Journal of Autism and Developmental Disorders
https://doi.org/10.1007/s10803-026-07249-9

ORIGINAL ARTICLE

Accuracy of Autism-Related TikTok Information in Italian: A Comparison Between Human Raters and Large Language Models

Alessandro Carollo[1] · Seraphina Fong[1,2,3] · Giovanni Belardinelli[1] · Silvia Perzolli[1] · Giacomo Vivanti[4] · Daniel S. Messinger[5] · Dagmara Dimitriou[6] · Gianluca Esposito[1]

Received: 15 December 2025 / Accepted: 27 January 2026

Abstract

Purpose Social networking sites are major channels for sharing information on neurodiversity, including autism spectrum disorder. TikTok has become a particularly influential platform for autism-related communication, yet concerns remain about the scientific accuracy of such content. Most prior studies have focused on English-language videos and have evaluated accuracy with limited granularity. Additionally, the difficulty of achieving consistent expert ratings underscores the need for automated reliability assessment.

Methods In this study, we examined 408 informational statements extracted from 148 TikTok videos posted under the hashtag *#Autismo* (Italian for *#Autism*). Three clinical experts independently classified each statement as inaccurate, overgeneralized, or accurate; their median ratings served as the human-derived ground truth and were compared with classifications from two large language models: ChatGPT 4.0 mini and Gemini 1.5 Flash.

Results Human raters showed moderate agreement ($\kappa_{mean} = 0.52$) and high specific agreement only for accurate statements, with lower agreement for overgeneralized and inaccurate content. ChatGPT achieved moderate agreement with human ratings ($\kappa = 0.58$), while Gemini reached only fair agreement ($\kappa = 0.29$). ChatGPT also exhibited a more conservative evaluation pattern (accurate information: precision = 0.89, recall = 0.82), whereas Gemini tended to overestimate accuracy (accurate information: precision = 0.76, recall = 0.93).

Conclusion These findings suggest that LLMs, particularly ChatGPT, may support cautious and assistive evaluation of online health content. Future research should assess their applicability across online communities and platforms and explore their integration into accuracy-based alert systems that provide users with contextual reliability cues.

Keywords Autism spectrum disorder · Social networking sites · Large language models · TikTok · Clinical information accuracy

Alessandro Carollo and Seraphina Fong have contributed equally to this work.

✉ Gianluca Esposito
gianluca.esposito@unitn.it

[1] Department of Psychology and Cognitive Science, University of Trento, 38068 Rovereto, Italy

[2] Department of Information Engineering and Computer Science, University of Trento, 38123 Povo, Italy

[3] Center for Augmented Intelligence, Fondazione Bruno Kessler, 38123 Povo, Italy

[4] A.J. Drexel Autism Institute, Drexel University, Philadelphia, PA 19104, USA

[5] Department of Psychology, University of Miami, Coral Gables, FL 33124, USA

[6] Sleep Education and Research Laboratory, Department of Psychology and Human Development, UCL Institute of Education, University College London, London WC1H 0AA, UK

Published online: 18 February 2026

Introduction

In recent years, media outlets and social networking sites have become a major platform for disseminating information about neurodevelopmental conditions, including autism spectrum disorder (Fong et al., 2025; Nordahl-Hansen et al., 2018). In particular, TikTok's short-form video format and algorithmic amplification make it a powerful tool for science communication and awareness raising, but also a potential source of misinformation. While media content can promote inclusion and neurodiversity advocacy, user-generated content frequently lacks scientific oversight, and messages about autism may reflect personal interpretations, stereotypes, or inaccurate claims (Brennan et al., 2025; Hungerford et al., 2025; Jones et al., 2023; Ononuju & Ujari, 2025).

Autism is a heterogeneous neurodevelopmental condition characterized by differences in communication, social interaction, and sensory processing style (American Psychiatric Association, 2022). Public understanding of autism has evolved from a predominantly medicalized view toward a neurodiversity perspective, which values individual variability rather than pathologizing it (Botha et al., 2024; Chapman, 2021). In relation to this point, social networking sites play an instrumental role for autistic individuals in building and consolidating their identity, raising awareness about autism, and affirming a neurodivergent perspective that challenges traditional communicative models (e.g., Alper et al., 2025; Fong et al., 2025; Guberman, 2023). However, the rapid circulation of user-generated content, particularly on platforms such as TikTok, poses challenges for ensuring information accuracy, especially when creators are not experts or when algorithmic mechanisms prioritize engagement over reliability (Brown et al., 2024).

Research has shown that autism-related content on TikTok often contains inaccurate or oversimplified information (e.g., Aragon-Guevara et al., 2025; Brennan et al., 2025; Brown et al., 2024). For instance, Aragon-Guevara et al. (2025) reported that only 27% of autism-related TikTok videos in English were accurate according to expert evaluation, while 41% contained false information and 32% presented overgeneralizations. Moreover, phenomena such as self-diagnosis, influencer-driven advocacy, and the commodification of neurodiversity discourse complicate how autism is represented and perceived online. To date, most accuracy assessments have evaluated videos as holistic units rather than the specific informational claims they contain, thus overlooking the granularity of factual content within each video. Despite growing interest in this topic, little is known about the accuracy of autism-related individual information in languages other than English, including Italian.

Additionally, until now, assessments of information accuracy have been primarily conducted by human raters. However, in fields such as clinical psychology, human evaluations may be influenced by subjective interpretation, disciplinary bias, or differences in expertise, which can lead to inconsistent judgments and reduce the replicability of findings. For this reason, the rapid advancement of large language models (LLMs), such as ChatGPT and Gemini, offers new possibilities for the automated evaluation of clinical information (e.g., Fong et al., 2024). These models have shown promising performance in identifying misinformation on social networking sites (e.g., Huang et al., 2025), yet their reliability in assessing the accuracy of complex, health-related content remains underexplored.

Building on previous work (e.g., Aragon-Guevara et al., 2025; Fong et al., 2024, 2025), the present study aims to evaluate the accuracy of autism-related information in Italian TikTok videos shared under the hashtag *#Autismo* ("#Autism" in English). Specifically, we compare evaluations conducted by three human experts and two LLMs (i.e., ChatGPT and Gemini) to examine their level of agreement and to assess whether artificial intelligence (AI) can approximate expert human judgment in identifying inaccurate, overgeneralized, and accurate information. Importantly, the present study adopts an ecological validity perspective. Rather than benchmarking the maximum theoretical performance of state-of-the-art LLMs, we focus on AI systems that are widely accessible and more likely to be encountered by lay users seeking information online. Accordingly, we selected ChatGPT-4.0 mini and Gemini 1.5 Flash, which prioritize speed and cost-efficiency and are representative of tools currently available to the general public. This choice allows us to evaluate how AI models that parents and non-experts may realistically use perform in assessing autism-related information on social media.

Methods

This study was approved by the Ethics Committee of the University of Trento (2024-24 ESA).

Data Collection

To minimize algorithmic bias, a new TikTok account was created exclusively for this study. Using existing accounts could have distorted search results based on prior interactions and personalization algorithms. On October 10, 2024, the term *"#Autismo"* (the Italian equivalent of *"#Autism"*) was entered into the TikTok search bar. In this way, we collected a sample of 172 videos, with the first one being published in December 2020.

All videos that were *(a)* not in Italian, *(b)* not related to autism, *(c)* duplicates were excluded. Twenty-four videos were excluded based on these criteria, resulting in a final dataset of 148 eligible videos. For each video, the following metadata were manually recorded: URL, number of likes, number of views, and date of publication. Transcriptions of the spoken, written, and captioned text were also manually generated.

Data Extraction and Coding

From the 148 eligible videos, a total of 408 distinct informational statements related to autism were extracted. A single video could contain one or more such statements. Informational statements were operationally defined as assertions about autism in general (such as its causes, diagnosis, treatment, or clinical presentation) rather than purely personal narratives. However, videos describing personal experiences (e.g., an autistic person discussing sensory tools, or a parent describing experiences with a child) were included as these could contain statements that users might interpret as generalizable information about autism.

Each statement was independently evaluated, in a blinded manner, by three independent human raters who have clinical expertise in autism and are co-authors (GB, SP, and GE). Statements were classified as inaccurate, overgeneralized, or accurate based on their consistency with current scientific knowledge regarding autism etiology, diagnostic criteria, and evidence-based interventions (as in Aragon-Guevara et al., 2025). Statements were considered accurate when they were factually correct, free from errors or misrepresentations, and did not overgeneralize individual experiences (for example, "Early interventions can improve communication skills in autistic children"). Statements were classified as overgeneralized when they extended specific individual experiences to the entire autistic population (for example, "Autistic people are afraid of going to concerts"). Finally, statements were deemed inaccurate when they contained factual errors, misconceptions, or misleading information (for example, "You can determine if a child is autistic using the WISC test").

Evaluation by Large Language Models

After manual coding, all 408 informational statements were independently submitted to two LLMs: (1) ChatGPT 4.0 mini and (2) Gemini 1.5 Flash for automated evaluation. The following Italian-language prompt was used for both models (English translation below):

Original (Italian):

> *"Ti fornirò delle informazioni ricavate da video TikTok rintracciati tramite l'hashtag 'autismo'. Valuta le seguenti trascrizioni come 'accurata', 'inaccurata' o 'sovra-generalizzazione', basandoti sui seguenti criteri*:
>
> *Le affermazioni sono accurate quando sono fattualmente corrette, prive di errori o distorsioni e non generalizzano eccessivamente esperienze individuali (ad esempio: "Gli interventi precoci possono migliorare le capacità comunicative nei bambini autistici"). Le affermazioni sono sovrageneralizzate quando estendono esperienze specifiche di singoli individui all'intera popolazione autistica (ad esempio: "Le persone autistiche hanno paura di andare a.i. concerti"). Infine, le affermazioni sono inaccurate quando contengono errori fattuali, misconcezioni o informazioni fuorvianti (ad esempio: "È possibile determinare se un bambino è autistico utilizzando il test WISC"). Ecco l'informazione: [specific informational statement]."*

English translation:

> *I will provide you with information taken from TikTok videos retrieved using the hashtag 'autism'. Evaluate the following statements as 'accurate', 'inaccurate', or 'overgeneralized', based on the following criteria: [criteria described above in section "Data extraction and coding"]. Here is the information: [specific informational statement]."*

To avoid potential bias due to context retention or order effects, each informational statement was evaluated in a separate chat session for both models. This procedure ensured that the models' assessments were independent of previous inputs and that no information from prior evaluations could influence subsequent responses.

Statistical Analysis

The statistical analysis proceeded in several steps.

First, descriptive analyses were conducted to examine the overall distribution of accuracy ratings provided by the three human raters. The proportion of statements classified as inaccurate, overgeneralized, and accurate was visualized using descriptive plots to provide an overview of the accuracy trends within the dataset. To evaluate inter-rater reliability among the three human experts, pairwise weighted Cohen's kappa coefficients (κ) were computed. Weighted κ was used to account for the ordinal nature of the three rating categories (inaccurate < overgeneralized < accurate), assigning partial credit to disagreements between adjacent

categories. In addition to the global reliability estimates provided by weighted κ, we computed the specific positive agreement for each evaluation category. This metric quantifies the probability that two raters consistently assign the same category when at least one of them selects it, offering a more fine-grained assessment of agreement patterns across inaccurate, overgeneralized, and accurate statements.

Next, a ground truth label for each statement was generated using the median rating across the three human raters. This approach allowed for an estimation of the general accuracy level of each informational statement, minimizing the influence of individual coder bias.

The performance of the two LLMs (ChatGPT 4.0 mini and Gemini 1.5 Flash) was assessed by comparing their classifications with the human-derived ground truth using weighted Cohen's κ. This allowed for the estimation of each model's level of agreement with expert human judgment. Along with the weighted Cohen's κ, we also computed the specific positive agreement between each LLM and the human-derived ground truth. Additionally, the classification performance of both human and artificial raters was quantified using standard information retrieval metrics. For each rater and for both LLMs (ChatGPT 4.0 mini and Gemini 1.5 Flash), we computed *precision* and *recall* scores across the three accuracy categories (inaccurate, overgeneralized, accurate). Precision was defined as the proportion of correctly identified instances within each predicted category, and recall as the proportion of correctly identified instances relative to the total number of true instances of that category.

κ values were interpreted following standard benchmarks proposed by Landis and Koch (1977), where $\kappa < 0.20$ indicates slight agreement, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 substantial, and > 0.80 almost perfect agreement.

Results

Inter-Rater Agreement

As outlined in the section "*Statistical analysis*", descriptive analyses and inter-rater reliability measures were first conducted to evaluate the consistency of human judgments. Weighted Cohen's κ coefficients were computed for each pair of raters to account for the ordinal nature of the three rating categories (inaccurate < overgeneralized < accurate). In this section, we report the results of these analyses, followed by the accuracy of the autism-related information and the performance of the LLMs compared to the human-derived ground truth.

Across all three human raters, the majority of informational statements were classified as accurate, followed by overgeneralized, and, to a lesser extent, inaccurate (see Fig. 1). This distribution indicates that most of the autism-related content in the analyzed TikTok videos aligned with current scientific understanding, while a smaller proportion exhibited either overly broad generalizations or factual inaccuracies.

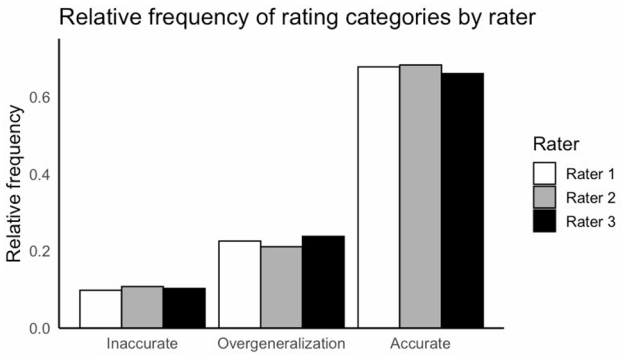

Fig. 1 Relative frequency of accuracy ratings assigned by each human rater. Each bar represents the proportion of informational statements classified as inaccurate, overgeneralized, or accurate out of a total of 408 informational statements

Table 1 Specific positive agreement computed between pairs of human raters across evaluation categories

Raters pair	Inaccurate	Overgeneralization	Accurate
Rater 1 vs. Rater 2	0.79	0.69	0.90
Rater 1 vs. Rater 3	0.17	0.46	0.81
Rater 2 vs. Rater 3	0.28	0.52	0.85

Inter-rater reliability among the three human experts was then assessed using pairwise weighted Cohen's κ coefficients. The weighted κ values ranged from 0.40 to 0.64, with a mean value of 0.52, indicating moderate agreement.

In addition to the weighted κ coefficients, we computed the specific positive agreement for each evaluation category to provide a more granular view of the agreement patterns among pairs of human raters (Table 1). Overall, specific agreement was highest for the accurate category, with values ranging from 0.81 to 0.90, indicating that raters consistently converged in identifying statements aligned with scientific evidence.

Agreement for the overgeneralization category ranged from moderate to substantial (0.46–0.69). In contrast, the inaccurate category showed the lowest levels of specific agreement, reflecting substantially lower consensus regarding which statements were factually incorrect. Taken together, these results suggest that disagreements among evaluators primarily arise when distinguishing between inaccurate and overgeneralized content, whereas the identification of accurate information appears to be more stable and consistent across raters.

Accuracy of Information on Autism

To obtain a rater-independent level of accuracy across information, we computed the median of the raters' evaluations. The resulting distribution followed the trend observed across individual raters, suggesting that most of the autism-related content retrieved through the hashtag *#Autismo* on TikTok was generally accurate (69.85%). Only a minority of statements was inaccurate (9.32%), while an intermediate proportion exhibited overgeneralized interpretations of individual experiences (20.83%).

Accuracy of Evaluations From Large Language Model

After establishing the human-derived ground truth based on the median ratings, we evaluated the performance of the two LLMs (ChatGPT 4.0 mini and Gemini 1.5 Flash) in classifying the same set of informational statements. Weighted Cohen's κ coefficients were computed to assess their level of agreement with the human median, accounting for the ordinal nature of the categories (inaccurate < overgeneralized < accurate). ChatGPT 4.0 mini demonstrated moderate agreement ($\kappa = 0.58$), whereas Gemini 1.5 Flash reached only fair agreement ($\kappa = 0.29$), indicating that ChatGPT's classifications more closely reflected expert judgments.

At a finer-grained level, specific positive agreement analyses revealed that both models were highly consistent with the human ground truth when identifying accurate statements, with values of 0.86 for ChatGPT and 0.84 for Gemini (Table 2). However, differences emerged for the other categories. ChatGPT showed substantial agreement for overgeneralized statements (0.64) but only fair agreement for inaccurate ones (0.29). Gemini, in contrast, exhibited only fair specific agreement for overgeneralized statements (0.35) and slight agreement for inaccurate statements (0.15).

The confusion matrices (Tables 3 and 4) show that most statements were correctly identified as accurate by both models. However, ChatGPT occasionally downgraded accurate statements to overgeneralized or inaccurate categories, while Gemini tended to overestimate accuracy by labeling as accurate a notable number of statements judged by experts as overgeneralized or inaccurate. Additionally, both LLMs' judgments of inaccuracy were mostly in disagreement with human ratings.

These patterns are consistent with the precision–recall analysis (Fig. 2; Table 5). Overall, ChatGPT demonstrated a more conservative and balanced evaluation profile,

Table 2 Specific positive agreement computed between pairs of large language models and human-derived ground truth across evaluation categories

Raters pair	Inaccurate	Overgeneralization	Accurate
ChatGPT vs. Human-derived ground truth	0.29	0.64	0.86
Gemini vs. Human-derived ground truth	0.15	0.35	0.84
ChatGPT vs. Gemini	0.13	0.38	0.80

Table 3 Confusion matrix comparing ChatGPT 4.0 mini classifications with human-derived median ratings

	Inaccurate (Human)	Over-general-ization (Human)	Accurate (Human)	Total
Inaccurate (ChatGPT)	**10**	5	15	30
Overgeneralization (ChatGPT)	16	**64**	36	116
Accurate (ChatGPT)	12	16	**234**	262
Total	38	85	285	408

Bold values along the diagonal indicate correct classifications

Table 4 Confusion matrix comparing ChatGPT 4.0 mini classifications with human-derived median ratings

	Inaccurate (Human)	Overgen-eralization (Human)	Accurate (Human)	Total
Inaccurate (Gemini)	**4**	1	10	15
Overgeneralization (Gemini)	13	**23**	11	47
Accurate (Gemini)	21	61	**264**	346
Total	38	85	285	408

Bold values along the diagonal indicate correct classifications

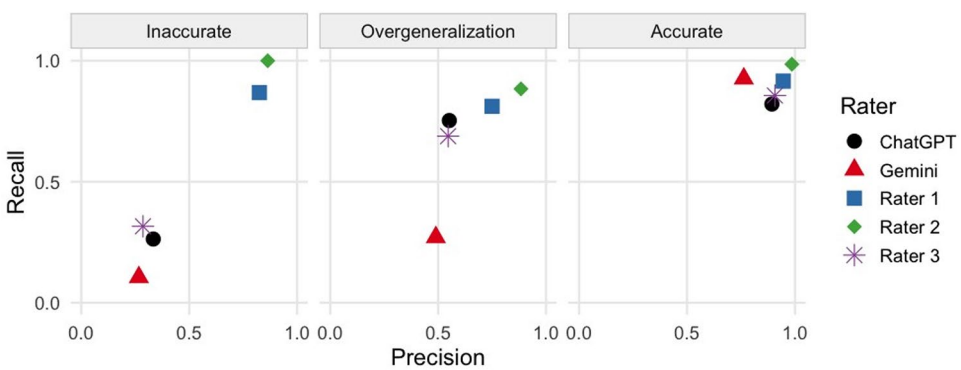

Fig. 2 Precision and recall scores illustrating the performance of human raters and large language models (ChatGPT 4.0 mini and Gemini 1.5 Flash) in classifying informational statements as inaccurate, overgeneralized, or accurate

Table 5 Precision and recall scores of human and artificial raters for inaccurate, overgeneralized, and accurate informational statements

Rater	Precision	Recall
Inaccurate information		
Rater 1	0.83	0.87
Rater 2	0.86	1.00
Rater 3	0.29	0.32
ChatGPT	0.33	0.26
Gemini	0.27	0.11
Overgeneralized information		
Rater 1	0.75	0.81
Rater 2	0.88	0.88
Rater 3	0.55	0.69
ChatGPT	0.55	0.75
Gemini	0.49	0.27
Accurate information		
Rater 1	0.95	0.92
Rater 2	0.99	0.99
Rater 3	0.91	0.86
ChatGPT	0.89	0.82
Gemini	0.76	0.93

achieving higher precision and comparable recall across categories, particularly for accurate information. In contrast, Gemini adopted a more liberal criterion, showing higher recall but lower precision for accurate statements and weaker performance in detecting overgeneralized or inaccurate content. For overgeneralized statements, ChatGPT displayed a moderate ability to capture nuanced generalizations, whereas Gemini appeared less sensitive to such distinctions. Both models struggled with inaccurate information; however, variability was also observed among human raters, suggesting that identifying misleading or false claims is inherently challenging, even for experts. Thus, while LLMs can approximate expert evaluations for accurate content, they remain limited in recognizing misinformation in user-generated material.

Discussion

This study investigated the accuracy of autism-related information shared on TikTok under the hashtag *#Autismo* (*#Autism*), comparing human expert evaluations with the judgments of two LLMs: ChatGPT 4.0 mini and Gemini 1.5 Flash. To minimize personalization bias, a new TikTok account was created for data collection, resulting in 408 distinct informational statements extracted from 148 Italian-language videos.

Three clinical experts independently classified each statement as inaccurate, overgeneralized, or accurate, and the median of their ratings was used as the human-derived ground truth.

Overall, human raters showed moderate agreement, and most statements were judged to be accurate (approximately 70%), followed by overgeneralized (21%) and inaccurate (9%). Hence, the majority of autism-related content retrieved through the hashtag *#Autismo* appears to align with current scientific understanding, although a meaningful proportion of overgeneralized or misleading claims persists. At the same time, the moderate agreement observed among human raters highlights the inherent subjectivity involved in assessing complex health-related information. This is further reflected in the category-level agreement patterns: while human raters showed high specific agreement when identifying accurate statements, their agreement was substantially lower for both overgeneralized and especially inaccurate content. Such discrepancies indicate that evaluators converged most readily on scientifically grounded claims, whereas determining whether a statement constituted misinformation proved more challenging. Evaluating accuracy in this context requires interpreting both clinical and experiential perspectives, making complete consensus unlikely even among experts. This finding supports the need for data-driven and replicable assessment frameworks capable of minimizing individual bias and promoting greater consistency across evaluators. In the long term, such frameworks could be integrated into social media ecosystems as automated alert systems, providing users with contextual cues or reliability indicators before engagement or sharing.

In contrast to previous studies reporting a predominance of inaccurate or misleading autism-related information on TikTok (e.g., Aragon-Guevara et al., 2025; Brennan et al., 2025; Ononuju & Ujari, 2025), the present results showed a higher proportion of accurate statements. This discrepancy is likely driven by methodological differences, and possibly by linguistic and cultural factors. First, whereas most prior work has examined English-language content, our study focused on Italian-language videos. There may therefore be differences in creator communities, advocacy practices, and audience expectations that might influence how autism-related information is produced and shared. However, a key factor is the shift from evaluating videos as holistic units to analyzing individual informational statements. This statement-level approach allows for a more fine-grained and context-sensitive assessment, capturing accurate information even within videos that may also contain overgeneralized or misleading elements. As a result, accuracy rates derived from statement-based coding may yield higher estimates of factual correctness.

Regarding the performance of the LLMs, ChatGPT's agreement with the human-derived ground truth was comparable to that observed among human raters, suggesting that advanced models can approximate expert-level evaluations of informational accuracy. The two LLMs displayed

distinct evaluative profiles. ChatGPT 4.0 mini adopted a conservative decision criterion, tending to underestimate accuracy rather than overstate it, and consequently produced fewer false positives for accurate information. In contrast, Gemini 1.5 Flash applied a more liberal threshold, showing high recall but lower precision for accurate statements, and frequently labeling as accurate content that human experts rated as overgeneralized or inaccurate. This tendency may partly reflect the effects of model alignment objectives, including safety features and guardrails, which are designed to favor inclusive, non-restrictive responses. Such training may bias the model toward accepting broadly framed or generalized statements as valid, particularly in sensitive health-related contexts, in order to avoid overly corrective or exclusionary judgments. While the internal mechanisms of proprietary models are not fully transparent, this interpretation offers a plausible explanation for Gemini's reduced sensitivity to overgeneralization. Overall, ChatGPT provided more balanced and cautious evaluations, suggesting that it may currently represent a more dependable tool for automated assessments of health-related information. An additional consideration concerns the class imbalance observed in the dataset, with only 9.32% of statements classified as inaccurate. This scarcity can affect the stability of agreement metrics and precision/recall estimates for this category, as Kappa-based measures and category-specific agreement are sensitive to low-prevalence classes (Ferri et al., 2009; Rácz, Bajusz, & Héberger, 2019). Consequently, the reduced agreement observed for inaccurate statements, particularly for Gemini, should be interpreted with caution, as it may partially reflect a statistical artifact of the unbalanced distribution rather than a clear limitation in reasoning ability. Future studies could address this issue by including larger or more balanced samples of inaccurate content.

Taken together, these findings contribute to the growing literature on the reliability of health-related information circulating on social media platforms such as TikTok (e.g., Aragon-Guevara et al., 2025; Brennan et al., 2025; Nordahl-Hansen et al., 2018). The relatively high proportion of accurate statements observed here may indicate a gradual shift toward more evidence-informed and neurodiversity-oriented communication about autism. This evolution could reflect both increased public awareness and the growing participation of neurodivergent creators and healthcare professionals in shaping the online discourse. At the same time, the moderate inter-rater and model–human agreement highlights that evaluating the quality of psychological and medical information remains a complex task, even for trained experts and advanced artificial intelligence systems. LLMs should therefore be viewed as complementary tools for supporting expert evaluation of health-related statements and for scalable detection of misinformation trends. However, most importantly, they should not be used as replacements for clinical expertise. Future research should explore whether model performance can be improved through domain-specific fine-tuning or prompt engineering grounded in evidence-based principles.

Finally, while the present findings are promising, several limitations should be acknowledged. The analysis was limited to Italian-language content and to a single platform (TikTok), which may constrain generalizability across languages and social media ecosystems. Moreover, our focus on verbal statements, though methodologically rigorous, did not capture paralinguistic or contextual features (e.g., tone, visuals, or engagement cues) that may influence viewers' perceptions of accuracy. In addition, the study relied on openly accessible LLMs rather than state-of-the-art models; while this choice enhances ecological validity, it may underestimate the upper-bound performance of current AI systems in this domain. Furthermore, the models were queried using a zero-shot prompting strategy. While outside the scope of the current work, prior literature (e.g., Ahmed et al., 2024; Hassanein et al., 2025; Yao et al., 2025) has shown that prompting design can modulate LLM performance in clinical tasks. Accordingly, the present results should be interpreted as a conservative estimate of LLM capabilities. Future studies could adopt multimodal and longitudinal designs and systematically examine the impact of prompt optimization to better capture to examine how LLM-based evaluations of autism-related information evolve alongside online discourses on neurodiversity.

Conclusion

As social media platforms increasingly shape public understanding of neurodevelopmental conditions (Fong et al., 2025), TikTok has become a major source of autism-related information, particularly among younger audiences. Investigating the accuracy of such content is therefore essential to understand how digital spaces contribute both to knowledge dissemination and to potential misinformation about autism (Aragon-Guevara et al., 2025; Brennan et al., 2025; Ononuju & Ujari, 2025).

The relatively high level of informational accuracy found in Italian TikTok videos suggests that online autism discourse may be evolving toward more evidence-informed communication. Nevertheless, the moderate agreement among raters and models alike underscores that assessing the quality of complex health information remains a complex and context-dependent task, particularly when identifying inaccurate or misleading content. By comparing expert human judgments with the outputs of two LLMs, we found that automated systems can partially approximate expert

evaluations of informational accuracy. Their performance was stronger for scientifically accurate content, but they still showed limitations in detecting inaccurate or misleading statements. Among the tested models, ChatGPT 4.0 mini demonstrated more conservative and reliable evaluations than Gemini 1.5 Flash, showing higher precision and fewer false positives for accurate information and more sensitivity for overgeneralizations. This suggests that ChatGPT may currently represent the more trustworthy and cautious LLM for the automated assessment of online health information.

Future research should extend these analyses across online communities, platforms, and clinical domains to further explore how LLMs can be responsibly integrated into health communication, as tools to support human expertise and to develop accuracy-based alert systems within media environments. Such systems could provide users with contextual indicators or reliability cues about online health information, promoting more critical and informed engagement with digital content.

Author Contributions Conceptualization: AC, SF, GV, DSM, DD, GE; Methodology: AC, SF, GB, GE; Formal analysis: AC, SF, GB; Investigation: SF, GB, SP, GE; Data curation AC, SF, GB; Writing-original draft preparation: AC, SF; Writing-review and editing: AC, SF, GB, SP, GV, DSM, DD, GE; Supervision: GE.

Funding Open access funding provided by Università degli Studi di Trento within the CRUI-CARE Agreement. No funding.

Declarations

Conflict of interest Giacomo Vivanti serves as an Associate Editor for *Journal of Autism and Developmental Disorders*. This author was not involved in the editorial handling, peer review, or decision-making process for this manuscript. The authors declare no other competing interests.

References

Ahmed, A., Hou, M., Xi, R., Zeng, X., & Shah, S. A. (2024). Prompteng: Healthcare prompt engineering: Revolutionizing healthcare applications with precision prompts. Companion proceedings of the acm web conference 2024 (pp. 1329–1337).

Alper, M., Rauchberg, J. S., Simpson, E., Guberman, J., & Feinberg, S. (2025). Tiktok as algorithmically mediated biographical illumination: Autism, self-discovery, and platformed diagnosis on# Autisktok. *New Media & Society*, *27*(3), 1378–1396.

American Psychiatric Association (2022). Diagnostic and statistical manual of mental disorders (5th ed., text rev. (DSM–5–TR) ed.). Washington, DC: American Psychiatric Publishing.

Aragon-Guevara, D., Castle, G., Sheridan, E., & Vivanti, G. (2025). The reach and accuracy of information on autism on Tiktok. *Journal of Autism and Developmental Disorders*, *55*(6), 1953–1958.

Botha, M., Chapman, R., Giwa Onaiwu, M., Kapp, S. K., Stannard Ashley, A., & Walker, N. (2024). The neurodiversity concept was developed collectively: An overdue correction on the origins of neurodiversity theory. *Autism*, *28*(6), 1591–1594.

Brennan, E., Lampinen, L. A., Paek, H., Wang, X., Romano, H., & Bal, V. H. (2025). Deconstructing information about autism diagnosis in adults on tiktok: A cross-sectional, descriptive content analysis. *Journal of Autism and Developmental Disorders*, 1–15.

Brown, E., Kuzmiak, F., Singh, A., Monga, V., Bell, T., Nolan, J., & Kashyap, R. (2024). A cross-sectional analysis of tiktok autism spectrum disorder content quality. Emerging Trends in Drugs, Addictions, and Health, 4, 100150.

Chapman, R. (2021). Neurodiversity and the social ecology of mental functions. *Perspectives on Psychological Science*, *16*(6), 1360–1372.

Ferri, C., Hernández-Orallo, J., & Modroiu, R. (2009). An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, *30*(1), 27–38.

Fong, S., Carollo, A., Dal Maso, M., Martinotti, G., Luciani, D., Khan, Y. S., & Esposito, G. (2024). Simple prompting enhances chatgpt's diagnostic accuracy in psychiatric cases. *NewAddictionsX*.

Fong, S., Carollo, A., Vivanti, G., Messinger, D. S., Dimitriou, D., & Esposito, G. (2025). Autism spectrum disorders discourse on social media platforms: A topic modeling study of Reddit posts. *Autism Research*, *18*(8), 1608–1619.

Guberman, J. (2023). # actuallyautistic Twitter as a site for epistemic resistance and Crip futurity. *ACM Transactions on Computer-Human Interaction*, *30*(3), 1–34.

Hassanein, F. E., Ahmed, Y., Maher, S., Barbary, A. E., & Abou-Bakr, A. (2025). Prompt-dependent performance of multimodal Ai model in oral diagnosis: A comprehensive analysis of accuracy, narrative quality, calibration, and latency versus human experts. *Scientific Reports*, *15*(1), 37932.

Huang, T., Yi, J., Yu, P., & Xu, X. (2025). Unmasking digital falsehoods: A comparative analysis of llm-based misinformation detection strategies. 2025 8th international conference on advanced algorithms and control engineering (icaace) (pp. 2470– 2476).

Hungerford, C., Kornhaber, R., West, S., & Cleary, M. (2025). Autism, stereotypes, and stigma: The impact of media representations. *Issues in Mental Health Nursing*, *46*(3), 254–260.

Jones, S. C., Gordon, C. S., & Mizzi, S. (2023). Representation of autism in fictional media: A systematic review of media content and its impact on viewer knowledge and understanding of autism. *Autism*, *27*(8), 2205–2217.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 159–174.

Nordahl-Hansen, A., Tøndevold, M., & Fletcher-Watson, S. (2018). Mental health on screen: A dsm-5 dissection of portrayals of autism spectrum disorders in film and Tv. *Psychiatry Research*, *262*, 351–353.

Ononuju, U. A., & Ujari, C. A. (2025). Stigma and misinformation about autism spectrum disorder (asd) on Tiktok and instagram: Content analysis using# asd,# autism and# Asdinfo. *Journal of Autism and Developmental Disorders*, 1–12.

Rácz, A., Bajusz, D., & Héberger, K. (2019). Multi-level comparison of machine learning classifiers and their performance metrics. *Molecules*, *24*(15), 2811.

Yao, G., Zhang, W., Zhu, Y., Wong, U., Zhang, Y., Yang, C., & Gao, H. (2025). Comparing the accuracy of large Language models and prompt engineering in diagnosing real-world cases. *International Journal of Medical Informatics*, 106026.