The Role of Generative Artificial Intelligence in Evaluating Adherence to Responsible Press Media Reports on Suicide: A Multi-Site, Three-Language Study

Z. Elyospeh, PhD[1Ψ], B. Nobile, PharmD, PhD[2,3Ψ], I. Levkovich, PhD[4], R. Chancel, MD[2], P. Courtet, MD, PhD[2,3], Y. Levi-Belz, PhD[5]

[1]University of Haifa, Mount Carmel, Haifa, Israel, zohar.j.a@gmail.com

[2] Department of Emergency Psychiatry and Acute Care, CHU Montpellier, France, benedicte.nobile@gmail.com, philippecourtet@gmail.com

[3] IGF, Univ. Montpellier, CNRS, INSERM, Montpellier, France

[4]Faculty of Education, Tel Hai College, Upper Galilee, Israel, levkovinb@telhai.ac.il

[5] Lior Tsfaty Center for Suicide and Mental Pain Studies, Ruppin Academic Center, Israel, yossil@ruppin.ac.il

Ψ These authors contributed equally, and are co-first author

https://doi.org/10.1192/j.eurpsy.2025.10037 Published online by Cambridge University Press

ABSTRACT

Background: Improving media adherence to World Health Organization (WHO) guidelines is crucial for preventing suicidal behaviors in the general population. However, there is currently no valid, rapid, and effective method to evaluate adherence to these guidelines.

Methods: This comparative effectiveness study (January–August 2024) evaluated the ability of two Artificial Intelligence (AI) models (Claude Opus 3 and GPT-4O) to assess adherence of media reports to WHO suicide reporting guidelines. A total of 120 suicide-related articles (40 in English, 40 in Hebrew, 40 in French) published within the past 5 years were sourced from prominent newspapers. Six trained human raters (two per language) independently evaluated articles based on a WHO guideline-based questionnaire addressing aspects such as prominence, sensationalism, and prevention. The same articles were also processed through the AI models. Intraclass correlation coefficients (ICC) and Spearman correlations were calculated to assess agreement between human raters and AI models.

Results: Overall adherence to WHO guidelines was approximately 50% across all languages. Both AI models demonstrated strong agreement with human raters, with GPT-4O showing the highest agreement (ICC = 0.793 [0.702; 0.855]). The combined evaluations of GPT-4O and Claude Opus 3 yielded the highest reliability (ICC = 0.812 [0.731; 0.869]).

Conclusions: AI models can replicate human judgment in evaluating media adherence to WHO guidelines. However, they have limitations and should be used alongside human oversight. These findings may suggest that AI tools has a potential to enhance and promote responsible reporting practices among journalists, and thus, may support suicide prevention efforts globally.

KEYWORDS: Suicide; Artificial Intelligence; Media; Suicide prevention ; Natural Langage Processing

https://doi.org/10.1192/j.eurpsy.2025.10037 Published online by Cambridge University Press

INTRODUCTION

With more than 800,000 deaths by suicide each year, preventing suicide is a global imperative [1]. Since it is a major cause of premature death, stronger prevention strategies must be developed to address it. While most studies and prevention efforts have focused on indicated and selected prevention (*i.e.* for specific high-risk group and for patients with previous suicide attempt or current suicidal ideation, respectively), growing evidence suggests that universal prevention (for the general population) strategies are promising for reducing suicide rates [2–4]. Among universal prevention efforts, media coverage of suicide and suicidal behavior is a critical area of focus.

Traditional media plays a key role in shaping public perception and has a significant influence on the general population. Consequently, the way suicide and suicidal behaviors are reported can have either a preventive effect (i.e., the "Papageno" effect) or a harmful one (i.e., the "Werther" effect) [5]. Numerous studies have demonstrated that irresponsible traditional media coverage of suicide (e.g., sensationalist reporting) leads to an increase in suicide rates and behaviors by triggering imitative or "copycat" suicides [2,5–10]. On the other hand, responsible traditional media coverage (e.g., providing information about available resources and avoiding details on suicide methods) has been shown to be effective not only for the general population but also for vulnerable groups such as youth [2,5,11,12]. Given the impact of traditional media on public behavior, the World Health Organization (WHO) published guidelines in 2008 for reporting suicide in traditional media (excluding social media), which were updated in 2017 [13]. However, adherence to these guidelines among journalists found to be poor [2]. For instance, a recent study reviewing 200 articles on suicide published in the last 10 years found an adherence of only approximately 49% to the WHO guidelines [14].

https://doi.org/10.1192/j.eurpsy.2025.10037 Published online by Cambridge University Press

Therefore, evaluating traditional media adherence to these guidelines and educating journalists is crucial for improving suicide prevention efforts at the primary level [5].

Manual screening and evaluation of every traditional media report on suicide is practically impossible due to the volume of reports and the variety of languages in which they are written. Thus, developing a simple valid tool which capable of screening and assessing whether traditional media reports on suicide comply with WHO guidelines is compelling. Such a tool could greatly enhance the monitoring and encourage journalists and traditional media organizations to adhere to guidelines more consistently. Artificial Intelligence (AI) offers valuable support in this regard [15,16]. Interest in the use of AI in the mental health field is growing, and it has shown promising results in various applications [17–20]. Notably, numerous studies are emerging on the use of AI for the prevention of suicidal behavior [21,22]. Most existing research on AI and suicidal behavior focuses on clinical applications, such as improving the detection of suicidality through automated language analysis, assisting in risk assessment and diagnosis, enhancing accessibility to crisis counseling, supporting training for mental health professionals, contributing to policy development, and facilitating public health surveillance and data annotation [22]. While some studies examine social media, particularly in the context of predicting suicide risk, no study to date has evaluated AI's ability to assess whether traditional media reports on suicide comply with WHO guidelines. Compared to conventional machine learning classifiers, which typically rely on manually engineered features and labeled training datasets, Large Language Models (LLMs) are better suited for assessing complex linguistic guidelines due to their advanced contextual understanding and ability to process unstructured text across multiple languages. Previous studies have demonstrated that LLMs can match or even outperform traditional classifiers in text classification tasks, particularly in domains requiring nuanced comprehension of natural language [23–25].

https://doi.org/10.1192/j.eurpsy.2025.10037 Published online by Cambridge University Press

In a preliminary study, we evaluated the use of generative artificial intelligence (GenAI) to assess suicide-related news articles in Hebrew according to WHO criteria. In that study, two independent human reviewers and two AI systems, Claude.AI and ChatGPT-4, were employed. The results demonstrated strong agreement between ChatGPT-4 and the human reviewers, suggesting that AI-based tools could be effective in this domain [26]. Building on these preliminary findings, the present study aimed to assess the capacity of AI, utilizing two different LLMs, to evaluate to what extent traditional media reports on suicide and suicidal behavior adhere to WHO guidelines. The evaluation was conducted in comparison with human raters and across three languages: English, Hebrew, and French. Specifically, we examined to what extent AI models could match the performance of human raters across multiple languages. If successful, such tools could serve as accessible and practical resources for journalists to screen their reports prior to publication, improving adherence to WHO guidelines and, ultimately, contributing to suicide prevention efforts.

To the best of our knowledge, no previous studies have attempted to evaluate traditional media adherence to WHO suicide reporting guidelines using GenAI or other computational methods. As mentioned, while some prior research has employed machine learning or rule-based systems to address related challenges in other domains of mental health (14–19), the novelty of this study lies in its application of AI to this specific and crucial aspect of suicide prevention. This study seeks to bridge an important gap in both mental health research and AI applications, while highlighting the potential for AI tools to make a meaningful impact in global suicide prevention efforts.

https://doi.org/10.1192/j.eurpsy.2025.10037 Published online by Cambridge University Press

METHODS

Data Collection

In this study, we systematically reviewed a corpus of 120 articles concerning suicide published in newspapers in three languages during the last 5 years: 40 articles in English, 40 in Hebrew, and 40 in French. The sample size was determined using G*Power software, assuming a minimum correlation of 0.8 between raters [14], a confidence level of 0.8, and an alpha level of 0.05. The results of the analysis indicated the need for a sample size of 40 articles by languages.

The selection process followed a structured approach to ensure the inclusion of widely read and influential sources. Newspapers were chosen based on the following criteria:

- High Readership & National/Regional Influence – We selected newspapers with significant circulation and impact on public discourse in their respective countries.

- Geographical & Political Diversity – To capture different reporting styles and perspectives, we included both national and regional newspapers.

- Availability of Online Archives – Only newspapers with accessible digital archives were included to ensure consistency in data collection.

Based on these criteria, the newspapers selected for each language were: English: The Guardian and The New York Times (representing internationally recognized, high-impact journalism); Hebrew: Israel Hayom and Yedioth Ahronoth (two of Israel's most widely read newspapers, offering different political perspectives); French: La Provence, Midi Libre, and La Dépêche (major regional daily newspapers in the south of France, where suicide rates are a significant public health concern).

https://doi.org/10.1192/j.eurpsy.2025.10037 Published online by Cambridge University Press

The selection process involved querying the electronic archives of these newspapers using relevant keywords for "suicide" (in the masculine, feminine, and plural forms), "self-destructive behavior," "attempted suicide," and "ended his/her life" in each respective language. Articles that employed any of these terms colloquially described suicide bombings in the context of terror attacks or used them metaphorically were excluded from the search results. Additionally, articles whose primary focus was not on suicide or self-destructive behavior but merely mentioned an individual's death by suicide in passing were also omitted. Furthermore, articles debating whether the described death constituted suicide or homicide were not included in the study.

Article Screening Criteria

The screening of articles was guided by criteria established by the WHO, as detailed in a study by Levi-Belz et al. (2023), which outlined 15 parameters for article screening. The criteria used are listed in Suppl Mat. Table 1. Two items (items 2 and 8) pertaining to the presence of images in articles were excluded from consideration given the current limitations in analyzing image content. The questionnaire's items assess different aspects of traditional media coverage of suicide such as: prominence (e.g., avoiding explicit mention of suicide in the headline, two items), complexity (e.g., avoiding speculation about a single cause of suicide, three items), sensationalism (e.g., avoiding glorifying the suicidal act, five items), and prevention (e.g., providing information about risk factors for suicide, three items) (Levi-Belz et al., 2023). Each criterion was assessed based on whether it was met or not.

Large Language Models

For this study, we employed two versions of LLMs, Claude.AI, using the Opus 3 model and ChatGPT-4o, each with a temperature setting of 0. This setting was chosen to minimize

https://doi.org/10.1192/j.eurpsy.2025.10037 Published online by Cambridge University Press

randomness in the output and ensure that the models produced consistent deterministic results in the analysis of the articles. The selection of these specific LLMs was informed by three methodological considerations. First, both models represent current computational approaches in natural language processing, as reflected by their commercial deployment status. Second, their established presence across research applications provides documented evidence of their capabilities. Third, and particularly relevant to this study's aims, both models have demonstrated effectiveness in multilingual processing, including documented performance with Hebrew text analysis, supporting their appropriateness for cross-linguistic evaluation tasks.

Claude.AI, created by Anthropic, was designed to generate beneficial, inoffensive, and truthful outputs by employing a constitutional approach. The Opus 3 version utilized in this study incorporates over 12 billion parameters and aims to ethically address linguistic complexity. This model was selected for its emphasis on educational data curation, alignment with human values, and safety considerations. A temperature setting of 0 was chosen to maximize the reliability of the model and reduce the variance in its assessments.

GPT-4o, developed by OpenAI, was configured similarly with a temperature setting of 0 for this study. The temperature setting was selected to enhance the model's accuracy and content policy adherence by reducing output variability. This configuration was applied uniformly across all three languages. Claude Opus 3 and GPT-4O were selected based on our empirical testing, which demonstrated these models' superior performance in Hebrew language processing—a critical requirement given our multilingual study design. From our experience, these were the only models at the time that could effectively analyze Hebrew content with sufficient accuracy for research purposes. Image analysis capabilities of AI models were

https://doi.org/10.1192/j.eurpsy.2025.10037 Published online by Cambridge University Press

relatively limited during the study period, and the inconsistent presence of images across articles further justified our text-only approach.

The prompt architecture integrated three methodological elements to ensure reliable guideline assessment. Role assignment positioned the AI model as both academic expert and traditional media editor, while a structured thought-chain protocol guided systematic evaluation of each WHO parameter. The implementation of binary scoring (0/1) with clear operational definitions enabled consistent cross-linguistic assessment. This framework aimed to maintain standardized evaluation while accommodating different linguistic contexts. The prompt used to analyze the 120 articles is available in supplementary materials (Suppl mat table 1).

Human benchmark

For English articles, the evaluation was conducted independently by a master's student in educational psychology (from Israel) and a resident in psychiatry (from France). Two trained psychology students, one pursuing a B.A. and the other an M.A., independently evaluated each of the 40 Hebrew articles, according to the screening criteria. The French articles were independently evaluated by one resident in psychiatry and one researcher specializing in suicide research. All evaluators were trained and supervised by researchers specializing in suicide research (one from Israel for Israelis students and one from France for French students). This dual-assessment approach was employed in each language group to enhance the reliability of the data through inter-rater agreement. The inter-rater agreement was calculated to ensure high consistency between human evaluators (see Results section).

Procedure

Evaluations were conducted from January 2024 to August 2024. Manual evaluations of the 120 articles were done by the six trained students. Following manual evaluation, all 120

https://doi.org/10.1192/j.eurpsy.2025.10037 Published online by Cambridge University Press

articles were processed through two LLMs, ChatGPT-4o and Claude.AI Opus, to document their respective assessments. This procedure was designed to compare the analytical capabilities of LLMs against human-coded data, thereby enabling an examination of the efficacy and consistency of automated text analysis in the context of psychological research on suicide reporting.

Statistical Analysis

The study employed a comprehensive analytical framework to assess the agreement between human evaluators and AI systems across multiple dimensions. The primary analysis focused on three complementary approaches to evaluate inter-rater reliability and agreement across the full corpus of 120 articles, with additional analyses performed separately for each language group (English, Hebrew, and French).

The first analytical component utilized Intraclass Correlation Coefficients (ICC) with 95% confidence intervals to assess the consistency and agreement between different rater combinations. This included examining the reliability between human evaluators, between AI models (Claude Opus 3 and GPT-4O), between individual AI models and human evaluators, and between combined AI evaluations and human ratings. The ICC analysis was particularly valuable for providing a comprehensive measure of rating reliability that accounts for both systematic and random variations in ratings.

The second analytical component employed Spearman correlation coefficients to examine the consistency of ranking patterns between different rater pairs. This non-parametric measure was selected to assess how well the relative ordering of articles aligned between human and AI evaluators, providing insight into the consistency of comparative judgments across raters. The analysis included correlations between individual AI models and human ratings, as well as between the combined AI ratings and human evaluations.

https://doi.org/10.1192/j.eurpsy.2025.10037 Published online by Cambridge University Press

The third component focused on examining absolute score differences between human raters and AI models through paired samples t-tests. This analysis was crucial for determining whether the AI models' evaluations showed systematic differences from human ratings in terms of their absolute magnitudes. The comparison specifically examined differences between mean scores of human raters and combined AI evaluations across the entire corpus of articles.

For language-specific analyses, the same analytical framework was applied separately to each subset of 40 articles in English, Hebrew, and French, with results reported in supplementary materials.

All statistical analysis were done with SPSS statistical software (version 28.0.1.1; IBM SPSS Statistics for Windows. Armonk, NY: IBM Corp). The significance level for all statistical tests was set at $p < .001$, and analyses were conducted using appropriate statistical software. This analytical approach provided a robust framework for evaluating both the overall reliability of AI evaluations and their specific performance characteristics across different languages and rating contexts.

Ethical Considerations

This study was exempt from ethical review since it only evaluated AI chatbots, and no human participants were involved.

RESULTS

The analysis presented here focused on the agreement between human evaluators and AI models (Claude Opus 3 and GPT-4O) across 120 articles, with additional breakdowns by language (English, Hebrew, and French). The results are structured to first present the ICC between human evaluators and AI models, followed by an analysis of the agreement between each AI model and the average human ratings, as well as the agreement between the combined

https://doi.org/10.1192/j.eurpsy.2025.10037 Published online by Cambridge University Press

AI models and human evaluators. The results are then separately detailed for each language group in the supplementary files (Suppl mat table 2).

Insert Table 1 here.

Assessing Consistency and Agreement Across All 120 Articles

The ICC between human evaluators across all 120 articles was .793, indicating a high level of consistency among human raters. Similarly, the ICC between the AI models (Claude Opus 3 and GPT-4O) was .812, reflecting strong agreement between the two AI systems when evaluating the same set of articles.

Claude Opus 3 vs. Human Evaluators

The average ICC between Claude Opus 3 and the average human evaluator across all 120 articles was r=.724. This ICC value indicates a good level of agreement between Claude Opus 3 and the human evaluators, suggesting that Claude Opus 3 provides evaluations that are consistent with human judgments.

The Spearman correlation between Claude Opus 3 and the average human evaluators was r=.636, which was statistically significant at $p < .001$. This positive correlation further supports the alignment between Claude Opus 3 and human evaluators in terms of the relative ranking of articles.

GPT-4O vs. Human Evaluators

https://doi.org/10.1192/j.eurpsy.2025.10037 Published online by Cambridge University Press

For GPT-4O, the average ICC with the average human evaluators was .793. This higher ICC value compared to that of Claude Opus 3 suggests that GPT-4O is more closely aligned with human evaluators.

The Spearman correlation between GPT-4O and the average human evaluator was r=.684, which was also statistically significant at $p < .001$. This strong correlation indicates that GPT-4O aligns well with human evaluators in terms of absolute ratings and the ranking of articles.

Combined AI Models vs. Human Evaluators

When considering the average ratings of both AI models combined (Claude Opus 3 and GPT-4O), the average measure ICC with the human evaluators was .812. This ICC suggests that combined AI models provide an even more robust measure of agreement with human evaluators.

The Spearman correlation coefficient between the combined AI models and human evaluators was .703, which was significant at $p < .001$ (Figure 1). This further confirms that the combined evaluations from both AI models are closely aligned with those of the human evaluators.

Comparison of Overall Evaluations Across All 120 Articles

The comparison between human raters and the combined LLMs (ChatGPT-4O and Claude Opus 3) across the 120 articles revealed no significant differences in the overall mean evaluations. The paired samples t-test indicated that the mean score for human raters was 7.00 (SD = 1.46), whereas the mean score for the AI evaluations was 7.12 (SD = 1.54). The mean difference was -0.12 (SD = 1.19), with a t-value of -1.09 and a two-sided p-value of .28,

https://doi.org/10.1192/j.eurpsy.2025.10037 Published online by Cambridge University Press

suggesting that the AI models generally align closely with human judgments in their evaluations (Figure 2).

Example of divergence between human and AI evaluations

Table 2 presents the ratings of a specific Hebrew-language article, comparing the evaluations of two human raters (Human Rater 1 and Human Rater 2) and two AI models (GPT-4o and Claude Opus 3) across the WHO guideline criteria.

Insert Table 2 here

This example demonstrates several interesting patterns of divergence:

1. **Headline interpretation (Item 1)**: Both AI models identified a mention of suicide in the headline, while both human raters did not.
2. **Causation and life events (Items 4-5)**: Claude Opus 3 did not identify single-cause reporting or links between specific life events and suicide, while the other three evaluators did.
3. **Prevention and intervention information (Items 14-15)**: Human Rater 2 determined that the article lacked prevention and intervention information, while both AI models and Human Rater 1 found that such information was present.

Despite the overall strong agreement observed in our statistical analysis, this example demonstrates that significant variation can exist in specific cases, both between human raters themselves and between AI and human evaluations.

DISCUSSION

Traditional media coverage significantly impacts public perception and suicide rates, making adherence to WHO guidelines crucial. This study main goal was to explore the potential of AI models to evaluate traditional media adherence to these guidelines in real-time across different languages. To our knowledge, this study is the first to assess AI's ability to evaluate

https://doi.org/10.1192/j.eurpsy.2025.10037 Published online by Cambridge University Press

the adherence of traditional media reports to WHO guidelines in comparison with human raters, across three languages: English, Hebrew, and French. The results showed that across all 120 articles, the AI models Claude Opus 3 and GPT-4O demonstrated strong consistency with human raters, as evidenced by the high ICC and Spearman correlation values, especially for GPT-4O. The combined evaluations from both AI models provided the highest level of agreement with the human raters. Language-specific analyses revealed that AI models performed best in Hebrew, followed by French and English. This variation may be attributed to linguistic complexity. Hebrew is a relatively direct language with simpler syntax and fewer ambiguities, which may allow AI models to interpret adherence criteria more effectively. In contrast, French tends to be more nuanced and context-dependent, potentially making it more challenging for AI to assess guideline compliance accurately. Regarding English-language articles, one possible explanation for the slightly lower AI agreement is that the human raters evaluating these articles were non-native speakers, which may have introduced variability in their assessments. Future advancements in language-based AI models are likely to enhance performance across all languages, including those with greater linguistic complexity. As models become more adept at handling nuance, ambiguity, and contextual variation, their ability to accurately assess guideline adherence is expected to improve accordingly.

Several studies already showed that adherence to WHO guidelines are essentials in related to suicide rates [11]. Unfortunately, as observed in other studies, there are poor adherence from traditional medias to these guidelines [14] and as mentioned in the goals of this study, we also found a poor adherence to the WHO guidelines in the different newspapers from which the 120 articles were taken. In fact, the overall mean score in our study, for each language, whether rated by humans or AI models, was around 7 out of a total score of 15 (with a higher score indicating worse adherence). These results suggest that adherence to WHO guidelines by the traditional media, whether in English, Hebrew, or French, is around 50%,

https://doi.org/10.1192/j.eurpsy.2025.10037 Published online by Cambridge University Press

reinforcing the need to improve compliance. Beyond individual media reports, the broader societal impact of suicide coverage must also be considered. Social network theory suggests that emotions, including distress and suicidal ideation, can spread through interpersonal connections, increasing vulnerability within communities [27]. Additionally, a shift in suicide prevention efforts is needed to move beyond psychiatric diagnoses and focus on emotional distress as a key risk factor [28]. Responsible media reporting can play a crucial role in this paradigm shift by promoting narratives of hope, coping, and available resources. Future research should explore how AI-driven assessments of media adherence to WHO guidelines can be integrated into broader suicide prevention strategies.

The main finding of our study is that our prompt shows high accuracy compared to human ratings, regardless of the language used in the traditional media reports, suggesting that this prompt could be applied globally. In addition, AI models analyze adherence to guidelines faster than human raters (around 2 minutes per article for AI models), facilitating the review of traditional media reports. Thus, this prompt could be easily used by journalists and editors before publishing articles on suicidal behavior to assess whether they comply with the WHO guidelines. Moving forward, the next step in our project is to improve our prompts by incorporating the automatic correction of articles. This would not only allow the prompt verification of whether an article adheres to the WHO guidelines but also correct problematic sentences. In this way, journalists and editors may be more likely to respect WHO guidelines by using a quick and easy tool to verify their articles, such as our prompt. To encourage adherence to these guidelines, regulatory bodies that oversee journalism should promote the use of such tools. For example, in France, the Journalistic Ethics and Mediation Council, a body responsible for regulating traditional media reporting, could help disseminate this tool to encourage journalists and editors to comply with the WHO guidelines on reporting suicide. To facilitate the integration of AI tools into journalistic workflows, AI could function as a pre-

https://doi.org/10.1192/j.eurpsy.2025.10037 Published online by Cambridge University Press

publication checker, assisting journalists and editors in evaluating adherence to WHO guidelines prior to publication. Collaboration between AI, researchers, media professionals, and policymakers is essential to align AI models with journalistic standards while maintaining editorial independence. Additionally, AI could assist regulatory bodies in tracking media compliance systematically, providing automated feedback to improve adherence. To ensure responsible implementation, governments and media organizations should establish clear ethical guidelines that support AI-assisted reporting without restricting journalistic freedom. However, the current monitoring process requires manual review of articles, making comparisons, and tracking changes - a labor-intensive process that rarely happens due to its complexity and resource requirements. Our proposed solution is to develop an automated system capable of collecting suicide-related articles from online sources (by screening and looking for the words suicide, suicide attempt and suicidal behavior, in the titles but also body texts of newspapers) and evaluating their compliance with WHO guidelines. This automation would enable us to generate a standardized index, allowing for both national and international comparisons. This system could assign each country a compliance score (ranging from 0-15) based on the average compliance of all relevant articles published within that country. The system would operate automatically and be language-independent, making it truly global in scope. By implementing such a measurement system, we could address one of the fundamental issues in improving traditional media coverage of suicide: the lack of systematic monitoring and comparison. Nevertheless, differences in journalistic practices across countries may also impact AI reliability and should be considered. For example, some countries have strict media regulations regarding suicide reporting (e.g., South Korea [29]), while others allow greater editorial freedom (e.g., India [30]), leading to variations in how suicide is framed in news reports. Additionally, cultural attitudes toward mental health and suicide may influence how journalists present such topics (e.g., current debate in India in the interpretation of suicide being

https://doi.org/10.1192/j.eurpsy.2025.10037 Published online by Cambridge University Press

punishable [30]), affecting AI models trained on global datasets. These factors suggest that AI tools may require further fine-tuning to adapt to country-specific journalistic norms, ensuring that adherence evaluations remain accurate across diverse reporting styles. However, our prompt has already demonstrated strong accuracy in evaluating traditional media from three different languages and countries, suggesting its robustness across various cultural contexts. Further refinements can enhance its adaptability, but its current performance indicates potential for broad application.

Our study has several limitations. While it concentrated on traditional media articles, it did not examine news shared on social networks, television serials or films, which host a substantial volume of reports. This study focused solely on textual content analysis and did not include evaluation of images accompanying media reports. This limitation stemmed from the limited capabilities of AI models in image processing at the time of the research and the absence of images in all examined articles. With recent technological advancements in models such as Claude 3.7 Sonnet and GPT-4.5, we are currently developing follow-up research specifically focused on analyzing visual aspects in media reports on suicide. This omission highlights a promising avenue for future research. While no prior automated methods have specifically assessed adherence to WHO guidelines, not allowing us to compare AI models with existing content analysis techniques, future research could perform such comparison to further evaluate their strengths and limitations. Additionally, the evaluators in this study came from diverse educational backgrounds; however, all of them received standardized criteria, specialized training on the topic, and guidance from a senior researcher in the field. Another limitation is the lower agreement between AI model predictions and human ratings for English articles compared with French and Hebrew articles. As mentioned before, this discrepancy may be explained by the fact that the individuals who rated the English articles were not native English speakers, whereas native speakers rated the French and Hebrew articles. This finding suggests

https://doi.org/10.1192/j.eurpsy.2025.10037 Published online by Cambridge University Press

that future assessments of English-language articles would benefit from the ratings provided by native English speakers to enhance their accuracy. However, it is important to note that the overall reliability of the study remains robust, as the agreement levels across all languages, including English, were sufficient to support the validity of the findings. Furthermore, the results indicate that the AI models can evaluate adherence to WHO guidelines consistently, regardless of minor variations in human rater performance. Despite these limitations, our study demonstrates a significant strength: a high alignment between AI models predictions and human ratings across all comparison methods. We evaluated this agreement using Intraclass Correlation Coefficients (ICC), Spearman correlations, and comparisons of global means. In each case, the AI models displayed strong accuracy relative to the human ratings.

While our findings demonstrate that LLMs can replicate human judgment in assessing adherence to WHO suicide reporting guidelines, it is essential to acknowledge the broader limitations of AI in mental health applications. AI models, including LLMs, rely on statistical language processing rather than true comprehension. As highlighted by Tononi & Raison (2024) [31], there is an ongoing debate about whether AI can ever possess human-like understanding or subjective awareness, with theories such as Integrated Information Theory (IIT) arguing that AI lacks the neural structures necessary for genuine consciousness. This distinction is particularly relevant in sensitive areas like suicide prevention, where human expertise remains critical for interpreting nuanced contexts and ethical considerations. Beyond issues of comprehension, generative AI models also raise important challenges related to privacy, reliability, and integration into mental health systems. While AI has the potential to enhance healthcare workflows and support tasks such as screening and risk assessment, concerns remain regarding data security, AI biases, and the risk of over-reliance on models that lack clinical validation [32]. The application of AI in mental health must therefore be accompanied by rigorous oversight, regulatory safeguards, and a complementary role for

https://doi.org/10.1192/j.eurpsy.2025.10037 Published online by Cambridge University Press

human professionals. This integration should be approached with caution and supported by empirical evidence to ensure both safety and effectiveness. These considerations are particularly relevant to our study, as AI-driven assessments of traditional media reports should be used to support rather than replace expert human evaluation since nuanced human interpretation remains essential. Additionally, AI misclassification poses a significant risk, as incorrect assessments may lead to harmful media reports being mistakenly deemed compliant or responsible articles being unnecessarily flagged. Such errors could reduce journalists' trust in AI-driven evaluations and, at scale, hinder suicide prevention efforts rather than support them. To mitigate these risks, AI models should always be used as an assistive tool rather than a replacement for expert human review, particularly in cases where guideline adherence is ambiguous or context dependent. Furthermore, as AI continues to be integrated into mental health applications, regulatory frameworks such as the WHO's "Key AI Principles" and the EU Artificial Intelligence Act (2024) [33,34] provide critical guidelines for ensuring transparency, accountability, and ethical AI deployment. These regulations emphasize the need for human supervision, fairness, and privacy protection, which are essential when applying AI in sensitive areas such as suicide prevention. Recent discussions, such as those by Elyoseph et al. [20], highlight the risks associated with AI's role in mental health, particularly its impact on human relationships and emotional well-being.

Improving traditional media adherence to WHO guidelines is crucial for preventing suicidal behaviors in the general population. Developing tools to facilitate adherence is a way to enhance compliance. Our results highlight the effectiveness of AI models in replicating human judgment across different languages and contexts. Therefore, the use of AI models can help assess and improve traditional media adherence to WHO guidelines. However, AI still faces limitations, particularly in identifying subtle linguistic nuances and adapting to regional variations in journalistic practices. Overcoming these challenges will require ongoing

https://doi.org/10.1192/j.eurpsy.2025.10037 Published online by Cambridge University Press

refinement of AI models and sustained human oversight, both of which are essential to ensuring the reliability of AI-assisted evaluations. Collaboration between technology and human expertise will be key.

COMPETITING INTERESTS

The authors declare no conflicts of interest related to this study.

FUNDING

This study did not receive any funding from any sources.

ACKNOWLEDGEMENTS

Authors of this manuscript thanks the students that rated the media's papers on suicide: Emma Sebti, Manon Malestroit, Tal Szpiler, Eden Ben Siimon, Gal Shemo.

AUTHOR CONTRIBUTIONS

Pr. Z. Elyoseph designed the prompt used in Claude Opus 3 and GPT-4O, contributed to the design of the study, to the supervision of the students that evaluate the articles and to the writing of the manuscript. Dr. B. Nobile contributed to the design of the study, to the supervision of the students that evaluate the articles and to the writing of the manuscript. Dr. I. Levkovich contributed to the writing of the manuscript. Dr. R. Chancel contributed to the supervision of the students that evaluate the articles. Pr. P. Courtet contributed to the supervision of the study and to the writing of the manuscript. Pr. Y. Levi-Belz contributed to the design of the study, the creation of the prompt used in AIs models, supervision of the study and to the writing of the manuscript. All authors have contributed to the manuscript and have accepted the final version of the paper.

Data Availability Statement

https://doi.org/10.1192/j.eurpsy.2025.10037 Published online by Cambridge University Press

The prompt used for AIs models is available in the supplementary materials file. On demand we can send articles used for this study as well as scores to the WHO guidelines found by human raters.

REFERENCES

1. An S, Lim S, Kim HW, Kim HS, Lee D, Son E, et al. Global prevalence of suicide by latitude: A systematic review and meta-analysis. Asian J Psychiatr 2023;81:103454.

2. Altavini CS, Asciutti APR, Solis ACO, Wang YP. Revisiting evidence of primary prevention of suicide among adult populations: A systematic overview. Journal of Affective Disorders 2022;297:641–56.

3. Mulder R. Suicide prevention: Time to change the paradigm. Aust N Z J Psychiatry 2020;54:559–60.

4. Sinyor M, Schaffer A. What would cardiology do? Lessons from other medical specialties should help guide suicide prevention research. Aust N Z J Psychiatry 2020;54:568–70.

5. Sufrate-Sorzano T, Di Nitto M, Garrote-Cámara ME, Molina-Luque F, Recio-Rodríguez JI, Asión-Polo P, et al. Media Exposure of Suicidal Behaviour: An Umbrella Review. Nursing Reports 2023;13:1486–99.

6. Zalsman G, Hawton K, Wasserman D, van Heeringen K, Arensman E, Sarchiapone M, et al. Suicide prevention strategies revisited: 10-year systematic review. Lancet Psychiatry 2016;3:646–59.

7. Vijayakumar L, Shastri M, Fernandes TN, Bagra Y, Pathare A, Patel A, et al. Application of a Scorecard Tool for Assessing and Engaging Media on Responsible Reporting of Suicide-Related News in India. IJERPH 2021;18:6206.

8. Pirkis J, Rossetto A, Nicholas A, Ftanou M, Robinson J, Reavley N. Suicide Prevention Media Campaigns: A Systematic Literature Review. Health Commun 2019;34:402–14.

9. Asharani P, Koh YS, Tan RHS, Tan YB, Gunasekaran S, Lim B, et al. The impact of media reporting of suicides on subsequent suicides in Asia: A systematic review. Ann Acad Med Singap 2024;53:152–69.

10. Ishimo MC, Sampasa-Kanyinga H, Olibris B, Chawla M, Berfeld N, Prince SA, et al. Universal interventions for suicide prevention in high-income Organisation for Economic Co-operation and Development (OECD) member countries: a systematic review. Inj Prev 2021;27:184–93.

11. Niederkrotenthaler T, Braun M, Pirkis J, Till B, Stack S, Sinyor M, et al. Association between suicide reporting in the media and suicide: systematic review and meta-analysis. BMJ 2020;368:m575.

https://doi.org/10.1192/j.eurpsy.2025.10037 Published online by Cambridge University Press

12. Niederkrotenthaler T, Voracek M, Herberth A, Till B, Strauss M, Etzersdorfer E, et al. Role of media reports in completed and prevented suicide: Werther v. Papageno effects. Br J Psychiatry 2010;197:234–43.

13. World Health Organization. Preventing suicide: a resource for media professionals, update 2017. Geneva: World Health Organization; [Internet]. 2017;Available from: https://www.who.int/publications/i/item/9789240076846

14. Levi-Belz Y, Starostintzki Malonek R, Hamdan S. Trends in Newspaper Coverage of Suicide in Israel: An 8-Year Longitudinal Study. Arch Suicide Res 2023;27:1191–206.

15. Shinan-Altman S, Elyoseph Z, Levkovich I. Integrating Previous Suicide Attempts, Gender, and Age Into Suicide Risk Assessment Using Advanced Artificial Intelligence Models. J. Clin. Psychiatry [Internet] 2024 [cited 2024 Oct 21];85. Available from: http://www.psychiatrist.com/jcp/Suicide-Risk-Evaluation-Advanced-AI-ChatGPT

16. Shinan-Altman S, Elyoseph Z, Levkovich I. The impact of history of depression and access to weapons on suicide risk assessment: a comparison of ChatGPT-3.5 and ChatGPT-4. PeerJ 2024;12:e17468.

17. Elyoseph Z, Levkovich I, Shinan-Altman S. Assessing prognosis in depression: comparing perspectives of AI models, mental health professionals and the general public. Fam Med Community Health 2024;12:e002583.

18. Elyoseph Z, Hadar-Shoval D, Asraf K, Lvovsky M. ChatGPT outperforms humans in emotional awareness evaluations. Front Psychol 2023;14:1199058.

19. Levkovich I, Elyoseph Z. Suicide Risk Assessments Through the Eyes of ChatGPT-3.5 Versus ChatGPT-4: Vignette Study. JMIR Ment Health 2023;10:e51232.

20. Elyoseph Z, Levkovich I. Beyond human expertise: the promise and limitations of ChatGPT in suicide risk assessment. Front Psychiatry 2023;14:1213141.

21. Kirtley OJ, Van Mens K, Hoogendoorn M, Kapur N, De Beurs D. Translating promise into practice: a review of machine learning in suicide research and prevention. The Lancet Psychiatry 2022;9:243–52.

22. Holmes G, Tang B, Gupta S, Venkatesh S, Christensen H, Whitton A. Applications of Large Language Models in the Field of Suicide Prevention: Scoping Review. J Med Internet Res 2025;27:e63126.

23. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature 2023;620:172–80.

24. Guo Y, Ovadje A, Al-Garadi MA, Sarker A. Evaluating large language models for health-related text classification tasks with public social media data. J Am Med Inform Assoc 2024;31:2181–9.

https://doi.org/10.1192/j.eurpsy.2025.10037 Published online by Cambridge University Press

554 25. Huang J, Yang DM, Rong R, Nezafati K, Treager C, Chi Z, et al. A critical assessment of
555 using ChatGPT for extracting structured data from clinical notes. NPJ Digit Med
556 2024;7:106.

557 26. Elyoseph Z, Levkovich I, Rabin E, Shemo G, Szpiler T, Shoval DH, et al. Applying
558 Language Models for Suicide Prevention: Evaluating News Article Adherence to WHO
559 Reporting Guidelines [Internet]. 2024 [cited 2024 Oct 21];Available from:
560 https://www.researchsquare.com/article/rs-4180591/v1

561 27. Bastiampillai T, Allison S, Perry SW, Licinio J. Social network theory and rising suicide
562 rates in the USA. The Lancet 2019;393:1801.

563 28. Pompili M. The increase of suicide rates: the need for a paradigm shift. The Lancet
564 2018;392:474–5.

565 29. Kang DH, Marques AH, Yang JH, Park CHK, Kim MJ, Rhee SJ, et al. Suicide prevention
566 strategies in South Korea: What we have learned and the way forward. Asian J Psychiatr
567 2025;104:104359.

568 30. Vijayakumar L, Chandra PS, Kumar MS, Pathare S, Banerjee D, Goswami T, et al. The
569 national suicide prevention strategy in India: context and considerations for urgent action.
570 Lancet Psychiatry 2022;9:160–8.

571 31. Tononi G, Raison C. Artificial intelligence, consciousness and psychiatry. World
572 Psychiatry 2024;23:309–10.

573 32. Torous J, Blease C. Generative artificial intelligence in mental health care: potential
574 benefits and current challenges. World Psychiatry 2024;23:1–2.

575 33. World Health Organization. Guidance on ethics and governance of artificial intelligence
576 for health. 2024;

577 34. European Union. Regulation (EU) 2024/XXX of the European Parliament and of the
578 Council on Artificial Intelligence (Artificial Intelligence Act) [Internet]. 2024;Available
579 from: https://www.artificial-intelligence-act.com/

580

581

https://doi.org/10.1192/j.eurpsy.2025.10037 Published online by Cambridge University Press

582 **Table 1. ICC (95%CI) and Spearman correlation between human evaluators and AI**
583 **models (n=120).**

	Human evaluators	
	ICC (95%CI)	Spearman
Claude Opus 3	0.724 (0.605; 0.808)	0.636 p<0.001
GPT-40	0.793 (0.702; 0.855)	0.684 p<0.001
Both	0.812 (0.731; 0.869)	0.703 p<0.001

584
585
586
587
588
589

590 **Table 2: Comparison of human and AI evaluations for a single article**

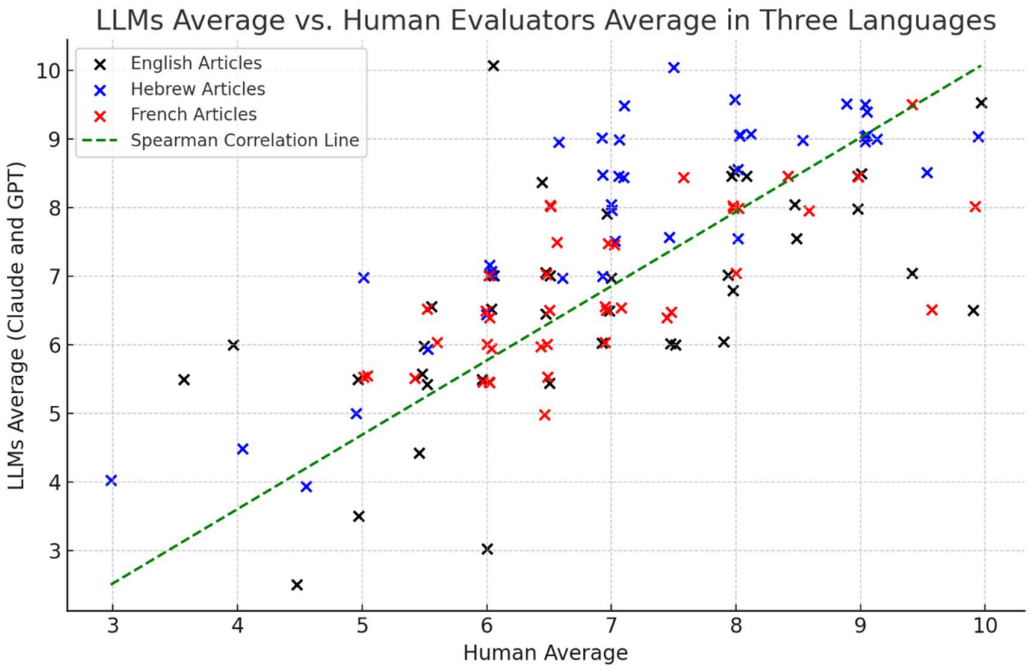
WHO Guideline Criterion	Human Rater 1	Human Rater 2	GPT-4O	Claude 3
(1) Is suicide mentioned in the headline?	0	0	1	1
(3) Is the person who died by suicide described as a celebrity?	0	0	0	0
(4) Does the article report on a single cause for suicide/suicidal behavior?	1	1	1	0
(5) Does the article imply a link between a specific life event and suicide/suicidal behavior?	1	1	1	0
(6) Does the article imply a link between social status and suicide/suicidal behavior?	0	1	1	1
(7) Does the article imply a link between mental state and suicide/suicidal behavior?	1	1	1	1
(9) Does the story present any myths about suicide/suicidal behavior?	0	0	0	0
(10) Does the story include glorifying descriptions of suicide/suicidal behavior?	0	0	0	0
(11) Is the method of suicide/suicidal behavior described in detail?	0	0	0	0
(12) Does the story describe the location of suicide/suicidal behavior?	0	0	0	0
(13) Does the story not inform the reader about warning signs for suicide/risk factors?	0	0	0	0
(14) Does the story not include any information about prevention?	0	1	0	0

https://doi.org/10.1192/j.eurpsy.2025.10037 Published online by Cambridge University Press

WHO Guideline Criterion	Human Rater 1	Human Rater 2	GPT-4O	Claude 3
(15) Does the story not include any information about intervention?	0	1	0	0
Total violations	**3**	**6**	**5**	**3**

Note: 1= adhere to the criterion, 0= not adhere to the criterion. Items are numbered according to the original WHO criteria numbering system. Items 2 (front page placement) and 8 (inappropriate images) were excluded from our analysis as explained in the Methods section.

https://doi.org/10.1192/j.eurpsy.2025.10037 Published online by Cambridge University Press



596

597

598 Figure 1: Average evaluations of large language models (LLMs) with human evaluators across
599 three languages: English (black × marks), Hebrew (blue × marks), and French (red × marks).
600 Notes: Each point represents an individual article evaluated by both human evaluators and
601 language models (Claude and GPT). The x-axis shows human average ratings (scale 1-10),
602 while the y-axis shows LLMs average ratings (scale 1-10). The green dashed line indicates
603 Spearman's correlation coefficient between these averages, demonstrating the overall alignment
604 between human and AI judgments across all three languages.

605

606

607

https://doi.org/10.1192/j.eurpsy.2025.10037 Published online by Cambridge University Press

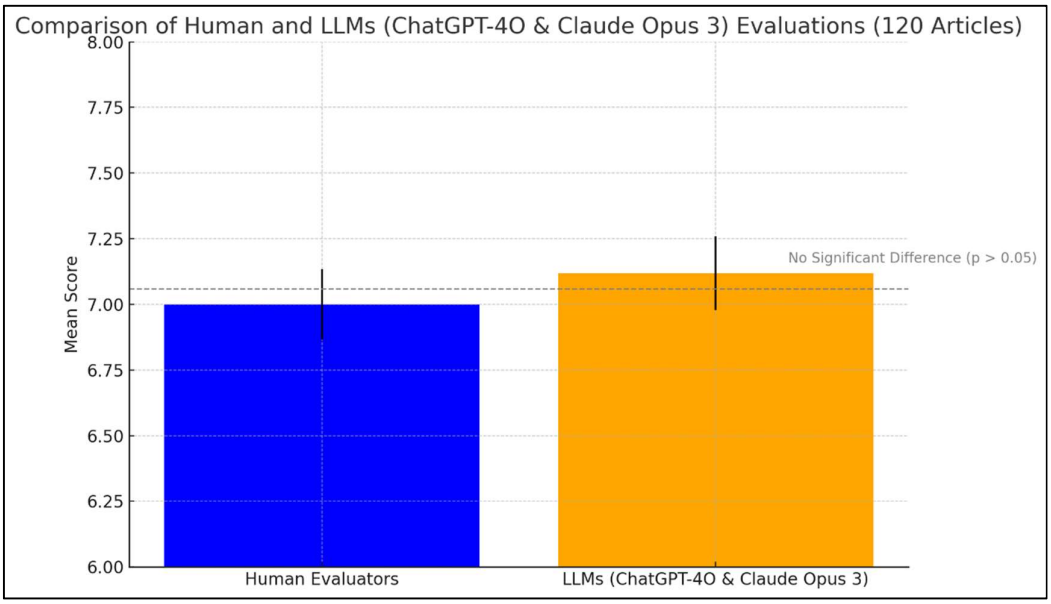


Figure 2: Comparison of mean scores between human evaluators and LLMs (ChatGPT-4O and Claude Opus 3) across 120 articles.
Notes: The bar chart illustrates that there was no significant difference in the evaluations between the two groups ($p > 0.05$). Error bars represent standard error of the mean.

https://doi.org/10.1192/j.eurpsy.2025.10037 Published online by Cambridge University Press