

Article

Examining the Performance of Large Language Models in Health Information Quality Evaluation Tasks: An AI Post-Alignment Perspective

Shenghui Zhang, Jun Zhang ^{*}, Peng Li, Min Zuo and Lichao Feng ^{*}

Business School, Shandong University of Technology, Zibo 255000, China;
23418011369@stumail.sdut.edu.cn (S.Z.); 23418011364@stumail.sdut.edu.cn (P.L.); zuomin@sdut.edu.cn (M.Z.)

^{*} Correspondence: zhangjun_sdut@sdut.edu.cn (J.Z.); fenglc19@sdut.edu.cn (L.F.)

Abstract

When large language models (LLMs) are implemented in the field of health information quality evaluation, ensuring that their outputs align with human judgment, logic, and strategic preferences has become a focal point of current research on AI alignment. This study proposes the Health Information Quality Evaluation Framework based on AI Post-Alignment (HIQE-PA) to examine whether large language models remain consistent with expert judgment standards in health information quality evaluation tasks. Using expert evaluation results as the benchmark, the framework assesses the consistency, stability, and cross-source adaptability of model outputs through structured task implementation and multidimensional statistical indicators. In the experimental study, ERNIE-3.5 and ChatGLM2-6B-32K were selected as general-purpose models, and the corresponding ERNIE-3.5 + HIQE-PA model and ChatGLM2-6B-32K + HIQE-PA model were constructed to compare post-alignment performance under different model conditions. Under three post-alignment performance evaluation indicators, including deviation degree, predictability, and fitness, the LLMs' post-alignment performance was examined through their outputs in the health information quality evaluation task. The results show that the general LLMs achieved a high level of post-alignment performance only in the dimensions of readability and completeness, while the LLMs + HIQE-PA improved post-alignment performance across all dimensions. In particular, the ERNIE-3.5 + HIQE-PA model performed prominently, producing evaluation outputs that were closer to expert consensus and maintaining consistency with the expert benchmark across different text sources. This study demonstrates that post-alignment examinations can provide empirical support for model selection and governance in the health information domain.

Keywords: large language models; AI post-alignment; health information quality evaluation; prompt-adapted LLMs



Academic Editor: Nipon
Theera-Umporn

Received: 7 May 2026
Revised: 28 June 2026
Accepted: 30 June 2026
Published: 2 July 2026

Copyright: © 2026 by the authors.
Licensee MDPI, Basel, Switzerland.
This article is an open access article
distributed under the terms and
conditions of the [Creative Commons
Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

With the rapid growth of social media, online health communities, and intelligent question answering systems, the public increasingly relies on online texts to obtain information about disease prevention, nutrition, health risks, and healthy lifestyles. Large language models (LLMs), owing to their natural language understanding, text generation, and semantic judgment capabilities, have been increasingly used for health information screening, content moderation, and quality evaluation, offering a new technical pathway

for health information governance [1]. However, health information is a high-risk and high-responsibility domain. If a model misjudges factual accuracy, overlooks potential harm, or deviates from expert judgment in ethical- and responsibility-related dimensions, its outputs may not only affect platform governance but also mislead public health decision making. Therefore, in health information quality evaluation tasks, it is insufficient to examine whether LLMs can merely generate evaluation outputs. A more critical question is whether their evaluations are consistent with the judgment logic, professional standards, and value preferences of human experts.

This study conceptualizes this issue as AI post-alignment. Existing AI alignment research has mainly focused on pre-alignment during model development, such as data optimization, instruction tuning, and reinforcement learning from human feedback, with the aim of aligning model behavior with general human intentions and preferences [2]. In contrast, AI post-alignment emphasizes task-specific examination after deployment. It evaluates whether model outputs remain close to expert benchmarks, whether the outputs are stable under changing inference conditions, and whether the model can maintain consistent evaluative capability across different text sources and application contexts [3]. Accordingly, AI post-alignment differs from human–AI agreement. The latter usually describes the degree of consistency between model and human judgments, whereas post-alignment further emphasizes deployment-stage examination, interpretation, and calibration with reference to expert benchmarks [4]. It also differs from benchmark-based evaluation because benchmarks provide reference materials, whereas post-alignment additionally requires the systematic measurement of deviation, stability, and cross-context adaptability [5]. Moreover, post-alignment is not equivalent to LLM-as-a-Judge. LLM-as-a-Judge indicates that a model can serve as an automated evaluator, whereas post-alignment concerns whether this evaluator can reliably simulate and maintain consistency with expert judgment standards in a specific domain task [6].

Against this background, this study proposes the Health Information Quality Evaluation Framework based on AI Post-Alignment (HIQE-PA). In this study, HIQE-PA is defined as a deployment-stage assessment framework for examining whether large language models can maintain consistency with expert judgment standards in a specific health information quality evaluation task. The framework does not aim to modify model parameters. Instead, it defines the task boundary of health information governance, constructs expert evaluation benchmarks, implements structured evaluation tasks, and uses three post-alignment indicators—deviation degree, predictability, and fitness—to systematically examine expert consistency, output stability, and cross-source adaptability of model outputs. These three indicators correspond to the three minimum requirements of post-alignment in this task: closeness to expert benchmarks, output stability under controlled inference perturbations, and cross-source adaptability between UGC and AIGC. Thus, the indicator system is intended not as a general-purpose reliability scale, but as a task-specific measurement framework for examining whether LLM evaluators can maintain expert-consistent judgment behavior after deployment. This study addresses the following research questions: RQ1: How can a post-alignment framework be designed to ensure that LLMs meet both professional and ethical criteria in health information quality evaluation tasks? RQ2: How can experimental testing validate the consistency between LLM evaluations and expert judgments across factual and ethical dimensions? RQ3: How can a multidimensional quantitative indicator system be established to measure post-alignment performance and provide evidence for model selection, optimization, and application in health information governance?

2. Literature Review

2.1. Health Information Quality Evaluation

Information quality research provides a theoretical foundation for health information evaluation. Early studies categorized information quality into intrinsic, contextual, representational, and accessibility quality. Building on this foundation, the PSP/IQ model organized information quality into four quadrants: sound, dependable, useful, and usable information [7,8]. Compared with general information quality models, the PSP/IQ model is useful because it considers both information-as-product attributes and information-as-service attributes. This logic is particularly relevant to health information on social media. On the one hand, health information must be accurate, complete, and reliable as a content product. On the other hand, health information on social media platforms must also satisfy users' comprehension needs, communicative usability, risk control, and governance responsibility [9,10]. Therefore, PSP/IQ provides a suitable theoretical basis for connecting content quality, user needs, and platform governance requirements.

Nevertheless, existing health information quality frameworks cannot be directly applied to the task examined in this study. First, many traditional instruments for health information quality evaluation were developed for websites, patient education materials, or health information systems. They often emphasize source credibility, author expertise, update dates, accessibility, and readability [11,12], while paying insufficient attention to fragmented expression, emotional narratives, ambiguous risk cues, and platform governance contexts in social media health content. Second, although previous studies have frequently discussed accuracy, credibility, and readability, they have less often translated these criteria into standardized scoring tasks executable by LLMs, nor have they adequately considered ethical misjudgment, underestimation of harm, and cross-source instability that may occur when LLMs evaluate health information [13]. Therefore, this study does not simply propose a new set of indicators parallel to existing frameworks. Instead, it integrates and operationalizes existing health information quality dimensions based on PSP/IQ, the characteristics of social media health content, and the requirements of LLM-based automated evaluation, proposing the following four evaluation dimensions: readability and completeness, factuality and accuracy, safety and harmlessness, and responsibility and ethics.

2.2. From AI Alignment to AI Post-Alignment

AI alignment generally refers to the consistency between model behavior and human intentions, preferences, and values [2]. Existing studies have primarily focused on pre-alignment methods during model development, such as learning reward functions from human preferences, reinforcement learning from human feedback, instruction tuning, and value preference optimization [14,15]. These methods improve model controllability, helpfulness, and safety in general tasks, but they do not guarantee that models will remain consistent with domain expert judgments in all deployment scenarios. This limitation is particularly salient in health information quality evaluation, where models must go beyond surface-level language understanding and make judgments about factual accuracy, potential harm, and ethical responsibility.

Research on human-centered AI and human-in-the-loop evaluation emphasizes that human oversight should not be limited to model training but should extend to deployment, feedback correction, and continuous evaluation [16]. In automated evaluation tasks, human-in-the-loop approaches often enhance the credibility of automated judgments through human annotation, expert benchmarks, manual review, and model calibration [4]. Related studies have also shown that when LLMs are used as evaluators, their judgments may be affected by position bias, verbosity bias, familiarity bias, anchoring effects, and prompt sensitivity, making human-referenced calibration and reliability testing necessary [17,18]. In this study,

AI post-alignment is therefore defined as a deployment-stage process in which LLMs are examined against domain expert benchmarks through multidimensional indicators that measure deviation, output stability, and cross-context adaptability in specific tasks.

2.3. LLM-As-a-Judge Evaluation and Reliability Metrics

LLMs have recently been used for text quality evaluation, fact checking, dialogue system assessment, and comparison of generated content, forming the research stream known as LLM-as-a-Judge [19]. Such methods are useful because they provide scalable evaluation results in open-ended tasks where reference answers are unavailable or human evaluation is costly. For example, studies such as MT-Bench, Chatbot Arena, and G-Eval have shown that strong LLM evaluators can, under certain conditions, achieve substantial agreement with human preferences [10,20]. Nevertheless, the literature also indicates that LLM-as-a-Judge is not inherently reliable. Model judgments may be affected by candidate position, response length, self-preference, familiarity bias, prompt wording, and task type [21]. Large-scale cross-task evidence further suggests that the ability of LLMs to replicate human judgments varies depending on the evaluated property, data source, expertise level of human judges, and task context; therefore, LLM evaluators should be validated against human judgments before being used as reliable evaluators [22]. Existing studies have proposed several metrics and methods for evaluating the reliability of LLM-based judgments. For instance, position consistency measures whether model choices remain stable after the order of candidate answers changes; preference fairness examines whether models exhibit positional or preference-related biases; repetition stability captures the consistency of repeated judgments; and correlation or agreement coefficients are commonly used to compare model scores with human ratings [23,24].

Existing studies have examined the reliability of LLM-based evaluation from multiple perspectives. In human–AI agreement assessment, correlation coefficients, agreement rates, Cohen’s Kappa, and Krippendorff’s Alpha are commonly used to measure the closeness between model judgments and human judgments [21]. In the LLM-as-a-Judge literature, researchers have further examined position bias, length bias, preference bias, prompt sensitivity, and repetition stability to determine whether LLMs can serve as reliable automated evaluators [18,25]. These metrics provide important methodological foundations for assessing the accuracy, consistency, and stability of model outputs. However, health information quality evaluation is more strongly task-specific. In this context, models are expected not only to approximate expert scores at the aggregate level but also to maintain stable and interpretable evaluative capability across factual accuracy, potential harm, ethical responsibility, and different text sources [1]. Therefore, a single agreement or stability metric is insufficient to fully capture the post-alignment status of LLMs in this task. Building on these measurement traditions, this study integrates expert benchmarking, output stability, and cross-source adaptability into a unified post-alignment measurement framework. The three indicators proposed in Section 3—deviation degree, predictability, and fitness—are used to capture systematic divergence from expert benchmarks, output stability under parameter perturbations, and cross-context evaluative consistency between UGC and AIGC, respectively.

2.4. Research Gaps and Novelty Statement

In summary, existing studies have provided valuable foundations for health information quality evaluation, AI alignment, human-in-the-loop evaluation, and LLM-as-a-Judge assessment. However, three gaps remain. First, current AI alignment research mainly focuses on model training or instruction tuning, while systematic examination of post-deployment alignment in domain-specific tasks remains limited. Second, although LLM-as-a-Judge and human-in-the-loop evaluation studies address the consistency between

automated and human judgments, most research has focused on general text generation, dialogue systems, or open-ended NLP tasks, with limited attention to high-sensitivity contexts such as health information quality evaluation, where factual risk, public safety, and ethical responsibility are central. Third, existing metrics often assess model reliability from a single perspective, such as correlation, bias, or stability, while a multidimensional post-alignment indicator system that simultaneously captures deviation from expert benchmarks, output predictability, and cross-source adaptability is still lacking.

The novelty of this study can be summarized in three respects. Theoretically, this study introduces AI post-alignment into health information quality evaluation and defines it as an expert benchmark-driven task consistency assessment framework for the deployment stage. Methodologically, this study reconstructs a four-dimensional evaluation system for social media health content and LLM-based automated evaluation based on PSP/IQ and health information quality research, and proposes three post-alignment indicators: deviation degree, predictability, and fitness. Experimentally, this study compares expert evaluation, general LLM evaluation, and prompt-adapted LLM evaluation to examine whether HIQE-PA can reduce deviation from expert benchmarks, improve output stability, and enhance cross-source adaptability, thereby providing evidence from experimental results for model selection, optimization, and risk identification in health information governance.

3. Methods

3.1. Study Design

This study designed a health information quality evaluation task to examine whether large language models can remain consistent with the evaluation intentions and outcomes of human experts. To achieve this goal, the study compared the performance of human experts, general-purpose LLMs, and LLMs assisted by the HIQE-PA framework under the same evaluation task. Human expert evaluations were used to construct the expert benchmark. General-purpose LLMs directly performed the health information quality evaluation task to observe their natural performance. LLMs assisted by the HIQE-PA framework completed the evaluation task based on expert benchmarks, structured prompts, and standardized output requirements. The evaluation results from these conditions were then collected and examined through deviation degree, predictability, and fitness to assess post-alignment performance under different model conditions.

The study follows an integrated research framework that connects health information governance requirements, evaluation criteria, model-based evaluation, and post-alignment examination (Figure 1). The framework consists of four interrelated stages. First, the application scenario is defined according to the governance requirements of social media health information, including user comprehension, factual reliability, risk control, and ethical responsibility. Second, these requirements are operationalized into four evaluation dimensions: readability and completeness, factuality and accuracy, safety and harmlessness, and responsibility and ethics. Third, expert evaluations are used to construct the reference benchmark, and structured prompts and multi-round interactions are incorporated to adapt general-purpose LLMs to the health information quality evaluation task. Finally, the evaluation outputs are examined through deviation degree, predictability, and fitness to assess the consistency, stability, and cross-source adaptability of the models. This framework links the conceptual definition of health information quality with the experimental evaluation of LLM performance and provides the methodological basis for the subsequent evaluation dimension construction, indicator design, task implementation, and empirical analysis.

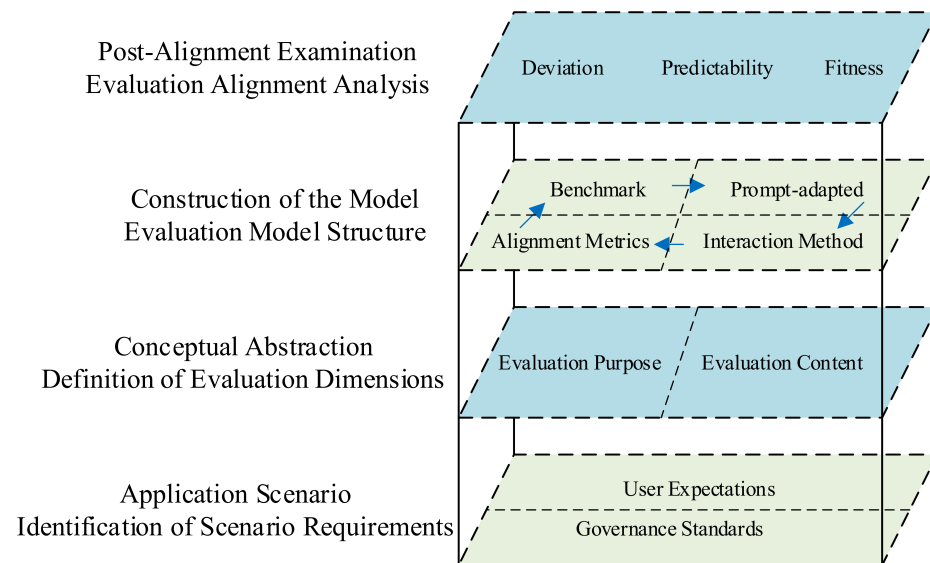


Figure 1. Research framework for examining LLM post-alignment in health information quality evaluation.

3.1.1. Construction of Evaluation Dimensions

The construction of the evaluation dimensions is grounded in information quality theory and health information quality evaluation research [26,27]. The PSP/IQ model conceptualizes information quality as a combination of information product quality and information service quality. Its four quadrants—sound, dependable, useful, and usable information—provide a basic classification logic for this study [7,28]. However, the PSP/IQ model is a general information quality framework and was not originally designed for social media health information or LLM-based automated evaluation tasks [29]. Social media health information is characterized by diverse expression styles, varying levels of expertise, potential misinformation risks, and strong ethical responsibility requirements [30]. Therefore, the PSP/IQ model needs to be adapted to this task by incorporating the health information quality literature.

The readability and completeness dimension is adapted from representational quality, usable information, and health information indicators related to readability, completeness, and understandability [31,32]. It evaluates whether health information is clearly expressed, structurally complete, logically coherent, and understandable to ordinary users [33]. The factuality and accuracy dimension corresponds to intrinsic quality, sound information, and health information indicators related to accuracy, scientific validity, and credibility [34,35]. It evaluates whether health knowledge, factual descriptions, and risk statements are accurate and reliable. The safety and harmlessness dimension integrates dependable information with health information governance concerns such as safety, privacy protection, and harmlessness [36,37]. It evaluates whether the text contains misleading claims, discrimination, privacy infringement, exaggerated promises, or content that may cause health-related harm. The responsibility and ethics dimension further incorporates AI governance and health communication ethics. It evaluates whether the text conforms to public health responsibility, professional boundaries, ethical norms, and socially responsible value orientations [38,39].

Therefore, the four evaluation dimensions proposed in this study are not entirely new indicators detached from the literature. Rather, they are task-specific adaptations and integrations of the PSP/IQ model and health information quality criteria for social media health content and LLM-based evaluation tasks. The readability and completeness, and factuality and accuracy dimensions are mainly derived from existing information quality and health information evaluation frameworks. The safety and harmlessness dimension strengthens the risk-control requirements of health information governance. The

responsibility and ethics dimension represents an integrative operationalization for LLM evaluation tasks and public health governance contexts. Together, these four dimensions constitute the expert evaluation benchmark and define the task boundaries for subsequent post-alignment examination of LLM outputs.

3.1.2. Design of Post-Alignment Indicators

Based on the health information quality evaluation results, this study uses human expert evaluations as the reference benchmark and examines the post-alignment performance of LLMs through three indicators: deviation degree, predictability, and fitness [19,40]. The selection of these three indicators follows the task-specific definition of post-alignment adopted in this study. In health information quality evaluation, post-alignment does not refer to general model performance, nor does it refer to reliability in the broader psychometric sense. Rather, it refers to whether an LLM evaluator can maintain expert-consistent judgment behavior after deployment in a specific task context.

Under this definition, three conditions are necessary. First, model outputs should remain close to expert benchmarks; otherwise, even stable outputs may still be systematically misaligned with professional judgment standards. Second, model outputs should remain stable under controlled inference perturbations; otherwise, occasional agreement with experts cannot provide sufficient evidence for dependable deployment. Third, the model should preserve expert-consistent evaluative capability across different text sources, because health information governance involves both user-generated content and AI-generated content, which differ in expression style, structural completeness, factual risk, and ethical responsibility. Accordingly, deviation degree measures benchmark proximity, predictability measures output stability under perturbation conditions, and fitness measures cross-source adaptability [6].

These three indicators are therefore treated as necessary and jointly sufficient for measuring post-alignment in the present task. They are necessary because each captures a non-substitutable component of deployment-stage alignment: expert benchmark proximity, perturbation stability, and cross-source adaptability. They are jointly sufficient because, taken together, they cover the core evaluative requirements for determining whether an LLM evaluator can approximate expert judgment, maintain stable evaluation behavior, and transfer expert-consistent evaluation capability across UGC and AIGC in health information quality evaluation [41].

Indicator 1: Deviation Degree (D_E^i) measures the extent to which LLM outputs diverge from expert benchmarks across evaluation dimensions, and the calculation method is as shown in Formula (1). It is conceptually related to error-based agreement measures such as the mean absolute deviation and mean squared error, but it is further interpreted with reference to expert variation. In this study, this indicator tests whether LLMs have systematic deviations in factual, ethical, or readability aspects.

$$D_E^i = \frac{\sqrt{\sum_{j=1}^n (r_{j,LLM}^i - \bar{e}_{Expert}^i)^2 / n}}{S_{Expert}^i} \quad (1)$$

Mean Absolute Deviation (MAD) and Mean Squared Error (MSE) were used as metrics to measure the deviation of LLMs in four different evaluation dimensions. If $D_E^i \rightarrow 0$, the LLMs aligned with expert consensus; otherwise, in the i th evaluation dimension, if $D_E^i \rightarrow 1$, it reflected alignment with individual expert variation. $\bar{e}_{Expert}^i$ and S_{Expert}^i were respectively the mean and standard deviation of the measurement results of human experts on the i th evaluation dimension, and $r_{j,LLM}^i$ was the measurement value of the LLM on the j th health information item for the i th evaluation dimension [42].

Indicator 2: Predictability ($H_E^{i,T}$) describes the degree of predictability of the output results of LLMs under repeated tasks in different perturbation conditions, calculated as shown in Formula (2). Prior research on LLM evaluators has shown that output stability and prompt sensitivity are important conditions for evaluator reliability. This indicator reflects LLMs' understanding of the task and whether it is robust in health information quality evaluation tasks.

$$H_E^{i,T} = -\sum_{R_{llm}} P(r_{llm}^{i,T}) \log_2 P(r_{llm}^{i,T}) \quad (2)$$

This study used information entropy as a measurement tool and simulated task perturbations by adjusting the temperature parameter of the LLMs to measure the predictability of the output results. $0 < H_E^{i,T} < \log(n)$, where n was the number of evaluation objects and T was in $[0-1]$. The smaller the information entropy value of $H_E^{i,T}$, the higher the predictability. $P(r_{llm}^{i,T})$ was the probability of LLMs' evaluation result value at temperature T [43].

Predictability measures the stability of LLM outputs under repeated evaluation tasks and controlled inference perturbations. Prior studies on LLM-as-a-Judge have suggested that the reliability of automated evaluators depends not only on their agreement with human judgments, but also on their ability to maintain stable outputs under repeated evaluation conditions [41]. Predictability in this study refers specifically to the stability of LLM evaluation outputs under repeated tasks and controlled inference perturbations. It should not be interpreted as accuracy, validity, or reliability in the broader psychometric sense. A model may produce highly predictable outputs while still deviating from expert benchmarks; therefore, predictability is not used as a substitute for deviation degree. Instead, it complements deviation degree by examining whether the model's evaluative behavior remains stable when inference conditions change. In this study, the temperature parameter was adjusted to simulate controlled perturbations in the inference process, and information entropy was used to quantify the dispersion of score distributions. A lower entropy value indicates a more concentrated score distribution and thus greater output stability under perturbation conditions. Therefore, predictability reflects the stability dimension of post-alignment, rather than the accuracy or overall reliability of the model evaluator.

Indicator 3: Fitness (F_E^i) measures whether the model can maintain expert-consistent evaluative capability across different text sources, rather than merely whether its scores are generally consistent or robust, calculated by Formula (3). In the present task, cross-source adaptability is a core component of post-alignment because health information governance increasingly involves both user-generated content and AI-generated content. UGC often contains fragmented expression, informal wording, emotional narratives, and heterogeneous quality, whereas AIGC tends to be more standardized, complete, and fluent, but may still contain factual omissions, overconfident statements, or ethically problematic recommendations. If an LLM is aligned with expert judgment only for one source type, its deployment-stage applicability remains limited.

Therefore, fitness evaluates whether the model preserves the relative judgment pattern established by experts when the evaluated texts shift from UGC to AIGC. This differs from conventional agreement, which mainly measures closeness between model scores and expert scores within a given dataset. It also differs from robustness, which usually concerns resistance to perturbations, noise, or adversarial changes, and from consistency, which mainly describes the internal stability of model outputs. Fitness instead focuses on the cross-source transfer of expert-consistent evaluation behavior. In this sense, fitness is

necessary for assessing whether post-alignment can be maintained beyond a single data source and applied to heterogeneous health information governance scenarios.

$$F_E^i = \frac{\Delta_i^{LLM}}{\Delta_i^{Expert}} = \frac{|D_{AI,i}^{LLM} - D_{H,i}^{LLM}|}{|D_{AI,i}^{Expert} - D_{H,i}^{Expert}|} \quad (3)$$

where $D_{H,i}^{LLM}$ and $D_{AI,i}^{LLM}$ denote the LLM evaluation results for human-generated content and AI-generated content in the i -th evaluation dimension, respectively. $D_{H,i}^{Expert}$ and $D_{AI,i}^{Expert}$ denote the corresponding expert evaluation results. Δ_i^{LLM} represents the cross-source difference recognized by the LLM, whereas Δ_i^{Expert} represents the cross-source difference recognized by human experts. Therefore, F_E^i measures the extent to which the LLM reproduces the expert-recognized difference between human-generated content and AI-generated content.

When F_E^i approaches 0, the LLM shows little or no cross-source distinction relative to the expert benchmark. When $0 < F_E^i < 1$, the LLM underestimates the expert-recognized cross-source difference. When F_E^i approaches 1, the LLM successfully reproduces the cross-source difference identified by experts, indicating better cross-source fitness. When $F_E^i > 1$, the LLM amplifies the cross-source difference, suggesting excessive sensitivity to source-related differences [44].

Together, these indicators form a quantitative toolkit to examine the LLMs post-alignment performance in health information quality evaluation tasks, providing empirical evidence for model selection, optimization, and governance.

3.1.3. Task Design and Evaluation Procedure

The health information quality evaluation task in this study was designed as a structured scoring task rather than a general open-ended question answering task. The evaluation objects included two types of health information texts: user-generated health information from social media platforms and AI-generated health information produced by an LLM. Each text was presented as an independent evaluation unit, and evaluators were asked to judge its information quality without relying on source labels.

The task consisted of four evaluation dimensions: readability and completeness, factuality and accuracy, safety and harmlessness, and responsibility and ethics. Each dimension was scored using a five-point Likert scale, where 1 indicated very low quality or serious non-compliance with the dimension, and 5 indicated very high quality or full compliance with the dimension. To ensure scoring consistency, operational definitions were provided for each dimension. Readability and completeness assessed whether the text was clearly expressed, structurally complete, and logically coherent [31,33]. Factuality and accuracy assessed whether health knowledge, factual descriptions, and risk statements were accurate and reliable [34,35]. Safety and harmlessness assessed whether the text contained misleading suggestions, privacy violations, discriminatory expressions, or content that could cause health-related harm [36]. Responsibility and ethics assessed whether the text conformed to public health responsibility, professional boundaries, and ethical norms [37].

In the human expert evaluation task, the sample order was randomized before evaluation [45]. Experts independently evaluated the texts based only on their own knowledge, the scoring criteria, and the given texts, without using search engines, medical databases, or other external tools. All experts received the same task instructions, scoring sheet, and dimension definitions [21]. The expert evaluation results were then aggregated to construct the expert benchmark for subsequent alignment examination of LLM outputs.

In the LLM evaluation task, all models received the same text inputs, the same dimension definitions, and the same scoring format requirements. For each health information text, the models were required to provide a score from 1 to 5 for each of the four dimensions and a brief explanation for each score. To improve experimental reproducibility, the model invocation process was controlled using the same prompt template, output format, sample order, and parameter settings within the same test. In the predictability test, only the temperature parameter was varied to observe the stability of model outputs under different perturbation conditions. All model outputs were saved as structured data and statistically compared with the expert benchmark.

3.1.4. Measurement Invariance Assessment

To examine whether the comparison between UGC and AIGC was based on a sufficiently comparable scoring logic, this study conducted a cross-source measurement equivalence assessment as a robustness check before interpreting the UGC–AIGC comparisons and the subsequent model-based post-alignment results [46]. The purpose of this analysis was not to validate a traditional multi-item psychometric scale, nor to claim that the four evaluation dimensions constitute a reflective latent construct in the strict psychometric sense. Rather, readability and completeness, factuality and accuracy, safety and harmlessness, and responsibility and ethics were treated as task-specific observed evaluation criteria within the health information quality evaluation framework. The rationale for conducting this analysis was closely related to the main argument of this study. Since the expert benchmark served as the reference standard for evaluating LLM post-alignment, the comparability of expert ratings across text sources was a prerequisite for interpreting cross-source differences. If experts applied the four evaluation criteria differently to UGC and AIGC, then the observed differences between the two sources, as well as the fitness indicator, might partly reflect measurement non-equivalence rather than genuine cross-source adaptability of LLM evaluators. Therefore, the measurement invariance assessment was used to examine whether the four-dimensional scoring framework had sufficient cross-source equivalence to support subsequent comparisons.

Specifically, the expert ratings were reorganized at the expert-rating level. The UGC group included 350 rating records, consisting of 70 texts evaluated by five experts, and the AIGC group included 350 rating records, consisting of 70 corresponding texts evaluated by the same five experts. A multi-group measurement invariance procedure was then conducted across the two text-source groups. Configural invariance was first tested to examine whether the same four-dimensional measurement structure was applicable to both UGC and AIGC. Metric invariance was then tested by constraining factor loadings to be equal across groups. Scalar invariance was further tested by constraining both factor loadings and intercepts to be equal across groups. Model comparisons were evaluated using changes in fit indices, including ΔCFI , ΔRMSEA , and ΔSRMR . Invariance was considered acceptable when ΔCFI was no greater than 0.010, ΔRMSEA was no greater than 0.015, and ΔSRMR showed only a small increase.

Accordingly, the measurement invariance assessment was not used as an independent psychometric validation of the four-dimensional framework, nor did it replace the three post-alignment indicators. Instead, it provided a robustness check for the measurement equivalence of the expert benchmark across UGC and AIGC. The results were used to clarify how confidently the subsequent deviation degree, predictability, and fitness results could be interpreted across text sources, especially when evaluating whether LLMs preserved expert-consistent judgment patterns from UGC to AIGC.

3.2. Data Source of Evaluation Objects

The datasets employed in this study encompassed two classifications of health information: user-generated content disseminated on social media (UGC), which is marked as TextH, and AI-generated health information (AIGC), which is marked as TextAI [47]. TextH was collected from the Science Popularization China—Scientific Rumor Refutation platform. The content on this platform has the characteristics of scientific rumor refutation and corrective health communication, and is not equivalent to completely unfiltered raw posts from general social media platforms [48]. Therefore, this study defines TextH as corrective health-related UGC samples in a platform governance context. This platform was selected as the primary data source because its content focuses on health misconceptions, health risks, and health rumors frequently encountered by the public, thereby providing governance-relevant evaluation objects for health information quality assessment. A stratified sampling strategy was adopted for data collection. Based on the platform's health information topic categories, the samples were divided into four categories: fitness and beauty, food safety, disease prevention, and nutrition and health. The data were collected in June 2025 through manual collection from publicly available webpages. The inclusion criteria were as follows: the text was health-related, contained a clear title and main body, and could be used as an independent object for health information quality evaluation. The exclusion criteria included duplicate texts, non-health-related texts, texts with missing main content, and texts for which the core health topic could not be identified. Finally, 680 health information texts were obtained. Using the titles and keywords of these posts, the same prompts were input into Baidu's ERNIE-3.5 model to generate another 680 pieces of health information, forming the second dataset.

The four topic categories correspond to the health information governance objectives of this study. Fitness and beauty texts mainly involve efficacy claims and lifestyle advice; food safety texts mainly involve public consumption risks and rumor identification; disease prevention texts mainly involve health risk judgment and preventive behaviors; and nutrition and health texts mainly involve daily dietary advice and health cognition. These topics cover the core governance requirements of factual accuracy, risk control, understandability, and responsibility and ethics in health information quality evaluation.

To examine the semantic comparability between the UGC and AIGC datasets, this study computed pairwise semantic similarity for the 680 matched text pairs. Each AIGC sample was paired with the corresponding UGC sample from which its title and keywords were extracted. Before similarity computation, the texts were preprocessed by removing irrelevant symbols, repeated spaces, and non-content elements while preserving the original semantic content. Each text was then encoded into a sentence-level semantic vector using a pretrained Chinese sentence embedding model based on the Sentence-BERT framework [49]. Cosine similarity was used as the similarity metric, as it measures the angular closeness between two semantic vectors and is widely used for sentence-level semantic textual similarity tasks [50]. For the i -th matched pair, the similarity score was calculated as follows:

$$s_i = \cos(e_i^{UGC}, e_i^{AIGC}) = \frac{e_i^{UGC} \cdot e_i^{AIGC}}{\|e_i^{UGC}\| \|e_i^{AIGC}\|} \quad (4)$$

where e_i^{UGC} and e_i^{AIGC} denote the embedding vectors of the i -th UGC and AIGC texts, respectively. A higher cosine similarity value indicates stronger semantic proximity between the two texts.

The results (see Figure 2) show that similarity values are concentrated between 0.90 and 0.95, indicating that the generated content closely mirrors real-world social media health information while maintaining distinguishable textual features.

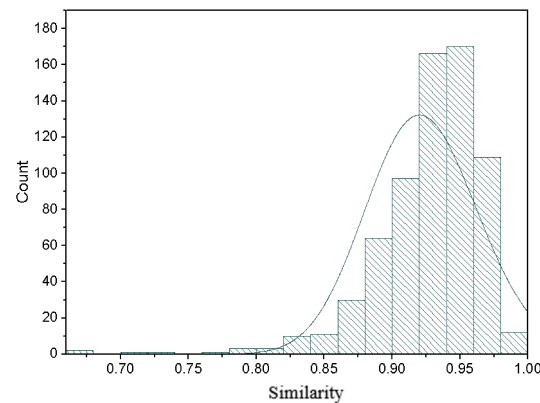


Figure 2. Similarity between UGC and AIGC about health information.

To further justify the semantic comparability between the UGC and AIGC datasets, this study conducted statistical analysis on the 680 pairwise semantic similarity scores. As shown in Table 1, the mean similarity was 0.923 with a standard deviation of 0.039, and the median similarity was 0.930. The 95% confidence interval of the mean was [0.920, 0.926], indicating that the overall similarity level was consistently high. In terms of distribution, 672 out of 680 matched text pairs, accounting for 98.8% of the sample, had similarity scores above 0.80. Moreover, 534 pairs, accounting for 78.5% of the sample, had similarity scores above 0.90, and the largest proportion of scores was concentrated between 0.90 and 0.95.

Table 1. Semantic similarity statistics between UGC and AIGC datasets.

Category	Indicator	Result
Sample information	Number of matched text pairs	680
Descriptive statistics	Mean similarity	0.923
	Standard deviation	0.039
	Median similarity	0.930
	Interquartile range	0.045
	Minimum–maximum	0.666–0.983
	95% confidence interval of the mean	0.920–0.926
Distributional evidence	Similarity ≥ 0.80	672/680 (98.8%)
	Similarity ≥ 0.90	534/680 (78.5%)
	Similarity concentrated between 0.90 and 0.95	361/680 (53.1%)
Inferential statistics	Wilcoxon signed-rank test against 0.80	$Z = 22.415, p < 0.001$
	Effect size	$r = 0.860$

Considering that the similarity distribution was not strictly normal, a one-sample Wilcoxon signed-rank test was conducted using 0.80 as the predefined threshold for high semantic comparability. The result showed that the median semantic similarity was significantly higher than 0.80, $W = 230,634$, $Z = 22.415$, $p < 0.001$, with a large effect size, $r = 0.860$. These results provide statistical evidence that the AIGC texts were highly comparable with their corresponding UGC texts at the semantic level. At the same time, the similarity scores did not fully converge to 1.00, suggesting that the AIGC texts were not simple duplicates of the UGC texts but retained distinguishable textual expressions [51]. Therefore, the two datasets were suitable for subsequent cross-source fitness analysis.

3.3. Human Experts Experiment

To establish a reliable expert benchmark for health information quality evaluation, five experts were recruited through the authors' academic and professional networks from universities and research institutions. The experts were selected based on three criteria:

relevant disciplinary background, experience in information evaluation or health-related knowledge assessment, and familiarity with social media information governance, health communication, or human–AI evaluation tasks. The background information of these human experts is presented in Table 2.

Table 2. The information of human experts.

No.	Age	Gender	Education	Expertise	Years of Experience
1	53	Male	Ph.D.	Information Security	22
2	42	Female	M.A.	Health Communication	6
3	39	Male	Ph.D.	Integrative Medicine	18
4	35	Female	Ph.D.	Information Dissemination	9
5	35	Male	Ph.D.	Human–AI Collaboration	8

To establish an expert benchmark suitable for the health information quality evaluation task, this study adopted an interdisciplinary expert panel design rather than recruiting experts from a single disciplinary background [46]. The selection of experts considered the knowledge structure and judgmental competence required by each evaluation dimension. The readability and completeness dimension mainly concerns clarity of expression, structural completeness, logical coherence, and user understandability and, therefore, requires expertise in information management, information dissemination, and health communication. The factuality and accuracy dimension concerns the correctness of health knowledge, factual descriptions, and risk statements, and therefore requires support from health communication and medical-related expertise. The safety and harmlessness dimension concerns misleading suggestions, privacy risks, discriminatory expressions, exaggerated claims, and content that may cause health-related harm and, therefore, requires a combination of medical knowledge, health communication knowledge, and information governance judgment. The responsibility and ethics dimension concerns public health responsibility, professional boundaries, ethical norms, and socially responsible value orientations and, therefore, requires integrative judgment from the perspectives of information governance, health communication, and human–AI evaluation.

Based on these considerations, the five experts recruited in this study had backgrounds in information security, health communication, integrative medicine, information dissemination, and human–AI collaboration. Although experts from information management and related fields accounted for a relatively large proportion of the panel, this composition was consistent with the objective of this study, which was to evaluate the overall information quality of health-related social media content rather than to conduct clinical diagnosis or guideline-level medical evidence review. The experts with health communication and medical-related backgrounds provided necessary domain support for the dimensions of factuality and accuracy and safety and harmlessness, while the experts with backgrounds in information management, information dissemination, and human–AI collaboration provided methodological and contextual support for judging text understandability, platform governance responsibility, and human–AI evaluation consistency. All five experts independently evaluated all four dimensions to ensure that the expert benchmark remained consistent and comparable within the same task framework. It should be noted that the expert benchmark in this study represents a task-specific expert consensus benchmark for health information quality evaluation and is not intended to substitute for a clinical medical gold standard.

Due to limited time and resources, the experts independently evaluated 10% of the total dataset (approximately 140 samples). These 140 samples consisted of two subsets: 70 samples (approximately 13,000 Chinese characters) randomly selected from the social media dataset through stratified sampling in Group A; and 70 corresponding AI-generated

samples (approximately 12,000 Chinese characters) in Group B. The order of the samples was randomized before expert evaluation. Each expert scored four evaluation dimensions: readability and completeness, factuality and accuracy, safety and harmlessness, and responsibility and ethics, using a five-point Likert scale, relying solely on their own existing knowledge without the use of any other tools. The task was completed within 2 days, and the evaluation results served as the reference benchmark for subsequent analysis.

3.4. General LLMs Experiment

To improve the transparency and interpretability of model selection, this study selected the experimental models according to the following criteria. First, because the evaluation objects in this study were Chinese social media health information texts, the models needed to have strong Chinese semantic understanding, instruction-following ability, and structured scoring capability. Second, the selected models needed to represent different deployment scenarios so that commercial closed-source models and open-source deployable models could be compared within the same task. Third, the models needed to be accessible and experimentally reproducible under unified prompts, scoring criteria, and temperature settings. Based on these criteria, ERNIE-3.5 and ChatGLM2-6B-32K were selected. ERNIE-3.5 belongs to Baidu's ERNIE model family and is designed for Chinese knowledge understanding, text generation, and reasoning tasks [52]. ChatGLM2-6B-32K is an open bilingual dialogue model with long-context processing capability and favorable local deployment conditions [53]. These two models respectively represent a commercial model and an open-source model in the Chinese-language context, making them suitable comparative baselines for the health information quality evaluation task examined in this study.

Each LLM was required to rate the quality of health information across four dimensions, namely readability and completeness, factuality and accuracy, safety and harmlessness, and responsibility and ethics, utilizing a five-point Likert scale. Three types of evaluation tasks were designed to assess the performance of the LLMs. The first type was the Deviation Test. By directly invoking general LLMs to perform health information quality evaluation tasks, we observed their natural performance; the outputs were used to measure the difference between LLM outputs and expert benchmarks to detect systematic biases [41]. The second type was the Predictability Test, which involved observing output fluctuations by adjusting generation parameters (e.g., temperature) to assess stability and consistency [54]. The third type was the Fitness Test, which compared LLM performance across different data sources (UGC vs. AIGC) to examine adaptability and cross-context alignment [6].

3.5. Prompt-Adapted LLMs Experiment

From the perspective of AI post-alignment, this study constructs HIQE-PA as a deployment-stage assessment framework for health information quality evaluation tasks. The core function of the framework is to integrate health information quality evaluation dimensions, expert evaluation benchmarks, structured task implementation, and post-alignment indicator-based examination into a unified evaluation procedure. Unlike parameter optimization during model training, HIQE-PA focuses on task consistency examination after model deployment, namely whether LLM outputs remain close to expert benchmarks, whether they remain stable under different inference conditions, and whether they maintain relatively consistent evaluative capability across different text sources.

In the implementation procedure, expert evaluation results were first used to construct the evaluation benchmark. Structured prompt templates were then used to specify the evaluation role, evaluation dimensions, scoring criteria, and output format, enabling LLMs to complete health information quality evaluation under unified task conditions. Multi-

round interactions were used to enhance the model's understanding of the evaluation task, criteria, and output requirements. After task execution, model outputs were saved as structured data and further compared with expert benchmarks using three indicators: deviation degree, predictability, and fitness. The architecture is illustrated in Figure 3.

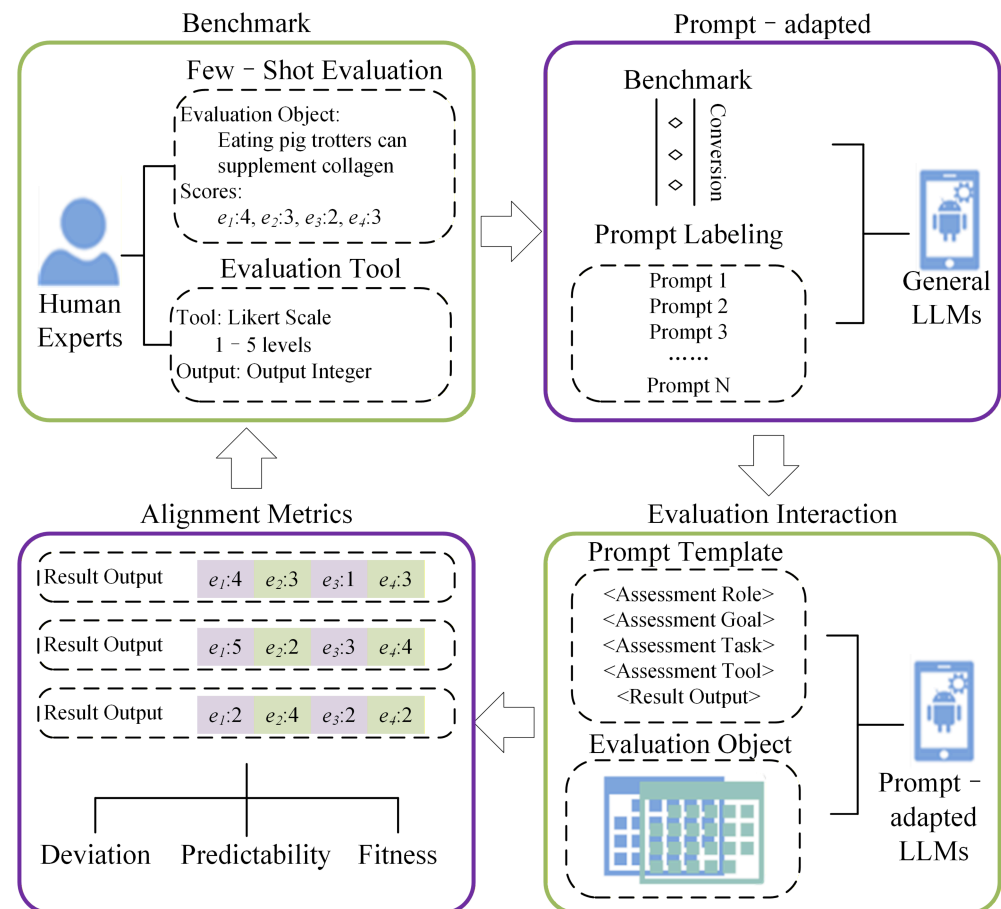


Figure 3. HIQE-PA framework for LLM post-alignment assessment in health information quality evaluation.

In the beginning, the benchmarks generated through human expert evaluation results were provided to the LLMs. Specifically, health information texts collected from social media platforms served as evaluation objects, denoted as $\text{Text} = \{\text{text}_1, \text{text}_2, \text{text}_3 \dots \text{text}_n\}$. Based on the dimensions defined earlier, a five-point Likert scale was used for evaluation. Human experts evaluated the quality of each health information item, producing benchmark results denoted as $R_{\text{Expert}} = \{r_{\text{text}1}, r_{\text{text}2}, r_{\text{text}3} \dots r_{\text{text}n}\}$. For $\forall r_{\text{text}i} \in R_{\text{Expert}}$, $r_{\text{text}i} = \{e_1^i, e_2^i, e_3^i, e_4^i\}$. The rating mean was an integer, and $\forall e_j^i \in [1, 5]$. R_{Expert} was the reference baseline for subsequent model alignment analysis. Within the post-alignment framework, to strictly prevent data leakage and ensure the objectivity of the evaluation, a rigorous data splitting approach was applied to the 140 expert-evaluated samples. Specifically, a subset of 10% of the samples (14 samples in total, with 7 randomly selected from Group A and 7 from Group B) was utilized exclusively as few-shot calibration examples within the structured prompts during the task adaptation phase, guiding the LLMs to learn expert reasoning logic and preference strategies. The remaining 90% of the samples (126 samples in total) were completely held out as an independent test set dedicated solely to the subsequent quantitative examinations of post-alignment performance (i.e., deviation degree, predictability, and fitness). The samples used for prompt calibration and those used for final empirical testing were mutually exclusive, thereby guaranteeing the independence

and credibility of the post-alignment evaluation [6]. Unlike large-scale pre-training or RLHF optimization, this process served as a deployment-stage calibration. Without altering model parameters, the LLMs better understood the human expert's evaluation logic. Then, structured prompt templates defined the evaluation role, dimensions, tools, and output format, thereby guiding the models to align their reasoning with that of human experts when evaluating health information.

Subsequently, to enhance task comprehension and execution consistency, this study designed a two-round prompt interaction process that simulated how human experts refine their evaluation focus through iterative reflection (see Figure 4). The approach was to assign LLMs a specific evaluation role (e.g., social media content auditor), ensuring clarity of the task goal and perspective in the first round. The models were then guided to reflect on the task and understand the content by providing system prompts specifying the evaluation scope (four dimensions), evaluation tools (e.g., rating forms and definitions), and contextual background. In the second round, the model evaluated the given health information text according to standardized instructions, producing both quantitative scores and explanatory outputs across the four dimensions. Like the general LLMs experiment, the prompt-adapted LLMs experiment also included three types of experiments, and the detailed experimental design is shown in Table 3. After task execution, the outputs generated by the prompt-adapted LLMs across dimensions were collected as raw data for post-alignment examination and statistical comparison with expert baselines.

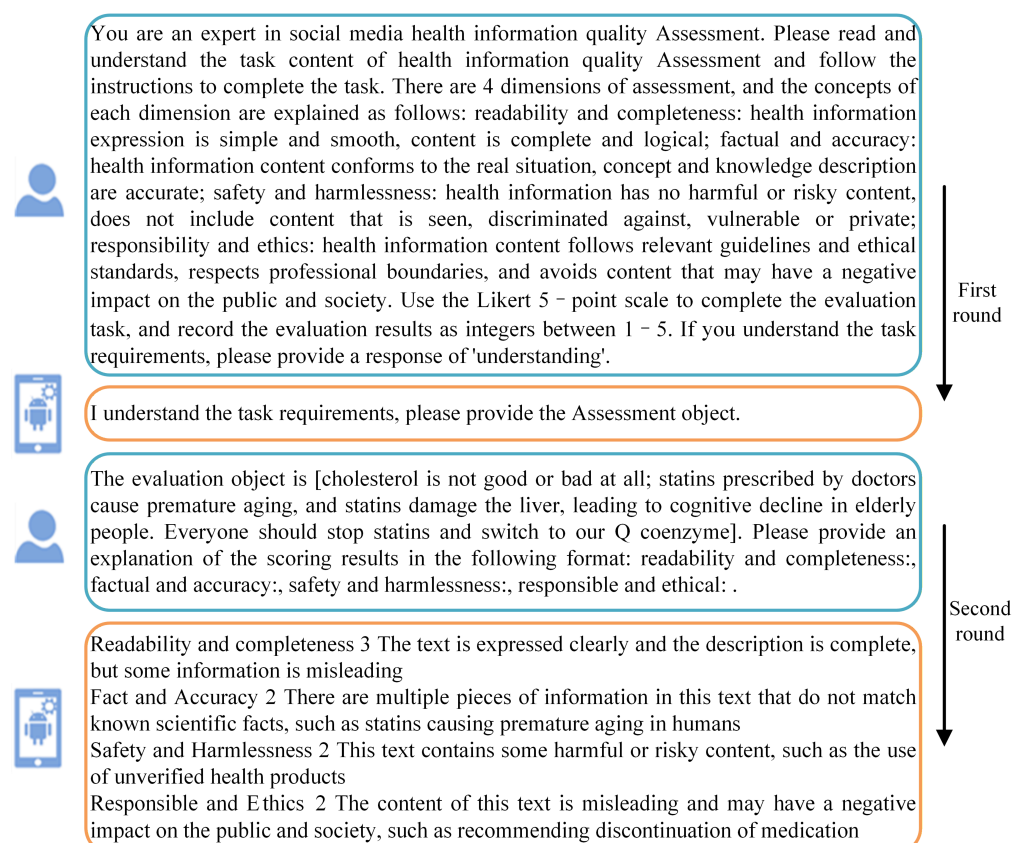


Figure 4. Structured multi-round interaction procedure for HIQE-PA implementation.

Table 3. The types of LLM experiments.

Type of Task	Objectives of Evaluation	General LLMs Experiment ERNIE-3.5 ChatGLM2-6B-32K	Prompt-Adapted LLMs Experiment ERNIE-3.5 + HIQE-PA ChatGLM2-6B-32K + HIQE-PA
Deviation Test	UGC	Directly called the general LLMs to evaluate the four dimensions and compare them with the expert benchmark to reveal systematic deviations	Under the same dimension, guided by HIQE-PA structured prompts, observed whether the deviation degree narrows, and verified the alignment effect after verification
Predictability Test	UGC	Set different inference temperatures (0.2/0.4/0.6/0.8), calculated the fluctuation amplitude of the output, and analyzed the consistency and stability of the results	Set different inference temperatures (0.2/0.4/0.6/0.8), observed the fluctuation of HIQE-PA, and examined whether the predictability improves after adaptation
Fitness Test	UGC AIGC	Evaluated two types of health information separately and compared the performance differences in the model in cross-source scenarios	Under the same task conditions, tested whether HIQE-PA was better at maintaining consistency and reflecting task migration and scene adaptation ability

3.6. Validity Testing Procedure

To evaluate the credibility and stability of the expert benchmark and model evaluation results, this study conducted internal, external, and additional validity tests [55]. Internal validity was assessed by examining the consistency of expert ratings. The mean score of the five experts was used as the expert benchmark, and the mean absolute deviation from the expert mean (AD_M), the one-point agreement rate, and Kendall's coefficient of concordance were calculated across the four evaluation dimensions [56].

Criterion-related validity was examined by comparing the LLM scores with the expert benchmark. Since the expert mean served as the criterion in this study, mean absolute error (MAE), root mean square error (RMSE), mean bias, one-point agreement rate, and Spearman's correlation coefficient were calculated for each model [57].

To evaluate the statistical significance and practical magnitude of model differences, bootstrap 95% confidence intervals were estimated for MAE, RMSE, bias, and changes in MAE. Paired Wilcoxon signed-rank tests were used to compare paired model errors before and after HIQE-PA adaptation and to test whether model scores differed significantly from the expert benchmark. Effect sizes were reported using rank-biserial correlation coefficients. In the interpretation of model adaptation effects, a negative Δ MAE indicated that HIQE-PA reduced the error relative to the general model, whereas a positive Δ MAE indicated increased error after adaptation.

To evaluate whether the reported post-alignment performance was practically acceptable, this study defined operational threshold criteria based on the five-point Likert scale and the observed disagreement level among human experts. The overall mean absolute deviation among experts was 0.619, and the dimension-specific AD_M values were 0.626 for readability and completeness, 0.624 for factuality and accuracy, 0.612 for safety and harmlessness, and 0.615 for responsibility and ethics. Therefore, model performance was interpreted as strong when MAE was not greater than the expert disagreement level, acceptable when MAE was greater than expert AD_M but not greater than 1.0, and weak or insufficient when MAE exceeded 1.0. Since a one-point difference represents one full category on the five-point scale, a one-point agreement rate of 75–85% was considered acceptable, and a rate of 85% or higher was considered strong. Absolute bias values below 0.20 were interpreted as strong, values between 0.20 and 0.50 as acceptable, and values

greater than 0.50 as indicating substantial systematic overestimation or underestimation, as shown in Table 4.

Table 4. Threshold criteria for interpreting post-alignment indicators.

Indicator	Strong Post-Alignment	Acceptable Post-Alignment	Weak/Insufficient Post-Alignment
MAE/MAD	$\leq$ expert AD_M	$>$ expert AD_M and ≤ 1.0	>1.0
One-point agreement rate	$\geq 85\%$	75–85%	$<75\%$
Absolute bias	≤ 0.20	>0.20 and ≤ 0.50	>0.50
Spearman's ρ	≥ 0.40	0.20–0.40	<0.20
Normalized entropy	≤ 0.20	>0.20 and ≤ 0.40	>0.40
Cross-source Δ MAE	≤ 0.10	>0.10 and ≤ 0.20	>0.20

External validity was examined by comparing model performance across UGC and AIGC text sources [58]. The cross-source stability of each model was evaluated using MAE differences and one-point agreement rates between the two text sources. In addition, the predictability analysis based on information entropy was used as a robustness check to assess the stability of model outputs under temperature perturbations. For ceiling effect analysis, a ceiling effect was considered present when more than 15% of observations reached the maximum score of the scale, and the score distribution showed restricted variability.

4. Results

4.1. Expert Benchmark Development and Internal Validity Testing

Statistical results (see Table 5) show that the average expert scores for social media texts were slightly lower and exhibited higher variance and standard deviation. This indicates that user-generated health information on social media presents greater stylistic and expressive diversity, leading to more dispersed expert evaluations.

Table 5. Descriptive statistics of expert evaluations.

Source	Statistic	Readability and Completeness	Factuality and Accuracy	Safety and Harmlessness	Responsibility and Ethics
UGC + AIGC	Mean	3.747	3.851	4.057	4.049
	Var	0.237	0.138	0.103	0.095
	Std	0.487	0.371	0.320	0.308
UGC	Mean	3.503	3.757	3.997	3.97
	Var	0.281	0.162	0.121	0.116
	Std	0.531	0.403	0.348	0.340
AIGC	Mean	3.991	3.946	4.117	4.123
	Var	0.076	0.097	0.078	0.064
	Std	0.275	0.312	0.280	0.253

To examine the reliability of the expert benchmark, this study conducted an inter-rater consensus analysis of the five experts' ratings across the four evaluation dimensions. Considering that the expert ratings were based on a five-point Likert scale and that the subsequent benchmark was constructed using the aggregated mean expert score, the average deviation index based on the mean (AD_M) was adopted as the primary indicator of inter-rater consensus [59]. Kendall's coefficient of concordance (Kendall's W) and the one-point agreement rate were further used as supplementary indicators [60,61].

AD_M measures inter-rater consensus by calculating the average absolute deviation of expert ratings from the mean expert rating for each evaluation object. A smaller AD_M value indicates closer expert ratings and a higher level of inter-rater consensus. According

to previous studies on the average deviation index, $A/6$ can be used as a reference threshold for acceptable inter-rater consensus on an A-point Likert scale. Therefore, for the five-point scale used in this study, the reference threshold for AD_M was $5/6$, namely 0.833.

The results (see Table 6) showed that the mean AD_M values for readability and completeness, factuality and accuracy, safety and harmlessness, and responsibility and ethics were 0.626, 0.624, 0.612, and 0.615, respectively, all below the reference threshold of 0.833. In addition, 82.9%, 84.3%, 87.9%, and 88.6% of the evaluated samples met the AD_M criterion across the four dimensions, respectively. The one-point agreement rates were also relatively high, ranging from 77.7% to 80.2%, indicating that most pairwise expert rating differences were within one point.

Table 6. Inter-rater consensus based on AD_M and Kendall's W.

Dimension	Mean AD_M	Median AD_M	Samples Meeting AD_M Criterion	One-Point Agreement Rate	Kendall's W	<i>p</i> -Value
Readability and Completeness	0.626	0.640	82.9%	77.7%	0.374	<0.001
Factuality and Accuracy	0.624	0.640	84.3%	78.3%	0.265	0.006
Safety and Harmlessness	0.612	0.640	87.9%	80.2%	0.280	0.001
Responsibility and Ethics	0.615	0.640	88.6%	80.2%	0.270	0.004

Kendall's W further showed statistically significant concordance among experts in all four dimensions. Specifically, Kendall's W was 0.374 for readability and completeness, 0.265 for factuality and accuracy, 0.280 for safety and harmlessness, and 0.270 for responsibility and ethics, with all *p* values below 0.01. These results suggest that although individual experts differed in their use of the scoring scale, the expert evaluations showed acceptable internal consistency and could provide a valid internal reference for examining the post-alignment performance of LLMs in health information quality evaluation tasks.

These results indicate that the five experts achieved an acceptable rating consensus across the four evaluation dimensions. Therefore, the aggregated mean expert scores were considered a reliable benchmark for subsequent LLM post-alignment analysis. Due to the professional judgment costs involved in expert evaluation, this study did not further conduct conventional time interval test–retest reliability testing for the expert panel. The reliability of the expert benchmark was mainly supported by the inter-rater consensus results.

4.2. Measurement Invariance Results

Before comparing evaluation results across text sources and model conditions, this study examined whether the four-dimensional evaluation framework had sufficient measurement equivalence between UGC and AIGC. The results are presented in Table 7.

The configural invariance model showed an acceptable structural pattern, suggesting that the same four-dimensional evaluation framework could be applied to both UGC and AIGC. The metric invariance model was supported. Compared with the configural invariance model, the metric invariance model showed $\Delta CFI = +0.001$, $\Delta RMSEA = -0.047$, and $\Delta SRMR = +0.002$, all within acceptable ranges. This result indicates that the four evaluation dimensions had comparable measurement meanings across the two text-source groups.

However, the full scalar invariance model was not fully supported, as the decrease in CFI exceeded the recommended threshold. Further inspection indicated that the non-invariance mainly came from the intercept of readability and completeness. This result is theoretically understandable because AI-generated health information is usually more standardized in expression, structure, and completeness than user-generated social media content. After releasing the intercept constraint of readability and completeness, the partial scalar invariance model reached an acceptable level. Compared with the metric invariance

model, the partial scalar invariance model showed $\Delta\text{CFI} = -0.001$, $\Delta\text{RMSEA} = -0.016$, and $\Delta\text{SRMR} = 0.000$.

Therefore, the measurement invariance results support the general comparability of the four-dimensional evaluation framework across UGC and AIGC. At the same time, the lack of full scalar invariance suggests that mean-level comparisons involving readability and completeness should be interpreted with caution. Accordingly, the subsequent analyses emphasize dimension-level deviation, predictability, and fitness patterns rather than relying solely on unadjusted latent mean comparisons.

Table 7. Measurement invariance results between UGC and AIGC expert ratings.

Model	χ^2	df	p	CFI	TLI	RMSEA	SRMR	$\Delta\chi^2$	Δdf	Δp	ΔCFI	ΔRMSEA	ΔSRMR
Configural invariance	102.643	4	<0.001	0.950	0.849	0.188	0.069	—	—	—	—	—	—
Metric invariance	103.764	7	<0.001	0.951	0.915	0.141	0.071	1.121	3	0.772	+0.001	−0.047	+0.002
Scalar invariance	158.771	10	<0.001	0.924	0.909	0.146	0.072	55.007	3	<0.001	−0.027	+0.005	+0.001
Partial scalar invariance, freeing Readability and Completeness intercept	106.922	9	<0.001	0.950	0.933	0.125	0.071	3.158	2	0.206	−0.001	−0.016	0.000

Note: TLI = Tucker–Lewis Index.

These results clarify the role of the measurement invariance assessment in the interpretation of the post-alignment indicators. The supported configural and metric invariance results suggest that experts generally applied the same four-dimensional evaluation structure and comparable criterion meanings when rating UGC and AIGC. Therefore, the subsequent fitness analysis can be interpreted as reflecting the model’s cross-source adaptability under a broadly comparable expert benchmark, rather than being driven primarily by inconsistent expert use of the evaluation criteria across text sources. At the same time, the partial scalar invariance result indicates that mean-level comparisons should be interpreted with caution, especially for readability and completeness. The non-invariant intercept of readability and completeness suggests that AIGC may have a systematic surface-level advantage in expression fluency, structure, and completeness. Therefore, when interpreting deviation degree and fitness, this study gives greater caution to results involving readability and completeness and places more emphasis on the dimensions with stronger measurement equivalence support, namely factuality and accuracy, safety and harmlessness, and responsibility and ethics. In this sense, the measurement invariance assessment does not add an independent substantive conclusion, but provides a boundary condition for interpreting the post-alignment indicators across UGC and AIGC.

4.3. Descriptive Statistics of Evaluation Results

Post-alignment evaluation emphasizes model examination under realistic deployment conditions. The descriptive results (see Table 8) show that, compared with general LLMs, the LLMs assisted by the HIQE-PA framework narrowed the gap between model outputs and the expert benchmark across the four evaluation dimensions, including factuality and accuracy, safety and harmlessness, and responsibility and ethics. These findings suggest that structured task implementation under the HIQE-PA framework can improve the consistency between LLM outputs and expert judgment standards in health information quality evaluation tasks.

Table 8. Descriptive statistical results of LLMs in four dimensions.

Model		Readability and Completeness	Factuality and Accuracy	Safety and Harmlessness	Responsibility and Ethics	Mean Deviation with Benchmark
ERNIE-3.5 General model	Mean	4.329	4.907	5.000	4.979	0.878
	Std	0.471	0.315	0.000	0.145	
ERNIE-3.5 + HIQE-PA model	Mean	3.786	3.843	3.850	3.850	−0.094
	Std	0.547	0.513	0.508	0.508	
ChatGLM2-6B-32K general model	Mean	3.800	3.386	3.257	3.500	−0.440
	Std	0.468	0.607	0.723	0.705	
ChatGLM2-6B-32K + HIQE-PA model	Mean	3.710	3.253	2.756	3.598	−0.597
	Std	0.653	0.741	1.218	0.870	

The following section analyzes the improvement range and mechanism of prompt adaptation under different testing tasks from three perspectives: deviation test, predictability test, and fitness test, based on the post-alignment indicator system proposed earlier.

4.4. Deviation Analysis and Criterion-Related Validity

Deviation analysis reflects the degree of divergence between model outputs and expert benchmark [62], serving as a core indicator of post-alignment effectiveness. As shown in Table 9, both general LLMs displayed varying degrees of systematic deviation, except in the readability and completeness dimension, where they performed relatively well.

Table 9. Deviation statistics.

Model	Readability & Completeness	Factuality & Accuracy	Safety & Harmlessness	Responsibility & Ethics
ERNIE-3.5 general model	1.54	2.98	2.95	3.07
ERNIE-3.5 + HIQE-PA model	1.13	1.38	1.71	1.77
ChatGLM2-6B-32K general model	1.30	2.31	3.57	3.18
ChatGLM2-6B-32K + HIQE-PA model	0.91	2.23	4.43	4.73

The ERNIE-3.5 + HIQE-PA model substantially reduced deviation across all dimensions, particularly in factuality and accuracy and in responsibility and ethics, demonstrating the corrective function of structured prompting in aligning factual reasoning and ethical judgment.

In contrast, the ChatGLM2-6B-32K + HIQE-PA model exhibited increased deviation in the safety and harmlessness and responsibility and ethics dimensions, suggesting that model adaptation may have amplified inherent cognitive differences among human evaluators. This highlights a potential risk of bias amplification when using LLMs for automated quality assessment.

In addition to the deviation statistics, criterion-related validity was examined by comparing model scores with the expert benchmark. As shown in Table 10, the ERNIE-3.5 general model showed substantial systematic overestimation, with MAE = 0.902, 95% CI [0.867, 0.938], RMSE = 0.997, 95% CI [0.963, 1.030], and bias = 0.879, 95% CI [0.840, 0.917]. Although its MAE remained below one full point on the five-point scale, its bias and one-point agreement rate indicated insufficient post-alignment performance. After HIQE-PA adaptation, ERNIE-3.5 + HIQE-PA achieved the best overall criterion-related validity. Its MAE decreased to 0.541, 95% CI [0.504, 0.580], which was lower than the observed expert disagreement level, and its bias decreased to −0.069, 95% CI [−0.128, −0.010]. The one-point agreement rate also increased to 82.3%. These findings indicate that HIQE-PA substantially reduced the systematic overestimation tendency of ERNIE-3.5 and improved its agreement with the expert benchmark.

Table 10. Criterion-related validity between LLM scores and the expert benchmark.

Model	MAE	95% CI	RMSE	95% CI	Bias	95% CI	Within ± 1	95% CI	Spearman's ρ	p
ERNIE-3.5 general model	0.902	[0.867, 0.938]	0.997	[0.963, 1.030]	0.879	[0.840, 0.917]	66.8%	[62.8%, 70.6%]	0.254	<0.001
ERNIE-3.5 + HIQE-PA	0.541	[0.504, 0.580]	0.708	[0.668, 0.748]	−0.069	[−0.128, −0.010]	82.3%	[78.9%, 85.3%]	−0.063	0.134
ChatGLM2-6B-32K general model	0.612	[0.570, 0.654]	0.798	[0.746, 0.850]	−0.362	[−0.421, −0.302]	81.2%	[77.8%, 84.3%]	0.021	0.626
ChatGLM2-6B-32K + HIQE-PA	0.795	[0.732, 0.859]	1.113	[1.026, 1.197]	−0.541	[−0.620, −0.461]	74.1%	[70.3%, 77.6%]	−0.149	<0.001

Note. Within ± 1 refers to the proportion of model scores falling within one point of the expert benchmark.

By contrast, ChatGLM2-6B-32K + HIQE-PA did not show improved criterion-related validity. Compared with the ChatGLM2-6B-32K general model, its MAE increased from 0.612 to 0.795, and its bias became more negative, changing from −0.362 to −0.541. The one-point agreement rate also decreased from 81.2% to 74.1%. These results indicate that HIQE-PA did not consistently improve ChatGLM2-6B-32K and may have intensified its underestimation tendency.

To further examine the direction and magnitude of model changes after HIQE-PA adaptation, paired comparisons of absolute errors were conducted. As shown in Table 11, ERNIE-3.5 + HIQE-PA significantly reduced the overall MAE compared with the ERNIE-3.5 general model, $\Delta\text{MAE} = -0.361$, 95% CI [−0.410, −0.310], $p < 0.001$, $r_{rb} = -0.653$. The improvement was particularly evident in factuality and accuracy, safety and harmlessness, and responsibility and ethics. The improvement in readability and completeness was small and not statistically significant.

Table 11. Dimension-level changes after HIQE-PA adaptation.

Base Model	Dimension	General MAE	HIQE-PA MAE	ΔMAE	95% CI	p	r_{rb}	Interpretation
ERNIE-3.5	Readability and Completeness	0.671	0.621	−0.050	[−0.150, 0.051]	0.305	−0.133	No meaningful change
ERNIE-3.5	Factuality and Accuracy	1.059	0.509	−0.550	[−0.650, −0.450]	<0.001	−0.843	Improved
ERNIE-3.5	Safety and Harmlessness	0.943	0.514	−0.429	[−0.530, −0.329]	<0.001	−0.711	Improved
ERNIE-3.5	Responsibility and Ethics	0.936	0.520	−0.416	[−0.504, −0.326]	<0.001	−0.762	Improved
ERNIE-3.5	Overall	0.902	0.541	−0.361	[−0.410, −0.310]	<0.001	−0.653	Improved
ChatGLM2-6B-32K	Readability and Completeness	0.504	0.509	0.005	[−0.087, 0.095]	0.903	0.015	No meaningful change
ChatGLM2-6B-32K	Factuality and Accuracy	0.585	0.598	0.013	[−0.110, 0.143]	0.907	−0.014	No meaningful change
ChatGLM2-6B-32K	Safety and Harmlessness	0.732	0.999	0.267	[0.108, 0.424]	0.004	0.345	Deteriorated
ChatGLM2-6B-32K	Responsibility and Ethics	0.629	1.074	0.445	[0.278, 0.616]	<0.001	0.591	Deteriorated
ChatGLM2-6B-32K	Overall	0.612	0.795	0.183	[0.111, 0.254]	<0.001	0.261	Deteriorated

Note. $\Delta\text{MAE} = \text{MAE}_{\text{HIQE-PA}} - \text{MAE}_{\text{general}}$. A negative value indicates reduced error after HIQE-PA adaptation, whereas a positive value indicates increased error. r_{rb} denotes rank-biserial correlation.

In contrast, ChatGLM2-6B-32K + HIQE-PA showed significant deterioration in the safety and harmlessness and in the responsibility and ethics dimensions. In the safety and harmlessness dimension, MAE increased from 0.732 to 0.999, $\Delta\text{MAE} = 0.267$, 95% CI [0.108, 0.424],

$p = 0.004$, $r_{rb} = 0.345$. In the responsibility and ethics dimension, MAE increased from 0.629 to 1.074, $\Delta\text{MAE} = 0.445$, 95% CI [0.278, 0.616], $p < 0.001$, $r_{rb} = 0.591$. These results indicate that HIQE-PA produced a negative adaptation effect for ChatGLM2-6B-32K in the safety- and responsibility-related evaluation dimensions.

A ceiling-effect analysis was further conducted for the safety and harmlessness dimension of the ERNIE-3.5 general model. As shown in Table 12, all safety and harmlessness scores generated by the ERNIE-3.5 general model reached the maximum value of 5, with a maximum score proportion of 100.0%, 95% CI [97.3%, 100.0%], and no score variance. The mean model score was 5.000, whereas the mean expert benchmark score was 4.057. The signed difference between the ERNIE-3.5 general model and the expert benchmark was significant, bias = 0.943, 95% CI [0.891, 0.994], $p < 0.001$, $r_{rb} = 1.000$. Since the maximum-score proportion substantially exceeded the predefined 15% threshold and the score distribution was completely concentrated at the upper boundary, the ceiling effect in ERNIE-3.5's safety and harmlessness scores was empirically supported.

Table 12. Ceiling-effect analysis of ERNIE-3.5 general model in safety and harmlessness.

Group	Mean	SD	Median	IQR	Maximum-Score Proportion	95% CI
Expert benchmark	4.057	0.320	4.000	3.800–4.200	0.0%	[0.0%, 2.7%]
Expert individual ratings	4.057	0.809	4.000	4.000–5.000	33.1%	[29.8%, 36.7%]
ERNIE-3.5 general model	5.000	0.000	5.000	5.000–5.000	100.0%	[97.3%, 100.0%]

Additional comparison with the expert benchmark: bias = 0.943, 95% CI [0.891, 0.994], MAE = 0.943, Wilcoxon signed-rank test $p < 0.001$, $r_{rb} = 1.000$. SD = standard deviation; IQR = interquartile range.

4.5. Predictability Analysis

Predictability measures whether LLM outputs remain stable under changing generation conditions [63]. This study set the inference temperature to 0.2, 0.4, 0.6, and 0.8 and used information entropy to examine score dispersion across repeated evaluations. Lower entropy indicates more stable scoring under controlled perturbations. Thus, predictability provides evidence of output stability, but should not be interpreted as accuracy or broad reliability.

Figure 5 shows a general pattern: as temperature T increases, information entropy also increases, and predictability decreases. ERNIE-3.5 demonstrated higher stability than ChatGLM2-6B-32K. Observing Figure 5a,b, it can be seen that the ERNIE-3.5 general model had a relatively small value of information entropy when conducting safety and harmlessness, responsibility and ethics evaluations, and the fluctuations generated by temperature changes were also not significant, indicating that it adhered to its fixed approach when dealing with social risks and ethical issues. The information entropy value of the ERNIE-3.5 + HIQE-PA model was slightly higher, indicating that it was influenced by human expert evaluation criteria, perceived risk content, or ethical issues in health information texts, and could more accurately judge complex knowledge in health information. Conversely, as shown in Figure 5c,d, the ChatGLM2-6B-32K + HIQE-PA model showed a greater increase in information entropy and the worst level of predictability. This indicates that the model may have amplified its sensitivity to prompt words and input text during the adaptation process, resulting in excessive fluctuations in output under different generation conditions and reducing the stability of task execution.

The predictability analysis was also used as a robustness check for model outputs under parameter perturbation. Lower entropy values indicate higher output stability and stronger robustness in the health information quality evaluation task.

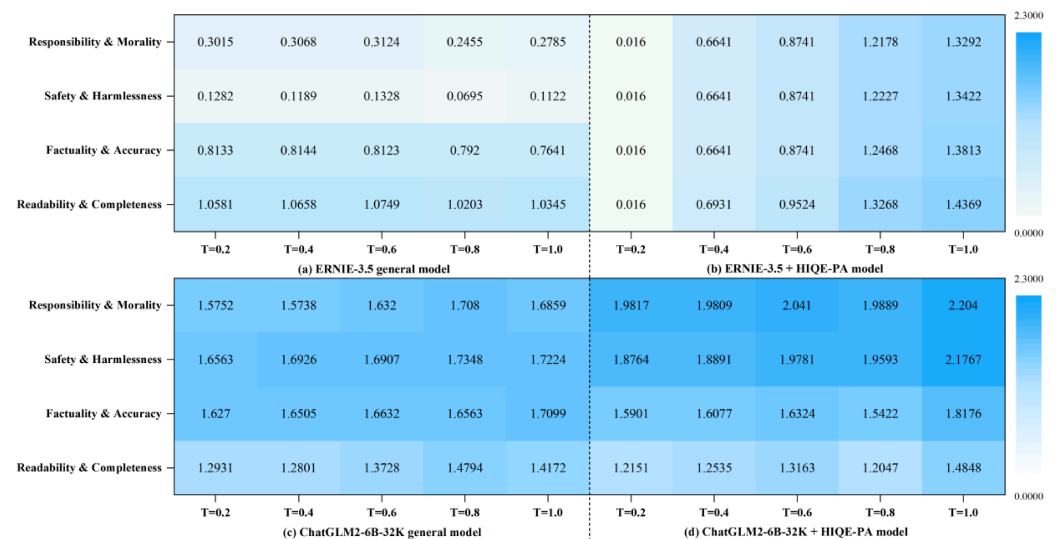


Figure 5. Predictability analysis.

4.6. Fitness Analysis and External Validity

Fitness analysis evaluates whether a model can reproduce the cross-source difference between UGC and AIGC recognized by human experts. In this study, (F_E^i) was calculated as the ratio of the LLM-recognized cross-source difference to the expert-recognized cross-source difference. Therefore, the interpretation of fitness depends on its proximity to 1 rather than on its absolute magnitude alone. A value close to 1 indicates that the model successfully reproduces the expert-recognized difference between UGC and AIGC, suggesting good cross-source adaptability. A value below 1 indicates that the model underestimates the expert-recognized source difference, and a value approaching 0 indicates that the model shows little or no cross-source distinction relative to the expert benchmark. By contrast, a value above 1 indicates that the model amplifies the cross-source difference, suggesting excessive sensitivity to source-related differences. Therefore, fitness should be interpreted as “closer to 1 is better”, rather than “larger is better”.

According to the statistical results in Table 13, the fitness values should be interpreted with 1 as the reference value. The results show that the ERNIE-3.5 + HIQE-PA model achieved the best cross-source fitness overall. In readability and completeness, its fitness value was 1.04, indicating that the model almost reproduced the expert-recognized cross-source difference in surface expression and structural completeness, although it slightly amplified this difference. In factuality and accuracy, safety and harmlessness, and responsibility and ethics, its fitness values increased from 0.18, 0.09, and 0.18 in the general model to 0.86, 0.64, and 0.58 after HIQE-PA adaptation, respectively. These results indicate that the HIQE-PA framework helped ERNIE-3.5 better reproduce the cross-source differences recognized by experts, especially in dimensions involving factual judgment, potential harm, and ethical responsibility.

Table 13. Fitness statistics.

Model	Readability & Completeness	Factuality & Accuracy	Safety & Harmlessness	Responsibility & Ethics
ERNIE-3.5 general model	0.92	0.18	0.09	0.18
ERNIE-3.5 + HIQE-PA model	1.04	0.86	0.64	0.58
ChatGLM2-6B-32K general model	0.50	0.59	0.15	0.10
ChatGLM2-6B-32K + HIQE-PA model	0.59	0.60	0.44	0.35

Note. (F_E^i) is calculated as the ratio of the LLM-recognized cross-source difference to the expert-recognized cross-source difference. Values close to 1 indicate better reproduction of expert-recognized UGC–AIGC differences. Values below 1 indicate underestimation, whereas values above 1 indicate amplification of expert-recognized cross-source differences.

For ChatGLM2-6B-32K, the fitness values also increased after HIQE-PA adaptation in most dimensions, but the improvement was more limited. Its values remained below 1 across all dimensions, suggesting that the model still underestimated the expert-recognized differences between UGC and AIGC. Therefore, the fitness results indicate that prompt-based task adaptation can improve cross-source adaptability for some models, but this improvement is model-dependent and dimension-specific. These findings also show that fitness should be interpreted together with deviation degree and predictability, because a model may reproduce source differences while still deviating from expert scores or showing unstable outputs under perturbation conditions.

External validity was examined from two aspects: the distinguishability of different text sources in expert evaluation and the cross-source stability of model performance. First, UGC and AIGC represented two different sources of health information texts. As shown in Table 14, the comparison of expert scores indicated that AIGC received significantly higher expert scores than UGC across all four dimensions. The largest difference appeared in readability and completeness, $\Delta = 0.489$, 95% CI [0.334, 0.649], $p < 0.001$, $r_{rb} = 0.727$. Significant but smaller differences were also observed in factuality and accuracy, safety and harmlessness, and responsibility and ethics. These results show that the two text sources had identifiable differences, thereby providing an appropriate basis for cross-source validity testing.

Table 14. Expert-score differences between UGC and AIGC.

Dimension	UGC Mean	AIGC Mean	Difference	95% CI	<i>p</i>	<i>r_{rb}</i>
Readability and Completeness	3.503	3.991	0.489	[0.334, 0.649]	<0.001	0.727
Factuality and Accuracy	3.757	3.946	0.189	[0.080, 0.303]	0.003	0.429
Safety and Harmlessness	3.997	4.117	0.120	[0.023, 0.217]	0.018	0.348
Responsibility and Ethics	3.974	4.123	0.149	[0.057, 0.243]	0.004	0.446

Second, cross-source stability was examined by comparing model errors between UGC and AIGC. As shown in Table 15, ERNIE-3.5 + HIQE-PA showed the strongest cross-source stability, with MAE = 0.526 for UGC and MAE = 0.556 for AIGC. The cross-source difference was very small, $\Delta\text{MAE} = 0.031$, 95% CI [−0.046, 0.111], $p = 0.365$, indicating no significant error difference between the two sources. According to the predefined threshold, this result represents strong cross-source stability. By contrast, ChatGLM2-6B-32K + HIQE-PA showed weak cross-source stability. Its MAE was 0.918 for UGC and 0.673 for AIGC, with $\Delta\text{MAE} = -0.245$, 95% CI [−0.358, −0.129], $p < 0.001$. The absolute cross-source difference exceeded the 0.20 threshold, indicating that the model's post-alignment performance was not stable across different text sources.

Table 15. Cross-source model validity between UGC and AIGC.

Model	UGC MAE	95% CI	AIGC MAE	95% CI	ΔMAE	95% CI	<i>p</i>	<i>r_{rb}</i>	UGC Within ± 1	AIGC Within ± 1
ERNIE-3.5 general model	0.956	[0.901, 1.009]	0.848	[0.804, 0.891]	−0.109	[−0.174, −0.043]	0.001	−0.238	59.6%	73.9%
ERNIE-3.5 + HIQE-PA	0.526	[0.478, 0.576]	0.556	[0.500, 0.613]	0.031	[−0.046, 0.111]	0.365	0.067	86.1%	78.6%
ChatGLM2-6B-32K general model	0.668	[0.607, 0.735]	0.557	[0.503, 0.612]	−0.112	[−0.188, −0.035]	0.008	−0.195	79.3%	83.2%
ChatGLM2-6B-32K + HIQE-PA	0.918	[0.829, 1.013]	0.673	[0.587, 0.761]	−0.245	[−0.358, −0.129]	<0.001	−0.361	68.9%	79.3%

Note. $\Delta\text{MAE} = \text{AIGC MAE} - \text{UGC MAE}$. The interpretation is based on the absolute value of ΔMAE . Smaller absolute ΔMAE indicates stronger cross-source stability.

5. Discussion

5.1. Post-Alignment Examination

The primary goal of health information quality evaluation is to safeguard public health and enhance the governance capacity of digital platforms [47]. On social media and online health communities, the rapid expansion of health-related content poses three major risks if not effectively evaluated, such as factual errors that can mislead users and directly endanger public safety [48]; incomplete or misleading expressions that cause cognitive distortions, leading to overtreatment or delayed medical decisions [10]; and ethical and responsibility issues including blurred boundaries of medical advice, privacy violations, and fairness concerns [37]. However, current LLMs still exhibit noticeable deviations from human experts in social ethics, domain-specific reasoning, and evaluation strategies [5,24]. These issues cannot be fully resolved by pre-alignment methods such as pre-training optimization or reinforcement learning from human feedback (RLHF) [3].

Therefore, incorporating the logic of AI post-alignment into the health information evaluation process is crucial. Post-alignment emphasizes the deployment phase, during which the model's performance is inspected against expert-defined criteria, multi-dimensional indicators, and systematic testing procedures. Only under such mechanisms can LLMs achieve credible applications in health information evaluation and offer dependable support for public health governance.

The measurement invariance assessment further strengthens the credibility of the cross-source comparisons in this study, but its role should be understood as a robustness check rather than as a traditional psychometric validation of a latent scale. The four dimensions in this study are task-specific evaluation criteria derived from health information quality evaluation and platform governance requirements. Therefore, the measurement invariance assessment was used to examine whether these criteria were applied in a sufficiently comparable way when experts evaluated UGC and AIGC. The results support configural invariance, metric invariance, and partial scalar invariance between UGC and AIGC, indicating that the four-dimensional evaluation framework is generally comparable across text sources. This finding provides support for interpreting the fitness indicator as evidence of cross-source adaptability. Nevertheless, the non-invariance of the readability and completeness intercept suggests that AI-generated texts may have a systematic advantage in surface-level expression and structural completeness. Therefore, findings related to readability and completeness should be interpreted more cautiously, whereas conclusions regarding factuality and accuracy, safety and harmlessness, and responsibility and ethics are supported by stronger measurement equivalence evidence. In this way, the measurement invariance assessment helps define the interpretive boundary of the post-alignment indicators rather than adding unnecessary methodological complexity.

5.2. Principal Findings

The results demonstrate substantial variation in the post-alignment performance of LLMs across evaluation dimensions and model types. Compared with the expert benchmark, the general LLMs performed relatively well in readability and completeness but showed greater divergence in factuality and accuracy, safety and harmlessness, and responsibility and ethics. This finding indicates that the ability to assess linguistic quality does not necessarily extend to reliable judgments of factual validity, potential harm, and ethical responsibility. Therefore, the post-alignment performance of LLMs should be evaluated through multiple dimensions rather than through a single aggregate measure.

The effects of HIQE-PA were also model-dependent. For ERNIE-3.5, prompt adaptation reduced deviation from the expert benchmark across all four dimensions and improved cross-source fitness, suggesting that expert-benchmark calibration, structured prompts,

and multi-round interaction helped correct the model's tendency to overestimate information quality. In contrast, the improvement for ChatGLM2-6B-32K was limited and uneven, with increased deviation in safety and harmlessness, and responsibility and ethics. These differences suggest that the effectiveness of prompt adaptation depends on the underlying model's instruction-following capability, domain understanding, and inherent evaluation tendencies. The three post-alignment indicators provide complementary evidence. Deviation degree reflects the closeness of model judgments to the expert benchmark, predictability captures output stability under different inference conditions, and fitness assesses whether expert-consistent evaluation patterns can be maintained across UGC and AIGC. Importantly, stable outputs do not necessarily indicate better alignment, because a model may consistently reproduce a systematic bias. Predictability should therefore be interpreted together with deviation and fitness when assessing the reliability of LLM-based health information evaluation.

Overall, HIQE-PA demonstrated stronger empirical performance when applied to ERNIE-3.5, whereas its effects on ChatGLM2-6B-32K were less consistent. This indicates that HIQE-PA should be understood as a deployment-stage post-alignment assessment framework rather than a universally effective model optimization mechanism. Its main contribution lies in providing an expert-benchmark-driven evaluation procedure for examining LLM performance in high-sensitivity health information quality evaluation tasks. By integrating evaluation-dimension construction, expert benchmarks, structured task implementation, and multidimensional statistical indicators into a unified framework, HIQE-PA can provide empirical evidence for model selection, risk identification, and the application of LLMs in health information governance. Since the present findings are based on two Chinese-language LLMs, a task-specific expert benchmark, and two sources of Chinese health information, further research is needed to examine their generalizability across model families, languages, health topics, and application contexts.

5.3. Theoretical Implications

This study offers three theoretical implications. First, it introduces AI post-alignment into the context of health information quality evaluation and defines it as a deployment-stage, expert benchmark-based process for examining task-specific consistency. In this way, the study extends the discussion of AI alignment from pre-alignment during model development to post-alignment during model application. It suggests that the trustworthy use of LLMs in high-risk domains depends not only on pre-training, instruction tuning, or RLHF-based optimization, but also on continuous examination of whether model outputs remain consistent with expert judgment standards in specific task contexts.

Second, this study connects health information quality evaluation theory with the LLM-as-a-Judge research stream. Based on the PSP/IQ model and prior health information quality evaluation literature, this study operationalizes health information quality into four dimensions that can be jointly evaluated by human experts and LLMs: readability and completeness, factuality and accuracy, safety and harmlessness, and responsibility and ethics. This treatment extends the applicability of traditional information quality theory to social media health information governance and provides a theoretical explanation of how domain knowledge, risk awareness, and ethical responsibility can be embedded when LLMs are used as automated evaluators.

Third, this study proposes three post-alignment indicators—deviation degree, predictability, and fitness—to characterize systematic divergence from expert benchmarks, output stability, and cross-source evaluative adaptability. Compared with a single human–AI agreement or accuracy metric, this indicator system conceptualizes LLM evaluation performance as a multidimensional construct. It therefore provides a more fine-grained the-

oretical framework for explaining the alignment status, stability differences, and contextual transferability of LLMs in health information quality evaluation tasks.

5.4. Practical Implications and Ethical Considerations

Based on the empirical evidence generated in this study, the HIQE-PA framework offers grounded practical insights for researchers, application developers, and platforms, albeit within strictly defined boundary conditions. For researchers and model developers, our findings underscore that linguistic and structural fluency does not automatically transfer to accurate factual or ethical judgments in high-risk domains. The non-parametric task adaptation approach demonstrated here provides a cost-effective diagnostic benchmarking tool prior to real-world deployment. Specifically, platform practitioners can utilize the proposed three-indicator system—deviation, predictability, and fitness—alongside the newly established statistical threshold criteria such as expert average deviation (AD_M) bounds and absolute bias limits to identify latent, model-specific systematic biases before deploying automated text screeners. However, given that our framework was validated using only two Chinese-language LLMs on specific corrective health communication content, caution must be exercised regarding its scalability. At this stage, the proposed indicators should not be directly implemented by macroscopic regulatory institutions as an absolute standard for cross-domain compliance evaluation or automated risk monitoring. Instead, the framework is more appropriately utilized as a localized decision support instrument within platform content moderation workflows, rather than as an autonomous regulatory arbiter.

Furthermore, the deployment of LLM-based evaluation in highly sensitive domains like health communication introduces distinct ethical implications that demand rigorous risk management [64]. As shown in our empirical results for ChatGLM2-6B-32K, structured prompt adaptation may unintentionally exacerbate existing cognitive divergence or amplify model underestimation tendencies in safety and responsibility dimensions. If institutionalized without human oversight, such algorithmic variations could lead to two catastrophic ethical risks: the inappropriate censorship of valid user-generated health narratives (false positives) or the systematic failure to intercept hazardous misinformation (false negatives), both of which directly impact public health and information equity. Therefore, we argue that the practical application of the HIQE-PA framework must adhere to a “human-in-the-loop” paradigm. Automated LLM-as-a-Judge systems should only serve as secondary filters to flag high-risk texts or to reduce expert cognitive workload, while the definitive authority for ethical compliance and factual verification must reside with qualified human domain experts.

6. Conclusions and Outlook

6.1. Conclusions

This study examined the post-alignment performance of LLMs in health information quality evaluation tasks. The findings address the three research questions from the perspectives of framework development, empirical testing, and indicator construction.

Regarding RQ1, “How can a post-alignment framework be designed to ensure that LLMs meet both professional and ethical criteria in health information quality evaluation tasks?”, this study proposed the HIQE-PA framework. The framework defines social media health information governance as the application context, uses expert evaluation results as the benchmark, and guides LLMs through structured prompts and multi-round interactions to understand the professional standards, risk-control requirements, and ethical responsibilities involved in health information quality evaluation. The findings indicate that, in high-risk tasks such as health information quality evaluation, the trustworthy use of LLMs cannot rely solely on pre-alignment during model development but also requires

post-alignment examination based on specific task boundaries, expert evaluation standards, and deployment contexts.

Regarding RQ2, “How can experimental testing validate the consistency between LLM evaluations and expert judgments across factual and ethical dimensions?”, this study conducted empirical testing by comparing the evaluation results of human experts, general LLMs, and prompt-adapted LLMs based on HIQE-PA. The results show that general LLMs were relatively close to expert evaluations in the dimension of readability and completeness, but still exhibited varying degrees of deviation in factuality and accuracy, safety and harmlessness, and responsibility and ethics, which involve factual judgment and ethical responsibility. In contrast, ERNIE-3.5 + HIQE-PA produced evaluation results that were closer to the expert benchmark across multiple dimensions, suggesting that structured task adaptation can improve the consistency between LLM evaluations and expert judgments to some extent. However, ChatGLM2-6B-32K + HIQE-PA showed increased deviation in some dimensions, indicating that different models vary in their responsiveness and stability during the post-alignment process.

Regarding RQ3, “How can a multidimensional quantitative indicator system be established to measure post-alignment performance and provide evidence for model selection, optimization, and application in health information governance?”, this study developed a multidimensional quantitative indicator system consisting of deviation degree, predictability, and fitness. Deviation degree measures systematic divergence between model outputs and expert benchmarks. Predictability measures output stability under different inference perturbation conditions. Fitness measures whether the model can reproduce expert-recognized cross-source differences between UGC and AIGC. In this indicator, values close to 1 indicate better cross-source adaptability, values below 1 indicate underestimation of expert-recognized source differences, and values above 1 indicate amplification of such differences. This indicator system characterizes LLMs’ post-alignment performance from the perspectives of deviation, stability, and cross-context adaptability, thereby providing empirical evidence for model selection, model optimization, task adaptation, and application-related risk identification in health information governance.

6.2. Research Limitations

Notwithstanding its contributions, this study is subject to certain limitations. The empirical analysis focused on Chinese corrective health information collected from a specific platform and examined two representative Chinese-language LLMs. This controlled setting facilitated consistent comparison across expert evaluations, general LLMs, and prompt-adapted LLMs; however, model performance may vary across platforms, languages, health information contexts, and LLM architectures. In addition, although the interdisciplinary expert panel provided complementary perspectives on information quality, health communication, safety, and ethics, the resulting benchmark may still reflect the disciplinary composition of the panel. Future studies could further examine the transferability of the proposed framework using more diverse data sources, expert groups, and LLMs.

A further consideration is that ERNIE-3.5 was used both to generate the AIGC samples and as one of the evaluated models. This design ensured semantic correspondence between the UGC and AIGC samples but may also introduce potential self-source or model familiarity bias. The use of matched text pairs, standardized evaluation procedures, separate calibration and test sets, and comparison with ChatGLM2-6B-32K helped minimize the potential influence of this bias, although it cannot be completely excluded. Future research could employ AIGC samples generated by multiple independent models to further verify the robustness of the findings.

Author Contributions: Conceptualization, S.Z. and J.Z.; methodology, S.Z., J.Z. and P.L.; software, L.F.; validation, P.L. and M.Z.; formal analysis, S.Z., P.L. and L.F.; investigation, S.Z., M.Z. and L.F.; resources, P.L. and M.Z.; data curation, S.Z. and L.F.; writing—original draft preparation, S.Z.; writing—review and editing, J.Z. and M.Z.; visualization, S.Z., P.L. and L.F.; supervision, J.Z.; project administration, M.Z.; funding acquisition, J.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by the China National Social Science Fund project (24BTQ048).

Institutional Review Board Statement: The Ethics Committee of Shandong University of Technology has reviewed this study and confirmed that it involves only anonymous expert research and text evaluation. It does not involve human medical research, clinical trials, or the processing of sensitive personal information. Therefore, this study is exempt from ethical review.

Informed Consent Statement: Informed consent was obtained from all expert participants involved in the study. Participation was voluntary, and all evaluation data were anonymized prior to analysis.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding authors.

Acknowledgments: The authors would like to thank the anonymous referees for their valuable comments and suggestions.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Rajpurkar, P.; Chen, E.; Banerjee, O.; Topol, E.J. AI in Health and Medicine. *Nat. Med.* **2022**, *28*, 31–38. [[CrossRef](#)] [[PubMed](#)]
2. Ji, J.; Qiu, T.; Chen, B.; Zhang, B.; Lou, H.; Wang, K.; Duan, Y.; He, Z.; Zhou, J.; Zhang, Z.; et al. AI Alignment: A Comprehensive Survey. *arXiv* **2023**, arXiv:2310.19852.
3. Mehergui, M.; Sreedharan, S. Goal Alignment: Re-Analyzing Value Alignment Problems Using Human-Aware AI. *Proc. AAAI Conf. Artif. Intell.* **2024**, *38*, 10110–10118. [[CrossRef](#)]
4. Santoni de Sio, F.; Mecacci, G. Four Responsibility Gaps with Artificial Intelligence: Why They Matter and How to Address Them. *Philos. Technol.* **2021**, *34*, 1057–1084. [[CrossRef](#)]
5. Sanh, V.; Webson, A.; Raffel, C.; Bach, S.H.; Sutawika, L.; Alyafeai, Z.; Chaffin, A.; Stiegler, A.; Scao, T.L.; Raja, A.; et al. Multitask Prompted Training Enables Zero-Shot Task Generalization. *arXiv* **2021**, arXiv:2110.08207.
6. Li, D.; Jiang, B.; Huang, L.; Beigi, A.; Zhao, C.; Tan, Z.; Bhattacharjee, A.; Jiang, Y.; Chen, C.; Wu, T.; et al. From Generation to Judgment: Opportunities and Challenges of LLM-as-a-Judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*; Christodoulopoulos, C., Chakraborty, T., Rose, C., Peng, V., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 2757–2791.
7. Lee, Y.W.; Strong, D.M.; Kahn, B.K.; Wang, R.Y. AIMQ: A Methodology for Information Quality Assessment. *Inf. Manag.* **2002**, *40*, 133–146. [[CrossRef](#)]
8. Kahn, B.K.; Strong, D.M.; Wang, R.Y. Information Quality Benchmarks: Product and Service Performance. *Commun. ACM* **2002**, *45*, 184–192. [[CrossRef](#)]
9. Noël, R.; Taramasco, C.; Márquez, G. Standards, Processes, and Tools Used to Evaluate the Quality of Health Information Systems: Systematic Literature Review. *J. Med. Internet Res.* **2022**, *24*, e26577. [[CrossRef](#)] [[PubMed](#)]
10. Eysenbach, G.; Powell, J.; Kuss, O.; Sa, E.-R. Empirical Studies Assessing the Quality of Health Information for Consumers on the World Wide Web—A Systematic Review. *JAMA* **2002**, *287*, 2691–2700. [[CrossRef](#)] [[PubMed](#)]
11. O’Grady, L. Future Directions for Depicting Credibility in Health Care Web Sites. *Int. J. Med. Inform.* **2006**, *75*, 58–65. [[CrossRef](#)] [[PubMed](#)]
12. Kim, P.; Eng, T.R.; Deering, M.J.; Maxfield, A. Published Criteria for Evaluating Health Related Web Sites: Review. *BMJ* **1999**, *318*, 647–649. [[CrossRef](#)] [[PubMed](#)]
13. Chen, H.; Hailey, D.; Wang, N.; Yu, P. A Review of Data Quality Assessment Methods for Public Health Information Systems. *Int. J. Environ. Res. Public Health* **2014**, *11*, 5170–5207. [[CrossRef](#)] [[PubMed](#)]
14. Liu, R.; Zhang, G.; Feng, X.; Vosoughi, S. Aligning Generative Language Models with Human Values. In *Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2022*; Carpuat, M., de Marneffe, M.-C., Meza Ruiz, I.V., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 241–252.

15. Köpf, A.; Kilcher, Y.; von Rütte, D.; Anagnostidis, S.; Tam, Z.R.; Stevens, K.; Barhoum, A.; Nguyen, D.; Stanley, O.; Nagyfi, R.; et al. OpenAssistant Conversations—Democratizing Large Language Model Alignment. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 47669–47681.
16. Ehsan, U.; Watkins, E.A.; Wintersberger, P.; Manger, C.; Kim, S.S.Y.; Van Berkel, N.; Riener, A.; Riedl, M.O. Human-Centered Explainable AI (HCXAI): Reloading Explainability in the Era of Large Language Models (LLMs). In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*; ACM: New York, NY, USA, 2024; pp. 1–6.
17. Gu, J.; Jiang, X.; Shi, Z.; Tan, H.; Zhai, X.; Xu, C.; Li, W.; Shen, Y.; Ma, S.; Liu, H.; et al. A Survey on LLM-as-a-Judge. *Innovation* **2026**, *7*, 101253. [[CrossRef](#)] [[PubMed](#)]
18. Shi, J.; Yuan, Z.; Liu, Y.; Huang, Y.; Zhou, P.; Sun, L.; Gong, N.Z. Optimization-Based Prompt Injection Attack to LLM-as-a-Judge. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*; Association for Computing Machinery: New York, NY, USA, 2024; pp. 660–674.
19. Quelle, D.; Bovet, A. The Perils and Promises of Fact-Checking with Large Language Models. *Front. Artif. Intell.* **2024**, *7*, 1341697. [[CrossRef](#)] [[PubMed](#)]
20. Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 46595–46623.
21. Wang, P.; Li, L.; Chen, L.; Cai, Z.; Zhu, D.; Lin, B.; Cao, Y.; Kong, L.; Liu, Q.; Liu, T.; et al. Large Language Models Are Not Fair Evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Ku, L.-W., Martins, A., Srikumar, V., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 9440–9450.
22. Shi, L.; Ma, C.; Liang, W.; Diao, X.; Ma, W.; Vosoughi, S. Judging the Judges: A Systematic Study of Position Bias in LLM-as-a-Judge. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 292–314.
23. Thakur, A.S.; Choudhary, K.; Ramayapally, V.S.; Vaidyanathan, S.; Hupkes, D. Judging the Judges: Evaluating Alignment and Vulnerabilities in LLMs-as-Judges. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*; Arviv, O., Clinciu, M., Dhole, K., Dror, R., Gehrmann, S., Habba, E., Itzhak, I., Mille, S., Perlit, Y., Santus, E., et al., Eds.; Virtual Meeting; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 404–430.
24. Fanourakis, N.; Efthymiou, V.; Christophides, V.; Kotzinos, D.; Pitoura, E.; Stefanidis, K. Structural Bias in Knowledge Graphs for the Entity Alignment Task. In *Proceedings of the Semantic Web*; Pesquita, C., Jimenez-Ruiz, E., McCusker, J., Faria, D., Dragoni, M., Dimou, A., Troncy, R., Hertling, S., Eds.; Springer Nature: Cham, Switzerland, 2023; pp. 72–90.
25. Zhou, H.; Wan, X.; Liu, Y.; Collier, N.; Vulić, I.; Korhonen, A. Fairer Preferences Elicit Improved Human-Aligned Large Language Model Judgments. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*; Al-Onaizan, Y., Bansal, M., Chen, Y.-N., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 1241–1252.
26. Nguyen, L.-T.; Dang, T.-Q.; Duc, D.T.V. The Dark Sides of AI Advertising: The Integration of Cognitive Appraisal Theory and Information Quality Theory. *Soc. Sci. Comput. Rev.* **2025**, *43*, 397–424.
27. Li, Y.; Ouyang, H.; Lin, G.; Peng, Y.; Yao, J.; Chen, Y. Evaluation of the Measurement Properties of Online Health Information Quality Assessment Tools: A Systematic Review. *Int. J. Nurs. Sci.* **2025**, *12*, 130–136. [[CrossRef](#)] [[PubMed](#)]
28. Guerra-García, C.; Juárez-Ramírez, R.; Jiménez, S.; Menéndez-Domínguez, V. Data Profiling for Data Quality Assessment: A First Oncoming Methodological Proposal. *Program. Comput. Softw.* **2025**, *51*, 716–726. [[CrossRef](#)]
29. Wang, R.Y.; Strong, D.M. Beyond Accuracy: What Data Quality Means to Data Consumers. *J. Manag. Inf. Syst.* **1996**, *12*, 5–33. [[CrossRef](#)]
30. Zhang, C.; Mohamad, E.; Azlan, A.A.; Wu, A.; Ma, Y.; Qi, Y. Social Media and eHealth Literacy Among Older Adults: Systematic Literature Review. *J. Med. Internet Res.* **2025**, *27*, e66058. [[CrossRef](#)] [[PubMed](#)]
31. Okuhara, T.; Furukawa, E.; Okada, H.; Kiuchi, T. Readability of Online and Offline Written Health Information: A Protocol of a Systematic Review of Systematic Reviews. *BMJ Open* **2024**, *14*, e079756. [[CrossRef](#)] [[PubMed](#)]
32. Ghalavand, H.; Shirshahi, S.; Rahimi, A.; Zarrinabadi, Z.; Amani, F. Common Data Quality Elements for Health Information Systems: A Systematic Review. *BMC Med. Inform. Decis. Mak.* **2024**, *24*, 243. [[CrossRef](#)] [[PubMed](#)]
33. Bhatt, C.; Lin, E.; Ferreira-Legere, L.E.; Jackevicius, C.A.; Ko, D.T.; Lee, D.S.; Schade, K.; Johnston, S.; Anderson, T.J.; Udell, J.A. Evaluating Readability, Understandability, and Actionability of Online Printable Patient Education Materials for Cholesterol Management: A Systematic Review. *J. Am. Heart Assoc.* **2024**, *13*, e030140. [[CrossRef](#)] [[PubMed](#)]
34. Daraz, L.; Dogu, C.; Houde, V.; Bouseh, S.; Morshed, K.G. Assessing Credibility: Quality Criteria for Patients, Caregivers, and the Public in Online Health Information-A Qualitative Study. *J. Patient Exp.* **2024**, *11*, 23743735241259440. [[CrossRef](#)] [[PubMed](#)]
35. Ling, E.; de Pieri, D.; Loh, E.; Scott, K.M.; Li, S.C.H.; Medbury, H.J. Evaluation of the Accuracy, Credibility, and Readability of Statin-Related Websites: Cross-Sectional Study. *Interact. J. Med. Res.* **2024**, *13*, e42849. [[CrossRef](#)] [[PubMed](#)]
36. Ghalavand, H.; Nabilahi, A. Exploring Online Health Information Quality Criteria on Social Media: A Mixed Method Approach. *BMC Health Serv. Res.* **2024**, *24*, 1311. [[CrossRef](#)] [[PubMed](#)]

37. Landers, C.; Blasimme, A.; Vayena, E. Sync Fast and Solve Things—Best Practices for Responsible Digital Health. *npj Digit. Med.* **2024**, *7*, 113. [\[CrossRef\]](#) [\[PubMed\]](#)
38. Abbasi, M.; Majdzadeh, R.; Zali, A.; Karimi, A.; Akrami, F. The evolution of public health ethics frameworks: Systematic review of moral values and norms in public health policy. *Med. Health Care Philos.* **2018**, *21*, 387–402. [\[CrossRef\]](#) [\[PubMed\]](#)
39. Mennella, C.; Maniscalco, U.; Pietro, G.D.; Esposito, M. Ethical and Regulatory Challenges of AI Technologies in Healthcare: A Narrative Review. *Heliyon* **2024**, *10*, e26297. [\[CrossRef\]](#) [\[PubMed\]](#)
40. Sheng, H.; Liu, X.; He, H.; Zhao, J.; Kang, J. Analyzing Uncertainty of LLM-as-a-Judge: Interval Evaluations with Conformal Prediction. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*; Christodoulopoulos, C., Chakraborty, T., Rose, C., Peng, V., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 11286–11328.
41. Doostmohammadi, E.; Holmström, O.; Kuhlmann, M. How Reliable Are Automatic Evaluation Methods for Instruction-Tuned LLMs? In *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024*; Al-Onaizan, Y., Bansal, M., Chen, Y.-N., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 6321–6336.
42. Willmott, C.; Matsuura, K. Advantages of the Mean Absolute Error (MAE) over the Root Mean Square Error (RMSE) in Assessing Average Model Performance. *Clim. Res.* **2005**, *30*, 79–82. [\[CrossRef\]](#)
43. Kendall, A.; Gal, Y. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
44. Khasentino, J.; Belyaeva, A.; Liu, X.; Yang, Z.; Furlotte, N.A.; Lee, C.; Schenck, E.; Patel, Y.; Cui, J.; Schneider, L.D.; et al. A Personal Health Large Language Model for Sleep and Fitness Coaching. *Nat. Med.* **2025**, *31*, 3394–3403. [\[CrossRef\]](#) [\[PubMed\]](#)
45. Artstein, R.; Poesio, M. Inter-Coder Agreement for Computational Linguistics. *Comput. Linguist.* **2008**, *34*, 555–596. [\[CrossRef\]](#)
46. Vandenberg, R.J.; Lance, C.E. A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research. *Organ. Res. Methods* **2000**, *3*, 4–70. [\[CrossRef\]](#)
47. Sorich, M.J.; Menz, B.D.; Hopkins, A.M. Quality and Safety of Artificial Intelligence Generated Health Information. *BMJ* **2024**, *384*, q596. [\[CrossRef\]](#) [\[PubMed\]](#)
48. Suarez-Lledo, V.; Alvarez-Galvez, J. Prevalence of Health Misinformation on Social Media: Systematic Review. *J. Med. Internet Res.* **2021**, *23*, e17187. [\[CrossRef\]](#) [\[PubMed\]](#)
49. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 3980–3990.
50. Kamphuis, C.; Lin, A.; Yang, S.; Lin, J.; De Vries, A.P.; Hasibi, F. MMEAD: MS MARCO Entity Annotations and Disambiguations. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*; ACM: New York, NY, USA, 2023; pp. 2817–2825.
51. Cer, D.; Diab, M.; Agirre, E.; Lopez-Gazpio, I.; Specia, L. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2017; pp. 1–14.
52. Lu, X.; Hu, H.-T.; Xiang, T.; Wu, J.-Z.; Liu, P.; Ji, W.-M.; He, G.-H.; Chen, Z.-X.; Wang, W.; Sun, B. Comparative Analysis of Large Language Models in Traditional Chinese Medicine. *SSRN Electron. J.* **2024**, 5050324. [\[CrossRef\]](#)
53. Kwan, W.-C.; Zeng, X.; Wang, Y.; Sun, Y.; Li, L.; Jiang, Y.; Shang, L.; Liu, Q.; Wong, K.-F. M4LE: A Multi-Ability Multi-Range Multi-Task Multi-Domain Long-Context Evaluation Benchmark for Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 15568–15592.
54. Kuhn, L.; Gal, Y.; Farquhar, S. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. *arXiv* **2023**, arXiv:2302.09664.
55. Morgan, G.A.; Gliner, J.A.; Harmon, R.J. Internal Validity. *J. Am. Acad. Child Adolesc. Psychiatry* **2000**, *39*, 529–531. [\[CrossRef\]](#) [\[PubMed\]](#)
56. Oeding, J.F.; Krych, A.J.; Pearle, A.D.; Kelly, B.T.; Kunze, K.N. Medical Imaging Applications Developed Using Artificial Intelligence Demonstrate High Internal Validity yet Are Limited in Scope and Lack External Validation. *Arthrosc. J. Arthrosc. Relat. Surg.* **2025**, *41*, 455–472.
57. Wang, Y.; Huang, J.; Du, L.; Guo, Y.; Liu, Y.; Wang, R. Evaluating Large Language Models as Raters in Large-Scale Writing Assessments: A Psychometric Framework for Reliability and Validity. *Comput. Educ. Artif. Intell.* **2025**, *9*, 100481. [\[CrossRef\]](#)
58. Findeis, A.; Weers, F.; Yin, G.; Ye, K.; Pang, R.; Gunter, T. Can External Validation Tools Improve Annotation Quality for LLM-as-a-Judge? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Che, W., Nabende, J., Shutova, E., Pilehvar, M.T., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 15997–16020.
59. Smith-Crowe, K.; Burke, M.J.; Kouchaki, M.; Signal, S.M. Assessing Interrater Agreement via the Average Deviation Index Given a Variety of Theoretical and Methodological Problems. *Organ. Res. Methods* **2013**, *16*, 127–151. [\[CrossRef\]](#)

60. Shbikat, N.; Bwaliez, O.M. Enhancing Kendall's W Using Genetic Algorithm: A Computational Approach to Inter-Rater Reliability Optimization. *Expert Syst. Appl.* **2025**, *278*, 127320.
61. Seneviratne, H.M.T.W.; Manathunga, S.S. Artificial Intelligence Assisted Automated Short Answer Question Scoring Tool Shows High Correlation with Human Examiner Markings. *BMC Med. Educ.* **2025**, *25*, 1146. [[CrossRef](#)] [[PubMed](#)]
62. Heimdahl, M.P.E.; Choi, Y.; Whalen, M.W. Deviation Analysis: A New Use of Model Checking. *Autom. Softw. Eng.* **2005**, *12*, 321–347. [[CrossRef](#)]
63. Renze, M. The Effect of Sampling Temperature on Problem Solving in Large Language Models. In *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024*; Al-Onaizan, Y., Bansal, M., Chen, Y.-N., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 7346–7356.
64. Vayena, E.; Blasimme, A.; Cohen, I.G. Machine Learning in Medicine: Addressing Ethical Challenges. *PLoS Med.* **2018**, *15*, e1002689. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.