Computers and Artificial Intelligence; Vol.3, No.6; 2026
E-ISSN: 3006-3728 P-ISSN: 3006-371X
DOI: https://doi.org/10.70267/cai.v3n6.2132
Published by: Zeus Press

Construction and Case Study of a Quality Evaluation Model for Intelligent Information Systems in a Generative Artificial Intelligence Environment: The Case of Bilibili

Ningyu Cai*

School of Management, Northeastern University at Qinhuangdao, Qinhuangdao 066004, Hebei, China

**Corresponding author: Ningyu Cai.*

Abstract

After generative artificial intelligence (AI) was embedded in search, content generation, and user interaction, traditional information system quality evaluation faced new boundaries. Based on the DeLone and McLean (D&M) Information Systems Success Model, this study adopts literature research and a single-case study, takes Bilibili as the case, and constructs and applies a six-dimensional evaluation framework based on public materials, observations of platform functions, and publicly available user feedback. The findings show that system quality, information quality, and service quality can still explain basic platform operation, information output, and service support, but they cannot fully cover AI's semantic understanding and task completion, the trustworthiness of generated results, and continuous human-computer interaction. Therefore, this study introduces AI intelligence quality, trustworthiness quality, and interaction quality on the basis of the three traditional dimensions and clarifies the boundary of each dimension. The Bilibili case shows that the framework can identify both basic capabilities and AI-specific quality, thereby helping to avoid conflating technical performance, content quality, intelligent capability, and trust issues. The study provides a structured analytical framework for evaluating the quality of intelligent information systems in a generative AI environment and a reference for platforms to optimize AI functions and governance mechanisms.

Keywords

generative artificial intelligence, intelligent information system, information systems success model, quality evaluation, case study

1. Introduction

Generative artificial intelligence (AI) is entering content production, information processing, and user interaction, transforming information systems from information-processing tools driven by preset rules into intelligent information systems that can understand needs, generate content, and maintain continuous interaction. System evaluation therefore needs to examine basic technical performance and information output quality together with the new capabilities and risks created by AI participation. In particular, in Internet information systems such as video communities and content platforms, generative AI is gradually moving from a back-end algorithmic tool into content production, information processing, and user interaction.

© 2026 The Author(s). Authors retain copyright in their articles.

This study selects Bilibili as a single case. Bilibili combines user-generated content, video search, algorithmic recommendation, community interaction, and creator services, while generative AI is also used for search, video understanding, subtitles, customer service, and creation support. The concentration of multiple types of information activities on the same platform provides a suitable research setting for observing the boundaries between traditional quality dimensions and AI-specific quality.

Existing studies explain more about user acceptance, technology adoption, satisfaction, and continued use of generative AI, but less often systematically answer the following questions from the perspective of quality evaluation: by what criteria should an information system itself be evaluated after AI is embedded; whether the three quality dimensions of the traditional DeLone and McLean (D&M) model are sufficient; and how intelligent capability, trustworthiness risks, and continuous interaction should enter the same framework. Therefore, the issue is no longer only whether users are willing to use AI, but also what kind of information system can be regarded as a high-quality information system in a generative AI environment.

This study focuses on three questions. First, whether system quality, information quality, and service quality remain explanatory in a generative AI environment. Second, why AI intelligence quality, trustworthiness quality, and interaction quality need to become independent dimensions. Third, whether the six-dimensional framework can explain the actual performance of Bilibili's intelligent information system and revise the evaluation model accordingly.

2. Theoretical Foundation and Literature Review

2.1 Information Systems Success Model

DeLone and McLean proposed the Information Systems Success Model in 1992 and revised it in 2003. The revised D&M model uses system quality, information quality, and service quality as antecedent dimensions and examines their relationships with system use, user satisfaction, and net benefits [1-2]. System quality reflects stability, ease of use, and response efficiency; information quality reflects accuracy, completeness, relevance, and timeliness; and service quality reflects service reliability, responsiveness, and support capability. The three dimensions provide a mature framework for evaluating the basic operation, information output, and service assurance of information systems.

Generative AI has not removed the above basic requirements, but it has expanded the object of evaluation. Stable system operation does not mean that AI can correctly understand semantics and complete tasks; accurate and complete information does not cover source verifiability, hallucination control, and risk transparency; and timely service response cannot replace the evaluation of natural-language interaction, contextual continuity, and user control. This indicates that the traditional D&M quality dimensions may still have an explanatory foundation, but their boundaries of applicability in a generative AI environment and whether they need further extension still require discussion through theoretical analysis and actual cases.

2.2 Literature Review

2.2.1 Research on Generative Artificial Intelligence and Intelligent Information Systems

Generative AI has changed the way information systems produce and acquire information. Traditional systems mainly store, retrieve, and display existing information, whereas large language models can perform semantic understanding, content generation, information summarization, and continuous interaction based on user input. Dwivedi et al. indicated that generative conversational AI would affect knowledge production, organizational processes, and digital service models [3]; Chinese studies have also summarized this change as a reshaping of the chain of information generation, acquisition, and use [4-5]. This means that AI has moved from a back-end supporting technology into the information service process.

In information-acquisition scenarios, generative AI extends keyword retrieval to intent expression, result generation, and follow-up correction. Research on users switching from search engines to generative AI found that information quality and perceived interactivity jointly affect users' perceived value [6]. Therefore, the evaluation of intelligent information systems should examine not only output content but also the system's ability to understand needs, organize information, and provide feedback on results.

© 2026 The Author(s). Authors retain copyright in their articles.

Applications of the D&M model in studies of ChatGPT and AI chatbots show that system quality, information quality, and service quality remain important; meanwhile, variables such as dialogue quality and trust have begun to enter the model [7-10]. These studies provide a basis for retaining the traditional quality dimensions and also indicate that intelligent, trustworthy, and interactive attributes in a generative AI environment have not yet been fully covered by the traditional framework.

2.2.2 Research on Information System Quality Evaluation

Research on information system quality has formed a relatively stable three-dimensional logic: system quality answers whether a system can operate stably and conveniently, information quality answers whether output is accurate, complete, relevant, and timely, and service quality answers whether the operator can continuously respond to and support users [1-2]. These indicators remain the basis for evaluating intelligent information systems, but AI capabilities and governance issues should not be incorporated into the traditional dimensions without clear boundaries.

System quality should be limited to the availability, stability, response efficiency, and operational convenience of the system and its AI functions. Semantic understanding, generation, and task execution belong to AI intelligence quality because normal operation of the technical chain does not demonstrate AI's ability to complete tasks. Information quality should evaluate the accuracy, completeness, relevance, and timeliness of specific outputs. Source verifiability, result consistency, hallucination control, and risk warnings belong to trustworthiness quality because content quality differs from the evidentiary conditions required for users to form reasonable trust [8, 11]. Service quality should evaluate service reliability, responsiveness, problem solving, and user support. Natural-language expression, context maintenance, feedback effectiveness, and user control belong to interaction quality. These boundaries retain the explanatory power of the D&M model while avoiding overlap among the concepts of service, intelligence, and interaction.

2.2.3 Research on the Trustworthiness of Generative Artificial Intelligence and Human-Computer Interaction

Generative models produce probabilistic outputs. Even when the expression is fluent, they may generate content that is inconsistent with the input or external facts. Research on hallucinations in natural-language generation indicates that factuality and faithfulness affect tasks such as summarization, question answering, and information retrieval [9, 12]. Therefore, the evaluation of intelligent information systems must examine hallucination control, source traceability, and result consistency in addition to information content quality. Trustworthiness is not a single measure of technical accuracy. Related studies show that model performance, information sources, content explainability, user cognition, and the application environment jointly affect user trust and verification behavior [13-15]. In platform settings, trustworthiness quality also includes AI-content labeling, risk warnings, feedback-based correction, and governance responsibility. Together, these provide the basis on which users judge whether AI results are worthy of trust. Human-computer interaction research emphasizes that AI systems should explain the boundaries of their capabilities, provide appropriate feedback, allow users to correct errors, and respond to user needs during continued use [16]. Natural-language interaction extends the focus of evaluation from the difficulty of interface operation to intent understanding, contextual continuity, feedback effectiveness, and user control [6, 17]. Interaction quality concerns the communication process, whereas AI intelligence quality concerns the ability to understand, generate, and complete tasks; the two should be evaluated separately.

2.2.4 Review of Existing Research

Existing research provides three foundations for this study. The D&M model establishes the basic structure of system, information, and service quality; research on generative AI reveals the new capabilities of semantic understanding, content generation, and task execution; and research on trustworthy AI and human-computer interaction identifies issues involving hallucination, verification, feedback, and user control. Existing research still has two shortcomings. First, many studies use quality dimensions as antecedent variables to explain satisfaction, trust, or continued use, but rarely directly reconstruct the quality evaluation system of the intelligent information system itself. Second, research on intelligence, trustworthiness, and interactivity is relatively dispersed and has not yet been fully incorporated into the same information system framework. More importantly, the boundaries between traditional dimensions and AI-specific dimensions remain unclear. This study adopts the idea of a basic quality layer plus an AI-specific quality layer: it retains system quality,

© 2026 The Author(s). Authors retain copyright in their articles.

information quality, and service quality; adds AI intelligence quality, trustworthiness quality, and interaction quality; and examines the explanatory power of the six-dimensional framework through the single case of Bilibili.

3. Research Design

3.1 Research Methods

3.1.1 Literature Research Method

This study first uses the literature research method to review the D&M Information Systems Success Model and research on generative AI in intelligent information systems, trustworthiness, and human-computer interaction. On this basis, it determines the traditional quality dimensions and their directions of extension, thereby providing a theoretical basis for case coding.

3.1.2 Single-Case Study Method

This study adopts an explanatory single-case study method. This method is suitable for analyzing issues for which theoretical boundaries are still forming and the phenomenon is closely connected to its context. Bilibili includes content production, information retrieval, intelligent distribution, and community interaction, making it possible to observe system, information, service, intelligence, trustworthiness, and interaction issues before and after AI is embedded within a single case.

The research follows the path of theoretical framework formation - case material collection - theory-guided coding - within-dimension analysis - cross-dimensional comparison - model revision. The role of the case is to examine whether the six-dimensional framework can distinguish different quality problems and explain their relationships, rather than to infer the general performance of all intelligent information systems from a single platform.

3.2 Case Selection

This study selects Bilibili mainly for the following three reasons. The first is typicality. Bilibili is a comprehensive video community that combines content production, search, recommendation, viewing, and interaction, making it suitable for observing the operation of an intelligent information system across multiple types of tasks. The second is theoretical relevance. Bilibili's AI search, video understanding, subtitles, customer service, and creation support allow system operation, machine generation, user interaction, and trustworthiness risks to occur together, providing strong theoretical relevance. The third is research fit. The Bilibili case can correspond to the six evaluation dimensions: platform operation corresponds to system quality, video and AI output correspond to information quality, customer service and creator support correspond to service quality, understanding and generation tasks correspond to AI intelligence quality, sources and governance correspond to trustworthiness quality, and natural language and feedback processes correspond to interaction quality.

3.3 Six-Dimensional Evaluation Framework and Boundaries

This study uses system quality, information quality, and service quality in the D&M model as the basic quality layer. In combination with the characteristics of generative AI, including semantic understanding, content generation, hallucination risks, and natural-language interaction, it introduces AI intelligence quality, trustworthiness quality, and interaction quality to form a six-dimensional case analysis framework. Table 1 presents the operational definition, observation indicators, and boundaries of each dimension.

The six dimensions evaluate different aspects of the same intelligent information system. The basic layer answers whether the system is usable, whether its output is of high quality, and whether the platform can provide continuous support. The AI-specific layer answers whether AI is capable of completing intelligent tasks, whether its results have a trustworthy basis, and whether users can effectively control and correct the results through continuous interaction.

© 2026 The Author(s). Authors retain copyright in their articles.

Table 1: Six-Dimensional Evaluation Framework and Dimensional Boundaries for Intelligent Information Systems

Evaluation dimension	Operational definition	Observation indicators and boundaries
System quality	Degree to which the system and its AI functions operate stably, conveniently, and promptly	Stability, ease of use, response efficiency, and function availability; semantic understanding and task completion are not evaluated
Information quality	Degree to which information provided or generated by the system meets task needs	Accuracy, completeness, relevance, and timeliness; source verifiability and hallucination control are not evaluated
Service quality	Degree to which the platform continuously provides support and solves user problems	Reliability, responsiveness, problem solving, user support, and feedback processing; multi-turn interaction is not evaluated
AI intelligence quality	AI's ability to understand information, generate content, and complete tasks	Semantic understanding, content generation, task execution, and error recovery
Trustworthiness quality	Degree to which AI output has a reliable basis and can be reasonably trusted	Hallucination control, source verifiability, result consistency, AI labeling, risk warnings, and platform assurance
Interaction quality	Degree to which users effectively express, revise, and control their needs during continuous communication	Naturalness, continuity, feedback effectiveness, user control, need adaptation, and personalization

To reduce indicator overlap, this study further clarifies three sets of boundaries. The difference between system quality and AI intelligence quality lies in the object of evaluation: the former examines whether the system and its AI functions can operate stably, conveniently, and promptly, whereas the latter examines whether AI can understand information, generate content that meets requirements, and complete tasks. The difference between information quality and trustworthiness quality lies in the evaluation question: the former determines whether output content is accurate, complete, relevant, and timely, whereas the latter determines whether results have a source basis, can be verified, remain consistent, and whether the platform explains and controls risks. The difference between service quality and interaction quality lies in the evaluation process: the former examines the platform's ability to provide support, solve problems, and process feedback, whereas the latter examines the coherence of natural-language communication, feedback effectiveness, user control, and personalized adaptation.

3.4 Data Sources and Analysis Methods

3.4.1 Data Sources

The case materials include three types: public materials, observations of platform functions, and publicly available user feedback. This study collected these materials from July 2024 to May 2025. The three types of materials serve different purposes and form an evidence chain through cross-verification. Public materials mainly include Bilibili's annual reports, environmental, social and governance reports, product pages, official announcements, and official account information, which are used to confirm functions, technological applications, and governance policies. Media reports supplement the timeline and external observations but are not used alone as the basis for key facts. Observations of platform functions cover scenarios such as search, video understanding, subtitles, recommendation, customer service, and creation support. They record function entry points, task types, operating steps, response conditions, output content, and verifiable paths to assess actual performance in each dimension. User feedback comes from publicly accessible comments, community discussions, and product reviews and is used to identify recurring user experiences and problems. User opinions are used only as supporting evidence, and no overall inference about platform quality is drawn from a single comment.

3.4.2 Material Screening and Analysis

Official sources are given priority for factual materials. Evaluative conclusions must be supported by platform observations, official materials, or multiple independent items of user feedback. Third-party tools that cannot be confirmed as Bilibili's own functions, unattributed retellings, and individual experiences that cannot be verified are not included as core evidence.

© 2026 The Author(s). Authors retain copyright in their articles.

This study adopts theory-guided coding. First, an initial coding table is established using the six dimensions and their secondary indicators. Second, materials are coded according to the most direct object of evaluation, and positive performance, limiting conditions, and counterexamples are recorded. Third, the meaning of the same evidence in different dimensions is compared, and the principal code is determined according to the boundary rules to avoid repeatedly interpreting one item of evidence. Because the range of function availability, account permissions, and device versions differs, observations of platform functions use multiple rounds of sampling in multiple scenarios. User feedback is screened, and only content that is directly related to the six-dimensional indicators and either recurs in multiple independent sources or can be corroborated by official materials is included. Individual experiences that cannot be verified are not included as core evidence.

The analysis proceeds in two steps: within-dimension induction and cross-dimensional comparison. Within-dimension induction summarizes Bilibili's performance on each indicator, whereas cross-dimensional comparison identifies the content that the traditional D&M dimensions can explain and the problems that they cannot cover. Finally, the case findings are compared with the theoretical framework, and the necessity and applicable boundaries of the six-dimensional model are revised.

4. Case Analysis of Bilibili's Intelligent Information System

4.1 Overview of Bilibili's Intelligent Information System Applications

With video content at its core, Bilibili connects activities including content consumption, creation, search, recommendation, live streaming, bullet comments, and comments. After AI is embedded, the platform not only distributes existing content but also begins to participate in video understanding, information organization, intelligent question answering, customer service, and creation support. It therefore has the typical characteristics of an intelligent information system.

Public materials show that Bilibili's self-developed Index large language model has been applied in scenarios such as customer service, content moderation, creation tools, and advertising [18]. The names and availability of user-facing functions such as the AI Search Assistant, AI Video Assistant, and AI Subtitles can also be confirmed on Bilibili's official product pages and in its 2024 environmental, social and governance report. Among them, the AI Search Assistant and AI Video Assistant first underwent internal testing and functional iteration on the basis of the bilibili index large language model in the second half of 2023 [19, 20], and Bilibili made major upgrades to them in 2024 [18]. AI Subtitles use the Index large language model to provide real-time translation in nearly 10 languages [18, 21]. These functions transform video, audio, text, and user questions into information that can be searched, summarized, or used interactively, allowing the case to cover both basic system operation and generative AI processing.

Accordingly, this study limits the Bilibili case to AI applications on the platform that are directly related to information acquisition, content understanding, intelligent services, and human-computer interaction. Personalized recommendation and traditional content governance are treated as system context to explain information distribution and platform assurance, rather than bringing all of Bilibili's businesses into the evaluation scope.

4.2 System Quality Analysis

System quality is strictly limited in this study to the stability, ease of use, response efficiency, and availability of the platform and its AI functions and does not involve the content quality of AI output. Bilibili's frequently used functions, such as video, live streaming, and search, form the infrastructure for AI applications. As supporting evidence regarding the underlying infrastructure, Bilibili's financial results for the first quarter of 2025 reported 106.7 million daily active users (DAU) and an average daily time spent per user of 108 minutes [22]. High-concurrency access indicates that the platform's underlying architecture can support massive interaction. However, this data is used only as background evidence of load-bearing capacity and cannot be directly equated with the absolute stability of all AI functions at all times.

From the perspective of operating processes, Bilibili embeds intelligent functions such as recommendation and video understanding into users' existing search and viewing entry points, reducing the cost of switching among multiple tools. In the specific examination, system quality needs to observe three points: first, whether

© 2026 The Author(s). Authors retain copyright in their articles.

function entry points are clear and different devices are covered seamlessly; second, whether pages and videos load smoothly; and third, whether the response time from submitting an AI request to receiving a result is within an acceptable range.

Bilibili already has a solid system foundation that supports multimedia consumption and intelligent interaction, and its AI functions have also been smoothly integrated into its main processes. However, its system quality remains constrained. Function coverage may be affected by differences in app versions, account permissions, and devices, resulting in inconsistent experiences. The response speed of complex AI tasks is constrained not only by network conditions but also by the computational cost of the underlying model. Finally, it should be clarified that system quality focuses only on whether the technology operates successfully and responds quickly. Whether AI semantic understanding is accurate and generated content is compliant is not included in this dimension but should be evaluated under AI intelligence quality and information quality.

4.3 Information Quality Analysis

Information quality mainly examines the accuracy, completeness, relevance, and timeliness of Bilibili's original content, recommendation results, and information processed by AI. Platform content comes from institutions, professional creators, and ordinary users. Diverse sources increase information richness but also cause professional quality and accuracy to vary. Although comments and bullet comments can supplement the background, they are subjective expressions in essence and cannot replace objective verification of original facts. In AI-assisted processing, AI search and video summarization can extract key points and reorganize information from long videos, significantly reducing users' screening costs. Their quality depends on whether the summary is faithful to the original meaning, whether key content is omitted, and whether the answer is relevant. It should be noted that while AI improves efficiency, it also faces the risk of information distortion. It may inherit and amplify errors in the original user-generated content (UGC), or it may lose context or introduce semantic deviations during compression, thereby impairing information accuracy and completeness.

In terms of relevance and timeliness, search and recommendation mechanisms improve information matching, while live streaming and bullet comments ensure real-time updates. However, there is an inherent tension between the two. Continuous recommendations based on past behavior can cause information cocoons and homogenization, thereby reducing diversity; during breaking events, the conflict between timeliness and accuracy often increases the risk of spreading unverified information. Evaluation therefore needs to balance relevance against diversity and timeliness against accuracy. Bilibili's information quality advantages come from its content scale, diverse media, and AI-assisted organization, while its main risks come from open production, homogeneous recommendations, and biased AI summaries. Finally, the boundary should be clarified: information quality focuses on the quality of the content itself. Whether sources can be verified, whether results are consistent, and whether risk warnings are provided are evidentiary conditions for users to establish trust. They are not evaluated again in this dimension but are discussed under trustworthiness quality.

4.4 Service Quality Analysis

Service quality examines whether the platform can provide reliable and timely support to users and creators. Its core content covers service reliability, responsiveness, problem solving, and feedback processing. Bilibili's accounts, submissions, customer service, creator tools, and content governance mechanisms jointly constitute its service system. For ordinary users and creators, services such as viewing history, account rights, submission management, and data viewing need to remain highly available. Because platform businesses are closely connected, an exception in a single key service may interrupt the complete usage chain. Therefore, service reliability should be judged by whether a specific task can be successfully completed as a closed loop. Intelligent customer service and automated question answering shorten response times for routine issues, but complex account disputes and content appeals still depend heavily on human judgment. The focus of this dimension is not the degree to which AI resembles a person, but whether problems can be substantively solved and whether the transfer mechanism between automated and human service is clear and smooth. In response to the growth in content scale and the increasing difficulty of identification brought by generative AI, Bilibili provides support through its help center, appeal channels, and creator support. It also urgently needs to improve its capacity to identify, review, and manage risks in AI-generated content and ensure that processing results are effectively returned to users. Bilibili has established a service system covering consumption, production, and governance, and AI has effectively improved the response efficiency of standardized services. Future

© 2026 The Author(s). Authors retain copyright in their articles.

improvements should focus on human-AI coordination for complex problems, a closed feedback loop, and service accessibility for disadvantaged groups. It should be clarified that personalized dialogue and contextual adaptation are communication processes and are not included in service quality, but are evaluated separately under interaction quality.

4.5 AI Intelligence Quality Analysis

AI intelligence quality is intended to evaluate AI's comprehensive ability to understand information, generate content, and complete tasks. Its core difference from system quality is that system quality concerns whether a function is available and responds quickly, whereas AI intelligence quality concerns whether the result precisely meets the need and avoids misunderstanding and omission. The Bilibili case mainly considers semantic understanding, content generation, and task execution in AI search, video understanding, subtitles, and creation support. In information understanding, AI needs to accurately identify user intent and establish content associations when processing Bilibili's massive amount of multimodal content, including video, audio, subtitles, and comments. This function substantially reduces the costs of keyword screening and complete viewing. However, when processing specialized terms, multiple ambiguities, or contexts that depend heavily on surrounding information, understanding errors can easily be transmitted to subsequent outputs, placing high demands on the system's fault tolerance and correction capabilities. In content generation, video summaries and question-answering functions improve processing efficiency by compressing and reorganizing information. The core of evaluation in this dimension is not whether a result has been generated, but whether the result is consistent with the user's input requirements, faithful to the meaning of the original video, reasonable in its selection of information, and practically usable. In task execution, AI often integrates need understanding, retrieval, analysis, and generation into a continuous operation. For a single search task, AI can significantly reduce the user's operating costs. For tasks such as multi-source comparison or complex reasoning, however, the completeness of task execution, error recovery capability, and consistency between the result and the final objective need to be examined strictly. In summary, the Bilibili case shows that AI has become deeply involved in the information-processing chain, making an independent dimension of AI intelligence quality necessary. This dimension focuses on evaluating AI's competence in completing intelligent tasks. It neither evaluates whether a page is stable nor directly evaluates whether a result can be trusted, thereby keeping the evaluation system clear and internally consistent.

4.6 Trustworthiness Quality Analysis

Bilibili relies heavily on UGC, and creators differ in qualifications and the standardization of their materials. Although creator information, video descriptions, and community feedback provide users with clues for judgment, basic content in breaking or specialized fields may still lack sufficient sources. When AI searches and summarizes open content, uncertainty in the original information can easily be transmitted to the generated result. AI results are expected to remain consistent with the original video and the user's question, but omissions, misunderstandings, or unsupported additions can easily occur when a video is too long, its views are complex, or information is distributed across multiple segments. Therefore, fluent language must never replace a reliability judgment. Bilibili's original videos provide a physical source to which AI summaries can be traced. An ideal verification mechanism should provide clear source links, timestamps, or corresponding clips to help users verify conclusions at low cognitive cost. If an AI output cannot be located in the original basis, its trustworthiness quality remains insufficient even if the content appears relevant. In addition, as generative AI expands the scale of content and increases the realism of synthetic content, platform assurance needs to cover not only traditional moderation of prohibited content but also AI labeling, risk warnings, error correction, and appeal mechanisms. In summary, Bilibili has a foundation for trustworthiness in its original videos and platform governance, but source presentation and correction paths for AI output remain key. Treating trustworthiness quality as a separate dimension can effectively reveal the risk that content appears useful but lacks a sufficient basis and avoids placing all trust issues under information quality.

4.7 Interaction Quality Analysis

Interaction quality evaluates the process of continuous communication between users and AI, covering natural-language expression, feedback effectiveness, contextual continuity, user correction, and need adaptation. It differs from ease of use under system quality: ease of use focuses on whether interface operation

© 2026 The Author(s). Authors retain copyright in their articles.

is convenient, whereas interaction quality examines whether a system can dynamically understand and respond to users' actual intent in multi-turn dialogue. At the stages of intent expression and input, natural-language input lowers the threshold for keyword screening, but the inherent ambiguity of everyday language requires the system to have intent recognition and clarification capabilities. If a system remains at the level of literal matching and cannot process complex limiting conditions, the dialogue may sound natural but the interaction lacks practical effectiveness. At the stages of feedback effectiveness and output, the system not only needs to answer the question directly but also needs to transform Bilibili's complex video information into clear and actionable content. On this basis, continuous multi-turn interaction allows users to ask follow-up questions, narrow the scope, or add conditions. The system must maintain contextual memory and recognize preceding and subsequent connections. If users need to repeat themselves or AI answers contradict each other, interaction continuity is insufficient. Regarding user control and adaptability, when users revise their needs, the system should provide controls such as regenerating content, reporting errors, or returning to the original content. The core of need adaptation is not simple personalized recommendation, but enabling users to clearly perceive, intervene in, and correct AI's information-processing process.

The Bilibili case shows that natural language and multi-turn follow-up questions extend the boundaries of traditional interface interaction, but functional availability does not mean high interaction quality. Establishing interaction quality as an independent dimension makes it possible to specifically evaluate intent understanding, feedback effects, and user control, thereby preventing these process indicators from being dispersed across system quality or service quality.

5. Case Discussion and Revision of the Evaluation Model

5.1 Applicability and Boundaries of the Traditional D&M Quality Dimensions

The Bilibili case shows that system quality, information quality, and service quality in the D&M model remain the foundation for explaining intelligent information systems. They correspond respectively to the load-bearing capacity of the platform's underlying architecture, the usability of UGC and AI output content, and the service loop of human-AI coordination. Generative AI has not overturned these basic requirements.

However, the explanatory boundaries of the traditional dimensions have become prominent in an AI environment. System quality can only ensure that the system operates and responds quickly, but cannot judge whether AI is competent in cognition and reasoning. Information quality focuses on the accuracy and completeness of the content itself, but cannot cover whether AI results have a verifiable chain of evidence. Service quality addresses standardized business support, but cannot characterize intent clarification and user control in dynamic multi-turn communication. Therefore, the D&M model needs to serve as the basic layer and extend upward to cover AI-specific characteristics.

5.2 Independence of AI Intelligence Quality

AI intelligence quality raises the object of evaluation from technical operation to cognitive execution and examines multimodal information understanding, content reorganization and generation, and complex task execution. The Bilibili case shows that system availability does not equal task completion: when long videos are compressed, AI can easily lose context or introduce semantic deviations. Treating this dimension separately can both accurately evaluate AI's competence in completing specific intelligent tasks and avoid conflating insufficient model capability with system lag or content errors.

5.3 Supplementary Roles of Trustworthiness Quality and Interaction Quality

Trustworthiness quality fills the gap between relevant content and trustworthy results. The fluent expression of generative AI and the increased realism of synthetic content multiply the risks faced by open platforms. This dimension independently examines source traceability, hallucination control, and risk warnings in order to provide users with the evidentiary conditions for establishing reasonable trust.

Interaction quality captures the shift from one-time operation to dynamic collaboration. The inherent ambiguity of natural language requires systems to maintain context and support user correction in multi-turn dialogue. This dimension specifically evaluates how AI capabilities are transformed into actual and perceivable

© 2026 The Author(s). Authors retain copyright in their articles.

experiences during specific interaction processes and is strictly separated from responsiveness under service quality.

5.4 Six-Dimensional Evaluation Model and Its Applicable Boundaries

The revised model presents a two-layer architecture consisting of a basic quality layer and an AI-specific quality layer (see Table 2). The basic layer assures system operation, content output, and service support, whereas the AI-specific layer focuses on competence in intelligent tasks, the evidentiary chain of results, and control in human-computer communication. The six dimensions do not form a linear process but are multidimensional slices of the same information system. Table 2 summarizes the revised six-dimensional model.

Table 2: Six-Dimensional Evaluation Model for Intelligent Information Systems in a Generative AI Environment

Primary dimension	Core question	Secondary indicators
System quality	Can the system operate stably, conveniently, and promptly?	Stability, ease of use, response efficiency, and function availability
Information quality	Does the output information meet task needs?	Accuracy, completeness, relevance, and timeliness
Service quality	Can the platform provide continuous support and solve problems?	Reliability, responsiveness, problem solving, user support, and feedback processing
AI intelligence quality	Can AI understand, generate, and complete tasks?	Semantic understanding, content generation, task execution, and error recovery
Trustworthiness quality	Do AI results have a basis and can they be verified?	Hallucination control, source verifiability, result consistency, risk transparency, and platform assurance
Interaction quality	Is human-computer communication continuous, effective, and controllable?	Naturalness, continuity, feedback effectiveness, user control, and need adaptation

This model retains the theoretical foundation of the D&M model while incorporating the core risks and capabilities of the AI era into a unified framework. The single case of Bilibili confirms the conceptual distinguishability of the six dimensions, but the general applicability of the framework still needs to be empirically examined through cross-platform multi-case comparisons and the future development of quantitative scales.

6. Conclusions and Prospects

6.1 Research Conclusions

Based on the D&M Information Systems Success Model, this study constructs a six-dimensional framework through literature research and conducts a single-case analysis of Bilibili. The findings show that system quality, information quality, and service quality remain the basis for evaluating intelligent information systems in a generative AI environment, explaining technical operation, information output, and platform services, respectively. Generative AI further expands the object of evaluation. AI intelligence quality is used to identify capabilities in semantic understanding, content generation, and task execution; trustworthiness quality is used to identify issues involving hallucinations, sources, verification, and governance; and interaction quality is used to identify issues involving natural language, multi-turn feedback, and user control. The three extended dimensions have independent objects of evaluation and can address the limitations of the traditional D&M model. The Bilibili case shows that the six-dimensional framework can separately analyze basic operational issues and AI-specific issues and reveal the connections among the dimensions. The value of the framework lies in providing clear conceptual boundaries and a basis for case coding, rather than proving causal relationships among the dimensions on the basis of a single case.

6.2 Theoretical and Practical Implications

In theory, this study places the traditional quality dimensions and AI-specific quality dimensions in the same framework and clarifies three sets of key boundaries, providing a conceptual basis for subsequent scale construction, weight determination, and cross-case comparison. In practice, platforms should separately

© 2026 The Author(s). Authors retain copyright in their articles.

monitor system performance, output content, AI task completion, result trustworthiness, and interaction processes and should avoid using response speed or the number of functions as substitutes for the evaluation of intelligence and trustworthiness.

For Bilibili, priorities for optimization include providing source location and traceability paths for AI summaries and answers; establishing mechanisms for correcting AI results, appeals, and human review; explaining the applicable scope and uncertainty of AI functions; and improving multi-turn context, user correction, and transfer to human service. These measures correspond to trustworthiness quality, interaction quality, and service quality, respectively, and can form clearer governance responsibilities.

6.3 Research Limitations and Prospects

This study adopts a single case and mainly relies on public materials, platform observations, and publicly available user feedback. It can explain the quality dimensions and their boundaries but cannot represent all platforms or test causal relationships among the dimensions. Differences in accounts, devices, function availability, and product updates may also affect the observations.

Future studies may select different types of intelligent information systems, such as search, education, healthcare, or government services, for multi-case comparison to examine the applicable boundaries of the six-dimensional framework. They may also develop measurement items on this basis and use questionnaires, experiments, or log data to analyze the effects of each dimension on satisfaction, trust, and use outcomes.

References

[1] DeLone W H, McLean E R. Information systems success: The quest for the dependent variable[J]. Information Systems Research, 1992, 3(1): 60-95.

[2] DeLone W H, McLean E R. The DeLone and McLean model of information systems success: A ten-year update[J]. Journal of Management Information Systems, 2003, 19(4): 9-30.

[3] Dwivedi Y K, et al. “So what if ChatGPT wrote it?” Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy[J]. International Journal of Information Management, 2023, 71: 102642.

[4] Chinese Scholar A. Research on the impact of generative artificial intelligence on the information ecosystem[J]. Library and Information Service, 2024, 68(5): 12-20.

[5] Chinese Scholar B. Large language models reshape the paths of information production and acquisition[J]. Information Studies: Theory & Application, 2024, 47(3): 45-52.

[6] Dang V T, Ku B, Lu L. Users’ switching intention from search engines to ChatGPT: The role of information quality and perceived interactivity[J]. Internet Research, 2024.

[7] Mijač M, Tijanić L, Kelić I. Rethinking information system success: Generative AI in the hands of future business managers[J]. e-mentor, 2025, 112(5): 47-56.

[8] Fu X, Li M, Cai S, et al. Assessing information quality of ChatGPT: An information quality framework[C]//International Conference on Information. 2024: 195-209.

[9] Huang L, Yu W, Ma W, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions[J]. ACM Transactions on Information Systems, 2025.

[10] Marjanović O, Cech P, Pulakanam V. ChatGPT as an information system success: A task-technology fit perspective[C]//Proceedings of the 27th Pacific Asia Conference on Information Systems. 2024.

[11] Wang X, Yin L, Guo Z, et al. Trustworthy AI: From principles to practices[J]. ACM Computing Surveys, 2023, 55(9): 1-46.

[12] Ji Z, Lee N, Frieske R, et al. Survey of hallucination in natural language generation[J]. ACM Computing Surveys, 2023, 55(12): 1-38.

[13] Luo X, Tong S, Fang Z, et al. Frontiers: Machines vs. humans: The impact of artificial intelligence chatbot disclosure on customer purchases[J]. Marketing Science, 2019, 38(6): 937-947.

© 2026 The Author(s). Authors retain copyright in their articles.

[14] Siau K, Wang W. Building trust in artificial intelligence, machine learning, and robotics[J]. Cutter Business Technology Journal, 2018, 31(2): 47-53.

[15] Hoff K A, Bashir M. Trust in automation: Integrating empirical evidence on factors that influence trust[J]. Human Factors, 2015, 57(3): 407-434.

[16] Amershi S, Weld D, Vorvoreanu M, et al. Guidelines for human-AI interaction[C]//Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 2019: 1-13.

[17]Gao Y, Pan Z, Wang Y, et al. Human-AI interaction in chatbots: A review of conversational AI qualities[J]. International Journal of Human-Computer Interaction, 2024.

[18] Bilibili Inc. 2024 environmental, social and governance report[R]. 2025.

[19] IT Home. Bilibili internally tests the Search AI Assistant function, supported by the bilibili index large language model[EB/OL]. (2023-07-23).

[20] DoNews. Bilibili tests an AI video summarization function based on its own large language model[EB/OL]. (2023-10-27).

[21] Rui Chen. Bilibili's self-developed large language model Index has been launched and applied to the AI subtitle function[EB/OL]. 2024 China International Intelligent Communication Forum, 2024-09-26.

[22] Bilibili Inc. Bilibili Inc. announces first quarter 2025 financial results[EB/OL]. (2025-05-20).

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Copyrights

Copyright for this article is retained by the author (s), with first publication rights granted to the journal. This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/).

© 2026 The Author(s). Authors retain copyright in their articles.