

Evaluating Agreement Between Human Consensus and Large Language Model Ratings of Suicide Reporting

Andrew Sarkin

asarkin@health.ucsd.edu

University of California, San Diego

Bilal Malik

University of California, San Diego

Elizabeth Doery

University of California, San Diego

Research Article

Keywords:

Posted Date: August 12th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9904774/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Responsible reporting on suicide is an important public health strategy but monitoring whether news articles follow reporting recommendations is difficult at scale. While the Tool for Evaluating Media Portrayals of Suicide (TEMPOS) provides a structured approach of rating suicide-related media, human rating requires time, training, and repeated exposure to distressing content. Acknowledging these challenges, our research examines whether large language models (LLMs) can rate incident-focused suicide news articles using TEMPOS with agreement comparable to human raters. We conducted a reliability and agreement study using 42 news articles describing suicide deaths. Seven trained human raters and five LLMs independently rated each article using TEMPOS. Human interrater reliability was assessed using intraclass correlation coefficients (ICCs). A human consensus benchmark, HumanGold, was calculated as the mean TEMPOS score across human raters for each article. Agreement between individual AI models and HumanGold was examined to develop an aggregated AI score, termed AISilver. Individual AI models demonstrated moderate to strong correlations with HumanGold, ranging from $r = 0.489$ to $r = 0.812$. AISilver also showed strong agreement with HumanGold ($r = 0.797$), comparable to the highest performing model, Grok ($r = 0.812$). Overall, we found that LLMs showed alignment with trained human raters when applying TEMPOS. These findings suggest that AI may be a plausible tool to assist with continuous, large-scale monitoring of suicide-incident reporting, particularly if paired with human oversight. While AI ratings should be interpreted as consensus-aligned screening tools rather than definitive judgments, they may offer a practical path for expanding feedback on suicide-incident reporting.

Introduction

Responsible media reporting on suicide is a public health issue, and the way suicide is portrayed can have a significant impact on suicidal behavior. A large body of evidence, including ecological studies and systematic reviews, supports the existence of harmful effects following certain types of suicide coverage (often termed the Werther Effect) as well as potential protective effects when stories emphasize mastery of crisis, help-seeking, and recovery (often termed the Papageno Effect). For example, Cheng and colleagues [1] explored the impact of media reports on suicidal behavior among suicide survivors. The authors found that there was a negative influence of such media reporting, and that the influence was greatest among people experiencing severe depression. More recently, Kim and colleagues [2] explored suicide risk following reports across three Korean newspapers of the suicide of an entertainer and politician. This research revealed that an increase in suicide incidents following the death of both the entertainer and the politician, and that there was substantial media coverage on both deaths. Both these studies provide evidence for the Werther Effect.

There are also published reviews that explore the impact of reporting on suicide more generally, with this research mostly reporting that there are both positive and negative impacts of media reporting on suicide. For example, Domaradzki [3] reviewed literature on both the Werther Effect and Papageno Effect and suggested that the quality and quantity of media reporting can have an impact on subsequent

suicides, with the reporting of celebrity suicides having the most significant impact. While noting the negative impacts of suicide reporting, Domaradzki [3] also acknowledged that when media is framed through an education or prevention lens it can instead have a positive effect. Similar conclusions were made by researchers Calvo and colleagues [4] who conducted a literature review on 25 articles that explored the Werther Effect in response to media stories on suicide, reporting that suicide content can be both protective and harmful. This demonstrates the complexity of reporting on suicide and the need for this reporting to be done cautiously and with great consideration of the potential negative and positive consequences.

To reduce preventable harm, multiple organizations have issued media recommendations for reporting on suicide, including the World Health Organization [5]. These guidelines emphasize the importance of media articles containing minimal details on suicide method and of avoiding sensationalizing language, while also stressing the importance of increasing contextual information and including prevention resources. Other recommendations include the Best Practices and Recommendations for Reporting on Suicide, which were developed in collaboration with several suicide prevention experts and public health organizations from across the world [6].

Despite widespread guidance, monitoring real-world adherence at scale remains challenging. Manual content coding is time-intensive, requires training, and can be difficult to sustain for large volumes of incident-focused coverage (e.g., local news and repeated events). The Tool for Evaluating Media Portrayals of Suicide (TEMPOS) was developed through a collaboration with the County of Santa Clara Behavioral Health Services Department's Suicide Prevention Program and the Stanford Department of Psychiatry and Behavioral Sciences' Media and Mental Health Initiative [7]. The aim of the TEMPOS tool is to address a key gap by providing a standardized, user-friendly rating scale that operationalizes major reporting recommendations into scorable criteria. Through their research to develop the TEMPOS tool, Sorensen and colleagues [7] found that while there are clear negative consequences of reporting on suicide in the media, media can also play a role in suicide prevention. In doing so, Sorensen and colleagues [7] highlight the importance of communication around suicide and the need to report on and discuss suicide cautiously, and with consideration of the potential consequences. The TEMPOS tool aims to consider these nuances and complexities, and to help guide journalists through the processes of reporting on suicide.

Recent advances in large language models (LLMs) raise the possibility that Artificial Intelligence (AI) could help evaluate media content at a scale that is infeasible for humans. For example, Elyoseph and colleagues [8] found strong agreement between human reviewers and LLM (ChatGPT-4) in assessing adherence to the WHO reporting on suicide guidelines, supporting the plausibility of scalable, automated feedback, and surveillance approaches.

The overall objective of our research is to evaluate whether LLM "AI raters" can reliably assess incident-focused suicide news articles using the TEMPOS framework, achieving agreement comparable to trained human raters. This study also explores the feasibility of using AI both individually and in

aggregated panels, for scalable, real-world monitoring of suicide reporting, thereby contributing to the emerging evidence base on LLM-assisted public health evaluation. To achieve these objectives, we set out four key aims. We aim to:

1. Establish a human consensus benchmark for TEMPOS ratings by quantifying interrater reliability among trained human raters on TEMPOS total and item-level scores, and to define a human consensus benchmark (termed HumanGold) for each article.
2. Evaluate agreement between each AI model and HumanGold on TEMPOS total and item-level scores by comparing multiple AI raters against HumanGold, and to identify differences in performance across models.
3. Determine whether combining the top-performing AI raters into an aggregated panel score (termed AISilver) improves performance and alignment with HumanGold relative to individual AI models.
4. Examine whether disagreement between human and AI raters is concentrated in more interpretive TEMPOS domains, such as glamorization, speculation, risk-factor framing, and visual content, versus more concrete domains such as method detail, suicide note description, and inclusion of prevention resources.

To address these aims, we also set out four hypotheses. We hypothesized that:

1. Human panel scores will show substantially higher reliability than single-rater scores, supporting use of HumanGold as the primary benchmark.
2. At least one AI model will demonstrate agreement with HumanGold that is comparable to the range of agreement observed for individual human raters.
3. AI models will differ meaningfully in their alignment with HumanGold, indicating that model choice matters, and AISilver will show modestly stronger agreement with HumanGold than any single AI model alone
4. Agreement will be highest on more concrete, directly observable TEMPOS items and lower on more interpretive items.

Methods

Design

We conducted a reliability and agreement study comparing:

1. Agreement among human raters using TEMPOS;
2. Agreement among multiple AI raters using TEMPOS; and
3. Agreement between AI ratings and an averaged human benchmark, both individually and as an averaged benchmark of multiple AI models rating the same articles.

Article source and retrieval

Articles were drawn from publicly accessible media sources (online news articles). Inclusion criteria for articles were that they were published in the United States, were in English language and were news stories describing a specific incident of a death by suicide (incident-focused reporting rather than prevention content or general reporting on suicide rates), and freely accessible (e.g., not behind a paywall) for both human and AI raters. Exclusion criteria were being duplicates or near-duplicates of another article in the dataset or being opinion pieces that reported content or stories primarily about prevention programs rather than on a specific suicide incident.

Rating instrument: TEMPOS

The TEMPOS instrument is a rating tool that was designed by leading experts in the field of suicide prevention to be used by journalists, suicide prevention experts, and researchers to encourage and monitor adherence to recommendations around reporting on suicide [7]. Journalists can apply the TEMPOS criteria to evaluate existing articles and social media content and inform the development of new media outputs. TEMPOS operationalizes suicide reporting guidance into 10 scorable domains (see Table 1), each rated on an ordinal scale (implemented as 0–2 with higher scores indicating better adherence). Domain-level scores are summed to create a TEMPOS total score from 0 to 20. By using a structured criteria-based approach, TEMPOS supports more consistent application across rater types and contexts and has been applied in large-scale evaluations of news coverage [7].

Raters were provided with a standardized rubric (adapted from the TEMPOS scoring tool) [7] that defined anchors for each rating level. Human and AI raters were instructed to base each rating on the article's text and associated media content (e.g., images), and to document supporting evidence (e.g., short excerpts) to improve transparency and adjudication.

Table 1
TEMPOS Items and Descriptions

TEMPOS Item	Description
T1	How does the report frame suicide?
T2	Does the report include factual or speculative language?
T3	Does the report use appropriate/non-stigmatizing language?
T4	How does the report describe the suicide method and scene?
T5	How does the report describe the suicide note
T6	What visual content does the report include?
T7	How does the report describe risk factors and reasons for suicide?
T8	Does the report use sensational language?
T9	Does the report glamorize suicide
T10	Suicide prevention and mental health resources?

Abbreviations: TEMPOS, Tool for Evaluating Media Portrayals of Suicide. Item descriptions correspond to the standardized TEMPOS rubric domains

Human raters

Eight human raters began the study, however one rater withdrew prior to completion and was excluded from analyses, yielding a final sample of seven human raters. All human raters were trained on the TEMPOS rubric using (a) a brief orientation that included reading the instructions from the TEMPOS authors, (b) practicing on a set of example articles, and (c) discussing ambiguous cases from the example articles. Training emphasized applying TEMPOS item definitions consistently and using examples from the rubric to limit individual interpretation of the items as much as possible. Raters were supported throughout the process but ratings of specific articles were not discussed to preserve independence, and raters were discouraged from sharing their ratings with each other. Raters were regularly given the option to opt out if they found the materials too emotionally burdensome.

AI raters (LLMs) and prompting protocol

AI raters (LLMs) and prompting protocol

Multiple LLMs (N = 5) were evaluated as independent “AI raters.” The included AI models were ChatGPT 5.3 (OpenAI), Claude Opus 4.7 (Anthropic), Grok 4.3 (xAI), Gemini 3.1 Pro (Google), and DeepSeek-V3 (Hangzhou DeepSeek Artificial Intelligence Co., Ltd.).

Each model received a standardized set of instructions to read the full article text and assign a TEMPOS score for each of the 10 domains, while including brief justification and quotes as supporting evidence for each score (see Appendix A for the complete and final AI query). Models were run as separate raters, and each article query was conducted in a separate chat to avoid contamination.

Both human and AI raters received batches of five to eight articles with the first batch received on October 31, 2025. Batches of articles were provided at approximately two-week intervals to allow raters sufficient time to carefully rate each article and to reduce emotional burden. Raters received the last batch of articles on March 11, 2026. Articles were processed through AI during two phases. The first occurred at the same time intervals as human raters to provide insight into AI performance relative to human raters, and the second occurred on April 28, 2026, to use the most recent version of each model. Article ratings produced in the final run (April 28, 2026) were used in the analyses as they represented outputs from the most advanced versions available at the time. To reduce unsupported ratings and enable auditing, the AI prompt requested each model to provide a brief justification and to quote short supporting evidence (when available) for each domain. Models were run as separate raters with no cross-model communication.

Outputs were checked for completeness (all 10 domains present, valid response range per item). If the model produced a malformed response (missing items/out-of-range values), the run was repeated once using a standardized “format correction” instruction without changing substantive guidance. Ratings were stored in a long format dataset with one row per article per rater and columns for TEMPOS domain ratings, total score, and optional comments/evidence.

Statistical analysis

Interrater reliability framework

To quantify interrater agreement, we used intraclass correlation coefficients (ICCs), treating AI outputs as additional raters. This approach, which follows established guidance for interrater reliability studies, is useful descriptively but is interpreted cautiously, as it combines different types of raters. For ICC analyses that required a fully crossed rater-by-article matrix, data were reshaped to wide format. When designs were incomplete (e.g., if not all raters rated all articles), analyses used complete-case subsets or models appropriate for missingness. A two-way random-effects model with absolute agreement was applied, consistent with a design in which multiple raters evaluate the same targets. We report two forms of ICC aligned with the intended use of the ratings:

1. **ICC(2,1)**: two-way random effects, absolute agreement, single-rater reliability
2. **ICC(2,k)**: two-way random effects, absolute agreement, average reliability across all raters

We reported ICCs with confidence intervals because very small early batches could yield unstable ICC estimates (including low/negative ICCs when item variance is restricted). Therefore, early results were treated as indicators of feasibility and workflow checks rather than definitive performance claims.

Planned reliability comparisons

We conducted influence analyses using a Leave-One-Rater-Out (LORO) approach, in which ICCs were recalculated iteratively after removing each rater (both human and AI). This allowed us to identify systematic outliers and assess the sensitivity of results to any single rater or model. Further, to evaluate AI-to-human agreement, we defined a human benchmark as the mean total score across human raters for each article. Agreement between each AI rater and this benchmark was then assessed using ICCs.

HumanGold Standard

Individual ratings can be subjective due to differences in interpretation when evaluating the same content. While some TEMPOS items assess more concrete and explicit content found in articles (e.g., TEMPOS 10, suicide prevention resources), other items are more context-driven (e.g., TEMPOS 9, glamorize suicide), which can lead raters to assess content and grade articles more subjectively. To account for this, a HumanGold standard variable was calculated as the mean TEMPOS score across all human raters for each article, which reduced variability caused by individual interpretation and provided a more comprehensive assessment of article quality. Given this variable captures the central tendency across human raters and provides a stable human consensus benchmark, it is the most appropriate variable to use in comparison to AI raters.

AI Silver

Similar to human raters, individual LLMs may differ in how they interpret TEMPOS items. Reliance on a single model can therefore introduce model-specific bias or instability, especially because LLMs are updated over time. To examine this issue, we analyzed two sets of AI ratings. The first set was collected during the same two-week intervals as the human ratings; therefore, the specific model versions available at each rating point may have varied across the study period. The second set was collected concurrently on April 28, 2026, using the most current version of each model available at that time.

For each AI rating set, we created an aggregated AI score, termed AISilver. AISilver was calculated by first identifying the three AI models with the strongest overall performance, based on correlation with HumanGold and results from the LORO analysis. Models were considered stronger candidates if they showed high agreement with HumanGold and did not appear to reduce overall reliability when included in the rater set. For the initial rating set, AISilver included ChatGPT, Claude, and Gemini. For the updated rating set, AISilver included ChatGPT, Grok, and Claude. AISilver was calculated in the same manner as HumanGold, by averaging the TEMPOS total scores across the selected AI raters for each article.

Sample size planning

Given that the main value of additional articles is improved precision (narrower ICC confidence intervals), we planned the sample size of articles to be reviewed using established ICC sample size methods. This approach was informed by existing research. For example, Walter and colleagues [9] reported that for moderate-to-high reliabilities, a minimum of two to three observations per subject is optimal. Further,

Mondal and colleagues [10] emphasized the importance of sample size planning for achieving desired confidence interval widths for ICC estimates and balanced designs with equal ratings per participant. Based on this evidence, we used a sequential batched approach, with data collection conducted in batches of five to eight articles. After each batch, we analyzed human and AI performance, evaluating the following:

1. Whether AI outputs remain well-formed and auditable (evidence quotes present; no systematic rubric drift) to ensure that our query was being followed precisely.
2. Whether agreement metrics were within a pre-specified band (e.g., point estimate near target and confidence intervals (CIs) not wildly inconsistent).
3. Whether there was a clear human rater or LLM that was rating articles significantly different than the remaining cohort.
4. Whether a clearly less capable model emerged requiring exclusion or separate reporting because of poor agreement with other raters.

Results

Across 42 articles and 12 raters (7 human and 5 AI), the dataset included 504 article-level ratings and 5,040 item-level TEMPOS scores. Missing responses were examined at the item level with blank responses reviewed and recoded where appropriate. For example, for TEMPOS 6 (visual content) when an article contained no image and the corresponding item was left blank by a rater, the response was recoded as “2” in accordance with our interpretation of the TEMPOS scoring criteria. This recoding was limited to cases where the absence of an image could be objectively verified. All other missing data were handled using standard missing data procedures (as outlined above) with no item-level ratings left as missing in the final analytic dataset.

Interrater Reliability

Interrater agreement was quantified using ICCs. Single-rater reliability was moderate, $ICC(2,1) = 0.532$ (95% CI: 0.414–0.662). In contrast, averaged-rater reliability was excellent, $ICC(2,k) = 0.932$ (95% CI: 0.895–0.959), with $k = 12$ raters. The substantial increase from single-rater to averaged-rater reliability demonstrates the value of pooled rater estimates.

LORO Analysis

LORO analysis demonstrated that interrater reliability was stable across all iterations (see Table 2). When individual raters were removed, minimal fluctuations in $ICC(2,k)$ values occurred, indicating that no single rater disproportionately impacted overall agreement. Removal of individual AI raters produced reliability estimates comparable to those observed when human raters were excluded, demonstrating consistency in scoring across rater types. In LORO rows, k represents the number of raters remaining after one rater is removed. In the row labeled “None,” no raters have been removed to provide a benchmark ICC value for all raters.

Table 2
Interrater Reliability Scores After LORO Analysis (Human, AI, and Combined)

Rater Dropped	Human ICC		AI ICC		Combined ICC	
	(2,1)	(2,6)	(2,1)	(2,4)	(2,1)	(2,11)
None	0.564	0.901	0.686	0.916	0.532	0.932
A	0.550	0.880	-	-	0.541	0.928
B	0.568	0.887	-	-	0.521	0.923
C	0.556	0.883	-	-	0.545	0.930
D	0.567	0.887	-	-	0.536	0.927
E	0.559	0.884	-	-	0.533	0.926
F	0.545	0.878	-	-	0.525	0.924
G	0.606	0.902	-	-	0.537	0.927
ChatGPT	-	-	0.626	0.870	0.511	0.920
Grok	-	-	0.679	0.894	0.528	0.925
Claude	-	-	0.667	0.889	0.523	0.923
DeepSeek	-	-	0.801	0.942	0.560	0.933
Gemini	-	-	0.665	0.888	0.527	0.925

Abbreviations: ICC, intraclass correlation coefficient; LORO, leave-one-rater-out; AI, artificial intelligence. Human ICC values are based on the remaining six raters after exclusion of the specified rater; AI ICC values are based on the remaining four AI raters after exclusion of the specified model. Combined ICC values are based on the remaining eleven raters after exclusion of the specified rater or model. ICC(2,1) represents single-measure absolute-agreement reliability and ICC(2,k) represents average-measure absolute-agreement reliability.

HumanGold and AISilver

HumanGold correlation analysis (see Table 3) demonstrated that all individual AI models showed moderate to strong alignment with the aggregated human score. Based on the results of the initial AI ratings, we created AISilver, which was a combination of three models (ChatGPT, Claude, and Gemini) from the LORO and HumanGold analysis. The AISilver score showed stronger correlation ($r = 0.809$, 95% CI: 0.689–0.888) with HumanGold than any individual AI rater.

We subsequently reran all articles through the most recent versions of each LLM to generate updated ratings. Based on the correlation of the updated ratings with HumanGold, we created an AISilver score using ChatGPT, Claude, and Grok. This updated AISilver score still showed strong correlation with

HumanGold ($r = 0.797$), but not stronger than the individual correlation of Grok. To investigate this further, we conducted a bootstrap analysis to produce a 95% confidence interval around each. This analysis demonstrated that although Grok had the highest correlation to HumanGold, its confidence interval [0.720, 0.883] had high overlap with AISilver [0.694, 0.870], indicating that their performance was more comparable rather than Grok clearly outperforming the panel.

To capture the degree of disagreement from each AI with human consensus, we calculated mean bias as the average difference between each individual LLM and HumanGold. This revealed that all LLMs scored higher than HumanGold on average. Although Grok was highly correlated with HumanGold, it also had the highest bias, overestimating scores by an average of 2.39 points compared to human raters. In contrast, Claude showed the smallest bias amongst individual LLMs.

Table 3
HumanGold Correlation with AI Models

Model	Initial AI Ratings (95% bootstrap CI)	Updated AI Ratings (95% bootstrap CI)	Mean Bias	p-value
ChatGPT	0.768 (0.603–0.875)	0.761 (0.632–0.849)	1.01	0.002
Grok	0.739 (0.576–0.859)	0.812 (0.720–0.883)	2.39	< 0.001
Claude	0.703 (0.509–0.830)	0.708 (0.543–0.824)	0.82	0.009
DeepSeek	0.549 (0.320–0.731)	0.489 (0.190–0.727)	1.63	0.002
Gemini	0.745 (0.580–0.853)	0.677 (0.470–0.812)	1.51	< 0.001

Abbreviations: HumanGold, mean score across all human raters; AI, artificial intelligence. Correlations represent Pearson correlation coefficients between HumanGold total scores and each AI model’s total scores during the initial and final run. Mean bias represents the average signed difference between each AI and HumanGold total scores across articles. P-values were derived from one-sample t tests assessing whether mean AI bias differed significantly

TEMPOS

Item-level interrater reliability was calculated for each TEMPOS item to evaluate agreement across TEMPOS domains (see Table 4). By evaluating interrater reliability for each item, we aimed to determine which domains provided consistent ratings and which introduced more variability. This analysis also aided with the assessment of the TEMPOS framework itself.

Highest reliability scores were seen in TEMPOS items that addressed clearly observable aspects in the article, such as suicide prevention resources (T10), explicit description of suicide method (T4), and description or inclusion of a suicide note (T5). The exception to this was T6 which assesses article visual content, however this can partially be explained by the recoding of missing data as explained

earlier. In contrast, TEMPOS items requiring greater interpretation and judgement show lower consistency. These items included factual or speculative language (T2), description of risk factors (T7), and glamorization of suicide (T9). This pattern suggests that items based on explicit, easily observable content yield higher inter-rater agreement.

Table 4
Item Level Interrater Reliability for TEMPOS Items

TEMPOS Item	Domain	ICC(2,1) (95% CI)	ICC(2,k) (95% CI)
T1	Framing of Suicide	0.259 (0.170–0.386)	0.808 (0.711–0.883)
T2	Factual vs Speculative	0.120 (0.060–0.215)	0.621 (0.434–0.766)
T3	Non-Stigmatizing Language	0.366 (0.260–0.502)	0.874 (0.808–0.924)
T4	Suicide Method	0.662 (0.559–0.767)	0.959 (0.938–0.975)
T5	Suicide Note	0.740 (0.650–0.826)	0.972 (0.957–0.983)
T6	Visual Content	0.330 (0.230–0.464)	0.855 (0.782–0.912)
T7	Risk Factors	0.181 (0.107–0.293)	0.727 (0.591–0.833)
T8	Sensational Language	0.289 (0.194–0.420)	0.830 (0.743–0.897)
T9	Glamorization	0.103 (0.047–0.194)	0.579 (0.371–0.743)
T10	Suicide Prevention Resources	0.848 (0.782–0.904)	0.985 (0.977–0.991)

Abbreviations: TEMPOS, Tool for Evaluating Media Portrayals of Suicide; ICC, intraclass correlation coefficient. ICC estimates were calculated using a two-way random-effects model with absolute agreement. ICC(2,1) represents single-measure absolute-agreement reliability and ICC(2,k) represents average-measure absolute-agreement reliability. Values in parentheses represent 95% confidence intervals. Domain-level ICC estimates were calculated across 42 articles and 12 raters for each TEMPOS rubric item.

Discussion

In this study, we evaluated whether large language models (LLMs) can rate incident-focused suicide news articles using TEMPOS with agreement comparable to trained human raters, in a setting where no external gold standard exists beyond human judgment. Below we provide evidence to support all four of our proposed hypotheses. Firstly, individual raters (both human and AI) showed only moderate single-rater reliability on TEMPOS totals, while aggregated panel scores were highly stable. This result supports hypothesis one and suggests that while single AI and Human raters were only moderately consistent in how they scored each item, when we consider multiple perspectives, reliability improves. Secondly, our result that individual AI models showed moderate to strong alignment with the aggregated human score (HumanGold), supports our second hypothesis. These two findings support the practical use of a human

consensus reference (HumanGold) and further reinforce a central premise of rubric-based assessment whereby agreement becomes more robust when multiple independent ratings are averaged.

Our third hypothesis was partially supported. AI models differed meaningfully in their alignment with HumanGold, confirming that model choice matters. In the initial rating set, AISilver showed stronger agreement with HumanGold than any individual AI model. However, in the updated rating set, while AISilver remained strongly correlated with HumanGold and showed overlapping bootstrap confidence intervals with Grok, Grok had the strongest correlation. However, despite Grok's strong correlation, it had the largest difference in scores compared to HumanGold, while Claude, which had a lower correlation, had the least amount of difference.

These findings demonstrate the importance of aggregating AI models in future similar analyses. Individual LLMs vary in capabilities, with some models having stronger reasoning, and image detection compared to other models. Additionally, with the growing capabilities of AI, model versions are continuously changing and surpassing others in capabilities. Our findings demonstrate that LLMs are not always consistent performers, especially across versions. We further argue for the importance of an aggregated model (i.e., AISilver) given that multi-model consensus score is not skewed by version changes or other factors in contrast to an individual model. Thus, the updated findings refine rather than reject the AISilver concept. Overall, we propose that AISilver may be most useful as a stability strategy in a changing model environment, rather than as a method that will necessarily outperform the best individual model at every time point.

Finally, our finding that the highest reliability scores were seen in TEMPOS items that addressed clearly observable aspects in the article, supports our fourth hypothesis. As noted above, TEMPOS item 10 (inclusion of suicide prevention resources), TEMPOS item 4 (explicit description of suicide method), and TEMPOS item 5 (description or inclusion of a suicide note) are easier to observe and therefore were easier to rate among both our human and AI raters. These items require little to no context or clinical knowledge to rate and are less influenced by personal judgement or biases.

Our research builds on the research of Elyoseph and colleagues [8] who applied language models to evaluate news articles on suicide using the WHO suicide reporting guidelines. These researchers reported that there was strong agreement between one of the AI tools with human reviewers and good agreement between the other tool and the human reviewers. By conducting a similar study but looking at a different tool (TEMPOS rather than WHO guidelines) our study helps to develop our overall knowledge on using LLM to evaluate suicide related media articles. Further, while WHO provides guidelines on reporting, TEMPOS is a rating system that more accurately measures adherence to a set of recommendations. Finally, where Elyoseph and colleagues [8] only used two AI raters and two human raters, our study compared ratings among seven human raters and five AI raters. By including more raters than prior research, our study improves the reliability of the consensus scores and reduces the influence of individual differences in interpretation. This provides a more stable and robust reference point for evaluating model performance.

Given the time and resource demands associated with human rating, a key motivation for this work is feasibility. Continuous, large-scale monitoring of suicide-incident reporting is rarely resourced, not only because of labor costs but because of the cumulative emotional burden of reading large volumes of distressing content [11–12]. While we have some understanding of the emotional burden caused by engaging with content on suicide, much of the research focuses on the emotional burden for people who are suicidal [13]. Therefore, developing our understanding of the emotional burden associated with conducting research on suicide, and determining strategies to reduce this burden is necessary. These considerations, alongside the practical challenges of reviewing large volumes of content at scale, highlight the need to explore alternative approaches such as the use of AI in supporting or augmenting human review. Even when the intended response is low stakes (such as providing supportive guideline reminders rather than punitive action), the large volume of articles makes it impractical to rely exclusively on human reviewers to provide timely feedback to journalists. In this context, the appropriate performance target is not to demonstrate that AI can produce superior work to humans, but rather that it can perform at a similar level, i.e., whether the lower bound of AI to HumanGold agreement is adequate for the intended use. Our results suggest that at least one model can approximate HumanGold closely enough to be operationally useful, especially when paired with simple safeguards such as audit sampling and uncertainty flags for ambiguous cases.

The item-level results provide additional insight into where LLM scoring is most reliable and where caution is warranted. Agreement was strongest on criteria that are directly observable in text (e.g., TEMPOS 5, how does the report describe the suicide note), consistent with TEMPOS' design as an operationalization tool. In contrast, lower agreement was observed for more interpretive domains (e.g., TEMPOS 9, does the report glamorize suicide). Among human raters, agreement was also strongest on criteria that are directly observable in text. This suggests that while we cannot completely control for ambiguity in journalistic phrasing, there may be a need to develop more explicit criteria rules to ensure greater consistency among raters.

Visual content (TEMPOS Item 6) emerged as a recurring friction point due to variability in article images and missingness, as well as the practical limitation that some LLM workflows may not reliably access or interpret images/video. However, as mentioned in the method section, we dealt with missing data related to TEMPOS Item 6 by recoding missing data (for articles where there was no image) to 2, in line with our interpretation of the TEMPOS criteria. To overcome this challenge for LLMs, we suggest several solutions including (1) providing models with a structured human-coded visual label, (2) using multimodal LLMs that can view page media directly, or (3) reporting a text-only score as a secondary metric.

The updated AI ratings also highlight an important practical issue for using LLMs as raters: model performance can change over time. In the initial rating set, the AISilver panel showed stronger alignment with HumanGold than any individual AI model. In the updated rating set, Grok had the highest point estimate, while AISilver remained strongly correlated with HumanGold and showed substantial overlap with Grok in the bootstrap confidence intervals. This suggests that AISilver should be understood less as

a guaranteed improvement over the best individual model and more as a stability strategy. In real-world monitoring, averaging across several well-performing models may reduce dependence on any single model's current version, vendor-specific behavior, or temporary performance shift. At the same time, a single validated model may be sufficient for many low-stakes uses, particularly if paired with periodic revalidation and human review of flagged or ambiguous cases.

Outside of the research conducted by Elyoseph and colleagues [8] and some tangentially related studies that examine AI and suicide prevention more generally [14] the use of AI to evaluate suicide related content remains scarce. This is despite previous literature noting the importance of safe reporting on suicide [3]. Our study both provides novel results on using AI to rate media articles on suicide and contributes to the wider field of suicide prevention. While we must remain cautious of an overreliance on AI, particularly in a complex space such as suicide prevention, our results demonstrate that if used responsibly, AI might help to reduce some of the emotional burden of working in the field of suicide prevention. To further support this cautious approach, we recommend the use of multiple LLMs rather than reliance on a single system.

Limitations

This study has several limitations. Firstly, human consensus is an operational reference standard rather than a ground truth. This means that the “best” performance should be interpreted as consensus alignment rather than correctness. The article sample was restricted to publicly accessible incident-focused reporting, which improves real-world relevance for monitoring but may limit generalizability to other article types or paywalled sources. Model availability and behavior can change over time; therefore, documenting model identifiers, run dates, and prompting protocols is essential, and synchronized re-runs may be useful as a robustness check. Additionally, the study's intended use case is supportive feedback and surveillance rather than clinical decision-making. Knowing this, performance thresholds and acceptable error tolerances should be calibrated to that context. A further limitation relates to the TEMPOS tool itself. Raters reported some challenges with using the tool to objectively rate the articles. Raters found the biggest challenges with TEMPOS 6, which addressed visual content.

Implications and conclusions

Our findings provide an actionable pathway towards the monitoring of suicide reporting at a large scale. The focus of our research was not whether AI outperforms human raters, but whether AI agreement with HumanGold is sufficient to support low-stakes monitoring and feedback on suicide-related content. Once a level of agreement between AISilver and HumanGold is established, the best performing model or models could be used cautiously alongside limited human review for quality assurance. In practice, this would involve using a small, trained human panel to establish a stable consensus benchmark, which can then be used to validate LLM-based raters. Although this work focuses on suicide reporting, the HumanGold/AISilver framework offers a broader approach for evaluating AI as a “rater” in public health contexts where no true objective standard exists and human judgment remains the most appropriate reference standard.

At the same time, the practical implications are not limited to methodology. For example, communities could use a validated model to help monitor local suicide coverage without needing to average across three models. Further, the same approach could be incorporated into future tools that allow journalists to review their own articles before publication. Finally, this framework could support broader research on suicide reporting, including comparisons across time periods, geographic regions, media types, and publication formats, such as print newspapers versus online news sites.

Overall, our findings demonstrate that LLMs may be a plausible option to assist with the large-scale monitoring of suicide reporting. However, it should be paired with human oversight (from multiple people) and implemented transparently and under other safeguards including auditing procedures and clear documentation of model versions and prompts. By using LLMs to assist with the monitoring of suicide articles, we can both reduce the emotional and time burden on people, and provide more efficient feedback to journalists to encourage the safer reporting of suicide. While acknowledging these promising findings, as mentioned above we must continue to be cautious when using LLMs, particularly when covering sensitive topics such as suicide. Given the ongoing public health burden of suicide, ongoing development of suicide prevention strategies grounded in evidence remains critical. We have demonstrated that LLMs have potential to be used in the field of public health, however we must continue to develop our understanding of these tools and ensure that safeguards are in place when using these models.

Data Availability

The deidentified TEMPOS rating dataset supporting the findings of this study, including article-level human rater scores, AI model scores, HumanGold and AISilver variables, and analysis code, will be made available by the corresponding author upon reasonable request. The full text of the news articles analyzed in this study is not redistributed because the articles are copyrighted by their original publishers. Article source information, including publication names, dates, titles, and URLs where available, can be provided to qualified researchers for purposes of verification and replication. The standardized AI-rating prompt is provided in Appendix A.

Declarations

Ethics Statement

This study analyzed publicly available news articles and deidentified ratings generated by trained adult raters as part of a methodological evaluation of the TEMPOS rating framework. The study did not involve recruitment of patients or members of the public, intervention with human participants, collection of private identifiable information, or use of clinical, biological, or health-record data. Human raters gave informed consent to participate, were informed that the rating task involved suicide-related news content, were permitted to discontinue participation with no penalty, and were not identified in the analytic dataset. The Office of IRB Administration at the University of California San Diego reviewed the

protocol and determined that the project does not constitute human subjects research under the U.S. Code of Federal Regulations (45 CFR 46); therefore, Institutional Review Board review was not required.

Clinical Trial Number

Not Applicable

Consent to Publish

Declaration: Not Applicable. The manuscript does not report individually identifiable participant data, images, or clinical information.

Consent to Participate

Declaration: Human raters participated voluntarily, were informed that the rating task involved suicide-related news content, gave informed consent, were permitted to discontinue participation at no penalty, and were not identified in the analytic dataset.

Accordance Statement

The study was conducted in accordance with applicable institutional, national, and international ethical standards, including the ethical principles of the Declaration of Helsinki, where applicable. The Office of IRB Administration at the University of California San Diego reviewed the protocol and determined that the project did not constitute human subjects research under the U.S. Code of Federal Regulations (45 CFR 46); therefore, Institutional Review Board (IRB) review was not required. Human raters participated voluntarily, provided informed consent before participation, and were free to discontinue participation at any time without penalty.

Funding Statement

This study received no dedicated external funding.

Author Contribution

Study conception and development lead by AS with support from ED and BM. BM and AS lead the analysis and interpretation of all data, as well as the development of the results section. ED lead the development of the introduction and discussion sections. Overall all authors contributed in writing the manuscript and approved the final manuscript.

Acknowledgement

We would like to thank the following individuals who took the time to rate the articles. Without their assistance, this research could not have been conducted: Hayley Carroll, Jenny Chau, Michelle Cox, Carol Gonzalez, Marlin Hart, Richard Heller, Kate Neamsapaya, Briana Nguyen, Sheri Ortuno and Nick Panagon. We acknowledge your hard work and the challenging nature of the task. Also, many thanks to our colleague, Frances Reyes Yee for her input. This study received no dedicated external funding.

Data Availability

The deidentified TEMPOS rating dataset supporting the findings of this study, including article-level human rater scores, AI model scores, HumanGold and AISilver variables, and analysis code, will be made available by the corresponding author upon reasonable request. The full text of the news articles analyzed in this study is not redistributed because the articles are copyrighted by their original publishers. Article source information, including publication names, dates, titles, and URLs where available, can be provided to qualified researchers for purposes of verification and replication. The standardized AI-rating prompt is provided in Appendix A.

References

1. Cheng ATA, et al. The influence of media reporting of a celebrity suicide on suicidal behavior in patients with a history of depressive disorder. *J Affect Disord.* 2007;103(1–3):69–75. <https://doi.org/10.1016/j.jad.2007.01.021>.
2. Kim J, et al. The Werther effect of two celebrity suicides: an entertainer and a politician. *PLoS ONE.* 2013;8(12). <https://doi.org/10.1371/journal.pone.0084876>. e84876.
3. Domaradzki J. The Werther effect, the Papageno effect or no effect? A literature review. *Int J Environ Res Public Health.* 2021;18(5):2396. <https://doi.org/10.3390/ijerph18052396>.
4. Calvo S, Carrasco JP, Conde-Pumpido C, Esteve J, Aguilar EJ. Does suicide contagion (Werther effect) take place in response to social media? A systematic review. *Span J Psychiatry Ment Health.* 2026;19(1):55–65. <https://doi.org/10.1016/j.sjpmh.2024.05.003>.
5. World Health Organization, (WHO). (2023). *Preventing Suicide: A Resource for Media Professionals, Update 2023*, 2023). <https://www.who.int/publications/i/item/9789240076846>
6. SAVE. Best Practices and Recommendations for Reporting on Suicide. (2020); retrieved April 24, 2026, from <https://reportingonsuicide.org/recommendations/>
7. Sorensen CC, et al. The Tool for Evaluating Media Portrayals of Suicide (TEMPOS): development and application of a novel rating scale to reduce suicide contagion. *Int J Environ Res Public Health.* 2022;19(5):2994. <https://doi.org/10.3390/ijerph19052994>.
8. Elyoseph Z, et al. Applying language models for suicide prevention: Evaluating news article adherence to WHO reporting guidelines. *npj Mental Health Res.* 2025;4., Article 25 (<https://doi.org/10.1038/s44184-025-00139-5>).

9. Walter SD, Eliasziw M, Donner A. Sample size and optimal designs for reliability studies. *Stat Med*. 1998;17(1):101–10. [https://doi.org/10.1002/\(SICI\)1097-0258. ::AID-SIM727>3.0.CO;2-E](https://doi.org/10.1002/(SICI)1097-0258.::AID-SIM727>3.0.CO;2-E).
10. Mondal D, Vanbelle S, Cassese A, Candel MJJM. Review of sample size determination methods for the intraclass correlation coefficient in the one-way analysis of variance model. *Stat Methods Med Res*. 2024;33(3):532–53. <https://doi.org/10.1177/09622802231224657>.
11. Andriessen K, Reifels L, Krysinska K, Robinson J, Dempster G, Pirkis J. Ethical concerns in suicide research: Results of an international researcher survey. *J Empir Res Hum Res Ethics*. 2019;14(4):383–94. <https://doi.org/10.1177/1556264619859734>.
12. Woodby LL, Williams BR, Wittich AR, Burgio KL. Expanding the notion of researcher distress: the cumulative effects of coding. *Qual Health Res*. 2011;21(6):830–8. <https://doi.org/10.1177/1049732311402095>.
13. Lakeman R, Fitzgerald M. Ethical suicide research: a survey of researchers. *Int J Ment Health Nurs*. 2009;18(1):10–7. <https://doi.org/10.1111/j.1447-0349.2008.00569.x>.
14. Bernert RA, et al. Artificial Intelligence and suicide prevention: A systematic review of machine learning investigation. *Int J Environ Res Public Health*. 2020;17(16):5929. <https://doi.org/10.3390/ijerph17165929>.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [AppendixA.docx](#)