MAPPING PATIENT-PERCEIVED PHYSICIAN TRAITS FROM NATIONWIDE ONLINE REVIEWS WITH LLMS

Junjie Luo[1], **Rui Han**[3,4], **Arshana Welivita**[2], **Zeleikun Di**[3], **Jingfu Wu**[3], **Xuzhe Zhi**[2], **Tianli Xu**[3], **Ritu Agarwal**[3,4], **Gordon Gao**[3,4,*]

[1]Johns Hopkins School of Medicine, Baltimore, MD, USA.
[2]Johns Hopkins University, Baltimore, MD, USA.
[3]Carey Business School, Johns Hopkins University, Baltimore, MD, USA.
[4]Center for Digital Health Artificial Intelligence (CDHAI), Baltimore, MD, USA.

*Correspondence: gordon.gao@jhu.edu

ABSTRACT

Understanding how patients perceive their physicians is essential to improving trust, communication, and satisfaction. We present a large language model (LLM)-based pipeline that infers Big Five personality traits and five patient-oriented subjective judgments. The analysis encompasses 4.1 million patient reviews of 226,999 U.S. physicians from an initial pool of one million. We validate the method through multi-model comparison and human expert benchmarking, achieving strong agreement between human and LLM assessments (correlation coefficients 0.72-0.89) and external validity through correlations with patient satisfaction (r = 0.41-0.81, all $p < 0.001$). National-scale analysis reveals systematic patterns: male physicians receive higher ratings across all traits, with largest disparities in clinical competence perceptions; empathy-related traits predominate in pediatrics and psychiatry; and all traits positively predict overall satisfaction. Cluster analysis identifies four distinct physician archetypes, from "Well-Rounded Excellent" (33.8%, uniformly high traits) to "Underperforming" (22.6%, consistently low). These findings demonstrate that automated trait extraction from patient narratives can provide interpretable, validated metrics for understanding physician-patient relationships at scale, with implications for quality measurement, bias detection, and workforce development in healthcare.

1 Introduction

Interpersonal and professional qualities of physicians profoundly shape patient trust, communication, adherence, and health outcomes [1, 2]. Understanding these qualities from the patient's perspective is essential to advancing patient-centered care, yet current measurement tools—such as standardized surveys or aggregate star ratings—capture only a narrow view of the physician-patient relationship. In parallel, millions of online physician reviews now provide an abundant, patient-generated record of real-world experiences, offering an unprecedented opportunity to examine how physicians are perceived in everyday practice [3, 4, 5, 6].

Extracting clinically meaningful information from such narrative data remains challenging. Prior studies have typically relied on sentiment analysis or topic modeling, approaches that overlook the multidimensional nature of patient perceptions. Well-established frameworks from psychology, such as the Big Five personality traits [7], offer interpretable constructs for describing interpersonal style, but have rarely been operationalized at scale in healthcare settings [8]. Similarly, healthcare-specific qualities—communication effectiveness, perceived competence, attentiveness to outcomes, and trustworthiness—are widely recognized as central to care quality but are difficult to measure systematically. Manual coding of these traits is costly, inconsistent, and infeasible for national datasets.

Recent advances in large language models (LLMs) enable a new approach [9]. Modern LLMs can synthesize information across large volumes of unstructured text, identify nuanced patterns, and generate structured outputs with

reasoning transparency [10, 11, 12, 13]. In healthcare, LLMs have been applied to tasks such as clinical summarization, guideline generation, and patient query answering [14, 15], but their potential for mapping patient-perceived physician traits at population scale remains untapped and unvalidated.

Here, we present and validate an LLM-based pipeline that infers ten clinically relevant traits from millions of online physician reviews: the Big Five personality dimensions and five patient-oriented subjective judgments. Using multi-model comparisons, an LLM-as-a-judge evaluation framework, and human annotation benchmarking, we demonstrate the reliability of this method and apply it to more than 226,999 U.S. physicians [16]. We characterize national trait distributions, explore correlations between personality and patient-oriented judgments, examine differences by gender and specialty, and identify latent physician archetypes through cluster analysis.

By converting unstructured patient narratives into interpretable psychological profiles, this work introduces a scalable and transparent methodology for capturing the interpersonal and professional dimensions of care. These insights not only deepen our understanding of how physicians are perceived but also provide a foundation for advancing fairness, bias detection, and personalization in healthcare delivery. We now present the results of applying this methodology at national scale, beginning with dataset characteristics and trait extraction validation, followed by analysis of trait distributions, demographic patterns, specialty differences, and physician archetypes.

2 Results

2.1 Dataset Characteristics and Scope

Our analysis encompassed 226,999 U.S. physicians derived from a national corpus of online patient reviews across four major rating platforms (Healthgrades, Vitals, RateMDs, and Yelp), representing over 4 million patient-submitted reviews collected through August 2021 [4]. After applying systematic filtering criteria requiring 5-100 reviews per physician to ensure adequate textual signal while maintaining representativeness, the final cohort spanned all major medical specialties and geographic regions nationwide.

The dataset captured substantial diversity in physician characteristics, with approximately balanced gender representation and comprehensive coverage across primary care, specialty, and subspecialty practices. Review content reflected contemporary patient experiences, with the majority of feedback submitted within the preceding three years, providing current insights into patient-physician interactions across diverse healthcare settings.

2.2 Trait Targets and Definitions

To capture the nuanced ways in which patients perceive their physicians, we focused on extracting ten key subjective characteristics from online reviews. These characteristics fall into two conceptually distinct categories: the well-established Big Five personality dimensions from psychology, and five healthcare-specific subjective judgment traits developed for this domain. Together, these traits represent a broad and clinically meaningful spectrum of interpersonal, emotional, and professional qualities that patients routinely reference when evaluating their providers.

The first category consists of the Big Five personality traits: openness, conscientiousness, extraversion, agreeableness, and neuroticism. These five dimensions have been extensively validated in the psychological literature as stable personality constructs [17] and have been increasingly applied in healthcare to explain variations in communication style, empathy, and patient trust [18]. In this study, openness reflects a physician's intellectual curiosity, receptiveness to novel approaches, and ability to engage creatively in problem-solving. Conscientiousness captures organization, reliability, and adherence to plans—qualities that patients often associate with attentiveness and safety in care. Extraversion encompasses sociability, assertiveness, and enthusiasm, which may influence a patient's experience of warmth and engagement during clinical encounters. Agreeableness denotes cooperativeness, empathy, and emotional sensitivity—factors that are commonly cited in perceptions of kindness, compassion, and respectful communication. Finally, neuroticism, which reflects emotional instability or a tendency toward stress reactivity, may negatively impact how patients perceive a physician's composure, confidence, or clarity under pressure.

The second category comprises five subjective judgment traits specifically developed for interpreting the content of online physician reviews in a healthcare context [18, 19]. These dimensions reflect patient-facing perceptions that go beyond personality and tap into behavioral performance, communication quality, and trustworthiness [20]. Interpersonal Qualities and Communication (IQC) refers to the physician's tone, empathy, clarity, listening skills, and respectfulness as inferred from the patient narrative. This trait captures how the physician comes across in real-time clinical interaction. Perceived Clinical Competence (PCC) assesses the patient's impression of the physician's medical expertise and professionalism—whether the doctor is seen as knowledgeable, accurate, and confident in decision-making. Sensitivity to Patient Satisfaction (SPS) reflects the degree to which the physician appears to care about the patient's

Category	Trait	Definition and Clinical Interpretation
Big Five Personality	Openness	Reflects intellectual curiosity, creativity, and receptiveness to new ideas and approaches—traits that may influence diagnostic flexibility and problem-solving.
	Conscientiousness	Denotes organization, reliability, and attention to detail—qualities associated with thoroughness in care and adherence to plans.
	Extraversion	Captures sociability, assertiveness, and enthusiasm—potentially shaping a patient's experience of engagement and interpersonal energy.
	Agreeableness	Represents compassion, cooperativeness, and empathy—factors strongly linked to respectful, patient-centered communication.
	Emotional Stability (Neuroticism)	Indicates emotional instability and reactivity—higher levels may be perceived as anxiety or reduced confidence in clinical interactions.
Subjective Judgments	IQC (Interpersonal Qualities & Communication)	Encompasses the physician's tone, clarity, listening, and empathy—reflecting the moment-to-moment communication style during patient encounters.
	PCC (Perceived Clinical Competence)	Reflects the patient's subjective impression of the physician's medical expertise, professionalism, and confidence in decision-making.
	SPS (Sensitivity to Patient Satisfaction)	Captures whether the physician is attentive to the patient's preferences, comfort, and emotional well-being—core to the patient experience.
	SCO (Sensitivity to Clinical Outcome)	Indicates whether the physician expresses concern for recovery, treatment success, and long-term health—often cited in reviews mentioning follow-up or care continuity.
	STS (Social Trust Signals)	Refers to broader reputational indicators, such as being widely recommended, trusted by families, or praised by other patients—serving as social proof of trustworthiness.

Table 1: Categories, Traits, and Clinical Interpretations of Physician Characteristics

preferences, comfort, and emotional needs—indicating attentiveness to experience-based outcomes. Sensitivity to Clinical Outcome (SCO) relates to whether the physician expresses or is perceived to express concern about the patient's recovery, treatment success, or long-term health goals. Finally, Social Trust Signals (STS) capture broader indicators of trustworthiness—such as being recommended by other patients or used repeatedly by families—which can serve as social proof or reputational reinforcement. These ten traits were selected based on their interpretability, frequency in real-world patient narratives, and conceptual relevance to healthcare quality and patient-centered care. Together, they provide a rich multidimensional framework for examining patient perceptions of physicians at national scale. A summary of trait definitions and their conceptual grounding is provided in Table 1.

2.3 Automated LLM-based Trait Extraction Approach and Evaluation

We developed a large language model-based pipeline to systematically extract ten physician characteristics from patient review text: the Big Five personality traits and five patient-oriented subjective judgments. The approach employed structured prompting with chain-of-thought reasoning, requiring models to provide evidence-based assessments with explicit justification for each trait score [21].

For each physician, models analyzed the complete aggregated review corpus and generated four outputs per trait: standardized scores (0-1 scale), supporting textual evidence, consistency ratings across reviews, and sufficiency assessments indicating confidence in the trait inference. This methodology enabled scalable psychological profiling of nearly 226,999 physicians while maintaining transparency and interpretability through detailed reasoning traces.

We conducted comprehensive validation of the trait extraction methodology through multi-model comparison and human expert benchmarking. Five state-of-the-art language models (GPT-4o [22], GPT-4.1 [23], Claude 3.7 [24], Gemini-2.5 Flash [25], Gemini-2.5 Pro [26]) were evaluated on 1,200 physicians using an LLM-as-a-Judge framework, with Gemini-2.5 Pro [26] serving as the judge model.

	LLM-as-a-judge				Human-as-a-judge			
Model	**MAE**	**RMSE**	**High**	**Low**	**MAE**	**RMSE**	**High**	**Low**
gemini-2.5-pro	0.0685	0.1446	79.98%	80.21%	0.1851	0.2598	73.3%	87.2%
gemini-2.5-flash	0.1057	0.1885	91.84%	72.56%	0.1801	0.2625	78.5%	77.6%
gpt-4.1	0.1117	0.1855	88.21%	49.78%	0.1493	0.2174	81.8%	71.1%
claude-3.7-sonnet	0.1151	0.1929	89.36%	57.54%	0.1577	0.2389	80.2%	66.7%
gpt-4o	0.1153	0.1911	86.94%	49.78%	0.1552	0.2241	80.7%	69.1%
gemini-2.5-flash-lite	0.1517	0.2394	88.93%	49.72%	0.1837	0.2685	88.1%	65.2%

Table 2: Model Performance Evaluation: updated normalized results for LLM-as-a-judge and Human-as-a-judge.

Model performance varied substantially across quality dimensions (Table 2), with Gemini-2.5 Pro [26] achieving the lowest mean absolute error (MAE = 0.0685) and highest agreement with human evaluators in the LLM-as-a-Judge evaluation. Human validation on 300 physician profiles revealed strong convergence between automated and expert assessments, with correlation coefficients ranging from 0.72 to 0.89 across traits.

Notably, LLM-based assessments demonstrated superior consistency compared to human evaluators in several domains, particularly for patient-oriented subjective judgments where automated approaches achieved higher inter-rater reliability. Based on comprehensive performance evaluation considering accuracy, cost-effectiveness, and scalability (Table 2), Gemini-2.5 Flash [25] was selected for large-scale trait extraction across the full physician cohort, balancing high performance (MAE = 0.1057) with computational efficiency for population-level analysis.

2.4 Trait Distributions and Quality

We mapped online patient reviews to ten standardized traits—the Big Five personality dimensions and five patient-oriented subjective judgments—across 226,999 physicians, achieving unprecedented scale in automated psychological assessment. This analysis examines the resulting trait distributions, their interrelationships, and extraction quality metrics.

The distribution analysis of personality traits across 226,999 physicians revealed distinct patterns in both Big Five personality dimensions and patient-oriented subjective judgment measures. Among the Big Five traits, Emotional Stability demonstrated the highest mean expression (M = 0.714, SD = 0.252), followed by Conscientiousness (M = 0.631, SD = 0.389) and Agreeableness (M = 0.611, SD = 0.394), while Openness showed the lowest average scores (M = 0.518, SD = 0.362). The subjective judgment measures exhibited even more pronounced positive patterns, with Social Trust Signals (STS) achieving the highest overall scores (M = 0.761, SD = 0.350), followed by Perceived Clinical Competence (PCC; M = 0.712) and Interpersonal Qualities & Communication (IQC; M = 0.631). Sensitivity to Patient Satisfaction (SPS; M = 0.592) and Sensitivity to Clinical Outcome (SCO; M = 0.578) showed moderate levels of expression. All trait distributions displayed approximately normal distributions with varying degrees of positive skewness, particularly evident in the subjective judgment measures where scores clustered toward higher values, suggesting that physicians in the dataset generally received favorable assessments in patient reviews.

The correlation matrix revealed meaningful patterns illuminating how personality traits manifest in clinical practice and shape patient perceptions. The exceptionally high correlation between Interpersonal Qualities & Communication and Sensitivity to Patient Satisfaction (r = 0.928) suggests these may represent two facets of the same underlying construct—the physician's interpersonal attentiveness. The strong association between Agreeableness and both IQC (r = 0.860) and SPS (r = 0.853) validates that personality traits translate into observable clinical behaviors. Perceived Clinical Competence showed stronger correlations with Conscientiousness (r = 0.710) than with Openness (r = 0.568), suggesting patients interpret organized, reliable behavior as clinical expertise more readily than intellectual curiosity. The robust relationship between PCC and Sensitivity to Clinical Outcome (r = 0.778) indicates that physicians perceived as competent are also seen as more invested in treatment success. Social Trust Signals showed moderate correlations with all other subjective judgments (r = 0.675-0.760), functioning as a global reputational metric that integrates multiple aspects of patient experience.

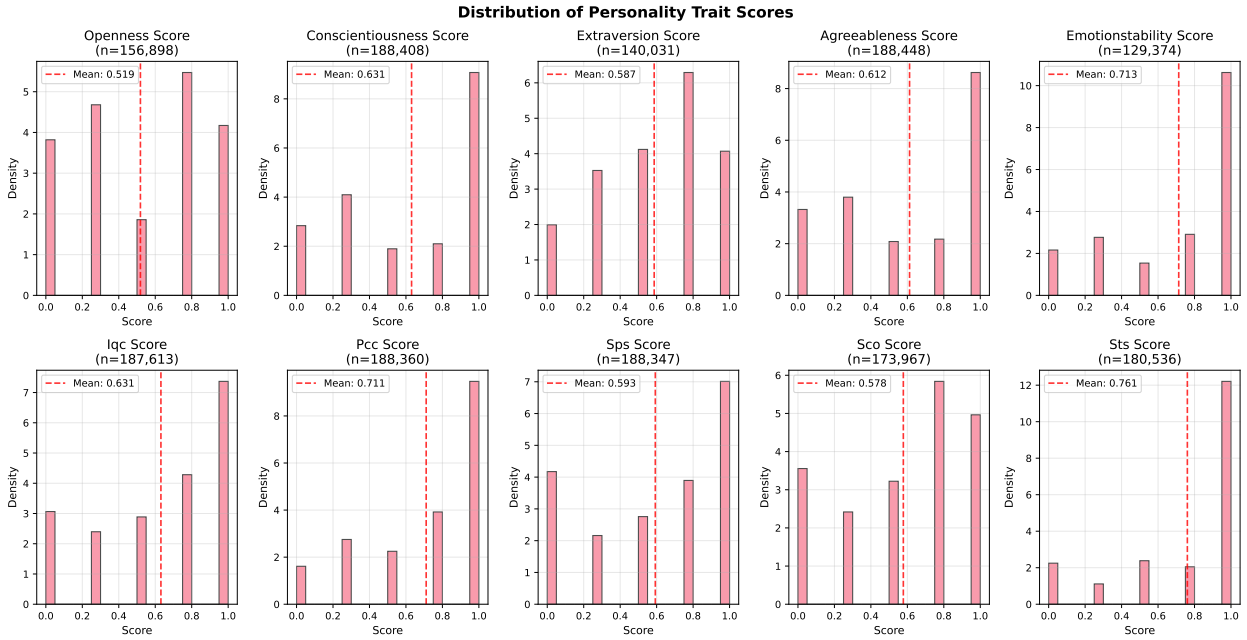


Figure 1: Distribution of personality trait scores. Each subplot shows the kernel density estimate for one of the 10 traits measured on a 0-1 scale.

The relationship between trait scores and quality metrics provides crucial validation for our automated extraction methodology. We evaluated two quality dimensions: Consistency, which measures whether multiple reviews of the same physician converge on similar trait assessments, and Sufficiency, which indicates whether the review corpus contains adequate evidence to support the extracted trait scores. Across all traits, sufficiency scores averaged 0.742, demonstrating that patient reviews generally provide rich enough descriptions for trait inference, while consistency scores averaged 0.627, reflecting the expected variability in how different patients express their observations. Most notably, the scatter plots revealed a distinctive pattern where extreme trait scores (approaching 0 or 1) exhibited markedly higher quality metrics for both consistency and sufficiency, creating a U-shaped relationship. This pattern is particularly interpretable: when physicians clearly exhibit or lack a trait, patients converge in their assessments—a highly empathetic doctor consistently receives praise for compassion, while one lacking empathy generates uniform complaints. In contrast, moderate trait scores (0.3-0.7) showed lower consistency, appropriately reflecting genuine ambiguity when physicians display traits inconsistently or when patient experiences vary.

Examining specific traits, most showed robust high consistency patterns. Openness (r = 0.258), Extraversion (r = 0.401), and Sensitivity to Clinical Outcome (r = 0.366) showed medium consistency levels. This suggests these dimensions are either expressed more variably in clinical encounters or perceived differently across patients. Similarly, while most traits demonstrated adequate sufficiency across score ranges, Openness, Extraversion, Emotional Stability, and SCO showed more sparse evidence in the medium score ranges, indicating that these particular dimensions may be less salient in patient reviews or require more nuanced textual cues for reliable extraction.

These patterns suggest that while interpersonal traits and clinical competencies are readily captured from patient narratives, certain personality dimensions and outcome-focused behaviors may be less consistently observable in routine clinical interactions.

Having established the reliability and validity of our trait extraction methodology, we now examine how these traits vary across physician demographics and specialties.

2.5 Subgroup Analysis — Gender & Specialty

Our analysis of gender differences across 153,816 to 224,044 physicians per trait revealed a systematic pattern: male physicians received higher ratings than female physicians across all ten assessment domains, with all differences reaching statistical significance ($p < 0.001$). While the effect sizes were generally small (Cohen's d ranging from 0.029 to 0.177), the consistency of this pattern across both personality traits and subjective judgments is striking. Most notably, the largest gender gaps emerged in patient-derived subjective judgments rather than personality traits,

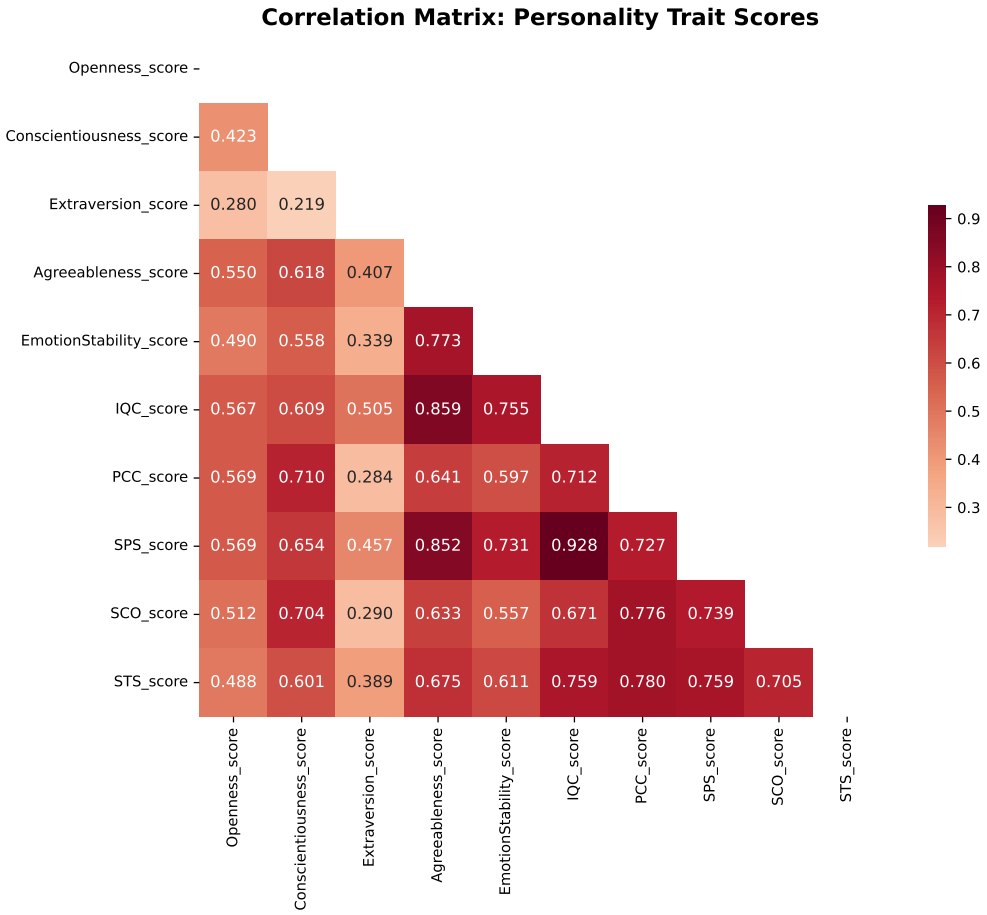
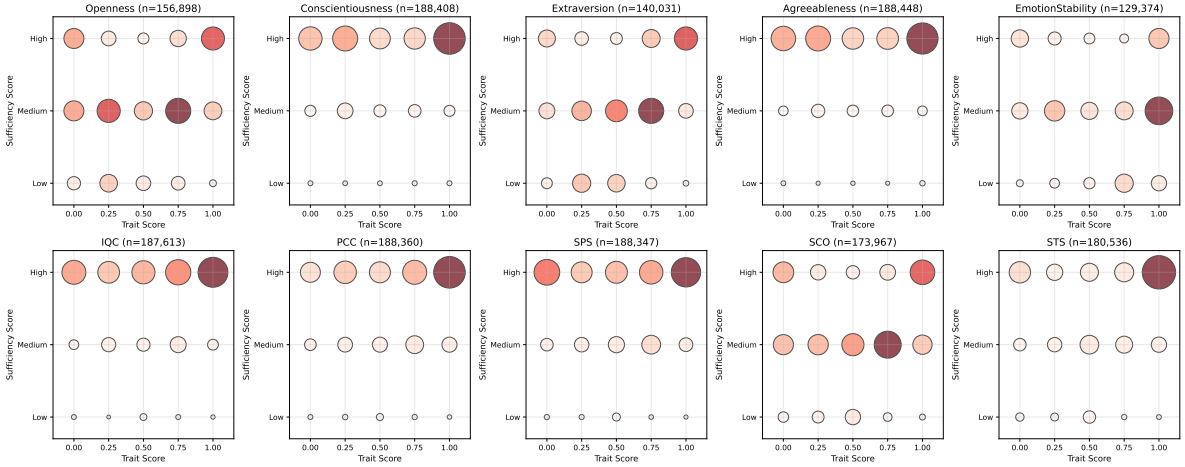


Figure 2: Correlation matrix of the 10 physician traits. Darker colors indicate stronger correlations. Notable patterns include high correlations between interpersonal traits (IQC-SPS: r=0.928) and between agreeableness and patient-oriented measures.

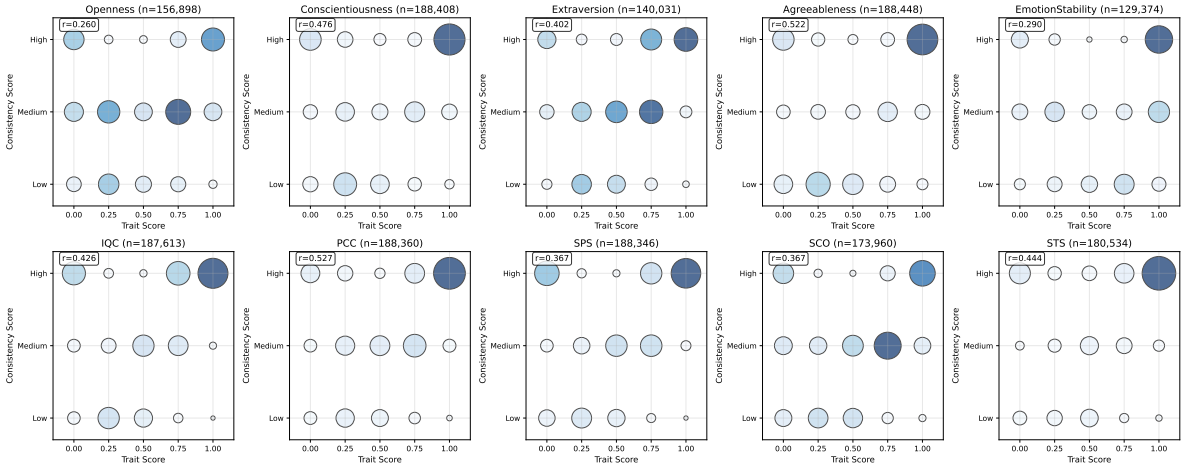
with Perceived Clinical Competence showing the largest disparity (d = 0.177), followed by Sensitivity to Clinical Outcome (d = 0.164) and Social Trust Signals (d = 0.151). In contrast, personality traits like Agreeableness (d = 0.029) and Extraversion (d = 0.045) showed minimal gender differences.

This pattern aligns with extensive literature documenting systematic undervaluation of female physicians in patient evaluations [18], suggesting that these disparities reflect patient perception biases rather than actual performance differences. However, the small effect sizes (Cohen's d = 0.029-0.177) indicate that while these differences are statistically significant due to our large sample size, their practical significance may be limited. Several mechanisms likely contribute to these findings, including implicit gender bias [27], societal stereotypes associating medical authority with masculinity [27, 28], the "double bind" faced by female physicians [29], and structural factors such as time constraints and gendered communication patterns [30]. Alternative explanations warrant consideration, such as the "excellence tax" whereby female physicians' additional emotional labor may paradoxically raise patient expectations, or selection effects in who writes online reviews. The fact that professional competency measures show larger gender gaps than personality traits particularly supports the bias hypothesis—patients appear more likely to question female physicians' clinical competence and expertise [27]. Future research employing mixed methods and controlling for physician experience, patient demographics, and consultation characteristics is needed to disentangle these complex dynamics and develop more equitable physician evaluation frameworks.

Analysis of trait distributions across nine major medical specialties and an aggregated "Others" category (total N ranging from 128,157 to 186,840 physicians per trait) revealed significant variations that reflect both selection effects and the distinct demands of different medical fields [31]. Surgical specialties demonstrated the highest scores across nearly all domains, with Diagnostic Radiology physicians showing particularly strong performance in Conscientiousness (0.794), Emotional Stability (0.839), and professional competencies (PCC: 0.896, SCO: 0.683, STS: 0.782),



(a) Trait scores vs. sufficiency



(b) Trait scores vs. consistency

Figure 3: Quality validation of trait extraction methodology. (a) Sufficiency scores show U-shaped patterns where extreme trait scores (0 or 1) demonstrate higher evidence adequacy compared to moderate scores. (b) Consistency patterns reveal that physicians with clearly defined traits (high or low scores) generate more convergent patient assessments, while moderate scores reflect appropriate uncertainty when traits are ambiguous. Data points represent mean values with standard error of the mean (SEM) calculated across all traits and physicians (n = 226,999). Trend lines show polynomial regression fits with 95% confidence bands.

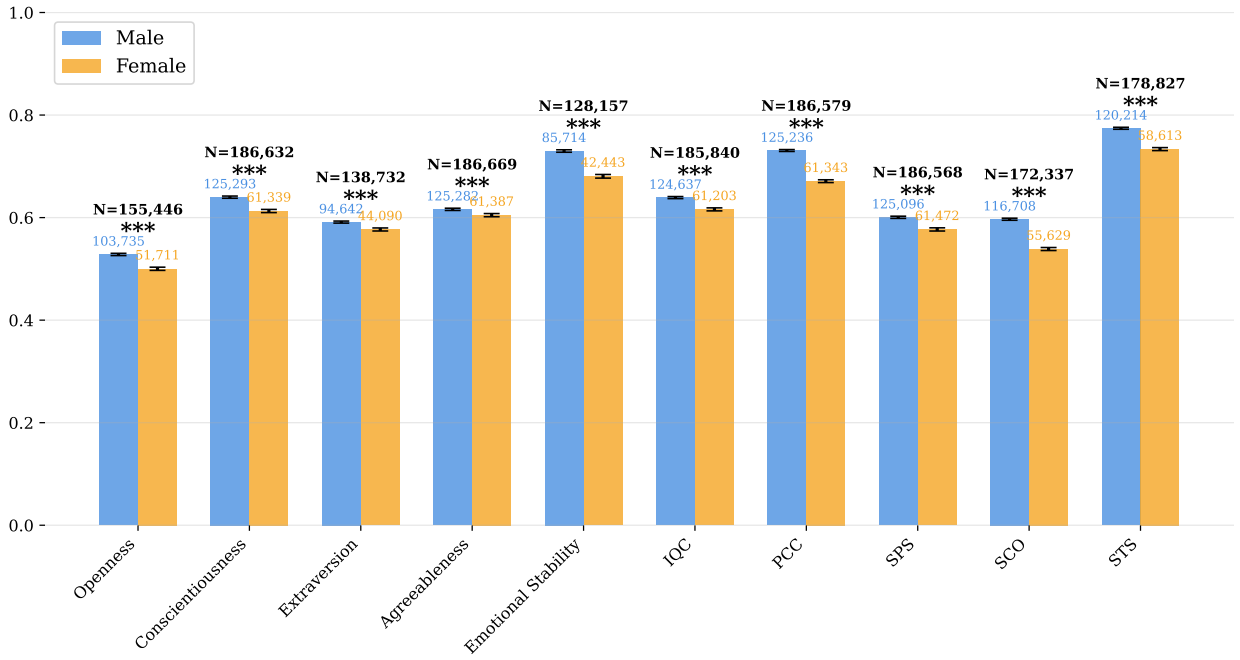
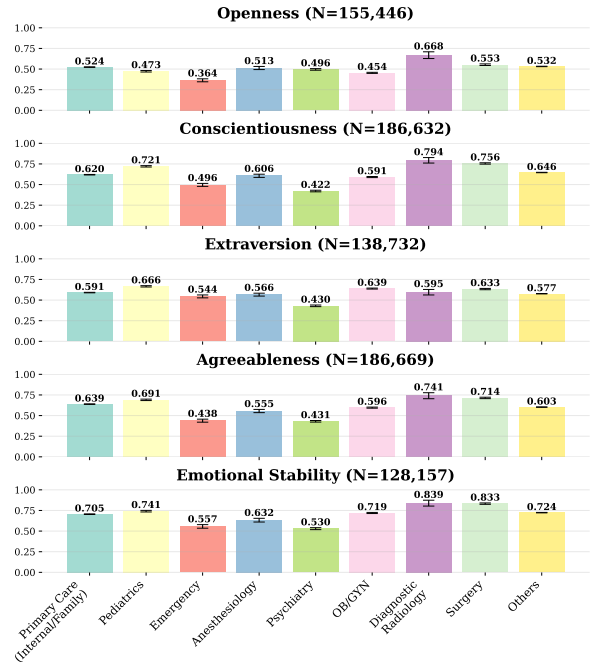
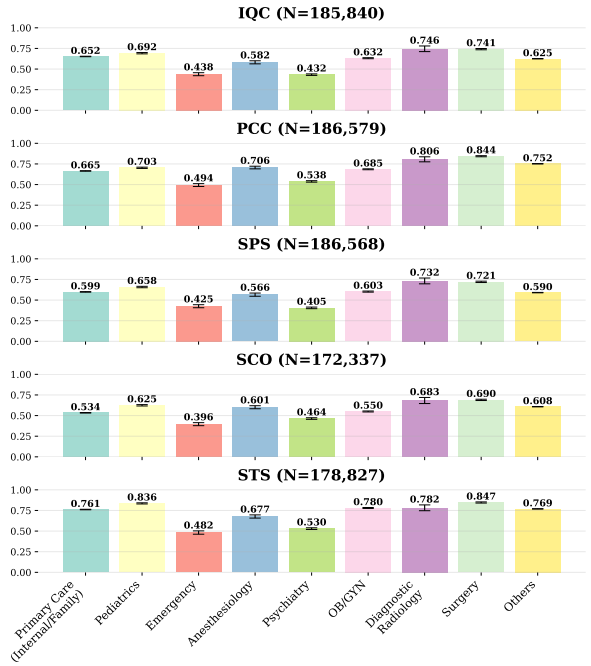


Figure 4: Gender differences in physician trait scores. Male physicians receive higher ratings across all 10 traits, with largest differences in patient-oriented subjective judgments (PCC, SCO, STS) rather than personality traits. Error bars represent 95% confidence intervals calculated from standard errors. Statistical significance was determined using Welch's t-test for unequal variances. Sample sizes vary by trait (n_{male} = 85,714–125,293; n_{female} = 42,443–61,472; total n = 128,157–186,669).



(a) BigFive personality traits by specialty



(b) Professional competencies by specialty

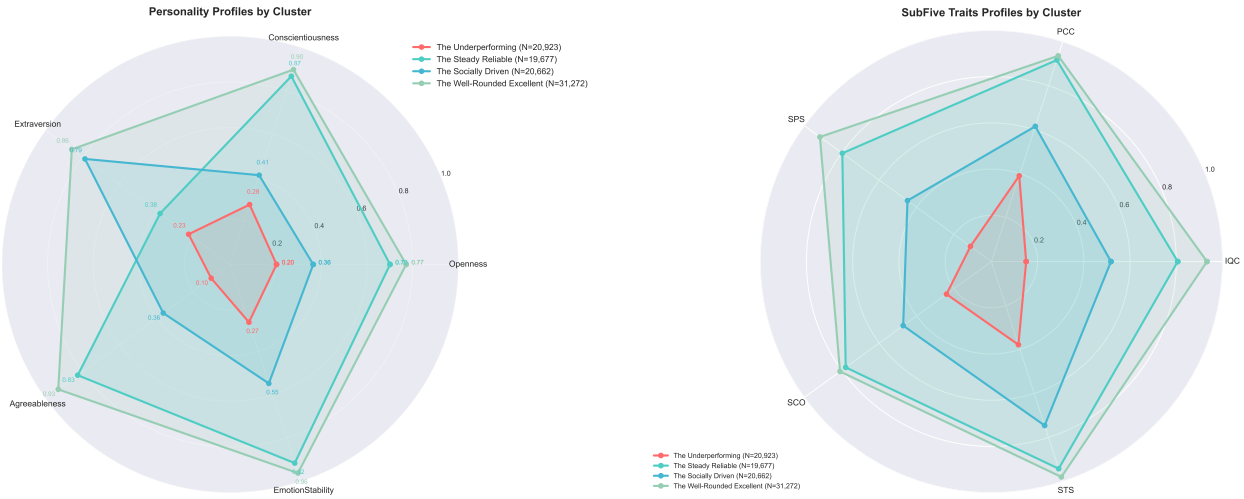
Figure 5: Specialty differences in physician traits. (a) BigFive personality trait scores by medical specialty show distinct patterns, with surgical specialties generally demonstrating higher conscientiousness and emotional stability. (b) Professional competency measures reveal specialty-specific strengths, with surgical specialties excelling in perceived clinical competence while patient-oriented specialties show higher interpersonal qualities.

while Surgery physicians excelled in patient-oriented measures (SPS: 0.721, IQC: 0.741). In stark contrast, Emergency Medicine and Psychiatry physicians consistently scored lowest across multiple domains. Emergency Medicine showed deficits in Openness (0.364), Conscientiousness (0.496), and Social Trust Signals (0.482). Psychiatry demonstrated notably low scores in Conscientiousness (0.422), Extraversion (0.430), and Social Trust Signals (0.530). Primary care specialties showed moderate profiles, with Pediatrics demonstrating higher Extraversion (0.666) and patient satisfaction sensitivity (SPS: 0.658) compared to Internal Medicine, suggesting alignment with the interpersonal demands of treating children and interacting with families. These findings likely reflect complex, bidirectional mechanisms. First, patient self-selection means individuals with specific conditions or personalities gravitate toward certain specialties, potentially creating more challenging evaluation contexts (particularly relevant for Psychiatry where patients with mental health conditions may have different evaluation patterns). Second, reciprocal shaping effects occur as prolonged exposure to particular patient populations and clinical scenarios may modify physician personality traits and professional behaviors over time. Third, structural factors such as time constraints in Emergency Medicine and stigma in Psychiatry influence interaction quality. Finally, specialty-specific evaluation biases arise when the nature of the medical encounter affects patient perceptions.

The consistently lower ratings for Emergency Medicine and Psychiatry suggest these specialties face unique challenges—Emergency Medicine dealing with acute, high-stress situations with limited relationship building, while Psychiatry manages complex mental health issues that may influence both patient satisfaction and the emotional toll on physicians themselves. This bidirectional relationship between patient populations and physician characteristics underscores the dynamic nature of medical practice, where specialties not only attract certain personality types but may also shape them through repeated patient interactions and environmental pressures.

Beyond examining trait differences by demographics and specialty, we sought to identify natural groupings of physicians based on their overall trait profiles, revealing distinct physician archetypes with important implications for patient care.

2.6 Cluster Analysis — Latent Profiles



(a) BigFive personality profiles

(b) Patient perception profiles

Figure 6: Physician personality archetypes based on K-means clustering. (a) Four distinct archetypes emerge based on BigFive personality traits: The Well-Rounded Excellent (highest across all traits), The Underperforming (lowest scores), The Socially Driven (high extraversion), and The Steady Reliable (balanced profile). (b) The same physician clusters display distinct patterns in patient-oriented traits, with personality-based clustering predicting patient satisfaction outcomes across clinical competency and interpersonal dimensions.

To understand how personality traits and patient perceptions interact to create distinct physician profiles, we performed K-means clustering analysis on physicians with complete trait assessments, revealing four meaningful archetypes that capture the heterogeneity of the physician workforce.

The clustering analysis identified four distinct physician profiles, with k=4 determined as optimal through elbow method analysis (see Supplementary Information SI-4 for clustering validation). The ”Well-Rounded Excellent” archetype (33.8% of physicians) demonstrated uniformly high scores across all Big Five dimensions (Emotional

Stability: 0.962, Agreeableness: 0.931, Conscientiousness: 0.898, Extraversion: 0.858, Openness: 0.772) and correspondingly exceptional patient perceptions, achieving the highest patient satisfaction scores. In stark contrast, "The Underperforming" archetype (22.6% of physicians) exhibited critically low scores across all dimensions, with particularly concerning deficits in Agreeableness (0.103), Extraversion (0.225), Openness (0.203), and Conscientiousness (0.276), resulting in the lowest patient satisfaction. Intriguingly, "The Socially Driven" archetype (22.3% of physicians) presented a paradox—despite high Extraversion (0.786), these physicians showed only moderate scores in other traits (Agreeableness: 0.363, Conscientiousness: 0.411, Emotional Stability: 0.549, Openness: 0.364), suggesting that social charisma alone does not translate to perceived clinical excellence. "The Steady Reliable" archetype (21.3% of physicians) demonstrated the importance of emotional stability and conscientiousness, achieving strong trait scores (Emotional Stability: 0.916, Conscientiousness: 0.868, Agreeableness: 0.826, Openness: 0.701) despite moderate extraversion (0.380), resulting in good overall performance. The perfect rank-order consistency between personality profiles and patient perceptions across all archetypes (Kruskal-Wallis H statistics ranging from 46,153 to 74,009, all p < 0.0001) provides compelling evidence that personality traits manifest directly in observable clinical behaviors.

This clustering reveals the natural diversity within the physician workforce, where different personality configurations lead to distinct patient interaction patterns, suggesting that patient-physician matching based on complementary personality profiles could potentially optimize healthcare experiences for both parties rather than assuming a one-size-fits-all approach to medical practice.

2.7 Trait-Rating Correlation

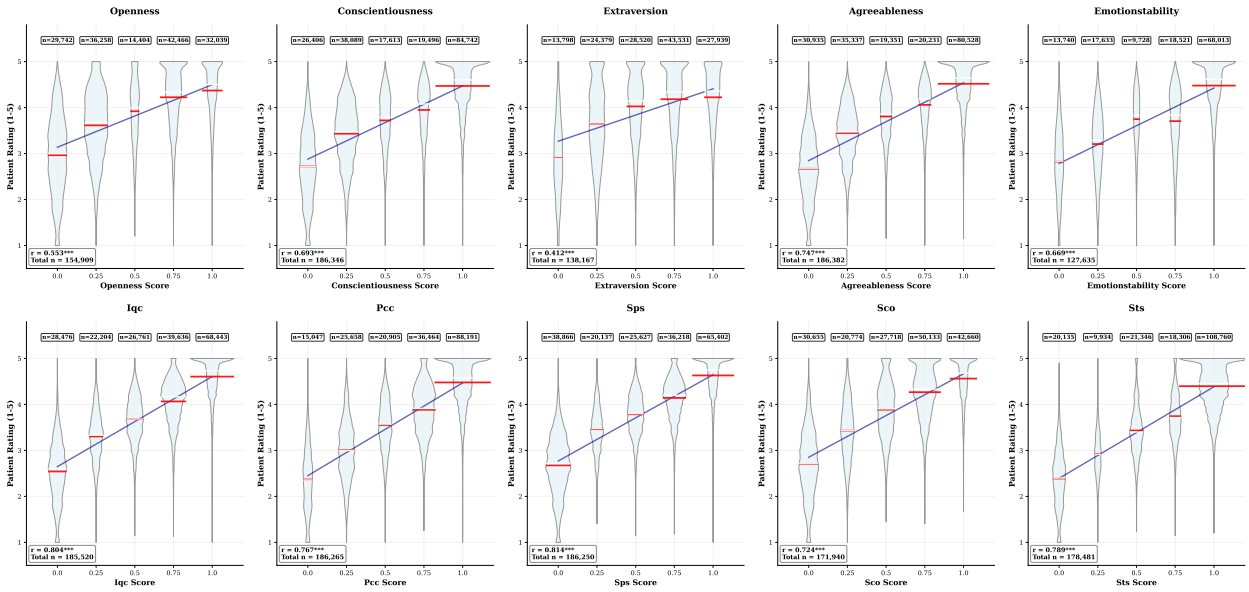


Figure 7: Relationship between LLM-inferred traits and overall patient satisfaction. Area-scaled violin plots show patient satisfaction distributions for discrete trait levels (0, 0.25, 0.5, 0.75, 1.0). Violin width is proportional to sample size. Clear monotonic increases in satisfaction with trait scores validate the external validity of our trait extraction methodology.

To establish the external validity of our LLM-inferred physician traits, we examined their relationship with overall patient satisfaction ratings across physicians, revealing remarkably strong associations that validate our trait extraction methodology.

The area-scaled violin plots demonstrate clear monotonic increases in patient satisfaction as trait scores rise from 0 to 1, with the most pronounced effects observed for patient-oriented subjective judgments. Sensitivity to Patient Satisfaction (SPS) showed the strongest correlation (r = 0.814, n = 223,479), followed closely by Interpersonal Qualities & Communication (r = 0.803) and Social Trust Signals (r = 0.789), all achieving $p < 0.001$. These exceptionally high correlations—rarely observed in behavioral research—suggest that our LLM-derived traits capture the core dimensions patients value in healthcare encounters.

The Big Five personality traits also demonstrated meaningful associations, with Agreeableness (r = 0.746) and Conscientiousness (r = 0.693) showing particularly strong relationships, while Extraversion exhibited the weakest though still

significant correlation (r = 0.413). The violin distributions reveal not only linear trends but also increasing consistency at higher trait levels. Physicians scoring 1.0 on traits show tighter satisfaction distributions around high ratings (4-5 stars), while those scoring 0 display broader distributions skewed toward lower ratings. Multiple regression analysis (R^2 = 0.804, n = 99,973) confirmed that these traits independently contribute to patient satisfaction. Social Trust Signals emerged as the strongest independent predictor (β = 0.214), followed by communication quality (β = 0.136) and conscientiousness (β = 0.132). The strong association between LLM-inferred traits and real-world patient outcomes shows that automated text analysis can reliably capture important physician characteristics. These results suggest that trait-based assessment could be a powerful tool for understanding and predicting patient experiences in healthcare.

3 Discussion

This large-scale analysis represents the first systematic application of large language models to extract standardized personality and professional competency assessments from online patient reviews at national scale. By analyzing over four million reviews for 226,999 US physicians, we demonstrate that LLM-based trait extraction can reliably capture clinically meaningful dimensions of physician behavior that strongly predict patient satisfaction and reveal distinct patterns across demographics, specialties, and practice contexts. Our multi-model validation framework establishes a new standard for automated trait extraction in healthcare settings. Convergence between LLM-as-a-judge evaluations and human expert assessments (Cohen's kappa 0.72-0.89) demonstrates that modern language models can reliably identify nuanced personality characteristics from unstructured patient narratives [13]. The exceptionally strong correlations between extracted traits and patient satisfaction (r = 0.413-0.814) provide compelling external validation, with effect sizes rarely observed in behavioral research suggesting our methodology captures the core dimensions patients value in healthcare encounters.

The dual framework combining established Big Five personality traits with patient-oriented subjective judgments (IQC, PCC, SPS, SCO, STS) offers a comprehensive assessment tool that bridges psychological theory with clinical practice. This approach addresses a critical gap in healthcare quality measurement by providing standardized, interpretable metrics that can be systematically applied across diverse clinical contexts while maintaining the rich, patient-centered perspective inherent in narrative feedback. The identification of four distinct physician archetypes—The Well-Rounded Excellent, The Underperforming, The Socially Driven, and The Steady Reliable—has immediate implications for healthcare workforce development and patient-provider matching. While 33.8% of physicians demonstrate consistently high performance across all trait dimensions, 22.6% show concerning deficits, suggesting significant heterogeneity in patient-perceived care quality that current evaluation systems may not adequately capture.

Of particular clinical relevance, "The Socially Driven" archetype (22.3% of physicians) presents a paradox whereby high extraversion and social engagement do not translate to superior patient satisfaction. This suggests that interpersonal charisma alone is insufficient for perceived clinical excellence and that authentic empathy, reliability, and clinical competence remain the primary drivers of patient-perceived quality. The demographic disparities revealed in our analysis—with male physicians receiving higher ratings across all ten traits despite similar objective qualifications—highlight systematic biases in patient evaluation systems that may perpetuate inequities in physician career advancement and compensation [32]. These findings align with extensive literature documenting evaluation bias against female and minority physicians [32] and suggest that current patient satisfaction metrics may inadvertently disadvantage qualified diverse physicians.

The specialty analysis reveals that physician traits reflect both selection effects and adaptation to practice contexts. Surgical specialties' consistently higher scores across personality and competency dimensions potentially reflect both self-selection of certain personality types into these fields and the structured, outcome-focused nature of surgical practice that facilitates positive patient perceptions. Conversely, the lower ratings for Emergency Medicine and Psychiatry physicians likely reflect the challenging contexts these specialties navigate. Emergency Medicine involves acute, high-stress encounters with limited relationship-building opportunities [33], while Psychiatry faces stigma and emotional intensity inherent in mental health care, where psychiatrists encounter occupational stigma linked to burnout and diminished professional value [34]. These patterns have important implications for specialty-specific quality metrics and professional development programs. Healthcare systems should consider context-adjusted assessments that account for the unique challenges and patient populations each specialty serves rather than applying uniform evaluation standards across fundamentally different practice environments.

The deployment of our LLM-based extraction pipeline across nearly a quarter million physicians demonstrates the scalability of automated trait assessment for healthcare quality monitoring. Its modular, model-agnostic architecture allows flexible implementation across different healthcare systems and geographic regions, while the transparent reasoning framework provides interpretable justifications that can be audited by clinical experts [12, 13]. Nonetheless, the reliance on online patient reviews introduces concerns about representativeness and selection bias, since individuals

who post reviews may differ systematically from the broader patient population in terms of demographics, digital literacy, and satisfaction levels [3, 6]. These limitations highlight the need for caution in interpreting findings and suggest that integrating LLM-based trait analysis with complementary data sources will be important for ensuring fairness and generalizability in future applications.

The automated nature of our trait extraction methodology also raises questions about algorithmic fairness and the potential for perpetuating existing biases present in training data or patient review patterns [27]. However, our validation framework demonstrates reliability across demographic groups, indicating that ongoing monitoring and bias assessment will be essential for responsible deployment of these tools in clinical practice [9, 15]. Several important research opportunities emerge from this work. Longitudinal studies tracking trait stability and evolution over physician careers could provide insights into professional development and the effects of practice experience on patient-perceived characteristics [1]. Intervention studies examining whether targeted feedback on specific trait dimensions can improve patient satisfaction and clinical outcomes could test the causal implications of our findings [20].

The methodology could be extended to examine trait-outcome relationships beyond patient satisfaction, including clinical quality metrics, safety outcomes, and physician well-being measures [20, 35]. Beyond medicine, the framework may be adapted for other healthcare contexts such as nursing, pharmacy, or telemedicine, where patient-provider relationships are similarly critical to care quality [36, 37]. Integration with electronic health records and clinical decision support systems could enable real-time trait monitoring and personalized feedback for continuous professional development [38], while future applications may also explore optimizing patient-provider matching based on complementary personality profiles [39].

Several important constraints should be considered when interpreting these findings. Online patient reviews, while offering unprecedented scale and real-world context, may not capture the full spectrum of patient-physician interactions or represent all demographic groups—a common limitation in digital health research. Similarly, the observational nature of our study precludes strong causal inference about the link between physician traits and patient outcomes. Our LLM-based approach demonstrates high reliability and validity, but the accuracy of automated trait assessments compared to comprehensive psychological evaluation by trained professionals requires further validation [10]. The focus on patient-perceived traits rather than objective behavioral measures means our findings reflect patient perceptions rather than independently verified physician characteristics [2, 20]. The cross-sectional design prevents examination of trait stability over time or the effects of career experience on patient-perceived physician characteristics. Longitudinal studies would be valuable for understanding how physician traits evolve throughout professional development and in response to different practice contexts [40]. Future implementations should consider strategies for broadening the patient feedback base and validating findings against more representative patient populations [3, 5] while carefully considering biases when implementing patient feedback systems for high-stakes decisions [27, 29].

4 Methods

We now describe the methodological details of our study, including data collection, LLM-based trait extraction, validation procedures, and statistical analyses.

4.1 Data Collection and Preprocessing

We constructed our analysis dataset from a national-scale corpus of online physician reviews previously aggregated from four major U.S. physician rating platforms: Healthgrades, Vitals, RateMDs, and Yelp, encompassing 587,562 physicians with 8,693,563 total reviews as of August 1, 2021 [4]. To ensure adequate textual signal while maintaining computational feasibility, we applied systematic filtering criteria: physicians must have between 5 and 100 reviews across all platforms, ensuring sufficient content for trait inference without bias toward extremely high-volume providers. This filtering process yielded 226,999 physicians for final analysis.

Patient reviews were aggregated at the physician level and concatenated into comprehensive textual profiles. We extracted structured physician metadata including gender, medical specialty, geographic location, and overall patient satisfaction ratings from the original dataset, supplementing missing information using publicly available provider directories. Reviews underwent minimal preprocessing, retaining original patient language and sentiment while removing only personally identifiable information and non-English content. Geographic and specialty classifications were from the NPPES directory data [41].

Data quality assurance involved systematic validation of physician identifiers and removal of duplicate entries. The final dataset represented physicians across all major medical specialties and geographic regions, ensuring national representativeness for population-level analysis.

4.2 Top-Down Annotation Protocol

We employed a top-down trait annotation strategy that processes the complete review corpus for each physician as a unified context, contrasting with bottom-up approaches that would analyze individual reviews separately and aggregate results. This design leverages the increasing contextual capacity of modern large language models to form holistic impressions across multiple patient narratives, enabling more consistent reasoning across long-form texts compared to sentence-level or review-level processing.

Formally, let $P = \{p_1, p_2, ..., p_n\}$ represent the set of physicians in our dataset, and for each physician p_i, let $R_i = \{r_{i,1}, r_{i,2}, ..., r_{i,k_i}\}$ denote their complete set of patient reviews, where k_i is the number of reviews for physician p_i. Our top-down annotation protocol can be mathematically expressed as:

$$T_i = f(R_i, \Theta)$$

where $T_i = \{t_{i,1}, t_{i,2}, ..., t_{i,m}\}$ represents the extracted trait vector for physician p_i across m traits, $f(\cdot)$ is the LLM-based extraction function, and Θ represents the model parameters and prompt specifications.

This differs from bottom-up approaches [19] that would compute:

$$T_i^{bottom-up} = g\left(\bigcup_{j=1}^{k_i} f(r_{i,j}, \Theta)\right)$$

where individual reviews are processed separately and subsequently aggregated through function $g(\cdot)$. The top-down architecture leverages principles from Chain-of-Thought prompting, wherein each physician is presented by name alongside their aggregated reviews and analyzed through structured psychological instructions requiring strict XML schema output. The approach explicitly encourages the model to reason before scoring rather than producing isolated assessments, thereby engaging in deeper, more interpretable reasoning over the text while yielding transparent justifications that can be audited by human experts [16, 21].

For each trait $t_{i,j}$, the extraction function produces a structured output $t_{i,j} = \{s_{i,j}, e_{i,j}, c_{i,j}, u_{i,j}\}$ where $s_{i,j}$ represents a five-point qualitative score, $e_{i,j}$ contains evidence grounded in review excerpts or paraphrases, $c_{i,j}$ measures consistency across multiple reviews, and $u_{i,j}$ assesses sufficiency of supporting evidence. This methodology enables the model to synthesize contradictory evidence across reviews, identify consistent patterns through multiple patient perspectives, resolve ambiguities using broader contextual understanding, and form holistic personality assessments that mimic expert psychological evaluation while supporting scalable trait inference from noisy, real-world review data.

4.3 LLM-Based Agent Implementation

We developed two complementary model-agnostic trait extraction agents supporting multiple large language model providers (OpenAI GPT-4/GPT-4o, Anthropic Claude 3.7, Google Gemini-2.5 Pro/Flash) through LangChain framework integration. The system employed standardized APIs with systematic rate limiting, exponential backoff strategies, and SQLite-based response caching to ensure reproducible results across providers while managing computational costs and API quotas.

The `PhysicianBigFiveExtractor` agent implemented a structured prompting framework positioning the model as an "expert psychologist" analyzing traditional Big Five personality dimensions (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism) from aggregated patient reviews. Complementing this approach, the `PhysicianSubFiveExtractor` agent employed an "expert analyst" framework specifically designed to evaluate patient perceptions along five healthcare-specific subjective dimensions: Interpersonal Qualities & Communication (IQC), Perceived Clinical Competence (PCC), Sensitivity to Patient Satisfaction (SPS), Sensitivity to Clinical Outcome (SCO), and Social Trust Signals (STS).

Both agents utilized identical two-message prompting structures with system-level analytical instructions and physician-specific review content, but with domain-appropriate role positioning and trait definitions optimized for their respective assessment frameworks. The BigFive agent emphasized psychological trait inference with detailed trait definitions and clinical interpretation guidelines, while the SubFive agent focused on patient decision-making factors with explicit behavioral examples and differentiation criteria between related dimensions (detailed prompt specifications for both agents provided in Supplementary Information).

Output formatting across both agents was enforced through strict XML schema validation requiring five components per trait: trait name validation, discrete scores [No Evidence, Low, Low to Moderate, Moderate, Moderate to High, High], evidence with direct quotes or paraphrased examples, consistency ratings across multiple reviews, and sufficiency assessments of evidence quality. The SubFive agent employed `<result>` XML tags for streamlined parsing, while the BigFive agent used `<personality>` tags for psychological framework alignment.

Quality assurance protocols included temperature control (0.0-0.1 for deterministic outputs), structured prompt templates preventing injection attacks, physician name validation for accurate trait attribution, and iterative prompt refinement with retry mechanisms featuring progressive attempt numbering for enhanced extraction reliability. This dual-agent implementation enabled comprehensive personality and patient perception assessment across nearly 227,000 physicians while maintaining methodological consistency and interpretability across both psychological and healthcare-specific evaluation domains.

4.4 LLM-as-a-Judge Methodology

To ensure rigorous and consistent evaluation of trait extraction quality across multiple models, we implemented a comprehensive LLM-as-a-Judge framework using the `PersonalityJudge` agent. This methodology addresses the challenge of scalable quality assessment in automated trait extraction by leveraging a high-performing language model (Gemini-2.5 Pro) as an expert evaluator, providing standardized assessment protocols that can be systematically applied across large datasets.

The LLM-as-a-Judge evaluation protocol follows a structured three-step assessment process for each physician-trait combination: (1) **Initial Independent Judgment**, where the judge model analyzes patient reviews without considering extraction model outputs, generating its own trait assessment with supporting evidence and reasoning; (2) **Comparative Model Evaluation**, where the judge systematically evaluates each extraction model's performance across five quality dimensions using integer scales (0-10); and (3) **Final Integrated Judgment**, where the judge synthesizes model assessments with its initial analysis to produce consensus ratings and reliability estimates. The five quality evaluation dimensions capture distinct aspects of trait extraction performance: *Evidence Quality* measures how well cited evidence supports the trait conclusion; *Reasoning Clarity* assesses the logical coherence and interpretability of the extraction rationale; *Trait Understanding* evaluates whether the model correctly interprets psychological construct definitions; *Evidence Specificity* rates the precision and detail of behavioral observations; and *Conclusion Accuracy* determines whether the final trait score aligns with the supporting evidence. Each dimension receives integer ratings (0-10) that are combined into weighted overall performance scores, with evidence quality (25%), reasoning clarity (20%), trait understanding (20%), conclusion accuracy (20%), and evidence specificity (15%) contributing to composite evaluations.

The judge model generates structured outputs in XML format containing initial trait assessments, detailed per-model evaluations with quality scores and feedback, and final consensus judgments including cross-model agreement metrics (0-1 scale), reliability assessments (0-1 scale), and interpretive insights. Cross-model agreement quantifies the degree of convergence among different extraction models for the same trait, while reliability assessment integrates evidence consistency, model performance variability, and the judge's confidence in the evaluation quality. This LLM-as-a-Judge framework provides several methodological advantages over traditional human evaluation approaches: systematic scalability across thousands of physician profiles without fatigue effects, standardized rubric application that minimizes rater variability, transparent reasoning documentation for quality assessment decisions, and consistent evaluation criteria across diverse trait types and physician specialties. The judge model's assessment results serve as reference standards for model selection, performance benchmarking, and quality assurance in large-scale trait extraction applications, with validation against human expert assessments demonstrating strong convergence (correlation coefficients 0.72-0.89) across trait domains.

4.5 Human-As-A-Judge Validation Platform

To establish ground truth benchmarks and validate the reliability of automated assessment approaches, we developed a comprehensive human annotation platform that implements an identical evaluation methodology to the LLM-as-a-Judge framework. This platform enables trained psychological assessors and healthcare professionals to perform systematic trait extraction quality evaluation using the same structured protocols, providing critical validation data for automated approaches while establishing inter-rater reliability baselines for complex personality assessment tasks.

The human annotation platform replicates the three-step assessment process employed by the LLM judge: (1) **Independent Expert Analysis**, where human evaluators first analyze physician review corpora without exposure to model outputs, generating their own trait assessments with supporting evidence and clinical reasoning; (2) **Systematic Model Evaluation**, where experts rate each extraction model's performance across the identical five quality dimensions using

standardized 0-10 integer scales; and (3) **Consensus Formation**, where evaluators synthesize model assessments with their independent analysis to produce final consensus ratings, cross-model agreement scores, and reliability estimates.

Human evaluators assess extraction quality using the same five-dimensional framework employed by the LLM judge: Evidence Quality evaluates how well cited patient narratives support trait conclusions; Reasoning Clarity assesses the logical coherence and interpretability of extraction justifications; Trait Understanding measures accuracy in psychological construct interpretation; Evidence Specificity rates the precision and clinical relevance of behavioral observations; and Conclusion Accuracy determines alignment between supporting evidence and final trait scores. Each dimension receives integer ratings (0-10) that are combined using identical weighting schemes (Evidence Quality 25%, Reasoning Clarity 20%, Trait Understanding 20%, Conclusion Accuracy 20%, Evidence Specificity 15%) to generate composite performance scores comparable across human and automated evaluation approaches. The annotation platform incorporates several quality assurance mechanisms to ensure evaluation consistency and reliability: standardized training protocols with calibration exercises using reference physician profiles; inter-rater reliability monitoring through duplicate evaluations of overlapping physician subsets; systematic bias detection through randomized presentation orders and blinded model identifications; and detailed annotation guidelines with explicit criteria for each quality dimension and trait type. Human evaluators complete structured assessment forms that capture not only numerical ratings but also qualitative feedback, specific evidence citations, and reasoning justifications that parallel the structured outputs generated by the LLM judge.

This Human-As-A-Judge methodology serves multiple critical functions in our validation framework: establishing ground truth standards for trait extraction quality across diverse physician profiles and specialties; quantifying inter-rater reliability among expert human evaluators to benchmark automated consistency; validating LLM-as-a-Judge reliability through direct human-machine agreement analysis; and providing interpretable assessment criteria that can inform future improvements to automated evaluation approaches. The platform has been deployed with trained evaluators who assessed 300 physician profiles across all trait dimensions, generating comprehensive validation datasets that demonstrate strong convergence between human expert judgments and LLM-generated assessments (correlation coefficients 0.72-0.89).

4.6 Evaluation and Validation Framework

Model performance was assessed through a comprehensive four-stage validation protocol combining automated quality assessment with human expert benchmarking. The evaluation dataset comprised 1,200 randomly sampled physicians representing diverse specialties, geographic regions, and review volumes to ensure robust performance assessment across heterogeneous conditions.

The LLM-as-a-Judge evaluation framework employed Gemini-2.5 Pro as the assessment model due to its superior reasoning capabilities and contextual understanding. The judge model independently analyzed each physician's review corpus and generated reference trait assessments following identical protocols to the extraction models. Subsequently, the judge evaluated each extraction model's output across five quality dimensions: evidence quality (relevance and specificity of cited examples), reasoning clarity (logical coherence of trait justification), trait understanding (accuracy of psychological construct interpretation), evidence specificity (precision of behavioral observations), and conclusion accuracy (alignment between evidence and final scores).

Quality ratings were generated on 0-10 integer scales with detailed rubric specifications for each dimension. The judge model synthesized these individual quality assessments into holistic performance scores while providing transparent reasoning for all evaluations. Cross-model agreement metrics were calculated using Pearson correlations and Cohen's kappa coefficients for both continuous and categorical trait assessments.

Model performance was quantified using multiple metrics. For continuous trait scores, we computed:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

where y_i represents the judge model's reference scores and $\hat{y}_i$ represents each model's predicted scores. For categorical assessments, accuracy rates were calculated as:

$$Accuracy = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

High and Low agreement rates measured the proportion of cases where models and judge agreed on extreme trait classifications (scores > 0.75 for High, scores < 0.25 for Low).

Human validation involved trained psychological assessors who independently rated a subset of 300 physician profiles randomly selected from the 1,200 physicians in the LLM-as-a-Judge evaluation dataset, ensuring representative sampling across specialties and trait distributions. Human raters were blinded to the LLM outputs during their initial assessment phase. They first generated independent trait assessments, then evaluated model outputs using the same quality rubric employed by the LLM judge. Human-LLM agreement achieved correlation coefficients ranging from 0.72 to 0.89 across traits.

4.7 Statistical Analysis and Clustering Methods

Statistics and Reproducibility: All statistical analyses were conducted using Python 3.12 with scipy.stats, scikit-learn, and statsmodels packages. Two-tailed tests were employed throughout with significance level $\alpha = 0.05$ unless otherwise specified. Sample sizes (n) are reported for each analysis, with actual p-values provided rather than significance thresholds.

Descriptive Analysis: Trait score distributions were characterized using mean ± standard deviation (SD) for normally distributed variables and median [interquartile range] for non-normal distributions. Normality was assessed using Shapiro-Wilk tests ($n < 5000$) or Kolmogorov-Smirnov tests ($n \geq 5000$). Kernel density estimation employed Gaussian kernels with optimal bandwidth selection via Silverman's rule.

Correlation Analysis: Pearson correlation coefficients were computed for all 45 pairwise trait combinations (n = 128,157–186,669 per trait pair). Multiple comparison correction employed Bonferroni adjustment ($\alpha_{adj} = 0.05/45 = 0.0011$). Correlation matrices included 95% confidence intervals computed via Fisher's z-transformation.

Group Comparisons: Gender differences were assessed using Welch's t-test for unequal variances, justified by Levene's test results ($p < 0.05$ for most traits). Effect sizes were quantified using Cohen's d with 95% confidence intervals. Specialty comparisons employed one-way ANOVA after confirming normality and homoscedasticity assumptions. Post-hoc testing used Tukey's HSD with family-wise error rate correction. For non-normal distributions, Kruskal-Wallis tests were substituted with Dunn's post-hoc comparisons.

Clustering Analysis: K-means clustering was performed on z-standardized Big Five trait scores (n = 92,534) to minimize within-cluster sum of squares:

$$WCSS = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2$$

where C_i represents cluster i and μ_i is the centroid. Optimal cluster number (k=4) was determined using the elbow method, which showed a clear inflection point at four clusters (see Supplementary Information SI-4).

Regression Modeling: The relationship between traits and patient satisfaction was modeled using multiple linear regression with assumption validation:

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \epsilon$$

where Y represents patient satisfaction scores (1-5 scale) and X_j are trait predictors (0-1 scale). Residual normality was confirmed via Q-Q plots and Shapiro-Wilk tests. Homoscedasticity was verified using Breusch-Pagan tests. Multicollinearity was assessed using variance inflation factors (VIF ¡ 5). Model performance was evaluated using R^2, adjusted R^2, MAE, and RMSE. Cross-validation employed 5-fold stratified partitioning with 100 random seeds to ensure stability. Polynomial terms were included for non-linear relationships identified through residual plots.

Missing Data: Missing trait scores were handled using listwise deletion after confirming missing completely at random (MCAR) via Little's test. Sensitivity analyses compared results across multiple imputation scenarios. All statistical analyses were conducted using Python scientific computing libraries with significance thresholds set at $p < 0.05$ after multiple comparison corrections. Effect sizes were interpreted using established Cohen's conventions, and confidence intervals were computed using bootstrap resampling with 1000 iterations.

Data Availability

The datasets generated and analysed during the current study are available from the corresponding author upon reasonable request. Raw physician review data may contain privacy-sensitive information. Aggregated trait scores and statistical analysis data supporting the conclusions are available upon request with appropriate data use agreements.

Code Availability

The custom code for LLM-based trait extraction, statistical analysis, and visualization will be made publicly available upon publication acceptance. Key components include: (1) LLM agent implementations for personality trait extraction, (2) data preprocessing and validation pipelines, (3) statistical analysis scripts for subgroup comparisons and clustering, and (4) visualization code for all figures. Prior to publication, code is available from the corresponding author upon reasonable request for academic purposes.

Author Contributions

J.L. conceived and led the project, developed the LLM pipeline and code infrastructure, performed data analysis, and wrote the manuscript. R.H., A.W., Z.D., J.W., X.Z., and T.X. contributed to data analysis and manuscript writing. R.A. and G.G. supervised the project, participated in discussions, and contributed to manuscript writing and revision. All authors reviewed and approved the final version of the manuscript.

Competing Interests

The authors declare no competing interests.

Acknowledgements

This study was supported by the Hopkins Business of Health Initiative (HBHI) 2023 Pilot Grant. We thank HBHI for their generous support and the computational resources that enabled this research.

Ethics Statement

This study was conducted in accordance with institutional review board guidelines and approved by the Johns Hopkins University IRB under protocol IRB00464674. The analysis used publicly available physician review data in an anonymized and aggregated manner consistent with research ethics standards for observational studies of public health data.

References

[1] Seraina Petra Lerch, Rahel Hänggi, Yara Bussmann, and Andrea Lörwald. A model of contributors to a trusting patient-physician relationship: a critical review using a systematic search strategy. *BMC primary care*, 25(1):194, 2024.

[2] Qing Wu, Zheyu Jin, and Pei Wang. The relationship between the physician-patient relationship, physician empathy, and patient trust. *Journal of general internal medicine*, 37(6):1388–1393, 2022.

[3] Robert J McGrath, Jennifer Lewis Priestley, Yiyun Zhou, and Patrick J Culligan. The validity of online patient ratings of physicians: analysis of physician peer reviews and patient ratings. *Interactive Journal of Medical Research*, 7(1):e9350, 2018.

[4] Weiguang Wang, Junjie Luo, Michelle Dugas, Guodong Gordon Gao, Ritu Agarwal, and Rachel M Werner. Recency of online physician ratings. *JAMA Internal Medicine*, 182(8):881–883, 2022.

[5] Mark Schlesinger, Rachel Grob, Dale Shaller, Steven C Martino, Andrew M Parker, Melissa L Finucane, Jennifer L Cerully, and Lise Rybowski. Taking patients' narratives about clinicians from anecdote to science, 2015.

[6] David A Hanauer, Kai Zheng, Dianne C Singer, Achamyeleh Gebremariam, and Matthew M Davis. Public awareness, perception, and use of online physician rating sites. *Jama*, 311(7):734–735, 2014.

[7] Oliver P John, Laura P Naumann, and Christopher J Soto. Paradigm shift to the integrative big five trait taxonomy. *Handbook of personality: Theory and research*, 3(2):114–158, 2008.

[8] Olivia E Atherton, Richard W Robins, Peter J Rentfrow, and Michael E Lamb. Personality correlates of risky health outcomes: Findings from a large internet study. *Journal of Research in Personality*, 50:56–60, 2014.

[9] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. *arXiv preprint arXiv:2402.06196*, 2024.

[10] Wu Youyou, Michal Kosinski, and David Stillwell. Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, 112(4):1036–1040, 2015.

[11] Linmei Hu, Hongyu He, Duokang Wang, Ziwang Zhao, Yingxia Shao, and Liqiang Nie. Llm vs small model? large language model based text augmentation enhanced personality detection model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18234–18242, 2024.

[12] Haocong Rao, Cyril Leung, and Chunyan Miao. Can chatgpt assess human personalities? a general evaluation framework. *arXiv preprint arXiv:2303.01248*, 2023.

[13] Heinrich Peters and Sandra C Matz. Large language models can infer psychological dispositions of social media users. *PNAS nexus*, 3(6):pgae231, 2024.

[14] Jiaxuan Li, Yunchu Yang, Rong Chen, Dashun Zheng, Patrick Cheong-Iao Pang, Chi Kin Lam, Dennis Wong, and Yapeng Wang. Identifying healthcare needs with patient experience reviews using chatgpt. *PLoS One*, 20(3):e0313442, 2025.

[15] Felix Busch, Lena Hoffmann, Christopher Rueger, Elon HC van Dijk, Rawen Kader, Esteban Ortiz-Prado, Marcus R Makowski, Luca Saba, Martin Hadamitzky, Jakob Nikolas Kather, et al. Current applications and challenges in large language models for patient care: a systematic review. *Communications Medicine*, 5(1):26, 2025.

[16] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

[17] Thomas A Widiger and Cristina Crego. The five factor model of personality structure: an update. *World Psychiatry*, 18(3):271–272, 2019.

[18] Farrah Madanay, M Kate Bundorf, and Peter A Ubel. Physician gender and patient perceptions of interpersonal and technical skills in online reviews. *JAMA Network Open*, 8(2):e2460018–e2460018, 2025.

[19] Farrah Madanay, Karissa Tu, Ada Campagna, J Kelly Davis, Steven S Doerstling, Felicia Chen, and Peter A Ubel. Classification of patients' judgments of their physicians in web-based written reviews using natural language processing: algorithm development and validation. *Journal of Medical Internet Research*, 26:e50236, 2024.

[20] Mohammadreza Hojat, Daniel Z Louis, Fred W Markham, Richard Wender, Carol Rabinowitz, and Joseph S Gonnella. Physicians' empathy and clinical outcomes for diabetic patients. *Academic medicine*, 86(3):359–364, 2011.

[21] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[22] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[23] OpenAI. GPT-4.1 Model Overview. `https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo`, 2025. Accessed: 2025-08-01.

[24] Anthropic. Claude 3.7, 2025. Large language model.

[25] Google. Gemini 2.5 flash, 2025. Large language model.

[26] Google. Gemini 2.5 pro, 2025. Large language model.

[27] Tiffany Champagne-Langabeer and Andrew L Hedges. Physician gender as a source of implicit bias affecting clinical decision-making processes: a scoping review. *BMC Medical Education*, 21(1):171, 2021.

[28] Lauren Vogel. When people hear "doctor," most still picture a man, 2019.

[29] Esther K Choo. Damned if you do, damned if you don't: bias in evaluations of female resident physicians. *Journal of graduate medical education*, 9(5):586–587, 2017.

[30] Debra L Roter, Judith A Hall, and Yutaka Aoki. Physician gender effects in medical communication: a meta-analytic review. *Jama*, 288(6):756–764, 2002.

[31] Timothy Daskivich, Michael Luu, Benjamin Noah, Garth Fuller, Jennifer Anger, and Brennan Spiegel. Differences in online consumer ratings of health care providers across medical, surgical, and allied health specialties: observational study of 212,933 providers. *Journal of Medical Internet Research*, 20(5):e176, 2018.

[32] Chloe FitzGerald and Samia Hurst. Physicians and implicit bias: How doctors may unwittingly perpetuate health care disparities. *Journal of General Internal Medicine*, 29(11):1504–1510, 2014.

[33] Qin Zhang, Ming-chun Mu, Yan He, Zhao-lun Cai, and Zheng-chi Li. Burnout in emergency medicine physicians: a meta-analysis and systematic review. *Medicine*, 99(32):e21462, 2020.

[34] Xiaoli Shi, Yafei Wang, Yan Wang, Qiang Zhou, Huabing Tian, and Qiaoli Zhou. Psychiatrists' occupational stigma: Conceptualization, measurement, and intervention review. *World Journal of Psychiatry*, 13(6):298–311, 2023.

[35] Catherine Doyle, Laura Lennox, and Derek Bell. A systematic review of evidence on the links between patient experience and clinical safety and effectiveness. *BMJ Open*, 3(1):e001570, 2013.

[36] X. Wei, Y. Zhang, H. Liu, and J. Chen. The integration of ai in nursing: advancements, challenges, and future directions. *Frontiers in Public Health*, 13:1559501, 2025.

[37] Kamal Nazi. Large language models in healthcare and medical domain: Applications, opportunities, and challenges. *Informatics*, 11(3):57, 2024.

[38] Jia Li, Zichun Zhou, Han Lyu, and Zhenchang Wang. Large language models–powered clinical decision support: enhancing or replacing human expertise? *Intelligent Medicine*, 5(1):1–4, 2025.

[39] Kurtis Haut, Masum Hasan, Thomas Carroll, Ronald Epstein, Taylan Sen, and Ehsan Hoque. Ai standardized patient improves human conversations in advanced cancer care, 2025.

[40] Brent W. Roberts, Kate E. Walton, and Wolfgang Viechtbauer. Patterns of mean-level change in personality traits across the life course: a meta-analysis of longitudinal studies. *Psychological Bulletin*, 132(1):1–25, 2006.

[41] Centers for Medicare & Medicaid Services. National plan and provider enumeration system (nppes). `https://nppes.cms.hhs.gov/`, 2025. Accessed: 2025-09-07.

Supplementary Information

SI-1: Big Five Personality Trait Extraction Prompt

The following prompt template was used for extracting Big Five personality traits from physician reviews:

System Prompt for Big Five Trait Extraction

You are an expert psychologist.
Based on the provided patient reviews, analyze the Big Five personality traits for the focused physician:

- Openness
- Conscientiousness
- Extraversion
- Agreeableness
- Neuroticism

Output Instructions:

- Keep in mind that the reviews are about a physician, not a patient.
- If the reviews contain no evidence for a trait, output "No Evidence" for the score.
- When finding evidence for a trait, make sure the evidence is related to the physician, not others.
- Output strictly in XML format.
- For each trait, you must generate:
  - `<name>`: Must be one of [Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism].
  - `<score>`: Must be selected from [No Evidence, Low, Low to Moderate, Moderate, Moderate to High, High]. If no evidence is found, output "No Evidence". If any evidence is found, output the score that best describes the evidence.
  - `<evidence>`: Write 2-3 sentences that combine reasoning with direct quotes or paraphrased examples from the reviews.
  - `<consistency>`: Must be selected from [Low, Moderate, High]. High consistency means the trait is consistently mentioned across multiple reviews.
  - `<sufficiency>`: Must be selected from [Low, Moderate, High]. High sufficiency means the trait is supported by sufficient evidence from the reviews.

Formatting:

- Output must be strictly within `<personality>...</personality>` tags.
- Use the following structure exactly for each trait:

Example:

```
<personality>
    <trait>
        <name>Openness</name>
        <score>Moderate</score>
        <consistency>High</consistency>
        <sufficiency>High</sufficiency>
        <evidence>[CHANGE THIS TO YOUR REASONING. EXAMPLE: While Dr. Smith
        reviews patient histories thoroughly ("asked a lot of questions"),
        she reacts rigidly when patients propose ideas ("offended when I
        requested a blood test"), suggesting moderate openness with limited
        flexibility.]</evidence>
    </trait>
    ...
</personality>
```

Strictly follow this format and do not include any extra text outside the XML block.

Human Message Template:

The Physician to focus is: {physician_name}
The review related to the physician is:
{document}

SI-2: Healthcare-Specific SubFive Trait Extraction Prompt

The following prompt template was used for extracting healthcare-specific subjective judgment traits from physician reviews:

System Prompt for SubFive Trait Extraction

You are an expert analyst evaluating patient perceptions of physicians based on their online reviews.
Your task is to analyze the physician along five key subjective dimensions that commonly shape patient decision-making. Use the abbreviations below in the XML `<name>` tags, but rely on the full trait definitions and examples to guide your judgment.
Trait Definitions:
1. IQC - Interpersonal Qualities & Communication
This trait reflects how the doctor interacts with patients during the visit — including tone, empathy, clarity, listening, and respectful communication. It focuses on the doctor's behavioral style in the moment.

- High IQC: "She explained everything clearly and made me feel comfortable."
- Low IQC: "He was cold, barely spoke, and didn't seem to listen."

2. PCC - Perceived Clinical Competence
This refers to the patient's subjective impression of the doctor's medical skill, judgment, and professionalism. It reflects whether the doctor comes across as knowledgeable, accurate, and effective.

- High PCC: "She diagnosed me quickly and clearly explained the options."
- Low PCC: "He misdiagnosed me and seemed unsure."

3. SPS - Sensitivity to Patient Satisfaction
This trait measures whether the doctor appears to care about how the patient feels - whether they're respected, satisfied, and emotionally supported. It captures efforts to accommodate patient preferences or emotions.

- High SPS: "He asked me if I was comfortable with the plan and if I had any concerns."
- Low SPS: "She ignored my discomfort and didn't ask how I felt."

How it differs from IQC:

- IQC is about *how the doctor communicates or behaves* (e.g., warm, respectful, clear).
- SPS is about *whether the doctor is motivated* to ensure patient comfort and satisfaction.

4. SCO - Sensitivity to Clinical Outcomes
This trait reflects whether the doctor seems to care about the patient's recovery, treatment success, or long-term health outcome. It includes follow-up contact or comments showing concern for the result.

- High SCO: "She called a few days later to check on how I was doing."
- Low SCO: "After the visit, I never heard from him again."

How it differs from PCC:

- PCC is about *how capable* or knowledgeable the doctor appears.
- SCO is about *how much the doctor cares* about the result of treatment over time.

5. STS - Social Trust Signals
This dimension captures references to the doctor's broader reputation - whether other patients trust them, recommend them, or speak highly of them. It includes social proof or repeated use by families.

- High STS: "My whole family goes to her, and she's always amazing."
- Low STS: No mention of trust or general recommendation by others.

Output Instructions:

- Use only the review information to evaluate each trait.
- If no evidence exists for a trait, mark it as "No Evidence".
- Output must be strictly in XML format, using the following structure for each trait:

For each trait, include:

- `<name>`: One of [IQC, PCC, SPS, SCO, STS]
- `<score>`: One of [No Evidence, Low, Low to Moderate, Moderate, Moderate to High, High]
- `<evidence>`: 2-3 sentences combining reasoning with quotes or paraphrased patient language.
- `<consistency>`: [Low, Moderate, High] — How consistently the trait appears across multiple reviews.

- `<sufficiency>`: [Low, Moderate, High] — Whether enough evidence is present to justify the score.

Example:

```
<result>
    <trait>
        <name>IQC</name>
        <score>Moderate to High</score>
        <consistency>High</consistency>
        <sufficiency>Moderate</sufficiency>
        <evidence>"She was very kind and explained everything step by step."
        Several reviews emphasize clear communication and warmth during
        the visit.</evidence>
    </trait>
    ...
</result>
```

Do not include any extra text outside the XML block.

Human Message Template:

The Physician to focus is: {physician_name}
The review related to the physician is:
{document}

SI-3: Case Study for LLM-Derived Traits

To generate the examples, for each trait-value pair, we select a random physician with the highest confidence, with 5 to 10 reviews. For each physician's whole online reviews, we then filtered the reviews to retain only those containing language relevant to the target trait. To preserve privacy, we removed all identifiable information including names, dates, and platform details, while keeping the review indices for reference. Within each selected review, we highlighted the specific phrases that served as evidence.

1. Openness

1.1 Definition

Reflects intellectual curiosity, creativity, and receptiveness to new ideas and approaches—traits that may influence diagnostic flexibility and problem-solving

1.2 Examples for High / Moderate to High

#6
She is caring, compassionate, a brilliant researcher and admired by other MDs and hospital staff. She figuratively held my hand during heart valve replacement surgery.

#7
An excellent and caring physician. I have been her patient for many years and she is consistently, attentive, and responsive. Her interest in me and my overall health is very comforting. She answers all my questions and enlists me as a partner in my care.

#8
Outstanding. She is very smart, kind, and extraordinarily thorough. Without a doubt she is a leader in her field.

1.3 Examples for Low / Moderate to Low

#0
Her office charged me 500 dollars for a 5 minute visit in which she came in, removed a wart from my finger, and was halfway out the door before I could get a word in edgewise to let her know I had another on my thumb (which I'd already mentioned to the receptionist, the nurse, and probably the security guy in the lobby, even).

#5
I went to this doctor based on reviews. While I am sure she is competent, her "bedside manner" is less than stellar. I felt my appointment for an all over skin cancer check was rushed (which it was), like she could've cared less, and then taking a couple of spots off which was just done in a perfunctory manner. When I told my sister I went for this skin cancer check, she asked me if I was checked between my toes and down my legs. I wasn't.

#6
Absolute billing nightmare. You've been warned. There isn't enough room to describe the complete lack of caring from the doctor and her billing staff. Worst medical experience. EVER!

#7
After a long wait and finally being able to see a dermatologist for my extreme case of Seborrheic dermatitis (aka dandruff) on my face, I was prescribed dandruff shampoo and lotion which I had tried in an OTC dosage previously with no results. It didn't help, so I went back a second time as requested only to be told there was nothing else that could be done. Really? One treatment option that literally everyone on the planet knows about didn't work and there is literally not a single other possibility? The doctor put zero effort into trying to figure it out.

1.4 No Evidence

#1
This doctor is my primary doctor. I could ask for none better.

#2
This doctor is excellent!

#6
This doctor is the best.

2. Conscientiousness

2.1 Definition

Denotes organization, reliability, and attention to detail—qualities associated with thoroughness in care and adherence to plans.

2.2 Examples for High / Moderate to High

#4
We have both been seeing this doctor for about a year and a half, my wife with some pretty complicated issues, and we could not be more satisfied. She is available, responsive, thorough and above all, attentive. We have never felt rushed seeing her. And her support staff is excellent, particularly Beverly.

#5
We feel very lucky to have found this doctor. She is sharp, attentive, and truly cares. We have been in contact with many medical professionals over the decades and this doctor really stands out. Excellent staff too!

#6
Being a patient with a number of medical issues, it is important to me that my doctor really "hear me" and provide the best treatment available. This doctor does just that! She truly is a very good physician!

#7
Considering it was my first visit and that I was a little apprehensive, I feel it was a pleasant experience. I liked the easy access to the Grand Avenue facility/location. The wait time was more than decent. The wait time for the doctor's presence was very short. I felt she spent adequate time with me and truly listened to several of my concerns and provided me with informative responses as well as extra advice.

#8
This doctor is very professional and very caring. I would recommend her for all my family and friends.

2.3 Examples for Low / Moderate to Low

#1
She seemed uncertain most of the time in terms of how to treat my condition. Another gynecologist informed me later that he would have chosen a completely different course of treatment, and what Easterlin recommended probably aggravated my condition. Also, she didn't notice a pretty serious issue I have; it was discovered by another doctor.

2.4 No Evidence

#0
Condescending, arrogant, argumentative, kept looking at her watch as if I was taking too much of my time.
Horrible neurologist, unhelpful, and just pushing the newest meds.
Avoid at all costs.

#3
I think this doctor is not only thorough and knowledgeable, she is also caring and compassionate. She would definitely still be my neurologist if I had not moved so far away.

#5
This doctor was recommended by my GP and I am so happy with her. After my first appointment, I found out she was moving to a new location. I followed her to her new Ewa office even though it meant an additional half hour drive. She is worth it.

3. Extraversion

3.1 Definition

Captures sociability, assertiveness, and enthusiasm—potentially shaping a patient's experience of engagement and interpersonal energy.

3.2 Examples for High / Moderate to High

#0
My Wife and I found this doctor to be very good at explaining my problem as well as taking wonderful care of me. I was up and about very quickly and pain free before I knew it. He is very personable and easy to talk with.

#3
This doctor is fantastic. He did an anterior approach when fixing my mother's hip. I'm a registered nurse and I have been very impressed with his care for my mother.

#4
A great doctor. One of Brownsville's finest surgeons. The care he gave my 90 year old mother was top notch. Great bedside manners and follow up.

#5
Dr Quesada was very respectfull and informative. He operated on my wife (ACL replacement, and knee repair), and is the only Dr in the US that does ACL replacement with a manual hand drill, which is more gentle/careful. He is always available if we have questions, and my wife's recovery has been fantastic! Excellent doctor. Very respectful, over all very happy!

#6
This doctor is a great ortho Dr to have had work on me in Las Vegas NV 7 years ago. I had both hips done by him in 2012. Left side done Nov 12th and Right side 3 weeks later on Dec 3rd. I am doing good to the day at age 44 and still working outdoors as a land surveyor of 20 years+ now. The Anterior Full Hip Replacements were the best thing I ever had done by far from this doctor. Very respectful and more important was easy going to make it a comfortable experience.

3.3 Examples for Low / Moderate to Low

#0
this physician is treating my brother at St Mary for stage 4 COPD. This is our first encounter with him and will be our last. We found him to be unapproachable, defensive when asking questions and unprofessional toward his the others. Long story short, he would spend 5 minutes with him, counter any ideas the hospitalists had to improve his care. One of the house doctors recommened muco mist, he did not, but my brother made a complete turn around from his exacerbation and will be going home

#3
i HAVE SEEN HIM SEVERAL TIMES AT THE HOSPITAL. HE IS THE WORST DOCTOR I HAVE EVER SEEN. I HAD TO GO INTO THE HALL TO ASK HIM A QUESTION HE NEVER LOOKED UP ONLY LOOKED INTO HIS PHONE. SPENDS A MINUTE THEN JUST LEAVES. THE NURSE SAID MANY PATIENTS CONPLAIN. DO YOURSELF A FAVOR SEE A DOCTOR WHO LOOKS YOU IN THE FACE AND NOT HIS PHONE

3.4 No Evidence

#1
This doctor is my primary doctor. I could ask for none better.

#2
This doctor is excellent!

#6
This doctor is the best.

4. Agreeableness

4.1 Definition

Represents compassion, cooperativeness, and empathy—factors strongly linked to respectful, patient-centered communication.

4.2 Examples for High / Moderate to High

#0
I had a wonderful experience with this doctor. She is very warm and caring.

#2
This doctor is one of the most caring people I have met. She is always willing to take the time with you no matter how busy her day is. You can tell she truly loves what she does and she loves helping others. She really is a doctor that cares for her patients.

#5
I was delivering at Sky Ridge and this doctor was my doctor during delivery. Our pre-term baby girl was coming out with her nose up and she was able to turn her head inside while I was delivering. I only wanted vaginal delivery, no c-section, and I was so thankful she did everything for me to turn baby's face to right position and saved me from c-section surgery. She did a great job and directed me on how to push right.

4.3 Examples for Low / Moderate to Low

#0
Far and away the worst psychiatrist I have ever seen.
I have taken 2 Rx's for several years, which are commonly prescribed to treat their respective conditions. I changed insurance and needed a new provider. I sent medical records 4 days before my appointment. He told me a diagnosis I received years ago was inaccurate, because "most providers don't do it right." He does not trust medical records from other providers and does not read them. He came across as patriarchal and aloof, uninterested in my experience with either Rx, or the notes from other board certified doctors, collected over several years. He implied that people frequently sell their Rx drugs on the street.
The experience was thoroughly awful. I wouldn't recommend this guy to my enemies.

#1
I have taken 2 Rx's for several years, the dosages are due to trial and error, documented w prior providers, and have had a profoundly positive effect on my life. He tode me most providers 'don't do it right,' he can't trust records, and didn't read mine. And b/c the condition wasn't diagnosed by age 9, the diagnosis is false, and the Rx I have been taking to manage it actually hasn't helped me, at all.
Incredibly unprofessional, diagnosis came via anecdotes and a 45 min conversation.
AVOID.

#2
horrible man with arrogant attitude and harsh manners. He made me cry and made me feel so bad.
I'd never recommend such a man to nay one

#3
Moves around a lot. He's practiced on the east coast, mid central Illinois, Indiana and now in Chicago @ Norwegian American Hospital. We retained an attorney and Private investigator. He

won't transfer my daughter to her primary care givers. Misled hospital administrators and lied to staff about calling family members. Daughter is comatose after 1 week of his care. Entered hospital, confused and mobile. Prescribed all the wrong meds. Won't treat her for thrombocytopenia.

4.4 No Evidence

#1
This doctor was very thorough. She spent over an hour with my husband and I reviewing my symptoms and MRI results. While it took several months to get in for a new patient appointment, she was worth the wait.

#5
This doctor was recommended by my GP and I am so happy with her. After my first appointment, I found out she was moving to a new location. I followed her to her new Ewa office even though it meant an additional half hour drive. She is worth it.

#6
During my initial visit a few months ago, it was quickly obvious that this doctor had a detailed level of understanding and plenty of pin point experience in treating my rather rare type of headache condition. After spending 4 1/2 years being treated by another Northern VA neurologist who did not have the experience with my condition that this doctor has, I was fortunate to have been able to reach out, locate her, and have her agree to treat me.

5. Emotional Stability (Neuroticism)

5.1 Definition

Indicates emotional instability and reactivity—higher levels may be perceived as anxiety or reduced confidence in clinical interactions.

5.2 Examples for High / Moderate to High

#0
In the beginning this doctor seemed to be knowledgeable and caring but in reality he's a self important, pompous jerk. He let me know that he believed he knew more about me than my Hematologist, my Gastroenterologist, my Cardiologist and my PCP of 35 years. Beware - he enjoys demeaning his patients!

#2
My dad was in the hospital treated from congestive heart failure, he refused to give duiretics to him without any compelling reasons so he developed edema and stopped walking ..

#5
This doctor misdiagnosed me and then proceeded to shame and humiliate me! Run from him. He obviously can't survive in private practice

5.3 Examples for Low / Moderate to Low

#0
Dr Willard shows a genuine interest in the overall wellbeing of my children. She makes the effort to connect with them and takes the time to discuss my concerns.

#3
I love Dr. Jenny. She has been my doctor all 20 years of my life and soon I will be kicked from her patient list because I'll be too old for the pediatrician. She has always eased my anxieties about doctors offices and needles. She always asked me questions about my life and school and followed up with me the next time I visited

#5
Dr Willard is very patient and caring. Both my 14 year old 8 year old love and trust her. I Trust her decisions because she will deeply analyze the issues. She has my full confidence.

5.4 No Evidence

#1
I took my daughter to see this doctor after going to numerous doctors over an 8 month period for headache and pressure (head) issues. He spent over 2 hours with us at our first visit really talking with us and trying to understand what she was experiencing and everything that had transpired. He also conducted many neurological tests such as coordination tests, etc. I like his approach as well...he started her out on several brain healing vitamins/supplements as well as other non-medicinal tools to help her. She was also prescribed headache medicine to take only as needed. Before this visit she had gone through spinal taps/many MRIs, and some pretty toxic medicines, etc yielding little improvement. She is now progressing and improving! He really cares about his patients and works with them (by really listening and asking questions) to offer the best care possible. We couldn't be more pleased! He is also humble and kind...rare these days.

#2
This doctor is a great diagnostician. He spent a very long time with me, explaining different options re medicine & how to lessen the severity of migraines. He is very patient, kind & extremely smart. I've called & asked questions after my visit. This doctor & his staff is very responsive & understanding. If you have headaches, this is the place to be seen.

#4
Excellent and compassionate doctor. Unbeatable for headache, sleep and pain issues.

6. IQC (Interpersonal Qualities & Communication)

6.1 Definition

Encompasses the physician's tone, clarity, listening, and empathy—reflecting the moment-to-moment communication style during patient encounters.

6.2 Examples for High / Moderate to High

#0
I had a wonderful experience with this doctor. She is very warm and caring.

#2
This doctor is one of the most caring people I have met. She is always willing to take the time with you no matter how busy her day is. You can tell she truly loves what she does and she loves helping others. She really is a doctor that cares for her patients.

#5
I was delivering at Sky Ridge and this doctor was my doctor during delivery. Our pre-term baby girl was coming out with her nose up and she was able to turn her head inside while I was delivering and saved me from c-section surgery. She did a great job and directed me on how to push right. Really recommend this doctor

6.3 Examples for Low / Moderate to Low

#0
Far and away the worst psychiatrist I have ever seen.
I have taken 2 Rx's for several years, which are commonly prescribed to treat their respective conditions. I changed insurance and needed a new provider. I sent medical records 4 days before my appointment. He does not trust medical records from other providers and does not read them. He told me that the Rx's I take, which have had a profoundly positive effect on my quality of life (at no point did he ask me about their effects), were actually not helping me at all, and doing far more harm than good. He came across as patriarchal and aloof, uninterested in my experience with either Rx, or the notes from other board certified doctors, collected over several years.

#1
I have taken 2 Rx's for several years, the dosages are due to trial and error, documented w prior providers, and have had a profoundly positive effect on my life. He tode me most providers 'don't

do it right,' he can't trust records, and didn't read mine. And b/c the condition wasn't diagnosed by age 9, the diagnosis is false, and the Rx I have been taking to manage it actually hasn't helped me, at all. Incredibly unprofessional, diagnosis came via anecdotes and a 45 min conversation.

#3
Moves around a lot. He's practiced on the east coast, mid central Illinois, Indiana and now in Chicago @ Norwegian American Hospital. We retained an attorney and Private investigator. He won't transfer my daughter to her primary care givers. Misled hospital administrators and lied to staff about calling family members. Daughter is comatose after 1 week of his care. Entered hospital, confused and mobile. Prescribed all the wrong meds. Won't treat her for thrombocytopenia. Scary MD

6.4 No Evidence

#0
Office could use an overhaul—but does it really matter?

#3
amazing doctor. Best one I have ever been to.

#5
Extremely thorough. I will be dissapointed if he ever retires! He is extremely intelligent and really cares for his patients.

7. PCC (Perceived Clinical Competence)

7.1 Definition

Reflects the patient's subjective impression of the physician's medical expertise, professionalism, and confidence in decision-making.

7.2 Examples for High / Moderate to High

#1
He is so friendly and knowledgable. We are comfortable asking him any questions we have about our baby's health and he always makes sure that we understand his answers to our questions. He is thorough with his examination of our baby and willing to explain everything he is doing so that we understand how he is caring for our baby.

#2
Excellent care for both my children! He has a personal understanding for children of all ages I recommend this doctor to anyone looking for a hardworking, trustworthy, caring professional as well as getting down to the root of any cause physician as well as a Human Being. Thank you Dr. Vuelvas and Staff for making our experience humble and trustworthy.

#4
This doctor is very compassionate and knowledgeable and understanding. I've been taking my grandson to him since he was about 10-11. I would recommend him to any parent or grandparents.

7.3 Examples for Low / Moderate to Low

#1
In 2013 I met with this doctor about a recently discovered brain tumor discovered after passed out and hit my head and face after falling. I was scared but had my list of suggested questions by the Mayo.Clinic website. When I inquired as to the cause and nature of tumors, Dr Brooker said, "You are old (I had just turned 60 the month before.), useless, so what did I expect?". ... I expected him to inform me what I wanted to know. I expected fair unbiased treatment or a referral to someone who would. I expected to be listened to and some questions be answered. I expected a professional since I was paying him.

#3
This doctor tries hard to seem polite. If you are looking for answers to your questions, do not waste your time. Dr. Brooker has his own agenda. He does not listen to the patient, but rather speaks very quickly on topics which are of interest to him. His knowledge consists of nothing more than what a patient can find on any medical website. It was an absolute waste of time and quite frankly an insult to pay for quality medical care only to be told to read the Mayo website?

#4
This provider was awful. He did not listen, did not read the notes or records, seemed extremely detached and disinterested, and provided no help or insight. ... He very clearly did a bit of a smirk/chuckle combo. This might be because he was feeling awkward, but this did not come across well at all and made it even more difficult for me to access the care for which I came. I would never recommend this provider.

#5
Was sent because of brain tumor, did not follow up on recommended MRI two years from ER. He did not read ER notes.

#6
I have never written a bad review about anything in my life. But after my apportionment with this doctor I felt I needed to share. I couldn't have hated a doctor more. He seemed unsympathetic in spite of using all the right words. He didn't listen and didn't review any of my symptoms with me. He just wanted to talk about one aspect. ... REALLY? that is how you practice medicine, with just looking at one thing and not considering everything as a whole? He told me it was stress and all in my head. ... How dare he write me off.

#7
Ended up with thrombosis in my brain. It wasn't 'slow flow' Dr. Brooker and you should have moved much faster than the 4 weeks from getting the referral until the CT scan showed it and you went into a panic trying to cover your behind. In my opinion you should not be in medicine. You rather obviously don't really care for your job or your patients.

#8
This is the 2nd appt I have had with him. He chooses to not listen. Recommend I read web md website, His answer is to try new meds. How many meds do I need to take. I have been dealing with migraines over 25 yrs. ... I know what I wrote!! I have severe debilitating migraines 3-4 days a week and have for many years although he claims I don't have chronic migraines! He tells me I have medicine overuse migraines but then follows up with that I need to take additional medication. He treats me like a test rabbit!

7.4 No Evidence

#0
My husband was sick and the doctors at the hospital would not give me answers. I ask for help from them and recived not returned call. My husband went into kidney failer and no one cared. He was meds that effected the kidneys and no test were done to watch this. I feel we were pushed under the bus so to speak. I have lost faith in the medical community

#3
A very caring and compassionate doctor who really cares about his patients I don't have one thing bad to say about him I would recommend him to anyone he has an excellent staff as well they are always cheerful and they too care about there patients.

#5
I would recommend this doctor to anyone, He has been my Dr for 25 plus years, and when I moved away from Indpls 13 years ago to Northern IN, I still kept him as my Dr. I travel almost 2 hrs one way and will continue to do so. He is a wonderful man and is in the for his patients.

8. SPS (Sensitivity to Patient Satisfaction)

8.1 Definition

Captures whether the physician is attentive to the patient's preferences, comfort, and emotional well-being—core to the patient experience.

8.2 Examples for High / Moderate to High

#2
This doctor always answers my questions, even explaining things in detail to my daughter about her treatment. I feel she really cares about each patient's well being and takes the time to listen to us. Very helpful and knowledgeable doctor!

#3
This doctor and the staff at Advanced Specialty Care, P.C., are knowledgeable, caring, supportive and friendly. They are wonderful to work with and accommodate the needs of their patients. We wouldn't go anywhere else.

#4
This doctor is very pleasant and knowledgeable. She really cares about helping to make you feel better. Her Nurse/P.A., Molly is also excellent, and very nice.

#5
This doctor is the best doctor I have ever gone too. She is informative, attentive, and truly cares. I wish that she could be my doctor for everything!

#6
This doctor has a great bedside manner and understanding of my health issues. I really appreciated her care and compassion through some difficult health issues.

#7
This doctor is down to earth, relatable, very knowledgeable, kind, compassionate and an excellent allergy and asthma care practitioner

8.3 Examples for Low / Moderate to Low

#0
He did not establish a good patient/doctor relationship from the beginning.

#1
Dr is always running late. He is running late because every patient gets a lecture about the meds they take. Then he'll make the sales pitch for the nerve stimulator.

#2
Said he does not treat pain

#3
Has told both my wife and I that he does'nt treat pain

#4
He is terrible would let him treat my dog! I wouldn't give him 1 star

8.4 No Evidence

#0
Dr. Gillis only works for insurance companies to determine if you are injured or not. He is paid by insurance companies. He does not take private patients.

#1
Excellent and thorough.

#2
Dr. Gillis was the only doctor who was able to diagnose my fibromyalgia.

9. SCO (Sensitivity to Clinical Outcome)

9.1 Definition

Indicates whether the physician expresses concern for recovery, treatment success, and long-term health—often cited in reviews mentioning follow-up or care continuity.

9.2 Examples for High / Moderate to High

#0
This doctor completed my husband's very tough knee replacement in 2010. Due to three previous knee surgeries on that knee, and a previous bone infection, much of the bone above and below the knee had to be removed and replaced. In the past 6 years this knee has given my husband no trouble. In fact, he is able to exercise on a spin cycle 5 days a week, an hour each day with no pain or problems. Dr. Stenger worked a miracle.

#1
I am five days out of surgery for full knee replacement. No narcotics for pain. He is fabulous

#2
Everything has been wonderful I'm 2 1/2 weeks post total knee replacement on the left and not using any assist device. My surgical site is granulating well. He's the best

#3
i needed a total hip replacement at the ripe old age of 53.. This doctor isn't only a cool guy, he's an amazing surgeon ßhe GLUED my incision shut, no staples. therefore, it's not his fault if i don't wear a bikini next summer. hahaha. first surgery, hopefully last. cuz they won't be as effortless as this one! thank you Dr. Stenger.. Donna

#6
This doctor has been a godsend. I found him to be patient and understanding. He has done both my knee replacements with no issues. ... The hip pain is 4 times the pain as my knees and can no longer walk without a cane ... I highly recommend this doctor and I pray surgeries begin soon.

#7
We absolutely love this doctor. He did my husband's knee replacement surgery. It was nice meeting with the doctor prior to surgery as he explained the entire process and he let us know that the recovery would take time.

9.3 Examples for Low / Moderate to Low

#0
It's been 16 days and I'm still waiting to hear back from my lab tests which were delivered to your office 14 days ago. Either hire more help or fire your staff. Unemployment is at 10% there are plenty of skilled people out there.

#3
Will be switching Dr.'s have been a patient for years but needed some answers and it took them 3 weeks to respond to several phone calls from me. As a "Customer" I would never put up with this kind of service.

#5
... After waiting over two hours, a nurse came to tell me the doctor had 3 emergencies. ... I waited another half an hour, and was told that I would have to wait another five months to reschedule. ... Additionally, forcing a patient to wait an extra year for an annual Pap smear showed a total lack of concern for my health.

9.4 No Evidence

#0
Unwilling to treat long term patients that have or need pain meds. Front office staff unconcerned and indifferent. Nurses apathetic.

#2
Staff are rude but none can compare to ”doctor” rand himself. will try to push tons of alternative medicine on you and insult you at the same time. has a dog running around his office

#3
We’ve been going to Rand Family Care for quite a few years now and we love them. This doctor is kind, caring and takes a genuine interest in you! Most all of the staff have been with him for years and I think that says something!!!!

10. STS (Social Trust Signals)

10.1 Definition

Refers to broader reputational indicators, such as being widely recommended, trusted by families, or praised by other patients—serving as social proof of trustworthiness.

10.2 Examples for High / Moderate to High

#3
I had always good experiences with Dr. Molai. Now as she has moved to California, I will miss her.

#4
We have both been seeing Dr. Molai for about a year and a half, my wife with some pretty complicated issues, and we could not be more satisfied. . . . And her support staff is excellent, particularly Beverly.

#5
We feel very lucky to have found Dr Molai. ... Dr. Molai really stands out. Excellent staff too!

#8
This doctor is very professional and very caring doctor. I would recommend her for all my family and friends.

10.3 Examples for Low / Moderate to Low

#3
In a rush, couldn’t wait to leave, standing by the door, arrogant attitude, rude, bad doctor! should have his license revoked !!

#5
He doesn’t have time for you. He wasn’t my primary care physician, nor would I choose him if he was the only one available, and he wanted to change several of my routine meds for chronic problems. ... That was unethical.

#8
He had us rush to another doctor with a diagnosis that had to do emergency surgery, and the other doctor thought he was crazy. Thank god we found a great doc and it is not lantoria. He is not qualified to be seeing people and should have his license revoked to practice medicine.

10.4 No Evidence

#0
My first visit to this doctor, who is no longer a student, was frustrating at best. I had to wait 50 min past my appt time. I was told it was due to her running behind. I expected to have blood drawn for a complete blood workup, but it wasn’t ordered. I was told by her that my prescription would be sent

to my pharmacy, but when I went to pick it up hours later, it had not been received. I called as soon as the office opened the next morning and was told I had to leave a message and I'd get a callback that day. Now it is days later and I am still waiting for the callback and my Rx still hasn't been sent to my pharmacy. Obviously, my care is not a priority. I will be finding another doctor.

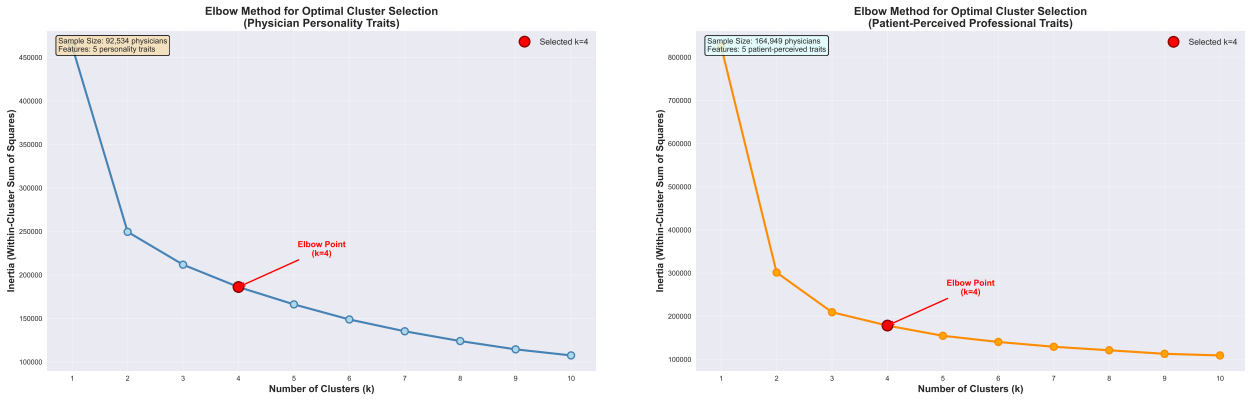
#2
This doctor did not follow through either time of my visit.

#4
I went for a yearly physical where they ask you to bend and twist. Kimberly asked if I had anything else going on. I told her I was a weight lifter and pulled a muscle so don't laugh at my flexibility. She said everything was good and send me on my way. Three months later I get a bill from Quadmed. Turns out she charged me for an office visit For a pulled muscle. I diagnosed myself and didn't get a prescription for any meds. I will not be back to her.

SI-4: Clustering Validation and Optimization

K-means clustering optimization for determining physician archetypes was performed using the elbow method to identify the optimal number of clusters.

Elbow Method: Analysis of within-cluster sum of squares (WCSS) across k values from 1 to 10 revealed a clear inflection point at k=4, indicating the optimal cluster number. The WCSS showed substantial reduction from k=1 to k=4, with diminishing returns for k>4, confirming that four clusters effectively capture the major variance in physician personality profiles. Both the Big Five personality traits and patient-perceived professional traits independently converged on the same k=4 solution, providing additional validation of this choice (Figure below).



(a) Elbow plot for Big Five personality traits

(b) Elbow plot for patient-perceived traits

Figure 8: Elbow method validation for optimal cluster selection. Both analyses show clear inflection points at k=4 (marked in red), with within-cluster sum of squares (inertia) showing diminishing returns beyond this point. (a) Clustering based on Big Five personality traits (n=92,534 physicians). (b) Validation using patient-perceived professional traits demonstrates consistent optimal k=4 solution.

The elbow method analysis provides clear evidence for the four-cluster solution, with both personality traits and patient-perceived professional traits independently converging on k=4 as optimal. The identified clusters (”Well-Rounded Excellent,” ”Underperforming,” ”Socially Driven,” and ”Steady Reliable”) represent distinct and coherent physician profiles that align with clinical intuition while being empirically grounded in the data.