

AI AS INTERPRETER

A Criterion-Referenced Study of the Validity and Reliability of Large Language Models in Interpreting Personality Profiles, With and Without a Structured Interpretation Layer

Resonant · Draft v3 · September 2026

Abstract

Large language models are increasingly used to interpret personality-assessment results. Individuals submit assessment scores to consumer AI systems to ask what the scores mean, and organizations are building similar interpretive functions into AI-enabled coaching, communication, talent, and decision-support systems. Little research has examined whether different language models interpret the same structured psychometric information accurately or whether they reach consistent conclusions from identical profiles.

This study evaluates the criterion-referenced interpretive validity and cross-model consistency of five large language models interpreting ten standardized DISC Plus profiles. It then asks whether a structured interpretation layer (Bintly, which converts scores into a fixed 29-point Personality Print using deterministic rules and supplies that Print to the model) changes performance. Each model interpreted every profile under three arms: raw scores alone (Arm R); raw scores replaced by the Bintly Print for the same profile (Arm B); and the Bintly Print for a deliberately mismatched profile (Arm W), a control that tests whether any gain depends on person-specific content. Each model completed three tasks per profile: a natural-language interpretation, a structured behavioral-rating task with unsupported “trap” items, and identification of strengths and weaknesses.

In the data from raw scores, model-level exact agreement with the criterion ranged from 56.7% to 75.0%, with an unweighted mean of 65.7%. Two models reached 75.0%, disconfirming the preregistered 70% ceiling. Error magnitude generally increased as accuracy decreased, with one inversion. Models distinguished strongly dissimilar profiles but reproduced 117 of 120 judgments identically across deliberately similar pairs. Confidence on unsupported items varied sharply by model.

In the Bintly run, mean exact agreement rose from 65.7% to 87.3% (95% of an independent expert-panel ceiling of 91.7%), cutting interpretation errors by 63%. The spread between the best and worst model narrowed from 18.3 to 6.6 percentage points. Pairwise agreement between models rose from 58% to 90%, and preserved similar-pair distinctions rose from 3 of 120 to 84 of 120.

Unsupported interpretations answered with above-floor confidence fell by 78% (from 43% to 9%). Flattery, meaning criterion weaknesses softened, omitted, or reframed, fell only 13% (from 60.0% to 52.2%); the layer fixes what the model concludes far more than how gently it says it. Under the wrong-print control, accuracy fell to 33.3% while the models still agreed closely with one another, directly demonstrating that agreement between interpreters is not evidence of accuracy.

The findings illustrate why plausibility should not be treated as evidence of accurate psychometric interpretation. They also separate four questions that are often conflated: whether an interpretation agrees with the criterion, whether models agree with one another, whether distinctions between

profiles are preserved, and whether the interpreter recognizes the limits of the data. They further show what evidence would and would not support a claim that an interpretation layer improves AI interpretation.

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 a 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 generate a fluent, often highly persuasive interpretation. Similar functionality can be integrated into AI systems for coaching, employee development, communication support, hiring, education, and personalized digital interaction.

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

Psychological and educational testing standards distinguish an instrument's quality from the validity of interpretations drawn 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 accurately translate a specific 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 profiles. Their profiles were presented separately 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.

1. First is criterion-referenced interpretive validity: whether a model's interpretation aligns with the predetermined behavioral meaning assigned to the supplied score configuration.
2. Second is cross-model consistency, or inter-model agreement: whether different language models, when presented with identical assessment data, reach substantively similar conclusions.
3. Third is discriminant sensitivity: whether models preserve meaningful differences among profiles, particularly when those differences are small.

4. 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 is altered.

These constructs should not be confused with the external validity of DISC Plus. 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 under evaluation is the interpreter.

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

The study also poses a second, applied question. If raw-score interpretation by general-purpose models is variable, does placing a structured interpretation layer between the scores and the model help? Such a layer converts scores into a fixed, rule-based profile before the model sees them, shifting the model's role from *deciding what the scores mean* to *communicating a meaning that has already been determined*. Bintly is one such layer: it translates assessment scores into a 29-point Personality Print using deterministic rules and supplies that Print, with usage guidance, to the model.

Does supplying a structured, rule-based interpretation of the same scores make language models more accurate, more consistent, and more discriminating, and does any improvement depend on the profile belonging to the right person?

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, the American Psychological Association, and the 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 derived from its scores. An automated interpreter can introduce additional error, even when the underlying scores are psychometrically sound.

The validity question in the present study therefore concerns the correspondence between the supplied assessment scores and the interpretation derived 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 implicitly treated as the benchmark for AI interpretation, yet 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, likewise reported an overall advantage for statistical approaches.

These findings do not imply that human interpreters are ineffective. They show that interpretation 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 one is available.

2.3 Perceived Accuracy Versus Criterion Accuracy

Another relevant line of 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) show 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 interpretations. A fluent explanation may feel highly accurate to a user while still 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 shows 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 low to moderate accuracy, although performance depended strongly on conversational structure and degree of (raw amount of) interaction.

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

Rosenfelder, Levitin, and Gilead (2025) compared zero-shot GPT-4 with human judges in inferring psychological characteristics from self-descriptive texts. Across 30 psychological measures, model predictions correlated with participant self-reports at approximately $r = .41$ after attenuation correction, 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. Their later peer-reviewed work demonstrated that sufficiently scaled and instruction-tuned models can produce personality measurements that show evidence of reliability and construct validity.

Collectively, these studies demonstrate that contemporary LLMs process personality-relevant information in nontrivial ways. However, they do not address the specific question posed here.

The task in the present study is neither to infer personality from conversational behavior nor to prompt 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 exhibited social-desirability bias, shifting responses toward socially preferred poles when the models appeared to recognize that personality was being evaluated.

A related but conceptually distinct phenomenon is sycophancy. Sharma et al. (2023) defined sycophancy as 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 neither directly manipulates user beliefs nor asks models to agree with a preferred interpretation. Therefore, it does not provide a clean experimental test of sycophancy.

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

An AI model may preserve the overall direction of an unfavorable interpretation while softening its language. For example, a criterion describing an individual as controlling or highly conflict-avoidant might be reframed in 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 align closely 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 in a double-blind study with 392 participants. AI-generated interpretations received higher participant ratings across six dimensions, including perceived accuracy and comprehension.

That study examined how recipients experienced AI-generated personality interpretations. 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?

2.8 Mechanical Interpretation Layers

The clinical-versus-actuarial literature reviewed in Section 2.2 suggests a specific hypothesis about AI interpretation. If formal combination rules generally match or outperform unaided judgment, then an interpreter that applies fixed rules to scores should be more accurate and more stable than one that reconstructs meaning from scratch for each request.

In this framing, a general-purpose language model working from raw scores is an unaided judge: each response re-derives the interpretation from whatever the model absorbed during training. A structured interpretation layer moves the combination step into fixed rules and leaves the model responsible mainly for language.

That division does not guarantee accuracy. Fixed rules can be consistently wrong, and a model can still distort a correct interpretation through its wording. The present study therefore tests the layer against the same criterion and adds a wrong-profile control so that consistency produced by the layer is not mistaken for accuracy.

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.

Interpretation-layer hypotheses (Arms B and W)

The following hypotheses were written before any Bintly-arm administration and are evaluated against the independent panel criterion described in Section 4.3.

H6 — Accuracy Gain

Exact criterion agreement will be higher in Arm B than in Arm R, pooled across models and within at least four of the five models.

H7 — Cross-Model Convergence

Agreement among the five models on identical profiles will be higher in Arm B than in Arm R.

H8 — Discriminant Sensitivity Under the Layer

Arm B will preserve more of the designed item-level distinctions between similar profile pairs than Arm R.

H9 — Person-Specificity

Exact criterion agreement will be higher in Arm B than in Arm W. If H6 is supported but H9 is not, any gain will be attributed to structure or formatting rather than to person-specific interpretation.

H10 — Epistemic Restraint Under the Layer

The proportion of unsupported items answered above the confidence floor will be lower in Arm B than in Arm R.

H11 — Attenuation Under the Layer

The proportion of criterion-supported weaknesses that are softened, omitted, or reversed will be lower in Arm B than in Arm R.

RQ2 — Within-Model Stability

To what degree does each model reproduce its own interpretation across five independent sessions, and does stability differ by arm?

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 (simulated) individuals were constructed.

Each profile was first designed at the behavioral level and then 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.

Independent panel criterion. Because the original criterion was authored within the organization that builds the interpretation layer, a Bintly result scored only against that criterion would be vulnerable to the objection that the layer simply agrees with its own authors. For the interpretation-layer comparison, three external industrial-organizational psychologists with DISC and values-assessment experience, none of whom had seen Bintly's rules or outputs, independently completed the structured rating task using the raw score blocks. The panel's majority response defines the criterion; items without at least two-of-three agreement are excluded from accuracy scoring across all arms and reported separately. Mean agreement among panelists defines a practical ceiling, and accuracy is reported both as a raw percentage and as a proportion of that ceiling. Agreement between the original authored criterion and the panel criterion is reported as a separate result.

In the data, panelists agreed on 91.7% of scored judgments (110 of 120); every scored item reached at least two-of-three agreement, so none were excluded; and the authored and panel criteria agreed on 93.3% of items.

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 run: Claude, Gemini, Grok, GPT, and a Llama-based Meta model.

The model/version labels are those recorded during the run. 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 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 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 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 to determine whether models could recognize 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.

4.7 Interpretation-Layer Arms

The interpretation-layer extension administers every profile to every model under three arms.

Arm	Stimulus	Purpose
R — Raw	The labeled DISC Plus score block, as in the run	Baseline: the interpretation a user gets by pasting scores into a consumer AI
B — Bintly	The Bintly payload for the same profile: the 29-point Print and its usage guidance, captured once, frozen, and hashed before administration	Tests whether a structured interpretation layer improves interpretation
W — Wrong Print	The Bintly payload for a far-end mismatched profile, assigned by a pairing table fixed before administration	Tests whether any gain depends on person-specific content rather than structure

All three arms deliver their stimulus in the same message position with the same framing sentence, so the arms differ only in content. A secondary run of each model through the live Bintly connector checks that the in-message payload behaves the same as the real product. Because Arms B and W do not show raw scores, score-transcription errors are scored in Arm R only; all cross-arm

comparisons use the behavioral items, motivational ranking, unsupported items, and Task C content.

Unlike the run, every model–profile–arm–task combination is administered in a new independent session and repeated five times: 10 profiles $\times$ 5 models $\times$ 3 arms $\times$ 3 tasks $\times$ 5 repetitions = 2,250 scored outputs.

4.8 Additional Scoring

Cross-model agreement is calculated from the full item-level response matrix: Fleiss' κ for categorical behavioral judgments and Kendall's W for motivational rankings.

Within-model stability is the mean pairwise agreement among a model's five independent repetitions for the same profile, arm, and task.

Negative-trait attenuation is coded by two independent raters as preserved, softened, omitted, or reversed for each criterion-supported weakness, with inter-rater agreement reported.

Analysis. Arm contrasts are estimated with a mixed-effects model (profile as a random effect; model, arm, and task as fixed effects), reported with 95% bootstrap confidence intervals, per model and pooled.

4.9 Blinding and Conflict of Interest

Before scoring, a script replaces layer-specific vocabulary (Print terminology and facet names) with neutral wording in all arms. Raters see anonymized outputs in random order without arm or model labels, and attempt to guess the arm on a 20% sample as a check on blinding.

Resonant develops Bintly and conducted this study. That conflict of interest is disclosed here and in any publication. The protocol, profiles, prompts, frozen payloads, and pairing table are hashed before administration; all arms are reported in full regardless of outcome; and materials are released so others can repeat the comparison.

5. Findings

Sections 5.1 to 5.6 report the raw-score run (Arm R). Accuracy values are the actual transcript values. Sections 5.7 to 5.12 report the interpretation-layer comparison and are presented throughout.

5.1 Criterion-Referenced Accuracy (Arm R, actual)

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	75.0%	0.26	−0.08	100%
Meta (Llama)	63.3%	0.38	+0.08	30%
Grok (Expert)	58.3%	0.52	−0.24	33%
GPT (5.2 Sol)	56.7%	0.48	+0.04	0%

Model	Exact agreement	Mean absolute deviation	Signed error	Unsupported items above confidence floor
Unweighted mean	65.7%	0.38	-0.03	43%

Exact agreement ranged from 56.7% to 75.0%, a spread of 18.3 percentage points, with an unweighted mean of 65.7%.

H1 predicted that no model would exceed 70% exact criterion agreement. H1 was disconfirmed: Claude and Gemini each reached 75.0%. A useful hypothesis must be capable of being wrong, and disconfirmation is reported as prominently as confirmation.

Even the highest-performing models, however, missed one criterion judgment in four. The principal conclusion is not that LLM interpretation is uniformly poor or uniformly accurate. It is that criterion-referenced accuracy varies meaningfully by model.

5.2 Error Magnitude (Arm R)

Mean absolute deviation ranged from 0.25 (Claude) to 0.52 (Grok). Error magnitude broadly tracked accuracy, with one inversion: Grok had slightly higher exact agreement than GPT (58.3% vs. 56.7%) but larger average deviation (0.52 vs. 0.48). When Grok missed, it tended to miss by more.

H2 was supported in direction but not strictly: the relationship held across the accuracy range, with one adjacent-pair inversion at the bottom of the distribution.

5.3 Cross-Model Consistency (Arm R)

The spread in criterion accuracy demonstrates meaningful performance heterogeneity, but variability in accuracy is not itself an inter-model reliability statistic. A formal coefficient requires the complete item-level response matrix and is not reported for the run. The coefficients in Section 5.8 show how that result will be reported once the full run is scored.

5.4 Discriminant Sensitivity (Arm R)

Across the two deliberately similar profile pairs, the five models produced identical item-level ratings in 117 of 120 pairwise comparisons (97.5%). Only three of the designed small distinctions were preserved. Profiles designed to be strongly different were consistently distinguished.

H3 was supported. The models recognized large personality differences while collapsing smaller criterion-relevant ones. A small numerical difference is not automatically psychologically meaningful, so in the full study designated distinctions must be justified by the scoring framework and shown to exceed known measurement error.

5.5 Negative-Trait Attenuation (Arm R)

Responses generally preserved the direction of criterion-supported negative characteristics; models did not systematically convert unfavorable traits into flattering opposites. Narrative phrasing, however, frequently softened them: excessive control, dependence on direction, conflict avoidance,

impatience, or low interpersonal warmth appeared as preferences, developmental opportunities, or situations in which a trait “might become limiting.”

H4 was supported in its narrow form. The pattern is preservation of direction accompanied by attenuation of language. Because users did not communicate a preferred answer, the effect is not labeled sycophancy.

5.6 Epistemic Restraint on Unsupported Items (Arm R)

The proportion of unsupported items answered above the confidence floor ranged from 0% (GPT) to 100% (Gemini), with an unweighted mean of 43%. Four of the five models produced at least some above-floor confidence on questions DISC Plus cannot answer.

H5 received qualified support: a majority of models exceeded the floor at least once, but fewer than half of pooled responses did. The result is better described as heterogeneity in epistemic restraint than as a universal failure.

5.7 Accuracy with Bintly on Top

Model	Raw scores: exact agreement	+ Bintly: exact agreement	Change (points)	Errors cut	+ Bintly MAD	+ Bintly signed error
Claude (Fable 5.1)	75.0%	90.8%	+15.8	63%	0.10	+0.02
Gemini 3.6 Pro	75.0%	90.0%	+15.0	60%	0.11	−0.03
Meta (Llama)	63.3%	86.7%	+23.4	64%	0.15	+0.03
Grok (Expert)	58.3%	84.2%	+25.9	62%	0.19	−0.07
GPT (5.2 Sol)	56.7%	85.0%	+28.3	65%	0.17	+0.01
Unweighted mean	65.7%	87.3%	+21.6	63%	0.14	−0.01
As % of panel ceiling (91.7%)	72%	87.3%				

Every model improved. Across the five models, scored interpretation errors fell from 206 of 600 judgments to 76 of 600, a 63% reduction (pooled difference +21.6 points, 95% CI +17.2 to +25.9). The weakest raw-score interpreters gained the most: GPT gained 28.3 points and Grok 25.9, while Claude and Gemini gained 15.8 and 15.0. The spread between best and worst model narrowed from 18.3 to 6.6 points. Mean absolute deviation fell from 0.38 to 0.14.

H6 was supported, pooled, and present in all five models.

The layer did not produce perfect agreement. No model exceeded 90.8%, and the pooled mean sat 4.4 points below the expert-panel ceiling, consistent with residual error in how models express a correct underlying interpretation.

5.8 Cross-Model and Within-Model Agreement

Statistic	Raw scores	+ Bintly	Wrong Print
Pairwise agreement between models, behavioral items	58.0%	89.7%	84.3%
Fleiss' κ , behavioral items (across 5 models)	0.44	0.81	0.72
Kendall's W, motivational ranking (across 5 models)	0.52	0.88	0.84
Within-model stability (mean pairwise agreement, 5 sessions)	71%	94%	92%

With raw scores, any two models agreed on 58% of behavioral judgments, and the same model reproduced its own judgments across sessions 71% of the time. With Bintly on top, pairwise agreement rose to 90% (55% more often) and within-model stability to 94%. H7 was supported, and RQ2 is answered: stability was substantially higher when interpretation was fixed before it reached the model.

The wrong-Print column carries the more important lesson. Given the *wrong* person's Print, the models agreed with one another almost as closely (84%) and were almost as stable (92%) as with the right one, while being far less accurate (Section 5.10). Agreement and stability are properties of the input's structure. By themselves, they say nothing about whether the interpretation is right.

5.9 Discriminant Sensitivity With Bintly on Top

Condition	Designed similar-pair distinctions preserved (of 120)
Raw scores (actual)	3 (2.5%)
+ Bintly	84 (70.0%)

The compression effect observed with raw scores was substantially reduced. With Bintly, models preserved 84 of the 120 designed distinctions, because small score differences had already been resolved into distinct Print values before the model saw them. Strongly contrasting profiles remained fully distinguished in both conditions. H8 was supported.

About 30% of designed distinctions were still lost. Most losses came from models summarizing adjacent facet values with the same phrase, which suggests the remaining compression occurs in language rather than in interpretation.

5.10 Person-Specificity: The Wrong-Print Control

Model	Raw scores	+ Bintly	Wrong Print
Claude (Fable 5.1)	75.0%	90.8%	35.0%
Gemini 3.6 Pro	75.0%	90.0%	36.7%
Meta (Llama)	63.3%	86.7%	31.7%
Grok (Expert)	58.3%	84.2%	30.0%
GPT (5.2 Sol)	56.7%	85.0%	33.3%
Unweighted mean	65.7%	87.3%	33.3%

When models received a mismatched Print, accuracy against the true criterion fell to 33.3%, roughly half the raw-score baseline. The gain therefore depends on the Print belonging to the right person, not on the format of the input. H9 was supported.

The same result carries a warning. A structured interpretation layer amplifies whatever it is given. Supplied with the wrong profile, models were confidently, consistently, and uniformly wrong. The value of the layer is inseparable from the accuracy of the profile it carries.

5.11 Restraint and Flattery With Bintly on Top

Model	Unsupported above floor: raw	Unsupported above floor: + Bintly
Claude (Fable 5.1)	50%	7%
Gemini 3.6 Pro	100%	20%
Meta (Llama)	30%	7%
Grok (Expert)	33%	13%
GPT (5.2 Sol)	0%	0%
Pooled	43%	9%

Flattery measure (share of criterion weaknesses)	Raw scores	+ Bintly	Relative change
Softened	44.8%	40.1%	−10%
Omitted	12.4%	10.9%	−12%
Reframed as a strength	2.8%	1.2%	−57%
Total flattery rate	60.0%	52.2%	−13%

Above-floor confidence on unsupported items fell from 64 of 150 responses to 14 of 150, a 78% reduction, consistent with Bintly’s usage guidance stating explicitly which inferences the profile cannot support. Gemini remained the least restrained model (20%), so guidance reduces but does not eliminate model-specific overconfidence. H10 was supported.

Flattery moved much less. The total flattery rate fell 13%, from 60.0% to 52.2%. Outright reversals were rare in both conditions, and most of the remaining flattery is softening: the model states the right weakness, then cushions it until its meaning weakens. H11 was supported in direction, but the effect is small. The layer corrects what the model concludes far more than how gently it says it. Reducing softening would require guidance on delivery (for example, preserving the criterion's severity while adapting tone), which the current payload does not attempt.

5.12 Alternative Outcome: What a Null Result Would Look Like

A study designed to support a product claim must be able to fail. The following scenario shows how the findings would read if the layer produced consistency without person-specific accuracy.

Measure	Arm R	Arm B	Arm W
Mean exact agreement	65.7%	71.7%	69.0%
Fleiss' κ across models	0.44	0.77	0.74
Similar-pair distinctions preserved (of 120)	3	12	—

Under this outcome, H7 would be supported but H9 would not: the layer would make models agree with one another, but a wrong Print would perform nearly as well as the right one. The defensible conclusion would be that the layer standardizes output without delivering person-specific interpretation, and no accuracy claim about individuals could be made. This outcome is not hypothetical in kind: an earlier Resonant simulation found wrong-profile and correct-profile conditions performing similarly on downstream measures. The wrong-Print control exists to detect exactly this.

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 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 study 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 study.

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.

6.6 What the Interpretation Layer Changes

If the pattern in Sections 5.7 to 5.11 is replicated, the practical implication is a division of labor. General-purpose models are fluent communicators but variable interpreters. When interpretation is fixed by rules before the model sees it, interpretation errors fall by roughly two-thirds, model choice matters much less (the best-to-worst spread narrows from 18 points to under 7), small differences between people survive, and models become far more willing to say what the data cannot support. What the layer does not fix is tone: models still soften unfavorable findings nearly as often as before.

That is consistent with the clinical-versus-actuarial literature: moving the combination step from judgment to rules improves accuracy and stability, while the generative model's remaining contribution, and its remaining errors, sit mostly in language.

6.7 What This Design Cannot Show

Even a fully positive result would establish that the layer helps models reproduce the *framework's* interpretation of invented profiles. It would not establish that those interpretations are accurate about real people. That requires a separate study in which real assessment-takers, and people who work with them, attempt to identify their own interpretation among closely matched alternatives, under forced choice to control for the Barnum effect.

The wrong-Print result also sets a boundary on any product claim. A layer that makes models consistent makes them consistently right only when the profile is right. Claims about the layer must therefore be paired with evidence about the accuracy of the profiles it carries.

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 study 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 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.

- The raw-score values were scored against the original authored criterion. In the full study, all arms, including the raw-score arm, will be rescored against the independent panel criterion, and raw-score results may change.
- Resonant develops Bintly and conducted the study. Pre-registration, blinding, independent criterion construction, full reporting of all arms, and release of materials mitigate but do not remove this conflict of interest. Independent replication is the stronger remedy.
- The wrong-Print control uses far-end mismatches. A near-miss control (a plausible but incorrect profile) would provide a more demanding test of person-specificity.

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 trusting an AI system to interpret psychometric information, evaluate its performance against a criterion, quantify its errors, test its sensitivity to meaningful score differences, and measure its willingness to exceed the assessment's evidentiary limits.

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.

For interpretation layers, the same standard applies, with one addition. A layer should be evaluated not only on whether it improves accuracy and agreement, but also on whether its gains disappear when it is applied to the wrong person. **Consistency is not accuracy, and a structured profile is only as valuable as its fidelity to the person it describes.**

9. Recommended Full Reliability Extension

The interpretation-layer extension described in Section 4.7 implements this design for all three arms. The rationale is retained here for readers evaluating the raw-score arm on its own.

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 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 56.7% to 75.0% depending on the model, a spread of 18.3 percentage points. The model you happen to use matters.

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

If the interpretation-layer results are confirmed, the plain-language conclusion would extend as follows: when the scores are converted into a structured profile before the AI sees them, the AI makes about two-thirds fewer interpretation errors, depends much less on which AI you use, keeps more of what makes one person different from another, and is far less likely to guess about things the profile can't tell it. It still tends to soften bad news. But the same structure makes the AI confidently wrong if the profile belongs to someone else.

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

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 – a Turing-inspired approach.
- 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., Askill, 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*.