Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Can Large Language Models Judge Stroke Information on Social Media? A Cross-Sectional Comparison of Large Language Models Against Human Ratings of Health-Related Videos

Ali Aamir
Liaquat National Hospital and Medical College, Karachi, Pakistan https://orcid.org/0000-0001-8775-7353

Zaid Choudhry
New Jersey Institute of Technology, USA

Zain Choudhry
Methacton High School, USA

Maha Tariq
Ziauddin Medical University, Pakistan

Uzair Yaqoob
ozair_91393@hotmail.com

Civil Hospital, SMBBIT, Karachi, Pakistan https://orcid.org/0000-0002-5910-2875

Zia Choudhry
Syneos Health (United States) https://orcid.org/0000-0002-4219-4777

Research Article

Keywords: stroke, social media, TikTok, Instagram, large language models, health information quality, DISCERN, artificial intelligence

Posted Date: October 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11300710/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Objective

To evaluate the extent to which large language models (LLMs) reproduce human assessments of the quality of stroke-related information on social media.

Background

Social media is a major source of health information, yet content quality varies. The reliability of LLMs in evaluating stroke-related social media content against human assessments remains uncertain.

Methods

This cross-sectional study evaluated 50 TikTok and Instagram videos using human reference scores and three AI models: ChatGPT, Microsoft 365 Copilot, and Claude. Publicly unavailable videos and those featuring songs or lacking transcribable speech were excluded. Spoken audio was transcribed using a free tool, and transcripts were submitted to all three models with an identical standardized prompt. Scores were compared using Mann–Whitney *U* tests; Spearman's correlation assessed associations with human ratings, and linear regression quantified explained variance.

Results

The median human reference score was 37.00 (IQR, 35.00–45.00). All AI models scored significantly lower: ChatGPT, 34.50 ($p = 0.001$); Microsoft Copilot, 33.50 ($p < 0.001$); and Claude, 26.00 ($p < 0.001$). All models correlated significantly with human ratings, with the strongest association for Claude ($\rho = 0.593$; $p < 0.001$), followed by ChatGPT ($\rho = 0.550$; $p < 0.001$) and Copilot ($\rho = 0.396$; $p = 0.004$). Claude explained 32.1% of the variance in human scores ($R^2 = 0.321$), compared with 31.4% for ChatGPT and 14.8% for Copilot. Inter-model correlation was strongest between ChatGPT and Copilot ($\rho = 0.862$; $p < 0.001$).

Conclusions

All three LLMs showed significant positive correlations with human ratings but systematically assigned lower scores. Claude showed the strongest correlation, whereas Microsoft Copilot showed the weakest. These findings highlight the limitations of LLMs in reproducing human judgments and the need for further validation before their use as standalone evaluators of social media content.

INTRODUCTION

Stroke remains one of the leading causes of death and long-term disability worldwide, and its absolute burden continues to grow as populations age [1, 2]. Outcomes depend heavily on how quickly symptoms are recognized and treatment is sought, because the loss of neural tissue accumulates with every minute of delay [3]. Public knowledge of warning signs, risk factors, and appropriate action therefore matters for stroke outcomes, and the sources from which the public learns about stroke also matter.

Social media platforms have become a routine channel for health communication, offering reach and accessibility that traditional patient education cannot match [4]. Short-form video platforms such as TikTok and Instagram are especially influential because content is brief, visual, and algorithmically distributed with little editorial oversight. The same features that make these platforms effective also allow inaccurate, incomplete, or commercially motivated content to spread widely [5, 6]. Studies of stroke-related videos on YouTube, TikTok, and Bilibili have repeatedly found variable and often suboptimal quality, including limited reliability and few actionable recommendations [7–9].

The quality of online health information is commonly appraised with structured instruments, most notably DISCERN, which was developed to judge written consumer health information on treatment choices [10]. Applying such instruments to video content requires trained reviewers, is time-consuming, and is difficult to scale to the volume of content uploaded each day [11]. Automated approaches that could screen content at scale would be valuable to researchers, platforms, and clinicians who wish to direct patients toward reliable material.

Large language models (LLMs) are a natural candidate. They have shown strong performance on medical knowledge tasks and on patient-facing communication [12–14], and they can follow structured scoring instructions when given a transcript. However, LLMs are also prone to hallucination and inconsistent outputs [15], and using them as evaluators (“LLM-as-a-judge”) depends on how closely their judgments track human experts [16]. Early work applying LLMs to appraise health videos has reported variable agreement with expert raters. In one study of YouTube medical videos, agreement varied widely across models, and models tended to award higher scores than experts [17]. In another study, ChatGPT-4 ratings of tonsillectomy videos correlated with human ratings to varying degrees, depending on the dimension assessed [18].

To our knowledge, little is known about how closely commercially available LLMs reproduce human appraisals of stroke-related content on short-form social media platforms, or how different models compare when given identical inputs. We therefore aimed to evaluate how closely three widely used LLMs (ChatGPT, Microsoft 365 Copilot, and Claude) reproduce human assessments of the quality of stroke-related information in TikTok and Instagram videos, in terms of absolute score level, rank-order association, and explained variance.

METHODS

This cross-sectional study analyzed publicly available social media videos. The study is reported in accordance with the applicable items of the STROBE statement for cross-sectional studies [19].

Formal ethical approval was not obtained for this study, which involved the analysis of publicly available online content. No human participants were recruited, no direct interaction with content creators took place, and no private or personally identifiable information was collected. The analysis was restricted to material accessible to the general public, without obtaining or examining additional private or sensitive information. Ethical considerations relevant to the use of publicly available online content were taken into account throughout the study.

The search was performed using a comprehensive set of stroke-related hashtags, including #stroke, #strokeawareness, #strokeprevention, #strokesymptoms, and #strokerecovery. We used Instagram and TikTok. Videos were excluded if they were publicly unavailable, featured songs, or lacked transcribable speech. The final sample comprised 50 videos.

Each video was assigned a human reference score for the quality of its stroke-related information using the DISCERN instrument [10]. Scoring was performed by two qualified professionals who worked independently and resolved disagreements by consensus.

Spoken audio from each included video was transcribed using Clipto.AI. We manually checked or corrected the transcripts before submitting them to the models.

Three LLMs were evaluated: ChatGPT (GPT version 5.5), Microsoft 365 Copilot (based on GPT version 5.0), and Claude (Sonnet 5.5). Each transcript was submitted to each model with an identical standardized prompt. We provided each LLM with the scoring instrument beforehand. Each video was assessed in the same session, and each video was scored once. Model outputs were converted to total scores on the same scale as the human reference scores.

The primary outcome was the difference in score level between each LLM and the human reference. Secondary outcomes were the rank-order association between each LLM and the human reference score, the proportion of variance in human scores explained by each LLM, and the association between LLMs.

Scores are summarized as medians with interquartile ranges (IQR). Scores from each LLM were compared with the human reference scores using the Mann–Whitney U test. Associations between each LLM and the human reference, and between pairs of LLMs, were assessed using two-tailed Spearman's rank correlation coefficients. We used simple linear regression, with the human reference score as the dependent variable and each LLM's score as the independent variable, to quantify the proportion of variance explained (R^2). All tests were two-sided, and we considered $p < 0.05$ statistically significant. No correction for multiple comparisons was applied. Analyses were performed using IBM SPSS Statistics, Version 25.

RESULTS

A total of 50 videos were analyzed. The mean video duration was 1.12 ± 1,08 minutes.

The median human reference score was 37.00 (IQR 35.00–45.00). All three LLMs assigned lower scores than the human reference (Table 1). Median scores were 34.50 (IQR 30.00–41.00) for ChatGPT, 33.50 (IQR 29.75–40.00) for Microsoft Copilot, and 26.00 (IQR 23.75–29.75) for Claude. Mann–Whitney *U* tests showed that each model's scores differed significantly from the human reference (ChatGPT, $p = 0.001$; Microsoft Copilot, $p < 0.001$; Claude, $p < 0.001$) (Table 1). Claude's median was 11 points below the human median, compared with 2.5 points for ChatGPT and 3.5 points for Microsoft Copilot.

Table 1
Descriptive statistics and comparison of LLM scores with human reference scores (N = 50).

Evaluator	Median (IQR)	P-value vs. human reference
Human reference	37.00 (35.00–45.00)	—
ChatGPT	34.50 (30.00–41.00)	0.001
Microsoft Copilot	33.50 (29.75–40.00)	< 0.001
Claude	26.00 (23.75–29.75)	< 0.001

IQR, interquartile range; LLM, large language model; SD, standard deviation. P values are from Mann–Whitney *U* tests.

All three LLMs correlated positively and significantly with the human reference scores (Table 2). Claude showed the strongest correlation ($\rho = 0.593$, $p < 0.001$), followed by ChatGPT ($\rho = 0.550$, $p < 0.001$) and Microsoft Copilot ($\rho = 0.396$, $p = 0.004$). These values are moderate, with Microsoft Copilot at the boundary between weak and moderate [20].

The LLMs correlated strongly with one another. The strongest correlation was between Microsoft Copilot and ChatGPT ($\rho = 0.862$, $p < 0.001$), followed by ChatGPT and Claude ($\rho = 0.789$, $p < 0.001$).

Table 2
Spearman rank correlation coefficients between evaluators (N = 50).

Evaluator	Human reference	ChatGPT	Claude	Microsoft Copilot
Human reference	1.000	—	—	—
ChatGPT	0.550*	1.000	—	—
Claude	0.593*	0.789*	1.000	—
Microsoft Copilot	0.396*	0.862*	0.671*	1.000

*Statistically significant (two-tailed).

In a linear regression with the human reference score as the outcome, Claude explained the largest proportion of variance ($R^2 = 0.321$; 32.1%), followed closely by ChatGPT ($R^2 = 0.314$; 31.4%). Microsoft

Copilot explained 14.8% (R^2 = 0.148).

DISCUSSION

In this cross-sectional comparison of 50 stroke-related TikTok and Instagram videos, three widely used LLMs followed a similar pattern: they ranked videos broadly in the same direction as human raters. Still, they scored lower and tracked human scores only moderately. The correlations ranged from ρ = 0.396 to 0.593, and no model explained more than about one third of the variance in human scores. The LLMs therefore captured part of human judgment, but not most of it.

Score level and rank order are separate rater properties, and the models differed on each. Claude had the largest downward offset (11 points below the human median) yet the highest correlation. Its explained variance was almost identical to ChatGPT's (32.1% vs 31.4%), and we did not formally test differences between correlations in a sample of 50, so these data do not support ranking one model above the other. Microsoft Copilot performed worst on both rank association and explained variance.

The lower scores contrast with a study of YouTube medical videos, in which most LLMs scored higher than expert raters [17], and with a study of tonsillectomy videos, in which agreement with human ratings differed by the dimension assessed [18]. Those studies used different instruments, prompts, models, and content, so direct comparison is limited. Several explanations for our finding are plausible but were not tested. The models saw only transcripts, so they did not have access to on-screen text, captions, or visual cues available to human raters. Items on sources, balance, and uncertainty may be scored strictly when they are not stated aloud, and a one-minute video leaves little room for them. Single-run LLM output can also vary and may contain errors [15], adding noise and lowering correlations.

The strong correlations between models, especially ChatGPT and Microsoft Copilot (ρ = 0.862), suggest shared scoring tendencies, possibly because they share underlying models. Agreement between models should not be read as evidence that their scores are correct, since LLM judges share known biases [16] and may share blind spots.

These findings do not support using any of the three models as a standalone evaluator of stroke-related social media content. A consistent downward offset means raw LLM scores cannot be compared with human benchmarks or applied against fixed quality thresholds without calibration. A moderate correlation still leaves substantial room for misclassifying individual videos. For example, when LLMs are used to prioritize videos for review in large-scale studies, human verification should remain part of the process, and results should be validated against human ratings for the specific task, model version, and prompt.

Strengths include using an identical standardized prompt across three models and including both score-level and rank-order analyses. The study also has limitations. The sample was small and limited to two platforms. Only videos with transcribable speech were included, so the findings may not apply to visually driven or music-based content. We used a single prompt and a single scoring run per model, and model

versions change frequently. The human reference scores are an imperfect standard [raters and reliability to be added]. We assessed score-level and rank associations rather than formal agreement statistics such as Bland–Altman limits of agreement [21] or intraclass correlation coefficients [22], and [state whether multiple-comparison correction was applied]. Finally, the cross-sectional design captures one point in time.

Future work should use larger, more varied samples; multimodal input; repeated runs to measure within-model consistency; and formal agreement analyses.

CONCLUSIONS

Three commonly used LLMs showed moderate positive correlations with human ratings of stroke-related TikTok and Instagram videos but systematically assigned lower scores. Their ability to reproduce human judgment was limited, and they need further validation before use as standalone evaluators of social media health content.

References

1. GBD 2021 Stroke Risk Factor Collaborators (2024) Global, regional, and national burden of stroke and its risk factors, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021. Lancet Neurol 23(10):973–1003. 10.1016/S1474-4422(24)00369-7
2. Hilkens NA, Casolla B, Leung TW, de Leeuw FE (2024) Stroke Lancet 403(10446):2820–2836. 10.1016/S0140-6736(24)00642-1
3. Saver JL (2006) Time is brain—quantified. Stroke 37(1):263–266. 10.1161/01.STR.0000196957.55928.ab
4. Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C (2013) A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. J Med Internet Res 15(4):e85. 10.2196/jmir.1933
5. Suarez-Lledo V, Alvarez-Galvez J (2021) Prevalence of health misinformation on social media: systematic review. J Med Internet Res 23(1):e17187. 10.2196/17187
6. Swire-Thompson B, Lazer D (2020) Public health and online misinformation: challenges and recommendations. Annu Rev Public Health 41:433–451. 10.1146/annurev-publhealth-040119-094127
7. Szmuda T, Alkhater A, Albrahim M et al (2020) YouTube as a source of patient information for stroke: a content-quality and an audience engagement analysis. J Stroke Cerebrovasc Dis 29(9):105065. 10.1016/j.jstrokecerebrovasdis.2020.105065
8. Özdemir Z, Acar E (2024) YouTube as a source of recognizing acute stroke; progress in 2 years. BMC Public Health 24(1):2208. 10.1186/s12889-024-19710-4

9. Nie SJ, Ning SN, Ding QC, Hou J, Luo YY, Wang HY (2025) Quality analysis of stroke-related videos on video platforms: cross-sectional study. JMIR Form Res 9:e80458. 10.2196/80458
10. Charnock D, Shepperd S, Needham G, Gann R (1999) DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. J Epidemiol Community Health 53(2):105–111. 10.1136/jech.53.2.105
11. Osman W, Mohamed F, Elhassan M, Shoufan A (2022) Is YouTube a reliable source of health-related information? A systematic review. BMC Med Educ 22(1):382. 10.1186/s12909-022-03446-z
12. Singhal K, Azizi S, Tu T et al (2023) Large language models encode clinical knowledge. Nature 620(7972):172–180. 10.1038/s41586-023-06291-2
13. Ayers JW, Poliak A, Dredze M et al (2023) Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 183(6):589–596. 10.1001/jamainternmed.2023.1838
14. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW (2023) Large language models in medicine. Nat Med 29(8):1930–1940. 10.1038/s41591-023-02448-8
15. Ji Z, Lee N, Frieske R et al (2023) Survey of hallucination in natural language generation. ACM Comput Surv 55(12):Article248. 10.1145/3571730
16. Zheng L, Chiang WL, Sheng Y et al (2023) Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. arXiv 2306:05685. 10.48550/arXiv.2306.05685
17. Khalil M, Mohamed F, Shoufan A (2025) Evaluating the quality of medical content on YouTube using large language models. Sci Rep
18. Şerifler S, Gül F (2025) Evaluating tonsillectomy-related YouTube videos via a human expert review and the ChatGPT-4: a multi-method quality analysis. BMC Med Educ 25(1). 10.1186/s12909-025-07739-x
19. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP (2007) The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet 370(9596):1453–1457. 10.1016/S0140-6736(07)61602-X
20. Schober P, Boer C, Schwarte LA (2018) Correlation coefficients: appropriate use and interpretation. Anesth Analg 126(5):1763–1768. 10.1213/ANE.0000000000002864
21. Bland JM, Altman DG (1986) Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1(8476):307–310. 10.1016/S0140-6736(86)90837-8
22. Koo TK, Li MY (2016) A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 15(2):155–163. 10.1016/j.jcm.2016.02.012

Figures

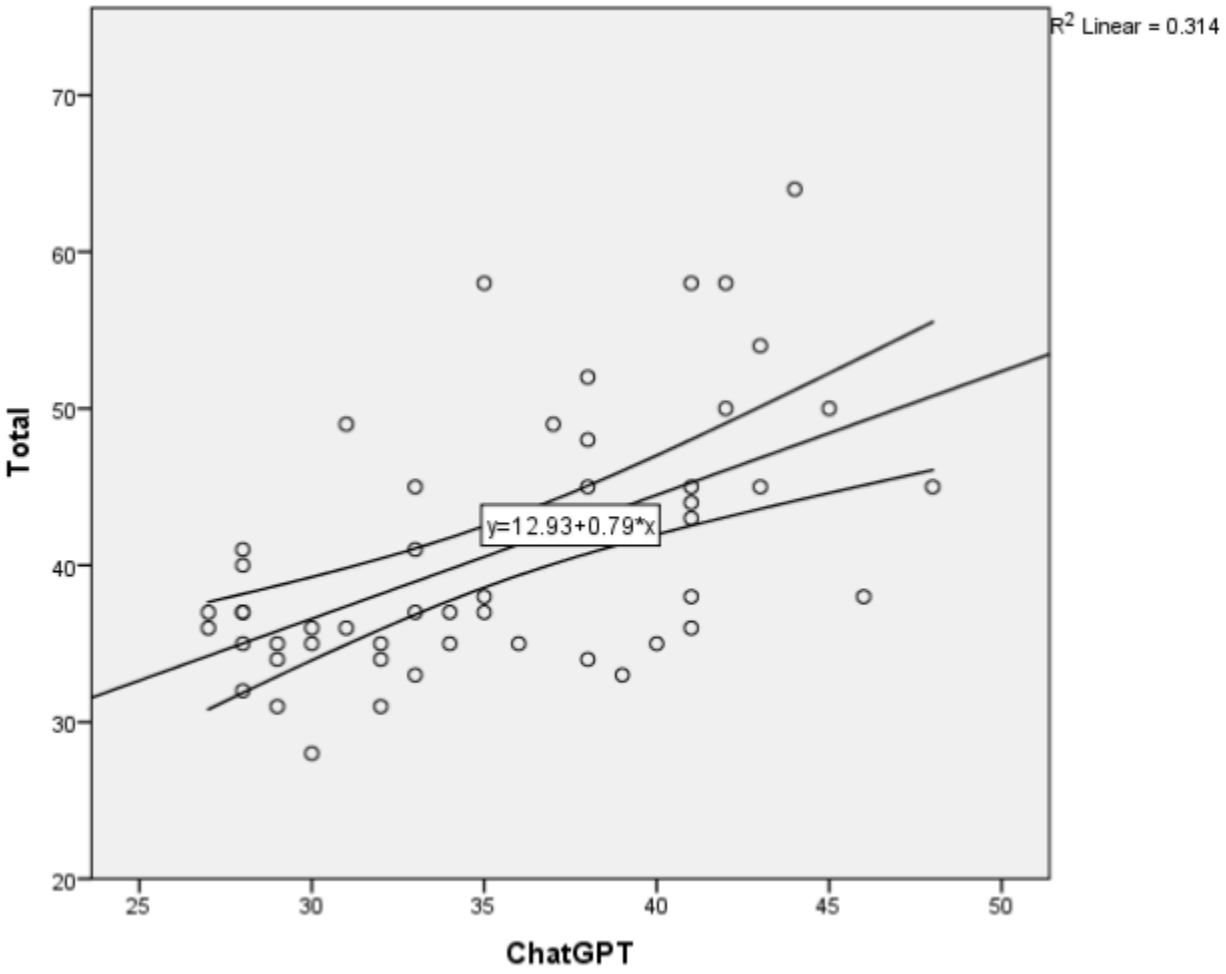


Figure 1

ChatGPT Linear model with human scoring vectors

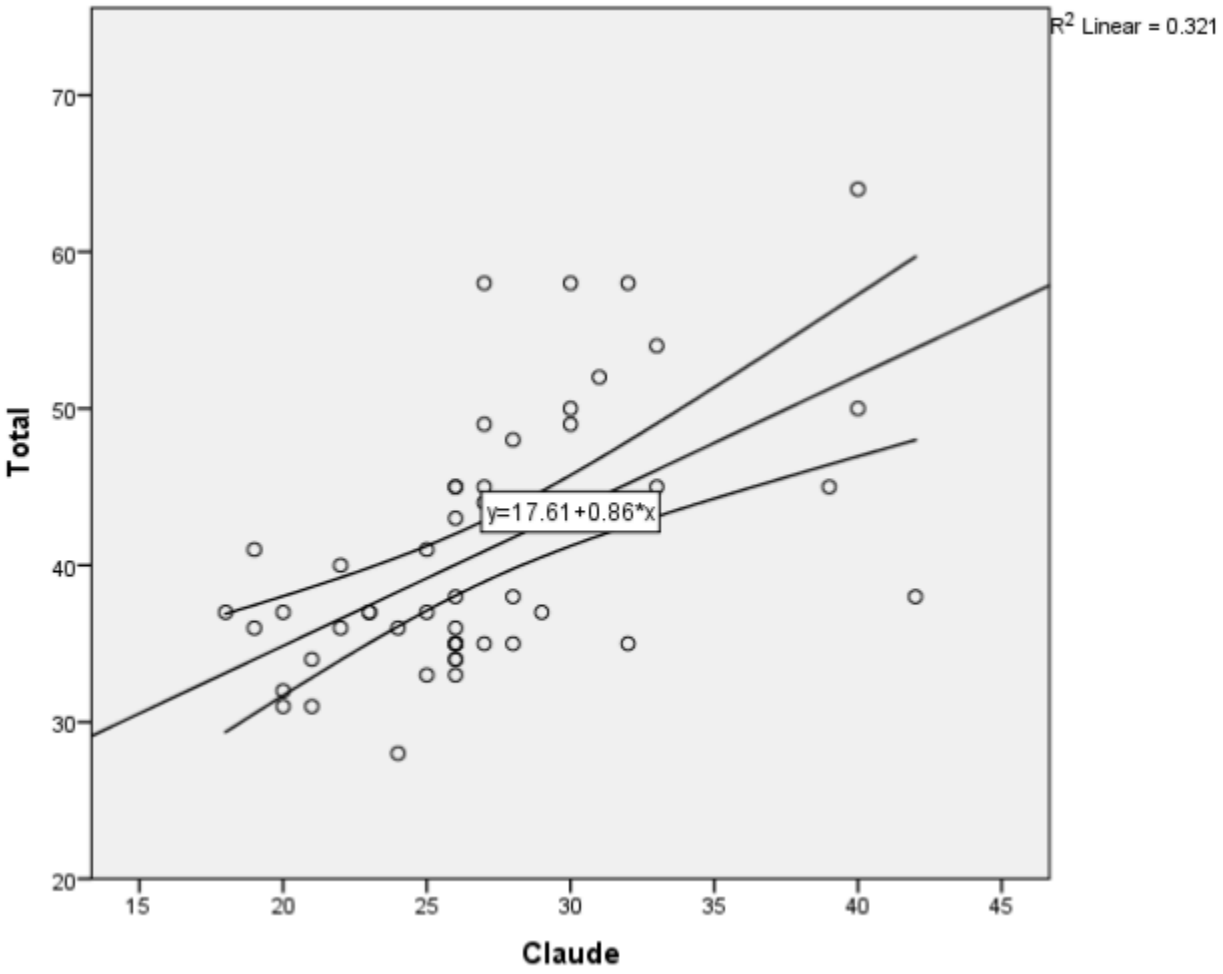


Figure 2

Claude Linear model with human scoring vectors

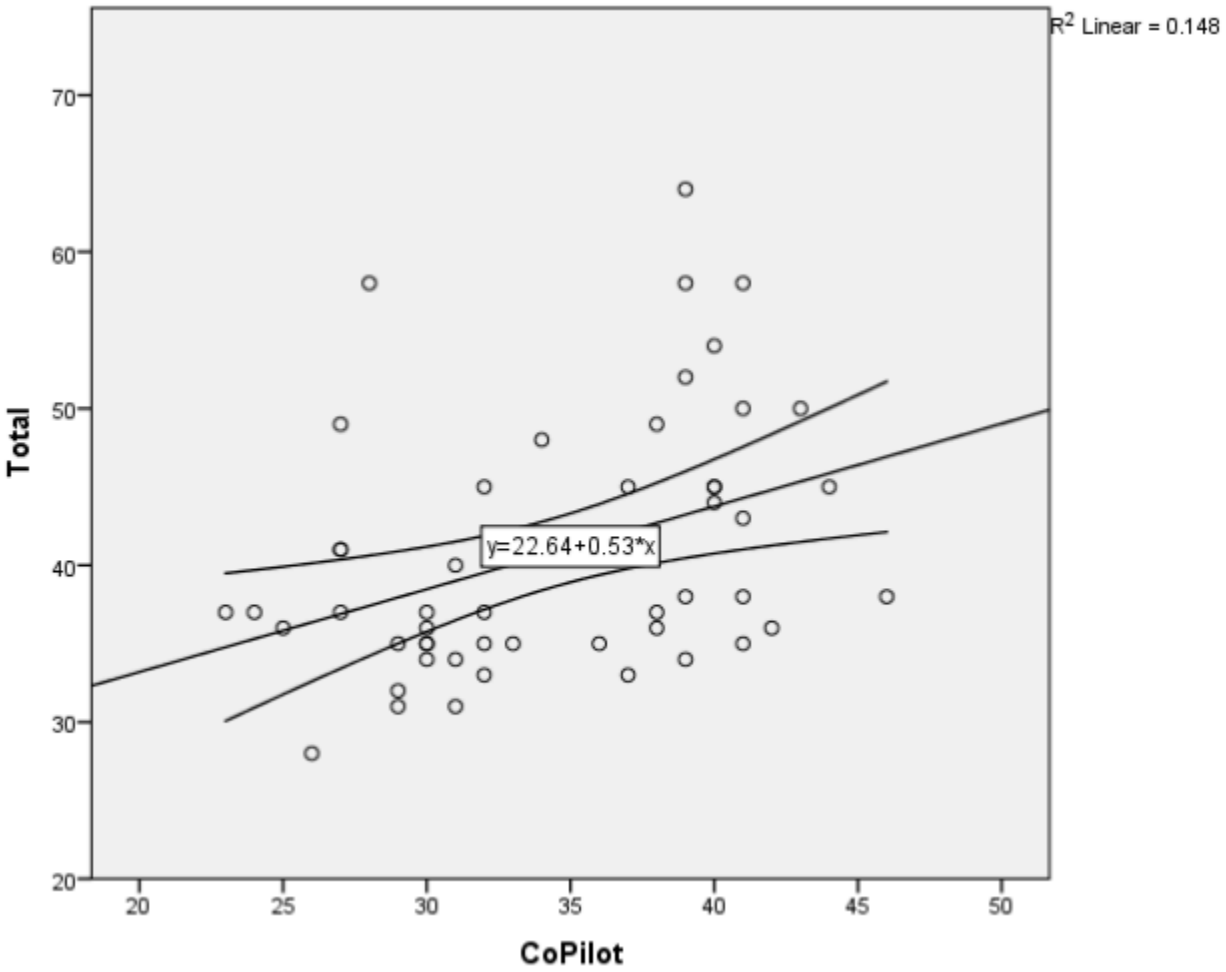


Figure 3

CoPilot Linear model with human scoring vectors