Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-024-03811-x

OPEN

Check for updates

AI hallucination: towards a comprehensive classification of distorted information in artificial intelligence-generated content

Yujie Sun[1], Dongfang Sheng[2]✉, Zihan Zhou[2] & Yifei Wu[2]

Amidst the burgeoning information age, the rapid development of artificial intelligence-generated content (AIGC) has brought forth challenges regarding information authenticity. The proliferation of distorted information significantly impacts users negatively. This study aims to systematically categorize distorted information within AIGC, delve into its internal characteristics, and provide theoretical guidance for its management. Utilizing ChatGPT as a case study, we conducted empirical content analysis on 243 instances of distorted information collected, comprising both questions and answers. Three coders meticulously interpreted each instance of distorted information, encoding error points based on a predefined coding scheme and categorizing them according to error type. Our objective was to refine and validate the distorted information category list derived from the review through multiple rounds of pre-coding and test coding, thereby yielding a comprehensive and clearly delineated category list of distorted information in AIGC. The findings identified 8 first-level error types: "Overfitting"; "Logic errors"; "Reasoning errors"; "Mathematical errors"; "Unfounded fabrication"; "Factual errors"; "Text output errors"; and "Other errors", further subdivided into 31 second-level error types. This classification list not only lays a solid foundation for studying risks associated with AIGC but also holds significant practical implications for helping users identify distorted information and enabling developers to enhance the quality of AI-generated tools.

[1] Shandong Normal University Library, Jinan, Shandong, China. [2] School of Management, Shandong University, Jinan, Shandong, China. ✉email: dfsheng@sdu.edu.cn

Introduction

With the rapid advancement of Generative Artificial Intelligence (GAI), the emergence of Artificial Intelligence Generated Content (AIGC) has become a reality. AIGC emerges as an innovative paradigm in content creation, succeeding Professional Generated Content (PGC) and User Generated Content (UGC), by capitalizing on artificial intelligence technology to automate the production of content (CAICT (2019)). As early as the 1950s, AIGC tried to become prominent, but due to technical limitations, research on AIGC was limited to a small range of experiments. It wasn't until the 2010s that AIGC underwent rapid development, transitioning from experimental stages to practical applications, accompanied by the emergence of various deep learning algorithms. Especially by 2022, the advent of pre-trained large-scale AI models significantly bolstered AIGC capabilities, fostering its robust development (Wang et al. 2023).

In recent years, leveraging increasingly mature big data technology and algorithmic models, AIGC has achieved significant advancements in generating text, images, video, audio, and multimodal outputs. With its inherent advantages of high efficiency and low cost, AIGC has been gradually applied in education, media, healthcare, finance, and entertainment (Rivas & Zhao 2023). For instance, in healthcare, AIGC has been employed for tasks like image processing and text generation, addressing critical challenges such as medical resource scarcity and complex procedures. The application of AIGC in healthcare has notably enhanced the quality and efficiency of medical processes (Shao et al. 2024). In the realm of art, the capabilities of AIGC in generating text and images have garnered increasing interest among designers. Its efficiency, precision, and creativity have substantially reduced the consumption of human and material resources, thereby enhancing production efficiency (Li, 2024(https://yc10.sdnu.edu.cn/s/cn/clarivate/webofscience/G.https/wos/author/record/36784494)). The advantages of high efficiency and low cost demonstrated by AIGC have attracted the attention of major industries, and its development potential has gradually garnered interest from various fields.

Based on big data technology and algorithmic models, AIGC has achieved the generation of diverse content types, including text, images, videos, audio, and multi-modal formats. Benefiting from its attributes of high efficiency and cost-effectiveness, AIGC has found gradual integration into various sectors such as education, media, healthcare, finance, and entertainment (Rivas & Zhao 2023). The efficiency improvement brought by AIGC to various industries has also attracted people's attention to the development potential of AIGC.

However, in the current era of digital information explosion, the rapid evolution of AIGC has posed challenges regarding the authenticity, compliance, and accuracy of information (Zhou & Zafarani 2020). Recently, a new term has emerged to describe the occurrence of disinformation within artificial intelligence systems: AI hallucination. Currently, research on AI hallucination has garnered significant attention among scholars. While exploring the beneficial impacts of AIGC on specific industries, some scholars have also raised concerns about the hallucination of AI. For instance, in an examination of ChatGPT's influence in education, scholars expressed apprehensions about potential hallucinatory responses from ChatGPT, emphasizing the need for caution when utilizing such AI systems (Huang 2023). Additionally, certain researchers have developed mathematical models to thoroughly assess the hallucinatory tendencies and creative capabilities of GPT. They argue that hallucination represents an inherent trait of the GPT model and suggest that completely eradicating hallucinations without compromising its high-quality performance is nearly impossible (Lee M 2023). Addressing the phenomenon of hallucination in AI presents a pressing challenge, as minimizing its adverse effects remains crucial, even if complete elimination is unfeasible. Scholars have proposed various strategies to mitigate these issues, yet these approaches are not without their limitations (Ji et al. 2023).

While AI hallucination continues to be a hot topic in scholarly discourse, there remains disagreement among researchers regarding the precise application of the term "hallucination". While the term "hallucination" has gained acceptance among certain scholars for characterizing the irrational behaviors exhibited by artificial intelligence systems, there remains significant dissatisfaction among others regarding this specific nomenclature. Moreover, "AI hallucination" has not yet solidified into a universally agreed-upon definition, with scholars independently exploring various interpretations of the term "hallucination" (Rawte et al. 2023). Some scholars define AI hallucination as "instances where an AI chatbot generates fictional, erroneous, or unsubstantiated information in response to queries" (Kumar et al. 2023). Moreover, a study posited that within the domain of large language models (LLMs), "hallucination" can be categorized into three distinct types: "Input-conflicting hallucination"; "Context-conflicting hallucination"; and "Fact-conflicting hallucination" (Liu et al. 2024). Alternatively, some scholars advocate for the term "AI fabrication" as a replacement for "AI hallucination" to denote instances where AI systems generate false information (Christensen 2024).

Regardless of the terminology employed to delineate this phenomenon of "hallucination" in artificial intelligence, its existence undeniably yields detrimental repercussions for people. On the one hand, from the users' perspective, this phenomenon has incited concerns over the veracity of information. On the other hand, distorted information generated by artificial intelligence tends to be more persuasive during transmission, potentially exacerbating issues of network security and online fraud. (Polyportis & Pahos 2024). This is because the proliferation of false information, made feasible by advancements in artificial intelligence, has significantly lowered the barriers to entry and heightened the deceptive potential (Casero-Ripollés et al. 2023). There have been numerous cases of "AI hallucination" being abused to commit illegal acts, leading to potential threats to the social economy and national order. For example, in the year 2023, an AI-generated image of an explosion near the Pentagon in the United States was widely circulated, resulting in a significant decline in the U.S. stock market due to the impact of the photo (Huanqiu 2023). Thus, the phenomenon of "hallucination" in artificial intelligence warrants thorough discussion.

In essence, "AI hallucination" refers to the phenomenon where artificial intelligence generates distorted information. Scholars have defined distorted information as "false or inaccurate information regardless of intentional authorship" (Chen 2023). When categorized based on the author's intent, distorted information bifurcates into two types: "disinformation", denoting intentionally falsified content (Dragomir et al. 2024), and "misinformation", which lacks deliberate fabrication (Komendantova et al. 2021). The differences of AI hallucination, distorted information, disinformation and misinformation are shown in Table 1. In this study, given our focus on artificial intelligence generation systems and acknowledging the unconscious nature of AI systems alongside the intentions of developers, we encompass both instances of distorted information within the purview of our research.

Research on "distorted information" has been a topic of significant interest among scholars from diverse disciplines for a considerable period. In the field of disinformation research, scholars have categorized their focus into several key areas. These

Table 1 The Differences of AI Hallucination, Distorted Information, Disinformation and Misinformation.

Concept	Explanation
AI Hallucination	"AI hallucination" has not yet solidified into a universally agreed-upon definition. In our research, "AI hallucination" refers to the phenomenon where artificial intelligence generates distorted information.
Distorted Information	Distorted information refers to false or inaccurate information regardless of intentional authorship.
Disinformation	Disinformation refers to deliberately fabricated distorted information.
Misinformation	Misinformation refers to inadvertently produced distorted information.

include typology research, fact-checking studies, analyses of disinformation on digital platforms and media literacy investigations. A significant emphasis has been placed on understanding the dissemination of disinformation and how it influences citizens' information behaviors (Salaverría & Cardoso 2023). For instance, some scholars have explored the impact of moral contagion on the spread and recognition of disinformation through empirical research (Brady et al. 2020). Some scholars have also discussed the impact of national response capabilities to disinformation on citizens' attitudes (Wilson & Wiysonge 2020). The research field of misinformation is similar to that of disinformation, but more attention is paid to the governance of misinformation. For instance, some scholars have shown that psychological inoculation improves the public's ability to guard against misinformation in social media (Roozenbeek et al. 2022). Scholars have also conducted categorizations of health misinformation circulating on social platforms, aiming to inform strategies for combating misinformation (Suarez-Lledo & Alvarez-Galvez 2021). In addition, the detection of disinformation and misinformation has also attracted wide attention, and automatic detection technology based on machine learning has been discussed by many scholars (Ahmad et al. 2020).

Currently, in the context of the rapid development of artificial intelligence, the phenomenon of distorted information in AIGC has become the focus of contemporary research. Particularly in the context of artificial intelligence, the phenomenon of distorted information in AIGC has emerged as a focal point of contemporary research. For example, in a study examining the accuracy of ophthalmic information provided by ChatGPT, Cappellani demonstrated that ChatGPT could provide incomplete, incorrect, and potentially harmful information about common ophthalmic diseases (Cappellani et al. 2024). Additionally, researchers found fabrications and errors in bibliographic citations generated by ChatGPT (Walters & Wilder, 2023). However, although many scholars have recognized the potential for AIGC to generate distorted information, few have delved into the inherent characteristics underlying this distorted information. Research investigating the specific types of distorted information in AIGC is also limited. Consequently, to address this research gap, we aim to conduct a study on the distorted information in AIGC using content analysis method. Our objective is to develop a relatively comprehensive list of the various forms of distorted information that can emerge from AIGC.

Currently, some scholars have conducted research on the classification of distorted information (see Table 2). For a detailed explanation of the categories in Table 2, see Supplementary Table S1. As previously mentioned, we have summarized prior studies into two categories of distorted information: "disinformation" and "misinformation". In the realm of "disinformation" categorization, a study conducted a literature review and identified 11 types of disinformation, including "Fabrication"; "Impostor"; "Conspiracy theory"; "Hoax"; "Biased or one-sided"; "Rumors"; "Clickbait"; "Misleading connections"; "Fake reviews"; "Trolling"; and "Pseudoscience" (Kapantai et al. 2021). Researchers provided 8 types of disinformation found on the Internet, namely "Fabricated"; "Propaganda"; "Conspiracy Theories"; "Hoaxes"; "Biased or one-sided"; "Rumors"; "Clickbait"; and "Satire News" (Zannettou et al. 2019). Moreover, a study mentioned 7 general categories of disinformation, namely "Fabrication"; "Manipulation"; "Misappropriation"; "Propaganda"; "Satire"; "Parody"; and "Advertising" (James et al. (2018)). Regarding the study of "misinformation", Carlos Carrasco-Farre analyzed 92,112 news articles were analyzed to explore the characteristics of misinformation content, and 6 types of misinformation was proposed: "Clickbait"; "Conspiracy theory"; "Fake news"; "Hate news"; "Junk science"; and "Rumor" (Carrasco-Farré 2022).

In particular, in the category study of AIGC distorted information, Borji categorized directly the erroneous outputs of ChatGPT into 11 aspects: "Reasoning"; "Logic"; "Math and arithmetic"; "Factual errors"; "Bias and discrimination"; "Wit and humor"; "Coding"; "Syntactic structure, Spelling, and Grammar"; "Self awareness"; "Ethics and morality"; "Other failures" (Borji 2023). While this study illustrates these 11 categories through definitions and examples, it's noted that these categories are not exhaustive. Moreover, researchers integrated the classification results from scholars and major databases. They extracted and reclassified ChatGPT misclassifications by establishing rules and eliminating supplementary methods. As a result, they identified 7 categories of errors: "Factual errors"; "Logic errors"; "Reasoning errors"; "Programming errors"; "Text output error"; "Overfitting"; "Synthesis problems". These error categories encompass a total of 23 error items (Fang, Tang (2023)). However, it should be noted that the establishment of rules in this study was done independently, resulting in a degree of subjectivity. Furthermore, the rules have not been validated with incorrect examples, which calls for caution when evaluating the comprehensiveness of the classification results. Mo' study conducted tests on the AIGC tools, acquired firsthand test data, and collected secondary data from social media platforms. They categorized the types of false information errors based on generation mechanism and manifestation form. They identified 5 types of factual errors: "Data errors"; "Author's work errors"; "Objective fact errors"; "Programming code errors"; "Machine translation errors". Additionally, they identified 4 categories of hallucinatory errors: "False news events"; "False academic information"; "False health information"; "Bias and discrimination" (Mo et al. 2023). Similar to Borji's work, this study demonstrates these categories through definitions and examples, without explaining the specific classification process.

Moreover, there are several databases available for compile instances of distorted information in AIGC, and some of them also classify the distorted information in AIGC. However, these databases often lack standardized classification principles. For example, NewsGuard extensively catalogs and monitors false narratives circulating online, which include instances of AIGC failures (NewsGuard 2023). Nevertheless, the platform only provides summaries of AIGC failures without categorizing them into specific distortion types. Similarly, the ChatGPT/LLM error tracker on the *Typeform.com* is accessible to all users, where they can report errors encountered when using ChatGPT and provide details of the original incorrect answers generated by ChatGPT

Table 2 Classification Types of Distortion Information from Current Related Research.

Classification Source	Category
Kapantai et al.	Fabrication; Impostor; Conspiracy theory; Hoax; Biased or one-sided; Rumors; Clickbait; Misleading connections; Fake reviews;Trolling; Pseudoscience
Zannettou et al.	Fabricated; Propaganda; Conspiracy Theories; Hoaxes; Biased or one-sided; Rumors; Clickbait; Satire News
Jame et al.	Fabrication; Manipulation; Misappropriation; Propaganda; Satire; Parody; Advertising
Carrasco-Farré	Clickbait; Conspiracy theory; Fake news; Hate news; Junk science; Rumor
Borji	Reasoning (Spatial reasoning; Temporal reasoning; Physical reasoning; Psychological reasoning; Commonsense reasoning); Logic; Math and arithmetic; Factual errors; Bias and discrimination; Wit and humor; Coding; Syntactic structure, Spelling, and Grammar; Self awareness; Ethics and morality; Other failures
Mo, Z. et al.	Data errors; Author's work errors; Objective fact errors; Programming code errors; Machine translation errors; False news events; False academic information; False health information; Bias and discrimination;
Fang, S. & Tang, Q.	Factual errors (Fabricated facts; Common sense mistakes); Logic errors (Causal relationships; Measurement units; Contradictions); Reasoning errors (Spatial; Physical; Temporal; Age-related; Metaphorical; Psychological); Programming errors (Mathematical formulas; Data errors); Text output errors (Code generation; Spelling and grammar; Repetition and redundancy); Overfitting (Illusions of confidence; Flattery; Imitation of attitudes; Falling into traps); Synthesis problems (Bias and discrimination; Ideology; Restrictive filtering)

(Typeform 2023). Nevertheless, the website relies on users' individual judgment to categorize distorted information, lacking a unified classification standard. By the end of 2023, it had accumulated approximately 40 types of distortion. Furthermore, the Giuven95 website includes error messages from ChatGPT and Microsoft Bing as of February 16, 2023 (GitHub (2023)). However, this site also lacks a unified classification principle, with some errors categorized by tool type and others classified by time. It has not achieved continuous updates and has ultimately recorded a total of 19 distortion categories.

After reviewing the aforementioned studies, it is evident that some scholars have attempted to categorize distorted information present in social media. However, it should be noted that a distinction exists between distorted information found in social media and that in AIGC. Consequently, the classification of distorted information in social media may not be directly applicable to AIGC. In particular, certain scholars and websites have made efforts to classify distorted information in AIGC. Nevertheless, there remain several shortcomings in both the methods and results of these classifications.

Primarily, the majority of scholars and websites tend to rely solely on distorted samples to define error types, without employing rigorous scientific methodologies. This approach is inherently subjective and lacks the rigor necessary for accurate classification. Secondly, there is a lack of standardized criteria and scientific frameworks for classifying distorted information in AIGC among most scholars and websites. Furthermore, the comprehensiveness of the classification results is often not adequately tested. Lastly, while some scholars have delineated categories of distorted information in AIGC, there remains a lack of clear definitions for each category, leading to difficulties in comprehension. These ambiguities hinder the effectiveness of classification efforts and impede further understanding of the phenomenon.

To address the subjective issues in classification methods and mitigate the lack of defined categories and conceptual fragmentation, our study aims to accomplish the following objectives: Firstly, by incorporating scholars' perspectives on the classification of distorted information in both social media and AIGC, we will construct an initial classification framework through systematic integration and refinement. Secondly, this initial classification framework will be refined and validated using precise samples of distorted information, resulting in the development of a comprehensive AIGC distortion information category list characterized by clear concepts, exhaustive categories, and standardized classification criteria. Finally, based on the detailed types of error messages in the architecture, we propose suggestions for users to identify or prevent error messages.

In this study, we will use ChatGPT as a case study to explore the distorted information of AIGC. However, it is noteworthy that the samples analyzed were collected between January 1, 2023 and December 31, 2023. A total of 243 valid samples were included, with 218 originating from GPT-3.5 and the remaining 25 from GPT-4. ChatGPT is a large-scale pre-trained artificial intelligence language model developed by OpenAI in 2022. Equipped with the capability to learn and comprehend human language for simulating human conversation, ChatGPT can process input from documents and images to produce text, translations, summaries, and question-and-answer responses (Bašić et al. 2023). Research has demonstrated that ChatGPT can exhibit human-like characteristics and can be humanized through algorithms (Abdulrahman et al. (2023)). Moreover, according to OpenAI, ChatGPT performs on par with humans across various professional and academic benchmarks. However, OpenAI also acknowledges that ChatGPT occasionally generates responses that may sound reasonable but are factually incorrect (OpenAI 2023). We gathered 243 original samples of chatGPT distortion information from *Typeform.com* and subjected them to content analysis. *Typeform.com* is a robust platform for form creation and data storage, allowing users to generate a wide range of forms and efficiently store and handle the collected data directly on the platform. The researcher utilized *Typeform.com* to develop a ChatGPT/LLM error tracker, designed to collect distorted information generated by ChatGPT (Typeform 2023). This information includes screenshots of the original conversations, submission timestamps, and other relevant data. The rationale for selecting the *Typeform.com* is threefold: Firstly, the number of samples collected in this website is the largest and most complete. Secondly, the samples acquired from this platform are the most exhaustive and original. Thirdly, the samples in this website are constantly updated, which can ensure the novelty of the samples. Specifically, we employed the initial classification framework obtained from the review as a coding scheme. Subsequently, we randomly coded 70% of the 243 samples to establish a category list tailored specifically to AIGC distorted information. It is worth noting that since a sample may contain multiple error points, we extracted the error points in the sample and took the error points as the analysis unit, finally obtaining 202 error points. Following the completion of pre-coding, we utilized the remaining 30% of the samples to scrutinize the adjusted category list, ensuring the absence of any new distorted categories and upholding the integrity of the category list.

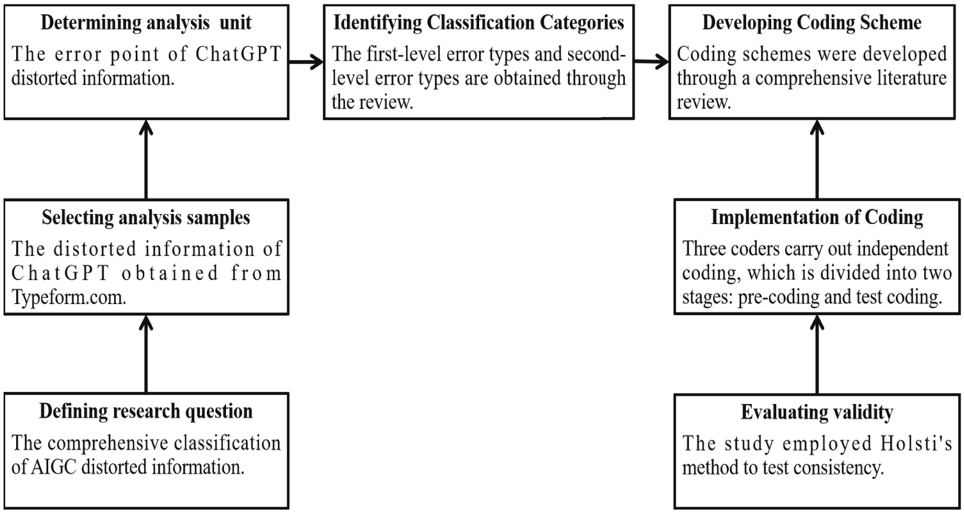
Determining analysis unit
The error point of ChatGPT distorted information.
Identifying Classification Categories
The first-level error types and second-level error types are obtained through the review.
Developing Coding Scheme
Coding schemes were developed through a comprehensive literature review.
Selecting analysis samples
The distorted information of ChatGPT obtained from Typeform.com.
Implementation of Coding
Three coders carry out independent coding, which is divided into two stages: pre-coding and test coding.
Defining research question
The comprehensive classification of AIGC distorted information.
Evaluating validity
The study employed Holsti's method to test consistency.

Fig. 1 The seven specific implementation steps of content analysis. The process of data analysis.

Methods

Content analysis. The research primarily employed the content analysis method. In 1952, Bernard Berelson first provided the definition of content analysis, viewing it as a research method that offers an objective, systematic, and quantitative description of explicit content (Berelson (1952)). With the development of content analysis, the basic steps followed in current research are as follows (Lee et al. (2020)):

(1) Defining research question. Content analysis is commonly employed to describe specific phenomena, presenting results in terms of concepts or categories. This method finds widespread use in fields such as communication, journalism, sociology, psychology, and business (Elo & Kyngäs 2008). For this study, our research question focused on comprehensively categorizing AIGC distorted information.

(2) Selecting analysis samples. Samples for content analysis can include books, papers, web content, social media posts, comments, speeches, photos, or videos (Anastasiei & Georgescu 2020). If the study scope is extensive, sampling techniques are used. In this study, we analyzed original text conversations of ChatGPT distorted information collected from *Typeform.com*.

(3) Determining analysis unit. The unit of analysis is crucial as the smallest element of content analyzed, such as individual words, phrases, or topics (Anastasiei & Georgescu 2020). The choice of analysis unit should align with the research objective. In our study, the unit of analysis was identified as the error points in ChatGPT distorted information.

(4) Identifying Classification Categories. The categories established in this study should be exhaustive and mutually exclusive (White & Marsh 2006). “Exhaustive” implies that each unit of analysis must fit into one category. “Mutual exclusion” means that each unit of analysis belongs exclusively to one category. In this research, we developed both first-level and second-level category catalogs based on prior studies of distorted information categories.

(5) Develop Coding Scheme. The coding scheme serves as a guiding document for the coding process, featuring a complete definition, straightforward instructions, and illustrative examples (White & Marsh 2006). This study’s coding scheme was crafted through a comprehensive literature review, ensuring its clarity and effectiveness.

(6) Implementation of Coding. Coding is a critical phase of the study typically carried out by two or more coders. The coding process comprises three main stages: pre-coding, modifying the coding scheme, and formal coding (Lee et al. (2020)). In this study, three coders implemented the coding scheme. Specifically, the entire coding process was divided into two key phases. The first phase, called pre-coding, aimed to validate the coding scheme with the majority of samples. After this process, we made changes to the coding scheme. This involved eliminating categories that were not relevant to ChatGPT’s distorted information and adding missing categories to the original coding scheme. The second phase, known as test coding, utilized the smaller remaining sample size to validate the modified category list of distorted information. This step ensured the thoroughness and completeness of the list, enhancing its reliability for subsequent analyses.

(7) Evaluating validity.This process mainly evaluates the degree of consistency between two or more coders’ independent coding (Liu et al. (2019)). Validity can be assessed using different methods, such as Cohen’s kappa and Holsti’s method. In this study, Holsti’s method was employed to test consistency.

Our research strictly followed the aforementioned steps, and the implementation process is illustrated in Fig. 1.

Sample selection of distorted information. The research selected samples from *Typeform.com* spanning from its inception to December 31, 2023, and identified samples requiring analysis in the final study through a screening process. Initially, we excluded samples lacking screenshots of the original human-computer conversation. While *Typeform.com* offers various content, such as the user’s own assessment of distortion categories, the user’s error reports, and screenshots of conversations with ChatGPT, certain submissions lacked screenshots of the original human-computer dialogue. Consequently, as the authenticity of these samples could not be assured, we opted to exclude them from the analysis. Subsequently, duplicate samples were addressed by retaining only one instance. Throughout the sample collection phase, we encountered numerous identical submissions originating from either the same user or different users. This duplication phenomenon may arise due to users submitting samples multiple times during the submission process or sharing sample information on social media platforms, resulting in duplicate submissions by other users. As a result, only one instance of duplicate samples was retained for analysis.

After screening the samples, 243 samples were included in the study process. We randomly sampled 70% of the samples (170 samples) for the pre-coding stage. It is important to highlight that our coding analysis is based on the error point as the unit of analysis. A single sample may contain multiple error points. Therefore, before the formal coding process, we extracted the error points from these 170 samples, resulting in a total of 202 analysis units. For the remaining 30% of the samples (73 samples), we extracted error points and obtained a total of 82 analysis units for the test coding phase.

Coding scheme. The encoding scheme employed in this study is derived from the integration and adjustment of distorted information categories obtained in existing scientific studies. As noted in the preceding review, several scholars have conducted categorized studies on distorted information in social media and AIGC. Kapantai, Zannettou, and Jame respectively identified 11, 8, and 7 categories of disinformation in social media. Carrasco-Farre categorized misinformation in social media into 6 distinct types. Additionally, Borji, Fang, and Mo respectively conducted a classification study on AIGC, delineating 11, 23, and 9 categories of distorted information. To ensure the integrity of our coding scheme, we compiled all 75 categories of distorted information identified by these scholars. Within these categories, duplicates and similar classifications proposed by different scholars were reconciled through discussions among three coders. For instance, categories like “Clickbait” proposed by Kapantai, Zannettou, and Carrasco-Farre, were merged into a single category under the label “Clickbait”. Similarly, the categories “Hoax” and “Hoaxes” initially identified separately by Kapantai and Zannettou, were consolidated and labeled as “Hoax”.Categories with similar labels differing only in grammatical form or quantity were treated as identical.Furthermore, categories that were conceptually akin but labeled differently underwent detailed discussions among the coders to determine appropriate mergers. Unanimously, representative and inclusive terms were selected as alternative labels for these categories.For instance, in the category of “Fake news”, Mo defines it as “False news events”, whereas Carrasco-Farre characterizes it as “Fake news”. However, both scholars’ descriptions of this concept are similar, prompting us to consolidate them under the unified label of “Fake news”. It is noteworthy that following the consolidation of all concepts, the three coders deliberated and decided to exclude the terms “Fabrication” and “Rumor” due to their broad and ambiguous nature (see Supplementary Table S2 for merged categories).

Drawing from the interpretations of distorted information categories in previous scholarly research, we provide a novel interpretation of the consolidated categories while endeavoring to maintain the original meaning as much as possible. These consolidated categories will serve as the second-level categories within the coding scheme. To enhance the generality and interpretability of AIGC distortion information categories, we conducted feature extraction on the second-level categories. Subsequently, we grouped categories with similar characteristics into first-level categories. This process culminated in the development of the coding scheme, comprising 12 first-level error types and 40 second-level error types (see Table 3). It is crucial to highlight that all categories within our coding scheme are grounded in findings from prior research. Apart from the categories detailed earlier, where we amalgamated duplicates and omitted two broad categories, the labels for the remaining categories originate from the original descriptions in cited studies. Hence, while certain categories may not perfectly align with our research objectives, we have opted to retain them all to uphold the scientific rigor of our study. Subsequent coding efforts will focus on refining the scheme further.

Coding process. The coding work was conducted by a team of three coders. Prior to commencing the formal coding process, all three coders underwent standardized training to ensure complete alignment in their comprehension of the coding scheme. Moreover, to prevent fatigue reaction among coders during the coding process from affecting the coding results, each coder was required to compile an analysis memo to sustain focus (Kleinheksel et al. 2020). Once consensus was reached among the coders, formal coding commenced, dividing the entire process into two distinct phases:

The first phase, termed pre-coding, involved using 70% of the samples randomly selected from the 243 samples included in the study for the pre-coding material, with the specific analysis units being the 202 error points extracted from them (Examples of the sample materials are provided in Supplementary Table S3). The primary objective of pre-coding was to test the initial coding scheme and adjust it according to the actual situation of the samples, in order to form a complete and scientifically sound distorted information category list. The pre-coding process involved three coders independently assessing the 202 analysis units in a “back-to-back” manner, meaning that each coder encoded without knowledge of the others’ encoding results. The minimum standard for consistency among the three coders should be between 70% and 80% (Boettger & Palmer 2010). Therefore, multiple rounds of encoding may be conducted during the pre-coding process to meet this standard. Once the standard is met, the coders will discuss the encoding results to reach a unanimous conclusion on all encoding outcomes and adjust the coding scheme accordingly.

The second phase, termed test coding, involves the utilization of the refined coding scheme established post pre-coding. Three coders are responsible for encoding the remaining 30% of the samples. Consistency in coding requirements is maintained as specified during the pre-coding phase, ensuring alignment and coherence throughout the coding process. The objective is to determine the completeness of the adjusted coding scheme. If all samples are successfully classified within the framework of the adjusted coding scheme during the verification process, it validates the scheme’s suitability as a comprehensive directory of distorted information generated by ChatGPT. However, if new categories of distorted information emerge during coding, it signals the need to revisit the pre-coding phase. In such instances, additional samples are sourced for testing in order to ensure the robustness and exhaustiveness of the coding scheme until no further new categories are identified. As a result, the category list for AIGC distortion information is established.

Results

Coding results. During the first round of pre-coding, all three coders identified additional types of distorted information that did not fit into existing category within the coding scheme. Furthermore, upon reviewing the coding results, it became evident that certain categories in the scheme were not applicable to the distorted information samples generated by ChatGPT. After completing the first round of pre-coding, the consistency of the coding results among the three coders was evaluated, resulting in a 68% match, below the expected threshold and indicated inadequate reliability. Consequently, the three coders convened to discuss the observed coding discrepancies and conducted a second round of pre-coding. During this phase, the coders re-

Table 3 Coding Scheme.

First-level error types	Second-level error types	Explanation	Source
Overfitting	Illusions of confidence	Overconfidence and over-reliance on one's own judgments and decisions, ignoring other information and possible errors, lead to wrong judgments and decisions.	Fang, S. & Tang, Q.
	Falling into traps	Fall into a trap or scheme set by the questioner.	Fang, S. & Tang, Q.
	Flattery	Generate false, exaggerated, or one-sided content to please or cater to the wishes and expectations of the audience.	Fang, S. & Tang, Q.
Logic errors	Causal uncorrelation	There is no clear causal relationship between the generated content and the associated problem or context, meaning that whether an event occurs or not does not impact the occurrence of another event.	Fang, S. & Tang, Q.; Borji
	Contradictions	Situations in which the generated content is self-contradictory or inconsistent.	Fang, S. & Tang, Q.; Borji
Reasoning errors	Spatial reasoning errors	Unable to understand and control the relationships between objects, people and locations in the physical space around us. It involves visualizing and mentally transforming objects in 2D or 3D space and recognizing patterns, transitions, and relationships between objects.	Fang, S. & Tang, Q.; Borji
	Temporal reasoning errors	Unable to reason and predict events and their chronological order. It involves understanding the temporal relationship between events, the duration of events, and the time of events relative to each other.	Fang, S. & Tang, Q.; Borji
	Physical reasoning errors	Unable to understand and control physical objects and their interactions in the real world. It involves applying physical laws and concepts to predict and explain the behavior of physical systems.	Fang, S. & Tang, Q.; Borji
	Psychological reasoning errors	Unable to understand and predict human behavior and mental processes. It involves applying psychological theories, models, and concepts to explain and predict human behavior and mental states.	Fang, S. & Tang, Q.; Borji
	Satire	Unable to comprehend the true meaning of satire, as well as using humor and exaggeration to present factual information in order to mock, reveal, and criticize individuals, narratives, or viewpoints.	Borji; Zannettou et al.; Jame et al.
	Metaphorical errors	Unable to understand the true meaning of metaphors, which involves using one thing to imply another.	Fang, S. & Tang, Q.
Mathematical errors	Conceptual errors	Unable to understand the meaning of certain mathematical concepts. This includes an inaccurate understanding of concepts such as fractions, decimals, positive and negative numbers, or an incorrect understanding of the properties of graphs when dealing with geometric problems.	Fang, S. & Tang, Q.
	Measurement units errors	Unable to understand the meanings represented by units of measurement.	Fang, S. & Tang, Q.
	Calculation errors	Unable to perform correct mathematical operations, including basic counting, comparison, addition, subtraction, multiplication and division, and complex mathematical operations.	Fang, S. & Tang, Q.; Borji
Unfounded fabrication	False health information	The system automatically generates health information that contains false opinions, false arguments, or false cases based on user prompts.	Mo, Z., et al.
	Fake reviews	The system generates any review that is not an actual consumer's honest and impartial opinion or that does not reflect a consumer's genuine experience of a product, service or business.	Kapantai et al.
	Fake news	The system completely fabricates information based on user questions, generates deceptive content, or grossly distorts actual news reports.	Mo, Z., et al.; Carrasco-Farré
	Pseudoscience	The system generates claims such as metaphysics, naturalistic fallacies, and other scientifically dubious claims.	Kapantai et al.; Carrasco-Farré
	False academic information	It mainly includes fictitious papers, apparently irrelevant fictitious references in reviews, and non-existent web links or irrelevant links.	Mo, Z., et al.
	Conspiracy theory	Stories without factual base as there is no established baseline for truth. They usually explain important events as secret plots by government or powerful individuals.	Kapantai et al.; Zannettou et al.; Carrasco-Farré

Table 3 (continued)

First-level error types	Second-level error types	Explanation	Source
Bias and discrimination	Hate news	The generated content is intentionally derogatory to certain ethnic groups, including promoting racism, misogyny, and homophobia.	Carrasco-Farré
	Discrimination	The generated content often reflects social and cultural biases, often based on race, gender, religion, ethnicity, social status, cultural background and other factors, ignoring the principles of equality, justice and inclusion.	Borji; Mo, Z., et al.; Fang, S. & Tang, Q.
	Biased or one-sided	The generated content are extremely one-sided or biased.It often occurs in political contexts,this type is known as hyperpartisan news and are stories that are extremely biased towards a person/party/situation/event.	Kapantai et al.; Zannettou et al.
Factual errors	Common sense mistakes	Contrary to common sense or generally accepted knowledge, leading to unreasonable or incorrect conclusions or judgments.	Fang, S. & Tang, Q.
	Objective fact errors	Errors occur in objective information such as time, place, person, and data during public events.	Mo, Z., et al.; Fang, S. & Tang, Q.
	Author's work errors	These include errors in judging the works and corresponding authors, errors in the relationship between different authors and characters, errors in the content of improvised works that are different from the original works, errors in the association of different works, and fabricating representative works of fictional authors.	Mo, Z., et al.
Text output errors	Repetition and redundancy	There are parts of the text that are repetitive, tedious, or unnecessary.	Fang, S. & Tang, Q.
	Programming code errors	Producing inaccurate or suboptimal code for programming problems. Including statement errors, patchwork code, etc.	Borji; Mo, Z., et al.; Fang, S. & Tang, Q.
	Translation errors	In the translation of different languages, the translated content is inconsistent with the meaning expressed in the original language.	Mo, Z., et al.
	Grammar errors	Errors that are grammatically incorrect, irregular, or do not conform to language rules. Including lexical errors, syntax errors, inaccurate expressions, tense errors and so on.	Borji
	Spelling errors	Spelling errors in the generated content or failure to generate the correct text based on user questions.	Fang, S. & Tang, Q.; Borji
Misleading error	Misleading connections	Individual parts of the information may be factual but presented using the wrong connection (context/content). This includes impersonating authentic sources of information and using false backgrounds and false connections.	Kapantai et al.; Jame et al.
	Propaganda	The information created with the purpose to influence public perception or public opinions to benefit a public figure, an organisation or a government. Propaganda stories are profoundly utilized in political contexts to mislead people about a particular political party or nation-state.	Zannettou et al.; Jame et al.
	Clickbait	The content is generally authentic but uses exaggerated, misleading or questionable headlines, social media descriptions or images to entice the public to click.	Kapantai et al.; Zannettou et al.; Carrasco-Farré; Jame et al.
	Manipulation	To support a false narrative, manipulate or transform real information to generate content that deceives the viewer, such as photoshopping the color of an item in a picture.	Jame et al.
	Parody	Parody builds on a shared understanding of the absurdity of its claims between the author and the audience. It is built upon the interplay between possibility and absurdity. Sometimes making it hard for audiences to distinguish parody from real information.	Jame et al.
	Trolling	The act of deliberately posting offensive or inflammatory content to an online community with the intent of provoking readers or disrupting conversation.	Kapantai et al.
Other errors	Hoax	Relatively complex and large-scale fabrications which may include deceptions that go beyond the scope of fun or scam and cause material loss or harm to the victim.It contains false or inaccurate facts and is presented as legitimate facts.	Kapantai et al.; Zannettou et al.
	Restrictive filtering	Automatically ignores certain words or statements in the question, or refuses to answer tautological questions.	Fang, S. & Tang, Q.
	Self awareness	The ability to see oneself as an individual separate from others and to understand one's own thoughts, feelings, personality, and identity.	Fang, S. & Tang, Q.; Borji

evaluated samples where coding results were inconsistent in the first round. Ultimately, the consistency among the three coders increased to 89%, meeting the anticipated reliability criteria and concluding the pre-coding process. After completing two rounds of coding, the three coders deliberated on samples where a consensus had not been reached during pre-coding. Upon reaching a consensus on all samples, the coding scheme was supplemented and revised accordingly.

By utilizing the revised coding scheme and using the remaining samples, none of the three coders identified any categories beyond those delineated in the coding scheme. As a result, the coding process concluded, leading to the establishment of the ChatGPT distorted information category list.

Categories of distorted information. After numerous iterations of coding and extensive discussion, this study refined the coding scheme by removing certain types of non-compliant ChatGPT distorted information samples and introducing four new categories: “Interpersonal reasoning error”; “Hypothetical reasoning error”; “False proof”; and “Harmful information”. Subsequently, corresponding explanations were provided for these categories. Within the first-level error types, no samples classified under “misleading error” were identified, leading to the removal of this category. In the error type of “Bias and discrimination”, only the subcategory of “discrimination” was retained, resulting in the adjustment of this category to the classification of “Other errors” within the first-level error types. Following the refinement and scrutiny of the coding scheme, the ChatGPT distorted information category list, comprising 8 first-level error types and 31 second-level error types, was ultimately established (see Table 4). For sample examples within each category, please refer to Supplementary Table S4. Furthermore, we also conducted statistical analysis on the coding results (Chiplot 2024). From the 234 ChatGPT distorted information samples we collected, 284 analysis units were extracted. The statistics for these 284 analysis units are presented in Fig. 2.

Discussion

We obtained a category table of distortion information through a comprehensive literature review. Subsequently, we refined and validated this category table using real samples of distorted information generated by ChatGPT. Finally, we established a specific category list that is tailored to the ChatGPT’s distortion information. We believe that this categorization effectively captures the range of distorted information produced by AIGC systems. It includes a total of 8 first-level error types and 31 second-level error types. Specifically, the first-level error types consist of “Overfitting”; “Logic errors”; “Reasoning errors”; “Mathematical errors”; “Unfounded fabrication”; “Factual errors”; “Text output errors”; and “Other errors”. In the following section, we provide detailed explanations of each of these eight first-level error types:

Overfitting. Overfitting is when a model performs so well on the training data that it perfectly fits the noise and outliers of the training data, rather than capturing only the fundamental trend. This phenomenon leads to suboptimal performance when applied to new, unseen data. The system typically employs a fixed speech technique to response to user questions and lacks flexibility in addressing specific question requirements. It is generally easier for users to identify such error messages because such errors usually occur during multiple rounds of conversations in which the user questions the information answered by the system or the user himself asks the system the wrong question. For instance, when users challenge the system, it readily acknowledges their viewpoint and apologizes without specifically indicating any errors in the user’s question. The system often initiates these conversations with a consistent phrase like “I’m sorry”. Consequently, inconsistencies or inaccuracies in the system’s information can be detected through simple verification of external data or by continuing the conversation, especially if the questioner already has suspicions.

Logic errors. Logic errors occur when the system provides responses that contradict objective laws and the principles of normal human reasoning. These errors can largely be attributed to the limitations of GPT series models, which rely on simplistic processing techniques for handling information, such as predicting the next possible word in a sentence, rather than accurately generating and modeling information (Wu et al. 2023). Since logical errors in generative artificial intelligence models often deviate from the typical reasoning logic of humans, they are relatively easier to detect. Users can identify these issues by carefully examining the information provided by the system. In cases where users perceive an error but struggle to pinpoint the specific issue, engaging in further dialogue with the system can help raise a challenge. By conducting multiple rounds of dialogue, the error becomes more easily identified.

Reasoning errors. Reasoning errors occur when the system fails to draw logical conclusions from one or more known premises, often resulting in inaccuracies, illogical outcomes, or factual inconsistencies. The deficiency of generative artificial intelligence models in semantic understanding and logical inference, along with the cognitive limitations of the real world, are the primary factors contributing to reasoning errors. This deficiency becomes more apparent, especially when facing the known conditions of more complex inference tasks. Reasoning errors are the most common type of error in AIGC distorted information, encompassing various types of inferences in both realistic and non-realistic scenarios. These errors typically arise during conversations where users provide specific conditions and expect the system to deduce unknown information based on the given input. To identify such errors, users should carefully examine the reasoning process provided by the system. Additionally, the authenticity of information can be verified using the backward inference method.

Mathematical errors. Mathematical errors occur when the system provides erroneous responses to mathematical queries. Most of these errors pertain to mathematical operations, and the system tends to make mistakes whether faced with simple size comparisons or complex calculation formulas. To address this issue, users are advised to exercise caution when seeking answers to mathematical problems. It is recommended to cross-check the results provided by the system with other reliable calculation tools, and users should refrain from solely relying on the calculation results generated by the system. In particular, the system may exhibit biases in its understanding of mathematical concepts, as it is unable to comprehend and apply the specific meaning of certain concepts. For instance, if the system provides a definition of prime numbers, it may still enumerate pairs of non-prime numbers as prime numbers. When such errors occur, it is essential for users to possess a basic understanding of the mathematical concepts mentioned in the response. Without a grasp of these relevant concepts, identifying errors within the system’s answers can prove challenging. Therefore, it is recommended that users continue to ask relevant mathematical questions after understanding the relevant concepts and describing them to the system. This approach can help mitigate the potential for misleading or fabricated responses from the system.

Table 4 Distortion Information Category List.

First-level error types	Second-level error types	Explanation
Overfitting	Illusions of confidence	Overconfidence and over-reliance on one's own judgments and decisions, ignoring other information and possible errors, lead to wrong judgments and decisions.
	Falling into traps	Fall into a trap or scheme set by the questioner.
	Flattery	Generate false, exaggerated, or one-sided content to please or cater to the wishes and expectations of the audience.
Logic errors	Causal uncorrelation	There is no clear causal relationship between the generated content and the associated problem or context, meaning that whether an event occurs or not does not impact the occurrence of another event.
	Contradictions	Situations in which the generated content is self-contradictory or inconsistent.
Reasoning errors	Spatial reasoning errors	Unable to understand and control the location of objects.
	Temporal reasoning errors	Unable to reason and predict events and their chronological order.
	Physical reasoning errors	Unable to understand the essential properties of physical objects and to control their interactions in the real world.
	Psychological reasoning errors	Unable to understand and predict human behavior and mental processes.
	Interpersonal reasoning error	Unable to understand the interpersonal relationship between people, including blood relationship, social relationship, etc.
	Hypothetical reasoning error	Unable to deduce the correct answer based on the substance of the unreal situation, including common logical reasoning problems, logic puzzles, etc.
	Satire	Unable to comprehend the true meaning of satire, as well as using humor and exaggeration to present factual information in order to mock, reveal, and criticize individuals, narratives, or viewpoints.
	Metaphorical errors	Unable to understand the true meaning of metaphors, which involves using one thing to imply another.
Mathematical errors	Conceptual errors	Unable to understand the meaning of certain mathematical concepts. This includes an inaccurate understanding of concepts such as fractions, decimals, positive and negative numbers, or an incorrect understanding of the properties of graphs when dealing with geometric problems.
	Measurement units errors	Unable to understand the meanings represented by units of measurement, including conversions between different units.
	Calculation errors	Unable to perform correct mathematical operations, including basic counting, comparison, addition, subtraction, multiplication and division, and complex mathematical operations.
Unfounded fabrication	False health information	The system automatically generates health information that contains false opinions, false arguments, or false cases based on user prompts.
	False proof	Fabricating the proof process for scientific theorems that have been proven or not yet proven.
	Pseudoscience	The system generates disproven hypotheses such as metaphysical, naturalistic fallacies, or other scientifically dubious claims.
	False academic information	It mainly includes fictitious papers, apparently irrelevant fictitious references in reviews, and non-existent web links or irrelevant links.
Factual errors	Common sense mistakes	Contrary to common sense or generally accepted knowledge, leading to unreasonable or incorrect conclusions or judgments.
	Objective fact errors	Errors occur in objective information such as time, place, person, and data during public events.
	Author's work errors	These include errors in judging the works and corresponding authors, errors in the relationship between different authors and characters, errors in the content of improvised works that are different from the original works, errors in the association of different works, and fabricating representative works of fictional authors.
Text output errors	Repetition and redundancy	There are parts of the text that are repetitive, tedious, or unnecessary.
	Programming code errors	Producing inaccurate or suboptimal code for programming problems. Including statement errors, patchwork code, etc.
	Translation errors	In the translation of different languages, the translated content is inconsistent with the meaning expressed in the original language.
	Grammar errors	Errors that are grammatically incorrect, irregular, or do not conform to language rules. Including lexical errors, syntax errors, inaccurate expressions, tense errors and so on.
	Spelling errors	Spelling errors in the generated content or failure to generate the correct text based on user questions.
Other errors	Discrimination	The generated content often contains unfair treatment to certain groups or individuals due to identity characteristics.
	Restrictive filtering	Automatically ignores certain words or statements in the question, or refuses to answer tautological questions.
	Harmful information	The system generates obscene, pornographic, vulgar and other content information.

Additionally, users are encouraged to be vigilant and critical in their interactions with the system, employing their own understanding and employing external verification methods to ensure accuracy in mathematical problem-solving.

Unfounded fabrication. Unfounded fabrication refers to the system creating facts, data, or opinions without adequate substantiation from evidence or references, contrary to the principles of authenticity and objectivity. This type of error often goes

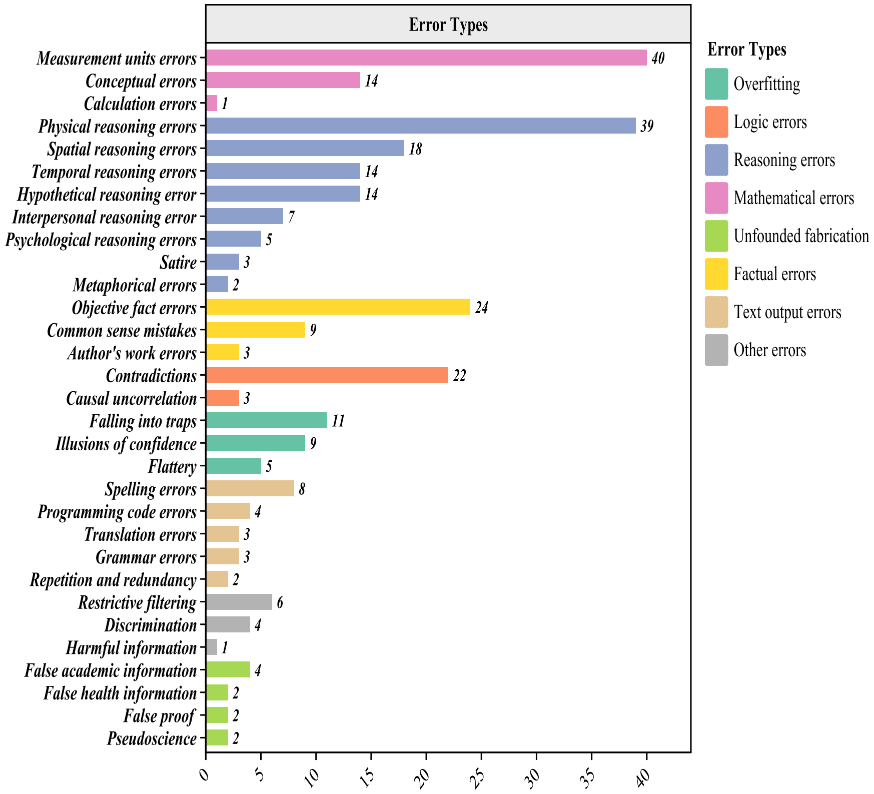

Fig. 2 The statistics for these 284 analysis units. Different colors represent different error types.

unnoticed, as the system generates responses that appear plausible but are ultimately incorrect, requiring further scrutiny by the user. It commonly occurs when users request the system to present an academic viewpoint or provide scientific evidence. In such cases, it is crucial for users to verify the answers provided by the system, particularly when attempting to locate the references cited. More often than not, these references turn out to be nonexistent. To address this issue, users planning to utilize generative artificial intelligence models to present a specific perspective can supply the system with a substantial amount of pre-collected data for analysis and request that the system generates a demonstration based on the provided reference materials. This approach also helps prevent the system from generating nonsensical responses. Moreover, the misinformation created by generative artificial intelligence models may surpass that of human propagandists in terms of writing quality, persuasion, and deceitfulness (Monteith et al. 2024). If deliberately used by certain illegal elements, it will cause great adverse consequences to society.

Factual errors. Factual errors pertain to inaccuracies in objective facts or actual data within system responses. Despite developers using vast amounts of data to pre-train generative artificial intelligence models, these models possess a substantial knowledge base. However, they lack the ability to identify errors and noise in the data, as well as retrieve information from external sources or databases. This limitation often results in inaccuracies in the generated information. Furthermore, generative artificial intelligence models rely on learning from human feedback, meaning they learn independently through constant feedback from humans on their recent actions (Koubaa et al. 2023). Unfortunately, human feedback tends to be subjective and inconsistent, making it difficult for the system to discern reliable feedback. Consequently, this can mislead the learning process of the system, leading it to incorporate and utilize inaccurate information, thereby providing users with misinformation. When users seek fact-related information, identifying incorrect information can be challenging if they lack understanding of the subject matter and do not verify it further. It is suggested that users verify information related to facts through other retrieval platforms to ensure its authenticity

Text output errors. Text output errors refer to unreasonable errors in the system's output related to the text itself. These errors commonly occurs during conversations involving translation, grammatical recognition, spelling, or the completion of innovative tasks, often resulting in the system's inability to meet user requirements. For instance, the system frequently makes mistakes in grammar recognition for languages other than English or performs poorly when tasked with generating innovative advertising slogans. This phenomenon can be attributed to the generative artificial intelligence model's limited comprehension of the relationships between words and the characters within them. In particular, the system may generate suboptimal or incorrect code, which requires the user to run the program further to verify the code's validity.

Other errors. Other errors encompass issues that cannot be categorized within the aforementioned first-level error types and lack similar characteristics to be grouped into a new first-level error types. In this study, this category encompasses 3 second-level error types: "Discrimination"; "Restrictive filtering"; and "Harmful information". It is important to note that in the field of social psychology, discrimination is defined as behavior that creates, maintains, or reinforces advantages for certain groups over other groups and their members (Bastos & Faerstein 2012). Based on the coding scheme and analysis of the original samples, it has been observed that instances belonging to this category indeed exist. However, our objective is to establish an objective and comprehensive list of categories for distorted information related to AIGC. Given the nuanced implications of "Discrimination", which involves ideological issues concerning multiple political factions, its subjectivity warrants careful consideration. The inclusion of "Discrimination" in the list of distorted information categories remains contentious. In summary, while the category of "discrimination" is included in our category table, its acceptance among readers may vary due to subjective considerations. Moreover, in adherence to research rigor, we cannot remove it without thorough discussion and supporting research evidence, which is also a major limitation in our research. Moving forward, further exploration by scholars across diverse fields may be necessary to ascertain whether "Discrimination" is appropriately categorized as a distorted category in this context.

Based on the categories of distorted information produced by AIGC, it becomes evident that AIGC is prone to generating various errors that deviate from factual accuracy and normal human logical thinking. It may even produce information that is challenging to discern as true or false. Such distorted information poses inevitable challenges for users and consequently raises awareness regarding the risks associated with AIGC.

This study holds significant theoretical implications for the exploration of distorted information within the realm of AIGC. Firstly, researchers can delve deeper into the causes and scenarios of AIGC errors using the distortion categories identified. The list of categories derived from this study can serve as a theoretical foundation for research on risks associated with AIGC. Secondly, by summarizing the categories of distorted information present in social media and AIGC, the findings of this study provides a comprehensive framework that can be utilized as a basis for investigating distorted information across different contexts. Thirdly, while the primary focus of this study is on distorted information generated by AI systems, its research methods and processes are transferable to the study of distorted information within social media. Therefore, this study can serve as a methodological reference for exploring distorted information in various platforms.

Moreover, this study has significant practical implications. Firstly, the distorted information category list extracted from ChatGPT in this research serves as a cautionary guide, advising individuals to approach artificial intelligence tools with a balanced perspective. It emphasizes the importance of avoiding excessive reliance on these tools and encourages users to exercise rational judgment based on prudent utilization. Secondly, while the distorted information category list is specific to ChatGPT, it is also applicable to other artificial intelligence generation tools similar to ChatGPT. This list becomes a guiding resource for users of artificial intelligence generation tools, helping them accurately assess the authenticity of information and enhance their overall information literacy. Users can utilize this resource to identify the locations of errors and the categories of distortion when employing ChatGPT, facilitating further verification and selection of obtained information. Thirdly, the distorted information category list is relevant for developers of artificial intelligence generation tools. It provides a theoretical foundation for the development and enhancement of future tools. Developers can leverage the insights gained from ChatGPT's distorted information category list to refine pre-training datasets, reduce error outputs, and effectively optimize the AI generation system.

In summary, this list of categories holds significant research implications both theoretically and practically. However, given the current state of technology, it is not feasible to completely eliminate the hallucination phenomenon in artificial intelligence generation tools while maintaining their high-performance standards. This category list can only serve to mitigate some of the risks associated with distorted information from AIGC. As previously noted, it can be leveraged to diminish the creation and dissemination of distorted information at both developer and user levels. Developers can utilize this list to analyze various types and causes of artificial intelligence hallucination phenomena, thereby enabling more precise model optimizations aimed at reducing the generation and impact of AIGC distorted information at its source. Simultaneously, users can develop a general psychological expectation regarding the characteristics and contexts of distorted information generated by artificial intelligence through this comprehensive classification tool. This empowers them to more accurately and promptly identify such distortions when encountered, thereby curtailing their propagation and use.

In addition, this study has several limitations. Firstly, since the distortion samples were exclusively generated from ChatGPT, it is evident that the distortion information category list of AIGC may lack comprehensive in terms of sample categories. While ChatGPT boasts the longest development history and a broad user base, other artificial intelligence generation tools may possess error categories that are not present in ChatGPT. Moreover, the sample of ChatGPT distorted information collected on *Typeform.com* does not include all cases of distortion, which may cause some categories to be missing to some extent. Therefore, future research should consider a broader selection of samples from various other artificial intelligence generation tools for analysis, and distortion information in AIGC could be gathered from multiple sources. Secondly, given OpenAI's continuous updates and enhancements to ChatGPT, the distortion categories identified in this study solely encompass error types specific to the original version of ChatGPT. With the rapid pace of technological advancement, certain error types associated with earlier versions may no longer be relevant to the current iteration. Therefore, the future research should continue to monitor, track the latest technological improvements and model performance, and constantly refine the obtained category list of distorted information. Thirdly, this study exclusively focuses on ChatGPT-like text-based artificial intelligence generation tools, overlooking the existing image, audio, video, and multi-modal artificial intelligence generation tools. Consequently, future studies could broaden their scope by investigating other forms of AIGC distortion information to provide a more comprehensive understanding of the phenomenon. Lastly, in the distorted information category list, the subjectivity of "Discrimination" is still controversial, and whether it is reasonable to include "Discrimination" in the distorted information category list can be discussed from more perspectives in future research. Additionally, it is important to note that our research is founded on conclusions drawn from previous studies and user-submitted samples. The resulting list of categories of distorted information may not be completely accepted, as perspectives among readers can vary.

Data availability

All relevant data are reflected in the article. The dataset used for sample encoding is available from the corresponding author on reasonable request.

Received: 26 March 2024; Accepted: 18 September 2024;
Published online: 27 September 2024

References

Abdulrahman E, Abdelrahim F, Fathi M, Firass A, Ali K (2023) ChatGPT and the rise of semi-humans. Humanit Soc Sci Commun 10(1):626. https://doi.org/10.1057/s41599-023-02154-3

Ahmad I, Yousaf M, Yousaf S, Ahmad M (2020) Fake News Detection Using Machine Learning Ensemble Methods. Complexity 2020:1–11. https://doi.org/10.1155/2020/8885861

Anastasiei I, Georgescu M (2020) Automated vs Manual Content Analysis – A Retrospective Look. Sci Ann Econ Bus 67:57–67. https://doi.org/10.47743/saeb-2020-0025

Bašić Ž, Banovac A, Kružić I, Jerković I (2023) ChatGPT-3.5 as writing assistance in students' essays. Humanit Soc Sci Commun 10(1):750. https://doi.org/10.1057/s41599-023-02269-7

Bastos J, Faerstein E (2012) Conceptual and methodological aspects of relations between discrimination and health in epidemiological studies. Cad de Saúde Pública 28(1):177–183. https://doi.org/10.1590/s0102-311x2012000100019

Berelson B (1952) Content analysis in communication research. Glencoe, Scotland

Boettger R, Palmer L (2010) Quantitative Content Analysis: Its Use in Technical Communication. IEEE Trans Professional Commun 53(4):346–357. https://doi.org/10.1109/TPC.2010.2077450

Borji A (2023) A Categorical Archive of ChatGPT Failures. arXiv. http://arxiv.org/abs/2302.03494

Brady W, Crockett M, Van Bavel J (2020) The MAD Model of Moral Contagion: The Role of Motivation, Attention, and Design in the Spread of Moralized Content Online. Perspect Psychol Sci 15(4):978–1010. https://doi.org/10.1177/1745691620917336

CAICT (2019) Artificial Intelligence Generated Content (AIGC) White Paper (2022) http://www.caict.ac.cn/kxyj/qwfb/bps/202209/t20220902_408420.htm. Accessed 25 Dec 2023

Cappellani F, Card K, Shields C, Pulido J, Haller J (2024) Reliability and accuracy of artificial intelligence ChatGPT in providing information on ophthalmic diseases and management to patients. EYE. https://doi.org/10.1038/s41433-023-02906-0

Carrasco-Farré C (2022) The fingerprints of misinformation: How deceptive content differs from reliable sources in terms of cognitive effort and appeal to emotions. Humanit Soc Sci Commun 9(1):162. https://doi.org/10.1057/s41599-022-01174-9

Casero-Ripollés A, Tuñón J, Bouza-García L (2023) The European approach to online disinformation: Geopolitical and regulatory dissonance. Humanit Soc Sci Commun 10(1):657. https://doi.org/10.1057/s41599-023-02179-8

Chen M (2023) A meta-analysis of third-person perception related to distorted information: Synthesizing the effect, antecedents, and consequences. Information Processing and Management. https://doi.org/10.1016/j.ipm.2023.103425

Chiplot (2024) Classification bar chart. https://www.chiplot.online/.Accessed 25 Jun 2024

Christensen J (2024) Understanding the role and impact of Generative Artificial Intelligence (AI) hallucination within consumers' tourism decision-making processes.Curr Issues in Tourism. https://doi.org/10.1080/13683500.2023.2300032

Dragomir M, Rúas-Araújo J, Horowitz M (2024) Beyond online disinformation: Assessing national information resilience in four European countries. Humanit Soc Sci Commun 11(1):1–10. https://doi.org/10.1057/s41599-024-02605-5

Elo S, Kyngäs H (2008) The qualitative content analysis process. J Adv Nurs 62(1):107–115. https://doi.org/10.1111/j.1365-2648.2007.04569.x

Fang S, Tang Q (2023) Typological analysis of ChatGPT error content generation. News and Writing. https://www.cnki.net/

GitHub (2023) LLM failure archive. https://github.com/giuven95/chatgpt-failures#llm-failure-archive-chatgpt-and-beyond.Accessed 25 Dec 2023

Huang H (2023) Performance of ChatGPT on Registered Nurse License Exam in Taiwan: A Descriptive Study. Healthcare 11(21):2855. https://doi.org/10.3390/healthcare11212855

Huanqi (2023) A fake picture causes stock market turmoil! AI "mischief" continues to appear, and countries have tightened supervision. https://www.huanqiu.com/article/4D2o3i2dY8w.Accessed 25 Dec 2023

James P, Howard N, Henrik A, Alicia F (2018) Countering Information Influence Activities : The State of the Art. MSB. https://www.msb.se/RibData/Filer/pdf/28697.pdf

Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, Ishii E, Bang Y, Chen D, Chan H, Dai W, Madotto A, Fung P (2023) Survey of Hallucination in Natural Language Generation. ACM Comput Surv 55(12):1–38. https://doi.org/10.1145/3571730

Kapantai E, Christopoulou A, Berberidis C, Peristeras V (2021) A systematic literature review on disinformation: Toward a unified taxonomical framework. N. Media Soc 23(5):1301–1326. https://doi.org/10.1177/1461444820959296

Kleinheksel A, Rockich-Winston N, Tawfik H, Wyatt, T (2020) Demystifying Content Analysis. Am J Pharma Edu. https://doi.org/10.5688/ajpe7113

Komendantova N, Ekenberg L, Svahn M, Larsson A, Shah S, Glinos M, Koulolias V, Danielson M (2021) A value-driven approach to addressing misinformation in social media. Humanit Soc Sci Commun 8(1):1–12. https://doi.org/10.1057/s41599-020-00702-9

Koubaa A, Boulila W, Alzahem A, Latif S (2023) Exploring ChatGPT Capabilities and Limitations: A Survey. IEEE Access 11:118698–118721. https://doi.org/10.1109/ACCESS.2023.3326474

Kumar M, Mani U, Tripathi P, Saalim M, Roy S, Kumar M, Mani U, Tripathi P, Saalim M, Sr S (2023) Artificial Hallucinations by Google Bard: Think Before You Leap. Cureus J Med Sci 15(8). https://doi.org/10.7759/cureus.43313

Lee L, Dabirian A, McCarthy I, Kietzmann J (2020) Making sense of text: Artificial intelligence-enabled content analysis. Eur J Mark 54(3):615–644. https://doi.org/10.1108/EJM-02-2019-0219

Lee M (2023) A Mathematical Investigation of Hallucination and Creativity in GPT Models. Mathematics 11(10):10. https://doi.org/10.3390/math11102320

Liu H, Xue W, Chen Y, Chen D, Zhao X, Wang K, Hou L, Li R, Peng W (2024) A Survey on Hallucination in Large Vision-Language Models. arXiv. http://arxiv.org/abs/2402.00253

Liu Y, Jacoby R, Jang H, Li D (2019) A Content Analysis of Adoption Articles in Counseling Journals: A 30-Year Review. Fam J 27(1):67–74. https://doi.org/10.1177/1066480718809424

Li W (2024) A Study on Factors Influencing Designers' Behavioral Intention in Using AI-Generated Content for Assisted Design: Perceived Anxiety, Perceived Risk, and UTAUT. International Journal of Human–Computer Interaction. https://doi.org/10.1080/10447318.2024.2310354

Monteith S, Glenn T, Geddes J, Whybrow P, Achtyes E, Bauer M (2024) Artificial intelligence and increasing misinformation. Br J Psychiatry 224(2):33–35. https://doi.org/10.1192/bjp.2023.136

Mo Z, Pang D, Liu H, Zhao Y (2023) Analysis on AIGC False Information Problem and Root Cause from the Perspective of Information Quality. Documentation,Inf Knowl 40(4):32–40. https://link.cnki.net/doi/10.13366/j.dik.2023.04.032

NewsGuard (2023) Transparent Reliability Ratings for News and Information Sources. https://www.newsguardtech.com/.Accessed 25 Dec 2023

OpenAI (2023) GPT-4. https://openai.com/research/gpt-4.Accessed 25 Jan 2024

Polyportis A, Pahos N (2024) Navigating the perils of artificial intelligence: A focused review on ChatGPT and responsible research and innovation. Humanit Soc Sci Commun 11(1):1–10. https://doi.org/10.1057/s41599-023-02464-6

Rawte V, Chakraborty S, Pathak A, Sarkar A (2023) The Troubling Emergence of Hallucination in Large Language Models – An Extensive Definition, Quantification, and Prescriptive Remediations. arXiv. https://doi.org/10.48550/arXiv.2310.04988

Rivas P, Zhao L (2023) Marketing with ChatGPT: Navigating the Ethical Terrain of GPT-Based Chatbot Technology. AI 4(2):375–384. https://doi.org/10.3390/ai4020019

Roozenbeek J, Van Der Linden S, Goldberg B, Rathje S, Lewandowsky S (2022) Psychological inoculation improves resilience against misinformation on social media. Sci Adv 8(34):6254. https://doi.org/10.1126/sciadv.abo6254

Salaverría R, Cardoso G (2023) Future of disinformation studies: Emerging research fields. El Profesional de La Información. https://doi.org/10.3145/epi.2023.sep.25

Shao L, Chen B, Zhang Z, Zhang Z, Chen X (2024) Artificial intelligence generated content (AIGC) in medicine: A narrative review. Math Biosci Eng 21(1):1672–1711. https://doi.org/10.3934/mbe.2024073

Suarez-Lledo V, Alvarez-Galvez J (2021) Prevalence of Health Misinformation on Social Media: Systematic Review. J Med Internet Res 23(1):e17187. https://doi.org/10.2196/17187

Typeform (2023) ChatGPT/LLM Errors Tracker. https://researchrabbit.typeform.com/llmerrors.Accessed 25 Dec 2023

Walters W, Wilder E (2023) Fabrication and errors in the bibliographic citations generated by ChatGPT. Sci Rep 13(1) https://doi.org/10.1038/s41598-023-41032-5

Wang Y, Pan Y, Yan M, Su Z, Luan T (2023) A Survey on ChatGPT: AI–Generated Contents, Challenges, and Solutions. IEEE Open J Computer Soc 4:280–302. https://doi.org/10.1109/OJCS.2023.3300321

White M, Marsh E (2006) Content Analysis: A Flexible Methodology. Libr Trends 55(1):22–45. https://doi.org/10.1353/lib.2006.0053

Wilson S, Wiysonge C (2020) Social media and vaccine hesitancy. BMJ Glob Health 5(10):e004206. https://doi.org/10.1136/bmjgh-2020-004206

Wu T, He S, Liu J, Sun S, Liu K, Han Q, Tang Y (2023) A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development. IEEE-CAA. J Autom Sin 10(5):1122–1136. https://doi.org/10.1109/JAS.2023.123618

Zannettou S, Sirivianos M, Blackburn J, Kourtellis N (2019) The Web of False Information: Rumors, Fake News, Hoaxes, Clickbait, and Various Other Shenanigans. J Data Inf Qual 11(3):1–37. https://doi.org/10.1145/3309699

Zhou X, Zafarani R (2020) A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. ACM Comput Surv 53(5):1–40. https://doi.org/10.1145/3395046

Acknowledgements

This study was partially supported in the collection, analysis, and interpretation of data by the National Social Science Fund of China under Grant No. 23BTQ086.

Author contributions

Yujie Sun collected and organized samples, led the coding work, wrote the first draft, and revised the first draft. Dongfang Sheng guided the research process and revised the first draft. Zihan Zhou and Yifei Wu implemented the coding work and adjusted the paper format.

Competing interests

The authors declare no competing interests.

Ethical approval

Ethical approval was not required as the study did not involve human participants.

Informed consent

Informed consent was not required as the study did not involve human participants.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1057/s41599-024-03811-x.

Correspondence and requests for materials should be addressed to Dongfang Sheng.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.