Computers in Human Behavior Reports 23 (2026) 101189

Contents lists available at ScienceDirect

Computers in Human Behavior Reports

journal homepage: www.sciencedirect.com/journal/computers-in-human-behavior-reports

# A scalable AI-Based system for evaluating and enhancing responsible media coverage of suicide: A multi-site, multi-language implementation study

Y. Levi-Belz [a,e,*], B. Nobile [b,e], I. Levkovich [d], P. Courtet [b,c], Z. Elyoseph [e]

[a] The Lior Tsfaty Center for Suicide and Mental Pain Studies, University of Haifa, Israel
[b] Department of Emergency Psychiatry and Acute Care, CHU Montpellier, France
[c] IGF, Univ. Montpellier, CNRS, INSERM, Montpellier, France
[d] Faculty of Education, Tel Hai University of Kiryat Shmona in the Galilee, Israel
[e] School of Therapy, Counseling and Human Development, University of Haifa, Israel

ARTICLE INFO

*Keywords:*
Suicide prevention
Media guideline adherence
Generative artificial intelligence
Large language models
Werther effect

ABSTRACT

*Background:* Irresponsible media coverage of suicide can increase suicide rates through the "Werther effect," while responsible reporting can reduce risk and stigma. Although the WHO has issued clear guidelines for responsible reporting, adherence remains inconsistent worldwide due to structural barriers and the lack of real-time tools for journalists. This study aimed to develop and rigorously evaluate a multilingual GenAI-based system that can automatically assess and enhance suicide-related news articles to align with WHO guidelines, offering a scalable pathway for more responsible coverage.

*Methods:* A total of 120 suicide-related articles (English, Hebrew, and French) were systematically selected and independently evaluated for guideline adherence by both expert human raters and the AI system. The articles were automatically revised using the AI correction module. Post-enhancement, all articles were re-evaluated by the system and by a new set of human raters to compare changes in adherence levels.

*Results:* The AI system demonstrated high agreement with human raters at baseline (ICC = .77), with slight conservative bias in some languages. Following the AI-based enhancement bot, both AI and human evaluations confirmed significant improvements in guideline adherence (mean scores increased from ~45% to ~85%), with no significant post-enhancement differences between human and AI ratings.

*Conclusions:* This study provides initial empirical support for using GenAI to automatically evaluate and enhance suicide-related media coverage across languages and contexts. By offering journalists a practical, real-time tool for producing accurate, ethically sound content, this approach can help mitigate suicide contagion, increase public awareness, and foster a global metric for responsible reporting.

## 1. Introduction

Suicide remains one of the most urgent global public health crises, with over 800,000 deaths annually and millions more attempts (World Health Organization, 2023). Beyond the human toll, suicide incurs substantial economic costs, both medical and societal, along with profound emotional suffering for those left behind (Shepard et al., 2016). Despite major prevention efforts, such as the "Zero Suicide" framework (Brodsky et al., 2018), progress remains limited and suicide rates continue to rise in many regions (e.g., Pompili, 2020). These trends highlight the urgent need for scalable, cost-effective, and innovative solutions for transforming the current prevention landscape.

### 1.1. Suicide and media coverage

It is well known that media plays a crucial role in shaping public perception and behavior regarding sensitive issues like suicide (e.g., Sisask & Värnik, 2012). Since Phillips (1974) introduced the concept of the "*Werther effect*"—referring to increased suicide rates following media coverage of suicide, particularly of celebrities—numerous studies have demonstrated that irresponsible newspaper reporting can lead to increased suicide rates across various countries and cultures (Stack, 1987; Chen et al., 2011; Yip et al., 2016).

Several comprehensive systematic reviews and meta-analysis (e.g., Niederkrotenthaler et al., 2020) revealed that media reports of celebrity suicide are associated with significant increase in total suicides in

* Corresponding author. The Lior Tsfaty Center for Suicide and Mental Pain Studies, University of Haifa, Israel.
*E-mail address:* Yossil@edu.haifa.ac.il (Y. Levi-Belz).

https://doi.org/10.1016/j.chbr.2026.101189
Received 30 November 2025; Received in revised form 25 June 2026; Accepted 27 June 2026

subsequent months. Furthermore, when specific suicide methods are reported, the rate of suicide using the cited method increases by 30% (Niederkrotenthaler et al., 2020). When the coverage of a suicide story is sensationalized or romanticized along with explicit location details, the suicide rate increases (Pirkis et al., 2006; Hawley et al., 2023). By contrast, responsible media coverage can reduce suicide risk by raising awareness and challenging stigma (World Health Organization, 2017). For example, stories portraying individuals who overcome suicidal crises, the "Papageno effect", have been linked to lower suicidal ideation. Several randomized controlled trials and meta-analyses have demonstrated that such content promotes help-seeking and reduces suicide risk (Niederkrotenthaler et al., 2010, 2020, 2022).

In response to the well-established impact of the media on suicide rates, the World Health Organization (WHO) and other public health bodies have issued clear guidelines for responsible suicide reporting. These include avoiding sensationalism, omitting explicit details of methods or locations, and refraining from romanticized or dramatic portrayals of suicide (Levi-Belz et al., 2023; Sinyor et al., 2025). Instead, they encourage framing suicide within broader psychosocial contexts, offer helpline resources, and highlighting recovery stories (Niederkrotenthaler et al., 2022). Adherence to these practices has been proven to be a cost-effective and impactful strategy in suicide prevention efforts worldwide (Niederkrotenthaler et al., 2007; Fu and Yip, 2008).

Despite widespread dissemination, compliance remains uneven. Some studies have shown progress following journalist training (Wu et al., 2018), while others have highlighted persistent gaps (Pirkis et al., 2019; McTernan et al., 2018; Tatum et al., 2010). A recent longitudinal study in Israel found that nearly half of suicide-related articles did not comply with WHO guidelines (Levi-Belz et al., 2023), underscoring the need for more proactive and scalable solutions that provide real-time feedback and promote sustained accountability.

Importantly, the contemporary suicide-related information environment extends beyond traditional news outlets. Digital and social media platforms have become central spaces in which individuals encounter, share, and interpret mental-health and suicide-related content. Recent studies suggest that problematic or addictive patterns of social media use are associated with poorer mental health and suicide-related outcomes. For example, longitudinal work has shown that Facebook Addiction Disorder is positively associated with suicide-related outcomes one year later, with positive mental health playing an important mediating role (Brailovskaia et al., 2020). More recently, the "Vicious Circle of addictive Social Media Use and Mental Health" model described how addictive social media use and psychological distress may mutually reinforce one another over time (Brailovskaia, 2024). In parallel, computational approaches have begun to examine whether machine-learning models can detect potentially harmful and protective suicide-related content in social media environments, further supporting the feasibility of scalable, technology-assisted approaches to suicide-prevention communication (Metzler et al., 2022). Although the present study focuses on suicide-related news articles rather than user-generated social media content, this broader literature highlights the growing importance of digital information environments in suicide prevention and the need for scalable, guideline-based tools for identifying and improving potentially harmful communication.

### 1.2. Generative AI and suicide prevention

In recent years, generative AI (GenAI) has emerged as a powerful tool for suicide prevention. Large language models (LLMs) such as ChatGPT, Claude, and Gemini have shown promising capabilities in emotion recognition, clinical decision support, and suicide risk assessment, often matching the performance of mental health professionals (e.g. Elyoseph and Levkovich, 2023 Levkovich & Omar, 2024; Elyoseph et al., 2026). These models are sensitive to key risk factors (e.g., age, gender, depression, prior attempts), demonstrate cross-cultural effectiveness, and offer real-time support to individuals in crisis (Shinan-Altman et al., 2024; Coppersmith et al., 2024). GenAI technologies have also been successfully used in training settings, including gatekeeper training using AI-based QPR simulations (Levkovich et al., 2025) and suicide risk assessment modules (Elyoseph, Levkovich, Rabin, et al., 2025). Taken together, these studies highlight GenAI's ability to expand access to clinical knowledge, enable the rapid deployment of culturally adapted interventions, and deliver scalable, cost-effective prevention strategies.

Recently, we conducted a two-stage pilot program on suicide-related media coverage. In the first study (Elyoseph, Levkovich, Haber, & Levi-Belz, 2025), we tested the ability of LLMs (ChatGPT-4 and Claude.AI) to evaluate Hebrew-language articles based on WHO guidelines. The AI's ratings showed strong agreement with human reviewers, demonstrating its potential to detect reporting violations. Building on this, a larger multilingual study (Elyoseph, Levkovich, Haber, & Levi-Belz, 2025) confirmed that GenAI can reliably identify such violations across English, Hebrew, and French, highlighting its scalability for real-time media monitoring. Taken together, these findings lay the groundwork for automated tools that can flag problematic articles and help journalists reduce suicide contagion through safer and more responsible reporting.

However, prior work has focused primarily on the detection and evaluation of reporting violations. It remains unclear whether GenAI can also generate guideline-concordant revisions that preserve the journalistic meaning of the original article while reducing potentially harmful reporting features. Moreover, few studies have tested whether such AI-generated enhancements are recognized as improvements by independent human raters across multiple languages. The present study addresses this gap by evaluating a full AI-supported pipeline: baseline assessment, automated enhancement, and blind post-enhancement re-evaluation by both AI and human judges.

### 1.3. The present study

Responsible media reporting on suicide is a vital yet underutilized public-health strategy because it targets the information environment through which suicide is framed, normalized, stigmatized, or presented as preventable. Building on the complementary logic of the Werther and Papageno effects, the present study conceptualizes responsible reporting not merely as the avoidance of harmful content, but as an active communicative practice that can reduce risk, promote help-seeking, and support more accurate public understanding of suicide.

Although the WHO guidelines aim to reduce suicide contagion and promote preventive narratives, adherence remains inconsistent, often due to newsroom turnover, time constraints, and the absence of real-time editorial feedback. In this study, we introduce and examine a scalable, potentially transformative strategy for suicide prevention: the systematic assessment and enhancement of suicide-related media content using generative AI.

Specifically, we conducted a multilingual, cross-cultural evaluation of a GenAI-based system applied to 120 news articles in English, Hebrew, and French. Each article was first assessed for adherence to WHO guidelines by both the AI system and trained human raters. After automated AI-based enhancement, the articles were reassessed by the system and a new set of independent human judges, offering a rigorous test of the system's ability to detect and improve compliance. By providing immediate, tailored corrections, this approach supports responsible journalism without infringing upon press freedom. The multilingual design enables real-time scalability across diverse cultural contexts. Ultimately, it offers media professionals a practical tool to improve reporting accuracy, reduce suicide risk, and combat stigma globally.

Our hypotheses were:

**Hypothesis 1. (Evaluation accuracy of original articles):** The GenAI-based system will exhibit high agreement with expert human raters in evaluating the adherence of suicide-related news articles to WHO reporting guidelines in all three languages.

**Hypothesis 2. (Enhancement Efficacy and evaluation of revised articles):** The AI-driven enhancement module will significantly improve articles' compliance with WHO guidelines, as confirmed by both the GenAI system and independent human raters during post-enhancement evaluations.

**Hypothesis 3. (Cross-Cultural Feasibility):** The system's performance—across both evaluation and enhancement tasks will remain consistent across English, Hebrew, and French articles, demonstrating its feasibility as a scalable, cross-cultural solution for promoting responsible suicide reporting.

## 2. Method

### 2.1. Study design

This study employed a multi-stage, multi-language experimental design to empirically test and evaluate a GenAI-based system for improving media coverage of suicide. The research framework comprised five core, sequential stages, as illustrated in Fig. 1. Because the primary aim was system evaluation rather than population-level estimation of media adherence, the study used a balanced multilingual corpus with equal numbers of articles in each language.

1. **Data Corpus Identification and Construction:** In the initial stage, a controlled data corpus was established to serve as the foundation for all subsequent analyses. A total of 120 suicide-related news articles were systematically identified and collected: 40 in English, 40 in Hebrew, and 40 in French.
2. **Baseline Evaluation:** Each 120 articles was independently evaluated to quantify the initial adherence to the WHO guidelines and establish an objective baseline. This evaluation was conducted by two parallel teams: a panel of two skilled human raters for each language (native French, Hebrew, and English speakers), and a dedicated AI-based evaluation bot.
3. **AI-based Enhancement:** Following the baseline evaluation, the original articles were processed using a dedicated AI enhancement module. This module operated based on a detailed prompt designed to revise each article to improve its adherence to WHO guidelines while preserving its core journalistic essence and narrative.
4. **Re-evaluation:** After the enhancement stage, the revised versions of the articles were re-evaluated to measure changes in guideline adherence. This was performed under "blind" conditions using both the same AI evaluation system and a new team of human raters who had not been exposed to the original articles, thereby preventing bias.
5. **Data Comparison:** In final stage, a comprehensive statistical analysis was conducted to compare the adherence scores of the original and revised articles. This comparison provided a quantitative measurement of the efficacy of the automated enhancement process.

### 2.2. Data corpus and selection criteria

The dataset includes 120 suicide-related news articles published between 2021 and 2024:40 in English, 40 in Hebrew, and 40 in French. The corpus was constructed specifically for the present multilingual experimental study and was designed to provide a balanced test set across three languages rather than to estimate the prevalence of suicide reporting in any single country or media system.

Sample size per language was determined using an a priori power/sensitivity analysis conducted in G*Power. Because the study had two main inferential goals—estimating human-AI agreement and testing within-article improvement following AI-based enhancement—we examined power for both correlation-based agreement analyses and paired pre-post comparisons. For language-specific analyses, a sample of 40 articles per language provided 80% power at $\alpha = .05$ to detect a correlation of approximately $r = .42$ between human and AI adherence scores, and a paired within-article effect of approximately $dz = .45$ in pre-to-post enhancement analyses. For the full corpus of 120 articles, the study was powered to detect smaller overall effects, approximately $r = .25$ for agreement analyses and $dz = .26$ for paired pre-post comparisons. Thus, 40 articles per language was selected to provide sufficient sensitivity for language-specific analyses while maintaining a balanced multilingual corpus. The sample size was intended for system evaluation and cross-language comparison, rather than for estimating the prevalence of suicide-reporting practices in the broader media ecosystem. Articles were systematically selected from leading newspapers in each language based on their readership and national exposure. Keyword-based searches were conducted in digital archives, with strict exclusion criteria to ensure relevance.

Articles were identified through systematic searches of digital news archives and online news databases from leading, high-readership news outlets in each language context. Source selection was guided by national exposure, readership, availability of searchable digital archives, and relevance to mainstream journalistic reporting. The search strategy followed the logic used in our previous work on suicide-related media coverage, but was adapted to the multilingual design of the present study. Searches were conducted using suicide-related keywords in each language. In English, these included terms such as "suicide," "suicidal behavior," "suicide attempt," "died by suicide," and "ended his/her/their life." In Hebrew, searches included the Hebrew equivalents of "suicide" in masculine, feminine, and plural forms, "suicidal behavior," "suicide attempt," and "ended his/her/their life." In French, searches

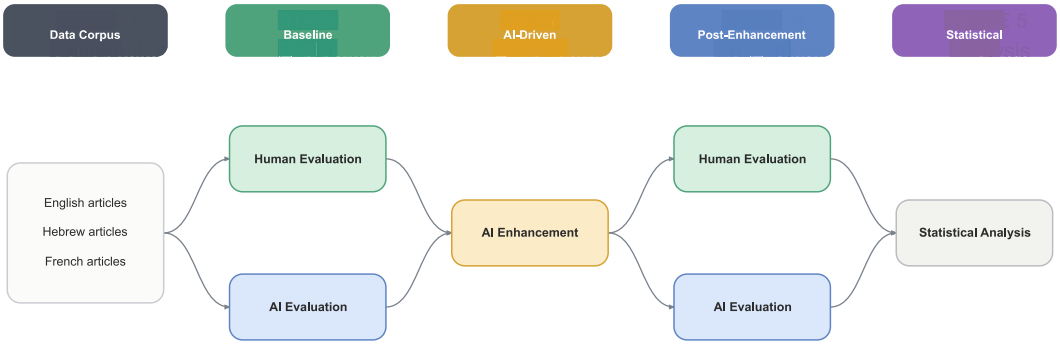


**Fig. 1.** Schematic representation of the study' stages.
*Note.* The figure illustrates the five sequential stages of the study, from data corpus construction (Stage 1), through parallel baseline evaluation (Stage 2), AI-driven enhancement (Stage 3), post-enhancement re-evaluation (Stage 4), and concluding with the final statistical analysis (Stage 5).

included terms such as "suicide," "comportement suicidaire," "tentative de suicide," "s'est donné la mort," and "mettre fin à ses jours."

To be eligible, articles had to focus substantively on suicide, suicide attempt, or suicidal behavior and to be written as journalistic news or feature reporting rather than academic, fictional, or purely opinion-based text. Articles were excluded if suicide-related terms were used only metaphorically or colloquially, if the article referred to suicide terrorism rather than suicidal behavior, if suicide was mentioned only briefly or incidentally, if the article focused only on general risk factors or bereavement without reporting suicide-related behavior, or if the circumstances of death were framed primarily as legally uncertain homicide rather than suicide. Duplicate articles, syndicated repetitions, and near-identical versions of the same story were removed.

After screening, eligible articles within each language were reviewed for relevance and clarity. From the eligible pool, 40 articles per language were selected to create a balanced corpus for cross-language comparison. When more than 40 eligible articles were available in a given language, articles were selected to ensure variation in publication year, outlet, article type, and reporting characteristics. This balancing procedure was intended to avoid overrepresentation of a single outlet, year, or highly similar suicide-related event. The final corpus therefore consisted of 120 unique articles, equally distributed across English, Hebrew, and French.

The present article-selection strategy differs from a prevalence-oriented content analysis. Its purpose was not to estimate the total volume of suicide-related reporting in each language, but to create a controlled multilingual corpus suitable for evaluating whether a GenAI-based system could assess and enhance adherence to WHO suicide-reporting guidelines across languages. For more detailed methodology and selection procedures, please refer to our previous study (Elyoseph, Levkovich, Haber, & Levi-Belz, 2025).

### 2.3. Article evaluation criteria

The evaluation of the articles was based on the WHO framework for responsible media reporting on suicide. This study utilized 13 of the original 15 parameters (e.g., Levi-Belz et al., 2023; Sinyor et al., 2025). Two items, concerning the placement of the article on the front page and the use of inappropriate images, were excluded from the analysis due to the current technological limitations of language models in analyzing visual content and graphic-journalistic context. Each article was assessed on a three levels basis (0 = does not adhere to the guideline; .5 = partly adheres to the guideline; 1 = adheres to the guideline) for each of the 13 criteria presented in supplement 1. This evaluation framework was applied uniformly and consistently by all raters, both human and artificial, before and after the enhancement stage to ensure a valid basis for comparison.

### 2.4. The GenAI system and language model

At the core of this study was a dual-component AI system developed through advanced prompt engineering by an expert in the field (Z.E.). The two components, evaluation bot and correction bot, were powered by the same language model. The evaluation and correction bots were powered by Claude 3.5 Sonnet, an advanced LLM developed by Anthropic. This model was selected for its high accuracy, speed, and efficiency, doubling the processing speed of the previous versions. Known for its nuanced understanding of language and complex instructions, Claude 3.5 Sonnet sets a new benchmark in language comprehension and is well suited to the analytical and enhancement tasks required in this study.

### 2.5. Evaluation and enhancement bots

The evaluation bot was specifically designed to evaluate the adherence of journalistic articles to the 13 WHO Health Organization criteria. It was used for both baseline evaluation and re-evaluation after correction, thereby serving as a consistent measurement tool throughout the study. The prompt guiding its operation was written entirely in English, applied to all articles across the three languages, and was based on advanced engineering techniques to ensure maximum accuracy:

- **Role Assignment:** The bot was instructed to adopt a dual persona of a "professor of psychology and communications" and a "newspaper editor." This definition was intended to ensure a sensitive and deep analysis combining clinical knowledge with an understanding of journalistic standards.
- **Chain-of-Thought and Few-Shot Learning:** The prompt explicitly instructed the bot to work systematically, step-by-step, for each of the 13 criteria. For each criterion, the prompt included clear definitions and examples illustrating when to mark a violation and when not to mark it, thereby calibrating the model's judgment.

All system runs were performed with a temperature setting of 0 to ensure that the results were deterministic and fully reproducible.

The enhancement bot was designed to revise suicide-related news articles in line with WHO guidelines. Its output was reviewed and approved by a panel of three experts on suicide prevention and media reporting. The process followed a structured prompt to ensure ethically sound corrections in both content and tone.

- **Systematic Identification:** The prompt instructed the bot to scan the original article and identify every violation of the 13 criteria.
- **Targeted Correction Application:** For violation identified, the prompt provided a built-in "rulebook" with specific correction instructions. For example, if an article included a detailed description of the suicide method, the instruction was to replace the description with general language and add a sentence directing readers to seek help.
- **Preservation of Journalistic Structure:** A key instruction in the prompt was to maintain the original article's length and journalistic narrative flow as much as possible. The goal was not to delete information, but to adapt its phrasing.
- **Generation of a Change Log:** At the end of the correction process, the bot was instructed to generate a report detailing all changes made, along with a brief clinical explanation of why each change was essential.

### 2.6. Human benchmark

The human evaluation process served as the gold standard for comparing the performance of AI. The human raters were psychology graduate students or post-graduates, who underwent comprehensive training by an expert in the field (Y.L.B.). The training process included evaluating sample articles until the raters achieved a high level of proficiency and good inter-rater agreement.

- **Pre-enhancement Evaluation:** In phase, six raters (two for each language) evaluated 120 original articles.
- **Post-enhancement Evaluation:** Three raters (one for each language) evaluated 120 corrected articles. A new set of raters was used at the post-enhancement stage to reduce memory effects and expectation bias that could arise if raters evaluated both the original and revised versions of the same articles.

All stages of human evaluation were conducted under "blind" conditions, meaning the raters were unaware of which study phase they were participating in to prevent bias.

### 2.7. Statistical analysis

The statistical analysis was conducted to quantitatively examine the

system's reliability and efficacy. The significance level for all tests was set at $p < .05$. To assess the reliability and agreement among different raters (human-human and human-AI), the Intraclass Correlation Coefficient (ICC) was used. ICC values below .5 are considered poor agreement, values between .5 and .75 are moderate, values between .75 and .9 are good, and values between .9 are excellent agreement. The choice of ICC was based on the study's aim of estimating agreement between human and AI adherence ratings, whereas paired-sample tests were used to examine within-article changes from baseline to post-enhancement. Wilcoxon signed-rank tests were included as non-parametric sensitivity analyses to ensure that the conclusions were not dependent on distributional assumptions. In addition, Spearman correlation was calculated to examine the degree of alignment in the ranking of articles between human and artificial systems. To evaluate the efficacy of the correction, the average adherence scores before and after the intervention were compared using a paired sample *t*-test. In parallel, the Wilcoxon signed-rank test was conducted as a non-parametric alternative. All analyses were conducted using IBM SPSS Statistics, version 29. The significance level for all statistical tests was established at $p < .05$.

### 2.8. Code availability

The underlying code for this study is not publicly available but may be made available to qualified researchers on reasonable request from the corresponding author.

## 3. Results

### 3.1. Inter-rater reliability of human raters in evaluation stage

To ensure the stability and consistency of the human evaluation, which serves as the gold standard in this study, the inter-rater reliability between the two primary human raters were assessed for the 40 original articles in each language.

The ICC was calculated using a two-way mixed-effects model for consistency. Across the entire sample (N = 120), the analysis revealed excellent agreement, with an ICC for average measures of .838 (95% CI [.768, .887], $F$ (119, 119) = 6.185, $p < .001$). This result signifies a high degree of consistency in the scores assigned by the two human raters in each language. This strong reliability was further examined within each language sub-sample. The agreement was excellent for Hebrew articles (ICC = .915, $p < .001$) and French articles (ICC = .872, $p < .001$), and good for English articles (ICC = .759, $p < .001$). These findings confirm that the human evaluation process was robust and reliable across all languages, providing a solid benchmark for comparison with the AI system.

### 3.2. Reliability between AI evaluation bot and human raters

Following the establishment of human rater reliability, the agreement between the GenAI evaluation bot's baseline scores and average human scores was assessed. As shown in Fig. 2, the ICC indicated good agreement across the entire sample (N = 120), with an average ICC of .774 (95% CI [.676, .843], $p < .001$). When analyzed by language, the agreement was good for Hebrew (ICC = .799) and English (ICC = .822) and moderate for French (ICC = .652).

A Pearson correlation analysis further supported this finding, with a strong, significant positive correlation for the entire sample ($r = .671$, $p < .001$, n = 120). This correlation remained strong and significant for Hebrew ($r = .703$, $p < .001$, n = 40) and English ($r = .738$, $p < .001$, n = 40), and moderate and significant for French ($r = .497$, $p = .001$, n = 40). Moreover, to examine whether the GenAI system systematically differed from human raters in its scoring at baseline, a repeated-measures ANOVA was conducted. There was a statistically significant difference between the mean scores of human raters (M = 6.02, SD = 1.43) and the AI bot (M = 5.43, SD = 2.02), F (1, 119) = 18.368, p <

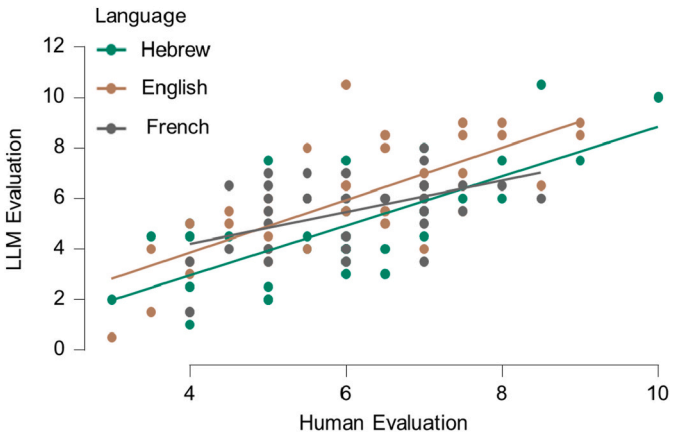


**Fig. 2.** Correlation Between Human and LLM Evaluations of WHO Guidelines Compliance by Language.
*Note.* Scatter plot showing the relationship between human evaluator scores (x-axis) and LLM (Large Language Model) evaluator scores (y-axis) for WHO guidelines compliance assessment across three languages: Hebrew (green), English (brown), and French (gray). Each point represents an individual article's paired evaluation scores. Trend lines indicate the linear relationship between human and LLM assessments for each language group. The plot demonstrates the degree of agreement between human and artificial intelligence evaluations of suicide-related news articles' adherence to WHO reporting guidelines (N = 120, 40 articles per language).

.001, with a large effect size ($\eta p^2 = .134$). This difference was driven by a very large effect in Hebrew articles ($\eta p^2 = .346$) and a large effect in French articles ($\eta p^2 = .164$), while the effect in English articles was negligible ($\eta p^2 = .003$). Thus, despite a slight tendency for AI to score more conservatively than humans in two of the languages, the strong correlations and good overall agreement demonstrate that the AI bot was well calibrated to the human evaluation standard.

### 3.3. Efficacy of the system enhancement and improving stage

***Efficacy of AI Enhancement: GenAI ratings.*** The efficacy of the AI correction bot was evaluated from the perspective of the AI evaluation bot. A Shapiro-Wilk test confirmed the normality of the baseline AI scores ($W$ (120) = .986, $p = .239$). A repeated measures ANOVA was then conducted to compare the AI scores before and after the correction, to all the articles, and to the articles in each language separately.

As shown in Fig. 3, a profound and statistically significant improvement was observed. The mean adherence score increased from $M = 5.43$ ($SD = 2.02$) to $M = 10.88$ ($SD = 1.68$), respectively. This improvement was highly significant, $F$ (1, 119) = 517.105, $p < .001$, with an extremely large effect size ($\eta p2 = .813$). In practical terms, this represents an increase in the articles' average adherence to WHO guidelines, as judged by AI, from 41.8% to 83.7%. This substantial effect was consistently observed and very large across all three languages: Hebrew ($\eta p2 = .898$), English ($\eta p2 = .689$), and French ($\eta p2 = .870$), with all changes being highly significant ($p < .001$). This demonstrates that the AI correction process was exceptionally effective in improving the articles' adherence to WHO guidelines according to the AI's own assessment.

***Efficacy of AI Enhancement: Human Evaluation.*** To provide external validation, human raters assessed the efficacy of AI correction. A repeated measures ANOVA compared the human-rated scores before and after AI correction for all articles and articles in each language separately. The results of human evaluation strongly corroborated the findings of AI. The mean score assigned by human raters rose significantly from $M = 6.02$ ($SD = 1.43$) for the original articles to $M = 11.20$ ($SD = 1.07$) for the corrected versions. This difference was highly significant, $F$ (1, 119) = 1120.284, $p < .001$, with an extremely large effect

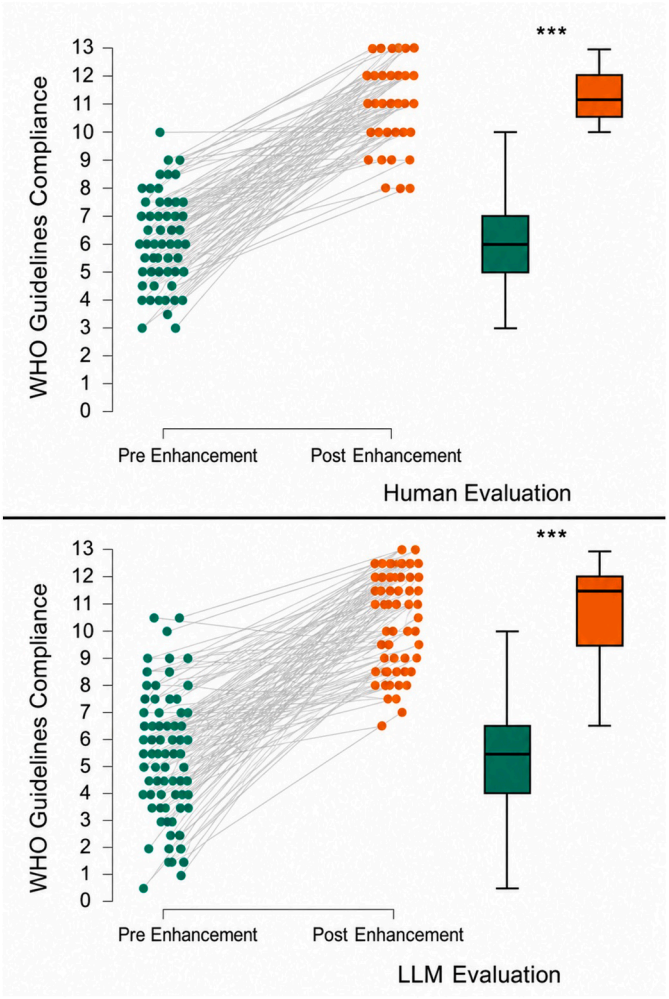


**Fig. 3.** Pre- and Post-LLM Enhancement WHO Guidelines Compliance: Human and LLM Evaluations.

*Note.* WHO guidelines compliance scores before and after LLM-based enhancement of suicide-related news articles (N = 120). Upper panel shows human evaluator ratings; lower panel shows LLM evaluator ratings. Gray lines connect paired data points showing individual article improvements from pre-enhancement (green) to post-enhancement (orange). Box plots display score distributions with individual data points. Asterisks (***) indicate significant differences ($p < .001$).

size ($\eta p2 = .904$). Practically, this signifies that the average adherence score, as judged by human experts, increased from 46.3% to 86.2%. This significant improvement and very large effect size were robust across all three language samples, Hebrew ($\eta p2 = .893$), English ($\eta p2 = .875$), and French ($\eta p2 = .945$), with all changes being highly significant ($p < .001$). This confirms that human experts recognize AI corrections as a substantial improvement in reporting quality.

***Comparison of evaluations between human raters and GenAI system after the enhancement stage.*** The final step of the analysis was to compare the evaluations of the corrected articles between the AI evaluation bots and human raters. A repeated-measures ANOVA was conducted on the post-enhancement scores from the AI ($M = 10.88$, $SD = 1.68$) and humans ($M = 11.20$, $SD = 1.07$). The analysis revealed no statistically significant differences in the mean scores assigned by the two groups, $F(1, 119) = 3.774$, $p = .054$, with a small effect size ($\eta p2 = .031$). This result indicates that the AI-enhanced articles reached a high-quality standard where the assessments of the AI system and human experts converged, showing no significant difference in their final evaluations.

## 4. Discussion

The primary objective of this study was to assess the feasibility and effectiveness of a dual-component AI-based system designed to evaluate media articles for adherence to WHO guidelines on responsible suicide reporting (World Health Organization, 2017), and to automatically enhance them for improved compliance. Our broader aim was to explore whether such a technological approach could serve as a scalable, multilingual tool for improving suicide-related media coverage, ultimately contributing to global suicide prevention.

Regarding the first study objective, the findings indicated that the AI-based evaluation bot accurately assessed article adherence to WHO guidelines, as reflected in its strong agreement with independent human raters across the three languages. These results replicate and extend our previous pilot studies (Elyoseph, Levkovich, Rabin, et al., 2025, Elyoseph, Levkovich, Rabin, et al., 2025), which validated similar prompt engineering approaches in multilingual contexts. Notably, the current study used a more advanced model (Claude 3.5 Sonnet), reinforcing the robustness of the underlying method by combining role assignment, few-shot learning, and chain-of-thought prompting. Moreover, the high correlations between system and human ratings even after enhancement suggest that the bot is not only effective at detecting violations but also reliably identifies high-quality guideline-compliant content. These findings align with broader research supporting the potential of AI to accurately evaluate complex content across domains (Islam et al., 2022; Miller et al., 2024; Refoua et al., 2025).

The second and third study objectives examined the system's ability to enhance article adherence to WHO guidelines and to generalize this capacity across languages and cultural contexts. Results showed a significant post-enhancement improvement in adherence, both across the full sample and within each language group. These findings offer a preliminary proof of concept for the dual role of GenAI in evaluating and improving suicide-related media content. However, further validation is required before broader implementation. Future research should include qualitative analyses to evaluate the reliability, tone, and factual integrity of AI-generated revisions, ensuring that enhanced articles preserve the original message and journalistic standards (Ji et al., 2023; Kreps, McCain, & Brundage, 2022). Such assessments are essential to establish this strategy as a trusted editorial tool for media professionals.

Taken together, the current findings suggest that combining automated evaluation with real-time enhancement of suicide-related articles may offer a practical bridge between WHO's (2017) suicide-reporting guidelines and their consistent, everyday application in fast-paced journalism settings. Strong agreement with human raters and significant post-enhancement improvements indicate that such a system can reliably identify reporting risks and provide guideline-aligned corrections with minimal human input (Niederkrotenthaler et al., 2020).

These findings align with previous proposals highlighting AI's potential of AI to overcome structural barriers to WHO guideline adherence in media reporting (Ransing et al., 2021). The dual-component GenAI system serves not only as an evaluator but also as an active editorial partner, complementing traditional training efforts that often yield limited long-term impact (Sisask & Värnik, 2012; Fu & Yip, 2008). Our results add to the growing evidence that responsible reporting, particularly narratives of hope and recovery can reduce suicide risk through the Papageno effect (Niederkrotenthaler et al., 2010, 2022). By embedding these protective principles into routine editorial workflows, GenAI may offer a scalable, cross-cultural solution for promoting sustained adherence. Nevertheless, real-time implementation studies and broader replications are essential to confirm lasting changes in journalistic practice.

This study has several important limitations that should be acknowledged. Conceptually, this study operationalized responsible suicide reporting as adherence to WHO-based reporting criteria. Although this framework is evidence-based and central to suicide-prevention communication, it does not capture all dimensions of

journalistic quality, ethical judgment, cultural meaning, narrative accuracy, audience trust, or reader impact. Thus, the findings should be interpreted as evidence that GenAI can improve guideline adherence in suicide-related news articles, but not as direct evidence that the revised articles would reduce stigma, increase help-seeking, or affect suicide-related outcomes among readers.

First, although the inclusion of three languages demonstrates the system's cross-cultural feasibility, the linguistic and cultural scope remains limited; expanding future testing to additional languages, media contexts, and cultural settings is essential for validating broader generalizability. In addition, the study was based on a controlled offline corpus of 120 articles and focused on traditional news media rather than social media, television, films, or other forms of suicide-related public communication. As this study focused on a controlled offline corpus, its findings must be replicated under real-world time-pressured newsroom conditions to fully assess practical implementation challenges and user acceptance, including editorial negotiations, legal constraints, audience considerations, and the willingness of journalists and editors to adopt AI-assisted feedback in routine practice.

Second, the use of a fixed temperature setting of zero was intended to ensure output stability and replicability. However, future research could incorporate multiple runs with varying model parameters to better assess the robustness and consistency of the AI-generated evaluations and corrections. Relatedly, the use of a single model and prompt architecture limits conclusions about robustness across different LLMs, model versions, parameter settings, languages, and future updates. Because proprietary models are continuously updated, future studies should also examine model drift and the stability of outputs over time.

Third, while the quantitative improvements in guideline adherence are encouraging, this study did not systematically evaluate the narrative and stylistic qualities of AI-enhanced articles. Rigorous qualitative analysis, possibly involving professional journalists and editors, is needed to examine whether the revisions preserve journalistic standards, readability, and audience engagement. Such evaluations should also assess factual integrity, journalistic voice, editorial independence, public trust, and whether the revised articles would be considered acceptable and publishable in real newsroom settings. In addition, because both the evaluation and enhancement modules were guided by the same WHO-derived framework, improvements may partly reflect optimization to the scoring rubric rather than broader improvements in reporting quality. The inclusion of independent human raters mitigates this concern, but future work should include external expert review and qualitative analysis of the revised texts.

Another methodological limitation concerns the text-only nature of the present system. This system currently only applies to traditional media sources. It did not evaluate images, layout, headline prominence in its full visual context, or other multimodal features that may shape the impact of suicide reporting. Given the increasing influence of social media, especially among younger populations at high risk for suicide, future adaptations should consider tools for detecting and improving suicide-related content in social media environments, as well as multimodal systems capable of evaluating visual and platform-specific features of suicide-related communication.

Finally, some limitations reflect broader challenges in the emerging field of AI-assisted suicide-prevention communication rather than weaknesses unique to the present study. LLMs do not possess human-like clinical, cultural, or journalistic understanding and may misclassify ambiguous or context-dependent content. In the context of suicide reporting, false reassurance may allow problematic reporting to pass uncorrected, whereas over-flagging may undermine journalists' trust and reduce adoption. AI-assisted systems also raise ethical and governance concerns related to transparency, bias, accountability, privacy, editorial autonomy, and public trust. For these reasons, the proposed system should be understood as an assistive editorial and public-health tool, not as a replacement for journalists, editors, suicide-prevention experts, or ethical oversight. Real-world implementation will require gradual testing, transparent governance, and continuous human supervision.

### 4.1. Conclusions and implications

This study demonstrates that GenAI can serve as a practical and effective bridge between global suicide reporting guidelines and their consistent application in real-world journalism. By validating a system that evaluates and enhances media content, the findings offer compelling initial evidence that AI can help translate complex ethical standards into actionable editorial practices, positioning it as an active partner in promoting responsible health-protective reporting.

Building on this proof of concept, the broader vision articulated here, illustrated in Fig. 4, maps a clear phased pathway for real-world deployment. The proposed system integrates automated topic detection, article evaluation, guided correction, blind reevaluation, and personalized communication with journalists and editors. Beyond its immediate function, this structure lays the groundwork for establishing a *living global index* of suicide-reporting quality: a dynamic, comparative metric that tracks adherence trends in real-time across outlets, countries, and languages. Such a benchmark, generated by the system's continuous monitoring—would enable stakeholders to identify compliance gaps, measure progress, and strategically direct resources to regions or sectors most in need of intervention (Niederkrotenthaler et al., 2020; Ransing et al., 2021; Metzler et al., 2022). In this sense, the system becomes not only a tool for individual article correction but also a foundation for global accountability and evidence-informed policy action.

Importantly, this AI-based system is not designed to replace frontline suicide prevention efforts, such as clinical care or crisis lines but to complement them by targeting a key upstream factor: the information environment. By detecting harmful media coverage and providing real-time, evidence-based corrections, the system helps mitigate the "Werther effect" and amplify the "Papageno effect" (Niederkrotenthaler et al., 2010; Pirkis et al., 2006). The automated email module (Fig. 4) illustrates this logic, offering journalists and editors personalized feedback with a clear rationale and suggested edits. This process reinforces learning, builds editorial capacity, and fosters routine adherence to the reporting standards. Evidence shows that such ongoing, feedback-based approaches improve media reporting more effectively than one-time trainings (Sisask & Värnik, 2012; Ransing et al., 2021). As such, the system operates not only as an editorial tool but also as a behavioral nudge that embeds guideline-aligned practices into everyday journalistic workflow.

Realizing this vision requires carefully staged implementation and continuous evaluation. Early versions of the system may operate as opt-in "editorial advisors," with the potential to evolve into semi-automated tools monitoring live content streams (Fig. 4). While this path offers substantial promise for scalable, cost-effective public health impacts, it also raises important concerns about editorial autonomy, algorithmic transparency, and public trust. To address these challenges, strong ethical governance frameworks are essential, balancing the urgency of saving lives with the safeguarding of journalistic freedom (Niederkrotenthaler et al., 2022; Sinyor et al., 2025). Future research should assess not only technical performance but also journalists' willingness to adopt the system, readers' trust in AI-assisted journalism, and cross-cultural adaptability. When developed and deployed responsibly, GenAI-based tools could become transformative assets in global suicide prevention, translating guidelines into daily practice and information into timely, life-saving interventions.

## CRediT authorship contribution statement

**Y. Levi-Belz:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Formal analysis, Data curation, Conceptualization. **B. Nobile:** Writing – review & editing,

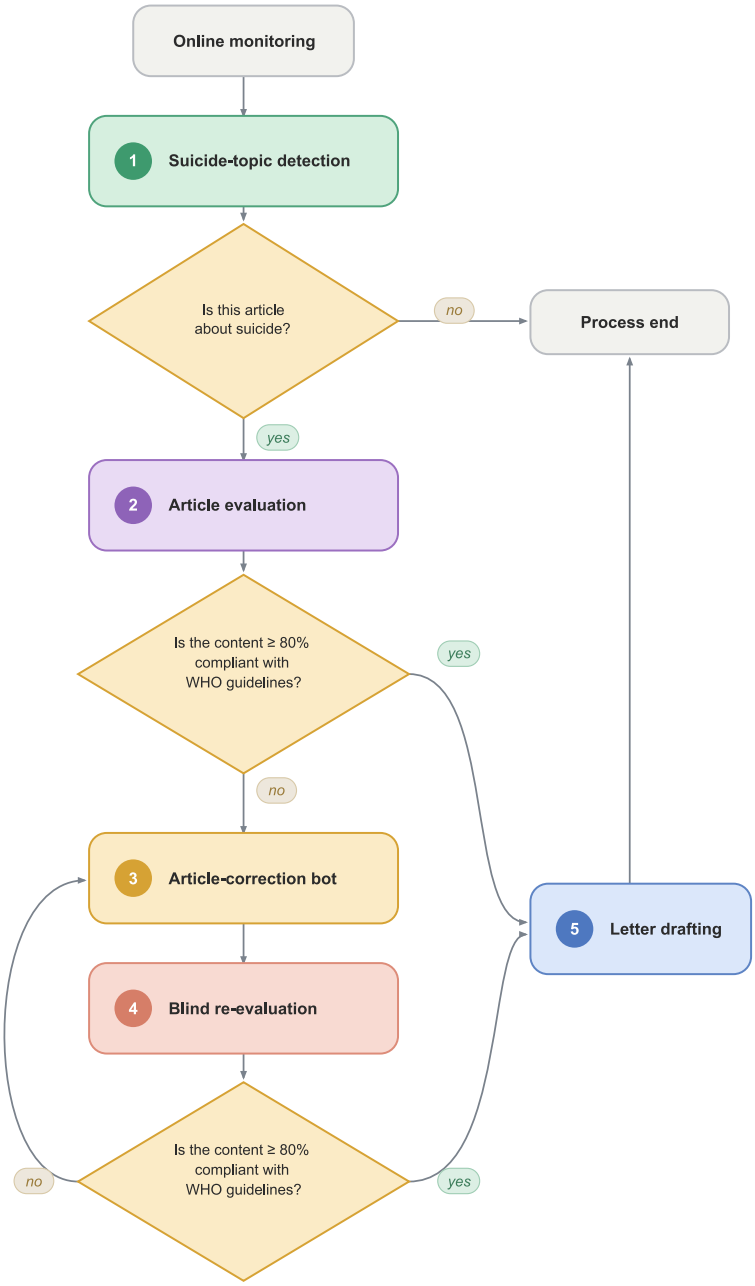


**Fig. 4. Proposed Workflow for Real-Time Evaluation and Enhancement of Suicide-Related News Coverage Using an AI-Based Multi-Agent System**
*Note.* This diagram illustrates a real-time, multi-agent system designed to monitor, evaluate, and improve suicide-related news articles according to WHO guidelines. The process begins with **Agent 1** (Suicide-topic detection), which identifies whether an article addresses suicide. If so, it proceeds to **Agent 2** (Article evaluation), which assesses guideline adherence. Articles scoring below the 80% threshold are passed to **Agent 3** (Article-correction bot), which generates an improved version. This version is then blindly re-evaluated by **Agent 4** to confirm that the revised article meets the compliance criteria. Articles that either pass the initial or post-correction evaluation are then forwarded to **Agent 5** (Letter drafting), which generates a tailored feedback letter to the relevant journalist or editor. This scalable workflow offers a structured and semi-automated approach to promoting responsible media coverage.

Writing – original draft, Resources, Conceptualization. **I. Levkovich:** Writing – review & editing, Resources, Conceptualization. **P. Courtet:** Writing – review & editing, Supervision, Data curation, Conceptualization. **Z. Elyoseph:** Writing – review & editing, Writing – original draft, Visualization, Resources, Project administration, Formal analysis, Data curation, Conceptualization.

## Data sharing

The data supporting the findings of this study are available from the corresponding author [YLB] upon reasonable request.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgement

This work was conducted in collaboration with the Suicide Prevention Unit of the Ministry of Health, Israel.

## Data availability

Data will be made available on request.

## References

Brailovskaia, J. (2024). The "Vicious Circle of addictive Social Media Use and Mental Health" model. *Acta Psychologica, 247*, Article 104306.

Brailovskaia, J., Teismann, T., & Margraf, J. (2020). Positive mental health mediates the relationship between Facebook addiction disorder and suicide-related outcomes: A longitudinal approach. *Cyberpsychology, Behavior, and Social Networking, 23*(5), 346–350.

Brodsky, B. S., Spruch-Feiner, A., & Stanley, B. (2018). The Zero suicide model: Applying evidence-based suicide prevention practices to clinical care. *Frontiers in Psychiatry, 9*, 33. https://doi.org/10.3389/fpsyt.2018.00033

Chen, Y. Y., Chen, F., & Yip, P. S. (2011). The impact of media reporting of suicide on actual suicides in Taiwan, 2002-05. *Journal of Epidemiology & Community Health, 65*(10), 934–940. .

Elyoseph, Z., & Levkovich, I. (2023). Beyond human expertise: The promise and limitations of ChatGPT in suicide risk assessment. *Frontiers in Psychiatry, 14*, Article 1213141. .

Elyoseph, Z., Levkovich, I., Haber, Y., & Levi-Belz, Y. (2025b). Using GenAI to train mental health professionals in suicide risk assessment: Preliminary findings. *The Journal of Clinical Psychiatry, 86*(3), Article 24m15525. .

Elyoseph, Z., Levkovich, I., Rabin, E., Shemo, G., Szpiler, T., Shoval, D. H., & Levi-Belz, Y. (2025a). Applying language models for suicide prevention: Evaluating news article adherence to WHO reporting guidelines. *Npj Mental Health Research, 4*(1), 25. https://doi.org/10.1038/s44184-025-00139-5. .

Fu, K. W., & Yip, P. S. F. (2008). Changes in reporting of suicide news after the promotion of the WHO media recommendations. *Suicide and Life-Threatening Behavior, 38*(5), 631–636. .

Hawley, L. L., Niederkrotenthaler, T., Zaheer, R., Schaffer, A., Redelmeier, D. A., Levitt, A. J., … Sinyor, M. (2023). Is the narrative the message? The relationship between suicide-related narratives in media reports and subsequent suicides. *Australian and New Zealand Journal of Psychiatry, 57*(5), 758–766. .

Islam, M. R., Ahmed, M. U., Barua, S., & Begum, S. (2022). A systematic review of explainable artificial intelligence in terms of different application domains and tasks. *Applied Sciences, 12*(3), 1353. .

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., … Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), 1–38. .

Kreps, S., McCain, R. M., & Brundage, M. (2022). All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. *Journal of experimental political science, 9*(1), 104–117. .

Levi-Belz, Y., Starostintzki Malonek, R., & Hamdan, S. (2023). Trends in newspaper coverage of suicide in Israel: An 8-year longitudinal study. *Archives of Suicide Research, 27*(4), 1191–1206. .

Levkovich, I., Haber, Y., Levi-Belz, Y., & Elyoseph, Z. (2025). A step toward the future? Evaluating GenAI QPR simulation training for mental health gatekeepers. *Frontiers of Medicine, 12*, Article 1599900. .

Levkovich, I., & Omar, M. (2024). Evaluating of bert-based and large language mod for suicide detection, prevention, and risk assessment: A systematic review. *Journal of Medical Systems, 48*(1), 113. https://doi.org/10.1007/s10916-024-02134-3

McTernan, N., Spillane, A., Cully, G., Cusack, E., & Arensman, E. (2018). Media reporting of suicide and adherence to media guidelines: A content analysis of media coverage of suicide in Ireland. *International Journal of Social Psychiatry, 64*(6), 536–544.

Metzler, H., Baginski, H., Niederkrotenthaler, T., & Garcia, D. (2022). Detecting potentially harmful and protective suicide-related content on Twitter: Machine learning approach. *Journal of Medical Internet Research, 24*(8), Article e34705. .

Miller, T., Durlik, I., Łobodzińska, A., Dorobczyński, L., & Jasionowski, R. (2024). AI in context: Harnessing domain knowledge for smarter machine learning. *Applied Sciences, 14*(24), Article 11612. .

Niederkrotenthaler, T., Braun, M., Pirkis, J., Till, B., Stack, S., Sinyor, M., Tran, U. S., & Voracek, M. (2020). Association between suicide reporting and suicide: Systematic review and meta-analysis. *BMJ, 368*, Article m575. https://doi.org/10.1136/bmj.m575

Niederkrotenthaler, T., Till, B., Kirchner, S., Sinyor, M., Braun, M., Pirkis, J., … Spittal, M. J. (2022). Effects of media stories of hope and recovery on suicidal ideation and help-seeking attitudes and intentions: Systematic review and meta-analysis. *The Lancet Public Health, 7*(2), e156–e168. .

Niederkrotenthaler, T., Voracek, M., Herberth, A., Till, B., & Strauss, M. (2010). Role of media reports in completed and prevented suicide: Werther vs. Papageno effects. *The British Journal of Psychiatry, 197*(3), 234–243. https://doi.org/10.1192/bjp.bp.109.074633

Phillips, D. P. (1974). The influence of suggestion on suicide: Substantive and theoretical implications of the Werther effect. *American Sociological Review, 39*(3), 340–354. https://doi.org/10.2307/2094294

Pirkis, J., Blood, R. W., Beautrais, A., Burgess, P., & Skehan, J. (2006). Media guidelines on the reporting of suicide. *Crisis, 27*(2), 82–87.

Pirkis, J., Rossetto, A., Nicholas, A., Ftanou, M., Robinson, J., & Reavley, N. (2019). Suicide prevention media campaigns: A systematic literature review. *Health Communication, 34*(4), 402–414. .

Pompili, M. (2020). Increased suicide mortality amidst the COVID-19 pandemic: A global perspective. *Psychiatry Research, 295*, Article 113596. https://doi.org/10.1016/j.psychres.2020.113596

Ransing, R., Dashi, E., Rehman, S., Mehta, V., Ramalho, R., Adiukwu, F., & Ojeahere, M. (2021). Artificial intelligence-based models for augmenting media reporting of suicide: Challenges and opportunities. *Asian Journal of Psychiatry, 58*, Article 102611. https://doi.org/10.1016/j.ajp.2021.102611

Refoua, E., Elyoseph, Z., Wacker, R., Dziobek, I., Tsafrir, I., & Meinlschmidt, G. (2025). The next frontier in mindreading? Assessing generative artificial intelligence (GAI)'s social-cognitive capabilities using dynamic audiovisual stimuli. *Computers in Human Behavior Reports*, Article 100702. .

Shepard, D. S., Gurewich, D., Lwin, A. K., Reed, G. A., & Silverman, M. M. (2016). Suicide and suicidal attempts in the United States: Costs and policy implications. *Suicide and Life-Threatening Behavior, 46*(3), 352–362. https://doi.org/10.1111/sltb.12225

Shinan-Altman, S., Elyoseph, Z., & Levkovich, I. (2024). The impact of history of depression and access to weapons on suicide risk assessment: A comparison of ChatGPT-3.5 and ChatGPT-4. *PeerJ, 12*, Article e17468. https://doi.org/10.7717/peerj.17468

Sinyor, M., Ekstein, D., Ming Chan, P. P., Men, Y. V., Malonek, R. S., Schaffer, A., … Levi-Belz, Y. (2025). Investigating the impact of parallel media engagement initiatives on suicide reporting in Canada and Israel. *Social Psychiatry and Psychiatric Epidemiology*, 1–12. .

Sisask, M., & Värnik, A. (2012). Media roles in suicide prevention: A systematic review. *International Journal of Environmental Research and Public Health, 9*(1), 123–138.

Tatum, P. T., Canetto, S. S., & Slater, M. D. (2010). Suicide coverage in U.S. newspapers following the publication of the media guidelines. *Suicide and Life-Threatening Behavior, 40*(5), 524–534.

World Health Organization. (2017). *Preventing suicide: A resource for media professionals (Update 2017)*. WHO. https://www.who.int/publications/i/item/9789241513543.

World Health Organization. (2023). *Suicide worldwide in 2023: Global health estimates*. Geneva: World Health Organization. https://www.who.int/publications/i/item/9789240078845.

Yip, P. S., Fu, K. W., Yang, K. C., Ip, B. Y., Chan, C. L., Chen, E. Y., & Lee, D. T. (2016). The effects of a celebrity suicide on suicide rates in Hong Kong. *Journal of Affective Disorders, 192*, 50–56. https://doi.org/10.1016/j.jad.2015.12.060