly on difficult problems.

The present design tests a narrower construct.

DISC Plus personality scores do not independently establish a person's general cognitive ability, integrity, or future job performance. Those questions therefore provide an opportunity to evaluate epistemic restraint.

The appropriate response is not a confident personality-based prediction. The model should communicate that the supplied data do not support the inference.

### 2.7 Prior Work on AI-Generated Personality Interpretation

Niblick (2025) compared AI- and human-generated DISC interpretations using a double-blind design involving 392 participants. AI-generated interpretations received higher participant ratings across six dimensions, including perceived accuracy and comprehension.

That study addressed how recipients experienced AI-generated personality interpretation. It did not determine whether those interpretations objectively matched a predetermined criterion.

That distinction creates the central gap addressed by the current study.

**Prior study: Does an AI interpretation seem accurate to the person receiving it?**

**Present study: Is the AI interpretation actually consistent with the criterion interpretation associated with the scores it was given?**

## 3. Study Objectives, Research Questions, and Hypotheses

The primary purpose of the study is to evaluate the validity and consistency of LLM interpretation of standardized personality-assessment profiles.

### H1 — Criterion-Referenced Accuracy Ceiling

No evaluated model will exceed 70% exact agreement with the predetermined behavioral criteria across the structured interpretation task.

### H2 — Error–Accuracy Relationship

Models demonstrating lower exact-match accuracy will demonstrate correspondingly greater mean absolute deviation from criterion ratings.

### H3 — Discriminant Sensitivity

Models will distinguish profiles containing large criterion-relevant differences more successfully than profiles separated by smaller but deliberately meaningful differences.

### H4 — Preservation of Negatively Valenced Information

Models will generally retain the direction of criterion-supported unfavorable personality characteristics rather than systematically reversing them into favorable opposites. The degree to which models soften the language of those interpretations will be examined separately as an exploratory analysis of negative-trait attenuation.

### H5 — Epistemic Restraint

A majority of evaluated models will produce at least some above-floor confidence judgments on questions that cannot legitimately be answered from the supplied assessment data.

### RQ1 — Cross-Model Reliability

To what degree do different LLMs agree with one another when interpreting identical personality profiles?

Because agreement with one another does not establish correctness, cross-model reliability must be analyzed separately from criterion validity. Five models could theoretically agree perfectly while all producing the same incorrect interpretation.

# 4. Methodology

## 4.1 Study Design

The study uses a repeated benchmark design in which five language models interpret an identical panel of ten DISC Plus personality profiles.

The personality profiles function as standardized test cases rather than as a random sample of human participants.

The unit of substantive interest is the interpretation generated from a given score configuration.

## 4.2 Stimulus Profiles

Ten fictional individuals were constructed.

Each profile was designed first at the behavioral level and subsequently represented numerically using DISC Plus scores.

Each included natural DISC scores for Dominance, Influence, Steadiness, and Compliance; adaptive DISC scores for the same four dimensions; and Values Index scores for Aesthetic, Economic, Individualistic, Political, Altruistic, Regulatory, and Theoretical motivation.

The ten-profile panel intentionally sampled several interpretive conditions. It included pronounced or extreme profiles, profiles containing less socially desirable behavioral tendencies, two pairs designed to be highly similar but not identical, strongly contrasting profiles, and one relatively moderate profile near the center of the score distributions.

This was not intended to represent the prevalence of personality types in a population. It was a deliberately constructed criterion-referenced benchmark panel designed to challenge different aspects of interpretation.

## 4.3 Criterion Development

Before the models were queried, each profile was associated with a predetermined behavioral interpretation.

The criterion included behavioral descriptions, expected structured item ratings, motivational priorities, and designated strengths and weaknesses.

The criterion was fixed before model evaluation.

The resulting benchmark therefore measures whether an LLM can recover the intended interpretation of a predefined score configuration.

The term criterion-referenced accuracy is used deliberately. The benchmark should not be interpreted as proving that the authored description represents objective psychological truth about a real individual. Nor does agreement with the benchmark establish the external validity of DISC Plus.

Instead, the criterion represents the interpretation the scoring framework specifies for the information supplied to the model.

For a publication-grade extension, independent assessment experts should review or independently construct the criterion mappings before model administration. This would reduce dependence on a single author's interpretation and strengthen the evidentiary basis of the benchmark.

## 4.4 Model Tasks

**Task A: Naturalistic Interpretation**

The model received the labeled DISC Plus scores and was asked to explain how to understand and communicate effectively with the person represented by the profile. This condition approximated the type of request a consumer, coach, manager, or practitioner might naturally submit to a general-purpose AI assistant.

**Task B: Structured Interpretation**

The model completed a structured rating protocol containing twelve behavioral judgment items, a seven-item motivational ranking, and three unsupported-inference items. For each substantive judgment, the model also supplied a confidence estimate. The unsupported items concerned characteristics such as general cognitive ability, integrity, and predicted job performance that could not validly be established from the supplied DISC Plus data alone.

**Task C: Strengths and Limitations**

The model was asked to identify the individual's five strongest characteristics and five most important weaknesses, limitations, or developmental risks. This task was used to examine whether criterion-supported negative information was retained, omitted, euphemized, or reversed.

## 4.5 Models and Administration

Five major LLM families were evaluated in the demonstration run: Claude, Gemini, Grok, GPT, and a Llama-based Meta model.

The student-facing dataset uses the model/version labels recorded during the demonstration. A formal empirical manuscript should report the exact provider, model identifier, access method, date of administration, relevant system settings, and any configurable parameters available at the time of testing.

The demonstration run administered all ten profiles to each model in a continuous session for that model.

This design is sufficient to examine criterion accuracy, between-model variation, discrimination, unsupported inference, and qualitative interpretation patterns. It does not establish within-model test-retest reliability.

A full reliability study should administer each profile in a new independent session and repeat the complete procedure multiple times.

## 4.6 Scoring

**Criterion-Referenced Exact Agreement**

For each structured behavioral judgment, the model's response was compared with the predetermined target response. Exact agreement was expressed as the proportion of scored judgments matching the criterion.

**Mean Absolute Deviation**

Mean absolute deviation quantified the average magnitude of disagreement between model ratings and criterion ratings. A model could therefore miss an exact target while still remaining relatively close to it.

**Directional Error**

Signed error was calculated to determine whether deviations tended systematically toward one end of the response scale. Where response poles differed in social desirability, this statistic was examined descriptively for evidence of directional interpretive drift. It should not, by itself, be interpreted as a measure of sycophancy.

**Cross-Model Agreement**

Cross-model agreement concerns whether different models assign similar interpretations to identical profiles. The complete empirical study should calculate a formal inter-rater agreement statistic from the item-level response matrix. For ordered or continuous rating data, an intraclass correlation coefficient or comparable agreement statistic is appropriate. For categorical judgments, a multi-rater kappa statistic may be used. Motivational rankings may be evaluated with an appropriate rank-agreement statistic such as Kendall's W. The demonstration summary below does not invent such a coefficient where the necessary calculation has not yet been reported.

**Discriminant Sensitivity**

The two deliberately similar profile pairs were compared item by item to determine whether models preserved the small criterion-relevant differences between them. Strongly contrasting profiles provided a comparison condition for determining whether models were capable of recognizing large differences.

**Unsupported-Inference Confidence**

Confidence ratings on cognitive ability, integrity, and predicted performance were compared with the expected floor-confidence response. The purpose was not to measure general probabilistic calibration but to determine whether models recognized that these specific questions exceeded what the provided data could support.

**Negative-Trait Attenuation**

Task C responses were qualitatively coded to determine whether criterion-supported weaknesses were preserved, omitted, softened, or transformed into favorable interpretations. This construct replaces the broader and less defensible label of sycophancy.

# 5. Findings

**SAMPLE / SIMULATED-LABEL COURSE DATA — NOT STUDENT-GENERATED FINDINGS**

## 5.1 Criterion-Referenced Accuracy

| Model | Exact agreement | Mean absolute deviation | Signed error | Unsupported items above confidence floor |
|---|---|---|---|---|
| Claude (Fable 5.1) | 75.0% | 0.25 | +0.06 | 50% |
| Gemini 3.6 Pro | 72.0% | 0.29 | −0.09 | 100% |
| Grok (Expert) | 51.3% | 0.58 | −0.27 | 30% |
| GPT (5.2 Sol) | 56.7% | 0.48 | +0.04 | 0% |
| Meta (Llama) | 63.3% | 0.38 | +0.08 | 30% |
| **Unweighted mean across models** | **63.7%** | **0.40** | **−0.04** | **42%** |

Exact agreement varied substantially across the five models, ranging from 51.3% to 75.0%. The unweighted mean of the five model-level accuracy rates was 63.7%.

H1 predicted that no model would exceed 70% exact criterion agreement. H1 was disconfirmed. Claude reached 75.0% and Gemini reached 72.0%.

This is an important result rather than a flaw in the study. A useful hypothesis must be capable of being wrong, and disconfirmation should be reported as prominently as confirmation.

At the same time, the results indicate substantial remaining interpretive error even among the highest-performing models. The difference between the highest- and lowest-performing models was 23.7 percentage points.

The principal conclusion from this result is therefore not that LLM interpretation is uniformly poor, nor that it is uniformly accurate. It is that criterion-referenced interpretive accuracy varies meaningfully by model.

### 5.2 Error Magnitude

Mean absolute deviation ranged from 0.25 for Claude to 0.58 for Grok.

The ordering of models by error magnitude tracked the ordering by exact-match accuracy: models with higher exact agreement exhibited smaller average deviations from the criterion.

H2 was supported in the demonstration dataset.

This convergence is useful because it suggests that the exact-match statistic is not being driven merely by models narrowly missing arbitrary categorical boundaries. Lower-accuracy models were also farther from the target when they were wrong.

### 5.3 Cross-Model Consistency

The substantial spread in criterion accuracy demonstrates meaningful performance heterogeneity across models.

However, variability in accuracy is not itself an inter-model reliability statistic.

A model can disagree with the criterion while still agreeing closely with other models, or conversely can be highly accurate while differing from models that make different errors.

Accordingly, a formal claim regarding cross-model reliability requires calculation directly from the complete item-level model response matrix.

The current demonstration therefore supports the descriptive conclusion that the models differed materially in interpretive performance, but it does not report a fabricated or inferred inter-rater reliability coefficient.

That coefficient should be included in the final empirical analysis because cross-model consistency is one of the study's primary research questions.

### 5.4 Discriminant Sensitivity

The clearest cross-model pattern appeared in the deliberately similar profiles.

Across the two similar profile pairs, the five models produced identical item-level ratings in 117 of 120 available pairwise profile comparisons, a rate of 97.5%. Only three of the designed small distinctions were preserved.

By contrast, profiles designed to be strongly different from one another were consistently distinguished.

H3 was supported.

The models appeared capable of recognizing large personality differences while frequently collapsing smaller criterion-relevant differences.

This distinction is theoretically important. An interpreter that can tell an extreme high-D/high-I profile from an extreme high-S/high-C profile may still lack the sensitivity required to distinguish two individuals whose scores differ modestly but meaningfully.

The finding should nevertheless be interpreted cautiously. A small numerical score difference should not automatically be treated as psychologically meaningful. In a final empirical version of the study, differences designated as criterion-relevant should be justified by the scoring framework and, ideally, shown to exceed trivial variation or known measurement error.

### 5.5 Negative-Trait Attenuation

Models were also asked directly to identify weaknesses and limitations.

The responses generally preserved the direction of criterion-supported negative characteristics. Models did not systematically convert unfavorable traits into their flattering opposites.

However, narrative phrasing frequently softened those characteristics.

A direct criterion such as excessive control, dependence on direction, conflict avoidance, impatience, or low interpersonal warmth could appear in more coaching-oriented language emphasizing preferences, developmental opportunities, or situations in which the characteristic might become limiting.

H4 was supported in its narrow form.

The models did not systematically reverse negative criterion information. However, it would be incorrect to conclude from this result that models showed no positivity bias.

The demonstration suggests a more nuanced phenomenon: preservation of direction accompanied by attenuation of language.

The present design does not justify labeling that effect sycophancy because users did not communicate a preferred answer that the model could accommodate. Negative-trait attenuation is the more accurate construct.

A future study could quantify attenuation by having independent raters compare the severity or directness of criterion statements with corresponding model-generated language.

### 5.6 Epistemic Restraint on Unsupported Items

The five models differed sharply in their confidence when responding to questions that could not legitimately be answered from DISC Plus scores alone.

The proportion of unsupported items answered above the expected confidence floor ranged from 0% for GPT to 100% for Gemini. The unweighted mean across models was 42%. Four of the five models produced at least some above-floor confidence on unsupported items.

The findings provide qualified support for H5 under the model-level interpretation specified in the hypothesis: a majority of models produced at least one instance of above-floor confidence on unsupported questions.

They do not support the stronger statement that most unsupported responses were confident. Fewer than half of the pooled model-level responses exceeded the expected confidence floor.

The result is therefore better characterized as substantial model heterogeneity in epistemic restraint than as a universal LLM calibration failure.

## 6. Discussion

The present study was designed to answer a relatively simple question that becomes more important as AI-generated interpretation becomes commonplace: If several large language models receive exactly the same personality-assessment scores, do they interpret those scores correctly and consistently?

The demonstration findings suggest that the answer is mixed.

Models showed meaningful capacity to interpret structured personality data. The strongest model-level exact agreement reached 75%, and all models captured at least some of the criterion structure.

At the same time, no model approached perfect agreement, and performance differed substantially between models.

The practical implication is that the phrase “AI interpretation” conceals meaningful model-level differences. The quality of an interpretation cannot be assumed simply because it was generated by a capable frontier language model.

### 6.1 Validity and Reliability Are Separate Problems

One of the most important conceptual distinctions emerging from the study is the difference between validity and reliability.

A model may be consistently wrong. Multiple models may agree with one another yet collectively disagree with the criterion. Alternatively, different models may use different language or intermediate judgments while still reaching broadly correct conclusions.

Criterion-referenced accuracy therefore evaluates one question: Did the model produce the interpretation it should have produced from these scores?

Cross-model reliability evaluates another: Did different models produce substantially the same interpretation from identical data?

Within-model reliability asks a third: Would the same model produce substantially the same interpretation if the identical profile were submitted again in a clean independent session?

The present demonstration provides criterion-validity evidence and information about between-model variability. It does not yet establish within-model test-retest reliability.

### 6.2 Small Differences Appear Particularly Vulnerable

The discrimination result may be the most practically consequential finding in the demonstration.

Models successfully recognized strong contrasts while erasing 97.5% of the designed item-level differences across closely similar profile pairs.

This suggests a potential compression effect.

Language models may recognize broad personality archetypes yet map nearby score configurations into the same generalized narrative.

If replicated, such a finding would matter in applications where personalization depends precisely on distinguishing people who are broadly similar but meaningfully different.

A system that reliably distinguishes extremes but treats nearby profiles as interchangeable may produce interpretations that appear personalized while operating at a much coarser level than the assessment itself.

### 6.3 Softening Is Not the Same as Reversal

The negative-trait findings require similar precision.

The models generally did not erase or reverse unfavorable information. They did, however, often express it more gently.

There are legitimate reasons for such phrasing. Professional coaching and assessment feedback should not be unnecessarily pejorative.

The methodological question is not whether a model uses diplomatic language. It is whether diplomacy changes the substantive interpretation.

Future work should therefore distinguish linguistic tone from construct fidelity.

A tactful restatement that preserves meaning is not an error. A flattering reformulation that materially reduces or reverses the criterion is.

### 6.4 Knowing What the Data Cannot Say

The unsupported-item condition identifies a different type of interpretive competence.

An assessment interpreter should recognize construct boundaries.

DISC Plus scores can support inferences about behavioral tendencies and motivational patterns within the framework. They cannot, by themselves, establish intelligence, honesty, or future occupational success.

Some models demonstrated strong restraint in these situations. Others did not.

This variation argues against treating hallucination, overconfidence, or unsupported inference as uniform properties of all LLMs. They appear to vary substantially by model and possibly by prompting and administration context.

### 6.5 Relation to Prior Research

The findings complement rather than contradict previous research showing that LLMs can infer personality information or produce convincing personality interpretations.

An AI interpretation can be persuasive, useful, and preferred by users while still containing criterion-level interpretive errors.

Likewise, an LLM can demonstrate meaningful personality inference from language while performing differently when asked to interpret a structured assessment instrument. These are related but distinct capabilities.

Prior work, including Niblick (2025), examined how recipients perceive AI-generated personality interpretation. The present framework shifts the criterion from acceptance to agreement with a predefined interpretation standard.

Together, the two approaches address different parts of the same larger validity problem.

An interpretation can be experienced as accurate and also be criterion-accurate. It can also be experienced as accurate while being objectively inconsistent with the scoring framework.

That distinction becomes increasingly important as AI systems move from casual explanation into organizational and professional use.

## 7. Limitations

- The benchmark includes ten deliberately authored profiles rather than a representative sample of real people. The findings therefore describe performance on this benchmark panel and should not be generalized directly to the population of assessment users.
- The criterion is framework-based and authored. It establishes what the supplied score pattern is intended to mean within the study, not objective psychological truth about a real person. Independent expert validation of the criterion matrix would strengthen future versions.
- The study evaluates interpretation of DISC Plus but does not independently establish the external construct, criterion, or predictive validity of the underlying assessment system. The validity of the instrument and the validity of an AI interpretation of that instrument are separate questions.
- The demonstration used a single continuous session for each model. Responses within a model may therefore have been influenced by prior profiles appearing in the same context window.
- Because each model completed only one demonstration pass, the study cannot establish within-model test-retest reliability. Repeated administrations in independent sessions are required.
- A formal inter-model reliability statistic should be calculated from the complete item-level response matrix before making a quantitative claim about cross-model reliability.

- Model behavior is version-dependent. Production models can change through updates that are invisible or only partially documented to users. Exact model identifiers, access dates, provider settings, and administration procedures must therefore be reported.
- Structured exact-match scoring necessarily simplifies the interpretive task. Two differently worded interpretations may sometimes be substantively equivalent. For that reason, quantitative scoring should be supplemented by transparent coding rules and, where narrative judgments are involved, independent raters and inter-rater agreement estimates.

## 8. Implications

The study does not establish that large language models are incapable of interpreting personality assessments. Nor does it establish that they should replace trained human interpreters.

It establishes a framework for asking a question that should precede either conclusion.

Before an AI system is trusted to interpret psychometric information, its performance should be evaluated against a criterion, its errors should be quantified, its sensitivity to meaningful score differences should be tested, and its willingness to exceed the evidentiary limits of the assessment should be measured.

The same principle applies regardless of whether the interpreter is human, algorithmic, or generative.

**Fluency is not validity.**

**Agreement is not necessarily accuracy.**

**Personalization is not demonstrated merely because two reports contain different words.**

A defensible AI assessment interpreter should demonstrate that it can produce the appropriate interpretation from the supplied scores, preserve meaningful distinctions between profiles, communicate unfavorable information without materially distorting it, and decline conclusions the assessment cannot support.

## 9. Recommended Full Reliability Extension

The next phase of the research should repeat each profile independently with each model multiple times.

Each model/profile combination should begin in a clean context with identical system and user instructions.

A minimum of three repetitions would permit an initial stability analysis; five or more repetitions would provide a stronger basis for estimating within-model reliability.

The resulting design would permit simultaneous evaluation of three distinct properties:

- Criterion validity: Does the interpretation agree with the predefined target?
- Cross-model reliability: Do different models agree with one another?
- Within-model reliability: Does the same model reproduce its own interpretation across independent administrations?

This expanded design would make it possible to identify four practically different classes of interpreter: models that are accurate and stable, accurate but unstable, consistently inaccurate, or both inaccurate and unstable.

## 10. Plain-Language Interpretation for End Users

For a person considering using an LLM to interpret a DISC Plus profile, the present findings support a cautious but not dismissive conclusion.

An LLM can provide a useful explanation of the broad themes in a DISC Plus profile, but it should not automatically be treated as the authoritative interpretation. In this benchmark, exact agreement with the predetermined criterion ranged from 51.3% to 75.0% depending on the model, a spread of 23.7 percentage points. The highest-performing model was therefore materially more accurate than the lowest-performing model on the same underlying task.

The models also performed better at recognizing large profile differences than small ones. This means an LLM may identify the general personality pattern while missing subtler distinctions that separate one person from another with a similar profile.

The most defensible consumer message is therefore: "AI can help explain your personality profile, but do not confuse a confident explanation with a correct one."

The study does not suggest that LLM interpretation is useless. Rather, it suggests that current LLMs are meaningful but imperfect interpreters, that model choice matters, and that high-stakes conclusions should remain anchored to the assessment framework and to interpreters who understand its limits.

## References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. American Educational Research Association.

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., Nichols, C. N., Lampropoulos, G. K., Walker, B. S., Cohen, G., & Rush, J. D. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. The Counseling Psychologist, 34(3), 341–382.

Chhikara, P. (2025). Mind the confidence gap: Overconfidence, calibration, and distractor effects in large language models. Transactions on Machine Learning Research.

Dickson, D. H., & Kelly, I. W. (1985). The "Barnum effect" in personality assessment: A review of the literature. Psychological Reports, 57(2), 367–382.

Furnham, A., & Schofield, S. (1987). Accepting personality test feedback: A review of the Barnum effect. Current Psychology, 6(2), 162–178.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. Psychological Assessment, 12(1), 19–30.

Jiang, H., Zhang, X., Cao, X., Breazeal, C., Roy, D., & Kabbara, J. (2024). PersonaLLM: Investigating the ability of large language models to express personality traits. In Findings of the Association for Computational Linguistics: NAACL 2024 (pp. 3605–3627).

König, C. J., & Marcus, B. (2013). TBS-TK Rezension: "persolog-Persönlichkeitsprofil." Psychologische Rundschau, 64(3), 189–191.

Meehl, P. E. (1954). Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. University of Minnesota Press.

Michael, N., BenShushan, D., Bien, J., & Moore, D. A. (2026). Confidence calibration in large language models. arXiv:2605.23909.

Niblick, J. N. (2025). Comparative efficacy of artificial intelligence and human expertise in interpreting personality constructs [Doctoral dissertation, Touro University Worldwide].

Peters, H., Cerf, M., & Matz, S. C. (2024). Large language models can infer personality from free-form user interactions. arXiv:2405.13052.

Rosenfelder, A., Levitin, M. D., & Gilead, M. (2025). Towards social superintelligence? AI infers diverse psychological traits from text without specific training, outperforming human judges. Computers in Human Behavior: Artificial Humans, 6, 100228.

Salecha, A., Ireland, M. E., Subrahmanya, S., Sedoc, J., Ungar, L. H., & Eichstaedt, J. C. (2024). Large language models display human-like social desirability biases in Big Five personality surveys. PNAS Nexus, 3(12), pgae533.

Serapio-García, G., Safdari, M., Crepy, C., Sun, L., Fitz, S., Romero, P., Abdulhai, M., Faust, A., & Matarić, M. (2025). A psychometric framework for evaluating and shaping personality traits in large language models. Nature Machine Intelligence, 7, 1954–1968.

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., & Perez, E. (2023). Towards understanding sycophancy in language models. arXiv:2310.13548.

## Assignment Notes

Students completing their own version of this study should independently construct or validate a benchmark panel, freeze their criterion interpretations and scoring rules before querying any models, document the exact models and administration settings used, run profiles in independent sessions, repeat administrations when making within-model reliability claims, calculate formal cross-model agreement from the item-level data, disclose any contamination or protocol deviations, and report disconfirmed hypotheses with the same prominence as supported ones.

**Is the model correct?**

**Do different models agree?**

**Does the same model give the same answer again?**

Those are validity, cross-model reliability, and within-model reliability, respectively. They are related, but they are not interchangeable.