Construction and Empirical Study of a Quality Evaluation Model for AI-Driven Multimodal Language Services

Yongjun Zhang
[1]Henan College of Transportation, Zhengzhou 451460, Henan, China
Corresponding author: Yongjun Zhang (e-mail: zhangyj050305@163.com)

ABSTRACT The language service industry in China is experiencing increasingly vigorous development and facing unprecedented opportunities for growth, driven in part by the rising need for accurate translation and multimodal processing of technical specifications, simulation reports, and global communication materials in advanced engineering fields such as electromagnetic waves, antennas, and propagation. This paper aims to construct an evaluation model for assessing the quality of AI-driven multimodal language services. First, through a comprehensive review of existing literature, this study identifies the key dimensions influencing multimodal language service quality and selects appropriate quality evaluation methods. Data are collected through questionnaires, and the Fuzzy Comprehensive Evaluation method is used to conduct an empirical assessment of the platform's service quality, in order to verify the feasibility and effectiveness of the constructed evaluation model. The research results show that the evaluation model can objectively and accurately reflect the quality level of AI-driven multimodal language services, with accuracy and user satisfaction being the core factors affecting service quality. This study can provide theoretical references and practical guidance for multimodal language service providers to optimize service quality and enhance user experience, particularly benefiting service scenarios that involve specialized terminology, technical diagrams, and cross-language communication requirements in electromagnetic engineering applications.

INDEX TERMS artificial intelligence, multimodal language services, quality evaluation model, fuzzy comprehensive evaluation, electromagnetic engineering

I. INTRODUCTION

A. RESEARCH BACKGROUND

LANGUAGE is the most important carrier of information and the most crucial communication tool for human beings, serving as the primary means to articulate complex technical specifications and artistic concepts vital. As a key branch of the linguistics system, language services have undergone interdisciplinary evolution for over two decades, with their connotative boundaries continuously expanding and theoretical frameworks becoming increasingly refined. The academic concept of "language services" was first proposed by Chinese scholar Qu Shaobing in September 2005 at the "International Forum on Language Environment Construction for the World Expo" held in Shanghai [1]. After more than a decade of theoretical and practical development, it has now formed relatively rich connotations and extensions, and spawned a series of closely related derivative concepts such as "language service products", "language service capabilities", "language service market", and "language service industry" [2].

Against the backdrop of the "Belt and Road" Initiative, China's language service industry is experiencing increasingly vigorous development and facing unprecedented opportunities for growth, a trend significantly fueled by the expansion of international collaboration, technical communication, and cross-border engineering service requirements. Driven by the advancement of artificial intelligence (AI) and information technology, the language service industry has undergone rapid and continuous changes. Language service is a type of service activity aimed at helping people overcome language barriers in cross-linguistic information exchange. It achieves this by directly providing language information conversion services and products, or by offering the technologies, tools, knowledge, and skills required for language information conversion [3]. In the AI era, the "language service" industry is developing rapidly while also undergoing a series of transformations and challenges brought about by information technology [4-7].

In the context of big data, the creation and use of corpora have gained tremendous momentum. The development of corpora has also gone through four stages: Corpus 1.0, Corpus 2.0, Corpus 3.0, and Corpus 4.0 [8]. Currently, the multimodal parallel corpus, which aligns with the core characteristics of Corpus 4.0 (integrating scale expansion, real-time processing, AI integration, and multimodal fusion), has emerged, opening a new chapter in the language service industry. Multimodal parallel corpora are no longer limited to the collection and processing of pure text and audio; instead, they integrate text, images, audio, video, and other forms, featuring multiple functions such as visualization, audibility, and readability, with greater emphasis on user experience. According to the definition given by Foster et al. (2007), a Multimodal Corpus refers to a corpus that focuses not only on the collection of English-Chinese parallel texts but also on the collection of linguistic texts and audio-visual materials closely associated with the texts [9].

AI technology encompasses machine learning, natural language processing, computer vision, knowledge representation and reasoning, and other aspects [10], providing strong technical support for the development of multimodal language services. As the application scope of language services becomes increasingly extensive, the evaluation of their service quality has become a key focus of current research. As a comprehensive discipline, quality evaluation involves various aspects of technology and management.

Different evaluation scopes and indicators determine the differences in research methods and model establishment for quality evaluation. The Analytic Hierarchy Process (AHP) was proposed by the renowned American operations research expert Saaty in the 1970s. It is a multi-objective decision-making analysis method that combines qualitative and quantitative approaches [11]. The core of this method is to quantify the decision-maker's empirical judgments, thereby providing decision-makers with a quantitative basis for decision-making [12]. Currently, this method has been widely applied to the analysis of complex problems for which there is no unified measurement standard, solving decision-making analysis problems that are difficult to address using pure parametric mathematical models [13]. Fuzzy analysis is a prediction and evaluation method based on the collection of fuzzy sets. Its characteristic lies in that its evaluation method is very close to people's normal thinking mode, describing objects using degree-based language [14].

This paper proposes a quality evaluation model based on AHP and the Fuzzy Comprehensive Evaluation method to assess the quality of AI-driven multimodal language services, aiming to provide a clear direction for quality improvement of multimodal language services.

B. RESEARCH STATUS

Li Yuming (2020) pointed out that language service is not only a discipline but also a social practice activity and a force for promoting social progress. Those engaged in language services must attach importance to the practical nature of this discipline. Against the background of major global public health emergencies, language has played a prominent role in doctor-patient communication, international rescue, emergency assistance, the dissemination of epidemic information, and the spread of medical knowledge [15].

Yuan Jun (2014) and Li Yuming (2016) defined the concept of "language services" [3,16]. Wang Huashu and Zhang Chengzhi (2015) studied the establishment of translation project management systems in the language service industry [17]. The "Report on the Development of China's Language Service Industry (2017)" examined the development status of the industry from both macro and micro perspectives [18]. Wang Lifei et al. (2016) conducted an investigation and analysis of the language service needs of Chinese enterprises going global [19]. Shao Zhangminzi (2017) studied the development status and market demand of language services in foreign cultural trade [20]. Mu Lei (2017) researched the characteristics of translators' language competence in the international language service industry [21]. Si Xianzhu et al. (2018) used big data to capture recruitment information from major domestic recruitment websites (such as Zhaopin.com) and analyzed the market demand for language services [22].

The Analytic Hierarchy Process (AHP) was proposed by operations researchers such as L.T. Saaty from the University of Pittsburgh in the mid-1970s. It is a multi-attribute decision-making analysis method that combines qualitative and quantitative approaches, referred to as AHP for short. It is suitable for decision-making problems with complex decision objective structures and insufficient data [23,24]. AHP hierarchizes and quantifies human thinking processes and uses mathematical methods to provide a quantitative basis for analytical decision-making, forecasting, or control. The basic idea of AHP is to decompose complex problems into several levels and conduct step-by-step analysis at levels that are much simpler than the original system.

According to the basic principle of AHP, the evaluation index system is divided into several levels, requiring multi-level Fuzzy Comprehensive Evaluation [25]. At present, AHP and the Fuzzy Comprehensive Evaluation method have been widely applied in various evaluation models, and the models constructed based on these methods all have good practicality [26-28]. Based on this, this paper intends to adopt these two methods to construct a quality evaluation model for assessing the quality of multimodal language services.

C. RESEARCH CONTENT AND METHODS

1) Research Content

The research content of this paper mainly includes the following aspects:

(1) Identification of factors influencing the quality of AI-driven multimodal language services.

(2) Construction of a quality evaluation index system for multimodal language services.

(3) Construction of a quality evaluation model for multimodal language services.

(4) Verification of the feasibility and effectiveness of the model, and proposal of quality improvement suggestions.

2) Research Methods

(1) Literature Research Method.

By systematically collecting, organizing, interpreting, and mining the content of literature, this method aims to extract valuable information, patterns, or viewpoints from textual information, and clarify the theoretical basis and research direction of the study.

(2) Questionnaire Survey Method.

Design a questionnaire for users of multimodal language services to collect users' evaluation data on various aspects of service quality.

(3) Analytic Hierarchy Process (AHP).

Decompose the problem of evaluating the quality of multimodal language services into different levels, including the target level, criterion level, and indicator level, and calculate the weight of each indicator.

(4) Fuzzy Comprehensive Evaluation Method.

Convert qualitative evaluation into quantitative evaluation, and calculate the comprehensive score of service quality by establishing a fuzzy evaluation matrix and combining indicator weights.

3) Technical Route

(1) Define the research topic based on the development needs of the language service industry and the advanced engineering sectors.

(2) Review relevant literature to clarify the theoretical basis, research status, and key influencing factors of multimodal language service quality evaluation.

(3) Construct a three-level evaluation index system (Target Level - Criterion Level - Indicator Level) following scientific, systematic, operable, and practical principles.

(4) Determine the weight of each indicator using the Analytic Hierarchy Process (AHP) through expert scoring and consistency checking.

(5) Construct a quality evaluation model using the Fuzzy Comprehensive Evaluation method, including determining evaluation grades, establishing membership functions, and forming a fuzzy evaluation matrix.

(6) Conduct model application and result verification through multi-channel data collection and preprocessing, and calculate comprehensive evaluation results.

(7) Summarize research conclusions, analyze advantages and shortcomings of the evaluated platform, and propose quality improvement suggestions.

II. CONSTRUCTION OF A QUALITY EVALUATION INDEX SYSTEM FOR AI-DRIVEN MULTIMODAL LANGUAGE SERVICES

A. PRINCIPLES FOR INDEX CONSTRUCTION

To ensure the scientificity and practicality of the evaluation index system, this paper follows the four principles below:

1) Scientific Principle

Indicators must reflect the essential characteristics of multimodal language services and conform to the application rules of AI technology and the theory of service quality evaluation.

2) Systematic Principle

The index system must cover all dimensions affecting service quality and form an organic whole with a clear hierarchy and logical structure.

3) Operability Principle

Indicators must be measurable. Both quantitative and qualitative indicators require clear data sources and evaluation standards.

4) Practicality Principle

The index system must balance theoretical depth and practical application, making it easy for service providers, users, and regulatory authorities to understand and use.

B. FRAMEWORK OF THE INDEX SYSTEM

Based on the above principles and the results of factor identification, this paper constructs a three-level evaluation index system consisting of a "Target Level - Criterion Level - Indicator Level". The specific framework is as follows:

Target Level (A): Comprehensive Evaluation of AI-Driven Multimodal Language Service Quality; Criterion Level (B): Accuracy (B1), Timeliness (B2), Interactivity (B3), Multimodal Fusion Degree (B4), User Satisfaction (B5), Security (B6); Indicator Level (C): Each criterion level is subdivided into 3-4 specific indicators, with clear indicator definitions to form a complete index system (see Table 1).

TABLE 1. Quality Evaluation Index System for AI-Driven Multimodal Language Services

Target Level (A)	Criterion Level (B)	Indicator Level (C)	Indicator Definition	Measurement Standards
Comprehensive Evaluation of AI-Driven Multimodal Language Service Quality	B1 Accuracy	C11 Text Translation Accuracy Rate	Consistency between the translation result and the original text meaning under the text modality	Excellent: Error rate $\leq$ 3%; Good: 3% $<$ Error rate $\leq$ 5%; Qualified: 5% $<$ Error rate $\leq$ 8%; Unqualified: Error rate $>$ 8%
		C12 Speech Recognition Accuracy Rate	Matching degree between the recognition result and the original speech content under the speech modality	Excellent: Accuracy rate $\geq$ 95%; Good: 90% $\leq$ Accuracy rate $<$ 95%; Qualified: 85% $\leq$ Accuracy rate $<$ 90%; Unqualified: Accuracy rate $<$ 85%

Table 1 continued

Target Level (A)	Criterion Level (B)	Indicator Level (C)	Indicator Definition	Measurement Standards
		C13 Image Text Recognition Accuracy Rate	Consistency between the extracted text content and the actual text in the image under the image modality	Excellent: Accuracy rate $\geq$ 92%; Good: 88% $\leq$ Accuracy rate $<$ 92%; Qualified: 83% $\leq$ Accuracy rate $<$ 88%; Unqualified: Accuracy rate $<$ 83%
	B2 Timeliness	C21 Response Delay	Time interval from when a user initiates a request to when the service provides the first feedback	Excellent: $\leq$ 0.5s; Good: 0.5s $<$ Delay $\leq$ 1s; Qualified: 1s $<$ Delay $\leq$ 2s; Unqualified: Delay $>$ 2s
		C22 Multimodal Processing Time	Total time required to complete the fusion processing of multiple modal information	Excellent: $\leq$ 3s; Good: 3s $<$ Time $\leq$ 5s; Qualified: 5s $<$ Time $\leq$ 8s; Unqualified: Time $>$ 8s
		C23 Peak Response Speed	Average response time when the service concurrency reaches its peak	Excellent: $\leq$ 1.5s; Good: 1.5s $<$ Speed $\leq$ 2.5s; Qualified: 2.5s $<$ Speed $\leq$ 4s; Unqualified: Speed $>$ 4s
	B3 Interactivity	C31 Diversity of Interaction Methods	Types of interaction modalities supported by the service (text/speech/image/video)	Excellent: Supports 4 or more types; Good: Supports 3 types; Qualified: Supports 2 types; Unqualified: Supports 1 type
		C32 Interaction Fluency	Coherence between user operations and service feedback, with no freezes/interruptions	Excellent: No freezes; Good: Occasional freezes ($\leq$ 1 time/10 minutes); Qualified: Slight freezes ($\leq$ 3 times/10 minutes); Unqualified: Frequent freezes
		C33 Intent Recognition Accuracy Rate	Accuracy of the service in understanding the user's interaction intent	Excellent: Accuracy rate $\geq$ 93%; Good: 88% $\leq$ Accuracy rate $<$ 93%; Qualified: 83% $\leq$ Accuracy rate $<$ 88%; Unqualified: Accuracy rate $<$ 83%
	B4 Multimodal Fusion Degree	C41 Consistency of Modal Information	Matching degree of information transmitted by different modalities in terms of semantics and logic	Excellent: Completely consistent; Good: Basically consistent (deviations $\leq$ 2 points); Qualified: Partially consistent (deviations $\leq$ 5 points); Unqualified: Deviations $>$ 5 points
		C42 Completeness of Fusion Results	Whether the core information of each modality is fully covered after multimodal fusion	Excellent: 100% coverage; Good: 95% $\leq$ Coverage $<$ 100%; Qualified: 90% $\leq$ Coverage $<$ 95%; Unqualified: Coverage $<$ 90%
		C43 Naturalness of Cross-Modal Conversion	Naturalness of conversion between modalities (e.g., text-to-speech, image-to-text)	Excellent: Natural with no sense of incongruity; Good: Basically natural; Qualified: Slight sense of incongruity; Unqualified: Severe sense of incongruity
	B5 User Satisfaction	C51 Overall Satisfaction Score	User's overall evaluation of service quality	Excellent: Score $\geq$ 4.5 (on a 5-point scale); Good: 4.0 $\leq$ Score $<$ 4.5; Qualified: 3.5 $\leq$ Score $<$ 4.0; Unqualified: Score $<$ 3.5
		C52 Problem Resolution Rate	Proportion of user needs successfully met by the service	Excellent: $\geq$ 95%; Good: 90% $\leq$ Rate $<$ 95%; Qualified: 85% $\leq$ Rate $<$ 90%; Unqualified: Rate $<$ 85%
		C53 Repurchase/Reuse Intention	User's willingness to continue using the service in the future	Excellent: Willingness rate $\geq$ 90%; Good: 85% $\leq$ Willingness rate $<$ 90%; Qualified: 80% $\leq$ Willingness rate $<$ 85%; Unqualified: Willingness rate $<$ 80%
	B6 Security	C61 Data Encryption Strength	Encryption level for transmission and storage of user information (text/speech/image)	Excellent: Adopts AES-256 + RSA-2048; Good: Adopts AES-128 + RSA-1024; Qualified: Only adopts AES-128; Unqualified: No encryption
		C62 Privacy Leakage Rate	Probability of unauthorized access to or leakage of user information	Excellent: 0; Good: $\leq$ 0.01%; Qualified: $\leq$ 0.05%; Unqualified: $>$ 0.05%
		C63 Abnormal Access Interception Rate	Service's ability to intercept unauthorized access and malicious attacks	Excellent: $\geq$ 98%; Good: 95% $\leq$ Rate $<$ 98%; Qualified: 90% $\leq$ Rate $<$ 95%; Unqualified: Rate $<$ 90%

III. CONSTRUCTION OF A QUALITY EVALUATION MODEL FOR AI-DRIVEN MULTIMODAL LANGUAGE SERVICES

A. DETERMINATION OF INDICATOR WEIGHTS BASED ON AHP

1) Construction of Judgment Matrix

In accordance with the principles of the Analytic Hierarchy Process (AHP), 10 experts were invited to score the relative importance of indicators at the Criterion Level (B1-B6) and Indicator Level (C11-C63). The background of the experts is as follows: 3 experts have more than 15 years of research experience, focusing on international trade and technical documentation translation; 4 experts are engaged in AI multimodal language service research with over 10 years of experience; 3 experts have practical experience in language service platform operation and quality management. All experts hold senior professional titles or doctoral degrees, ensuring the authority and industry relevance of the scoring results. This study was approved by the Ethics Committee of Henan College of Transportation, and was performed in accordance with the principles of the Declaration of Helsinki. All eligible participants signed an informed consent form.

The 1-9 scaling method was adopted to form the judgment matrix. Taking the judgment matrix of the Criterion Level as an example (see Table 2):

TABLE 2. Judgment Matrix and Consistency Check of the Criterion Level (B)

Criterion Level	B1	B2	B3	B4	B5	B6	Weight (W)
B1 Accuracy	1	3	4	2	2	3	0.285
B2 Timeliness	1/3	1	2	1/2	1/2	1	0.121
B3 Interactivity	1/4	1/2	1	1/3	1/3	1/2	0.078
B4 Multimodal Fusion Degree	1/2	2	3	1	1	2	0.186
B5 User Satisfaction	1/2	2	3	1	1	2	0.186
B6 Security	1/3	1	2	1/2	1/2	1	0.144

Consistency Check: $\lambda_{\max} = 6.238$, CI=0.048, CR=0.038 (CR<0.1, passed the consistency check).

2) Calculation of Indicator Weights

The weights of indicators at each level were calculated using the eigenvalue method, and a consistency check was conducted (a result of CR<0.1 indicates passing the check). Finally, the weights of indicators at the Indicator Level were obtained (see Table 3):

TABLE 3. Weights and Consistency Check Results of the Indicator Level (C)

Criterion Level (B)	Indicator Level (C)	Weight (W)	Criterion Level Weight (B)	Combined Weight (B×C)	Consistency Check (CR)
B1 (0.285)	C11	0.42	0.285	0.1197	0.042 (<0.1)
	C12	0.35	0.285	0.0998	
	C13	0.23	0.285	0.0656	
B2 (0.121)	C21	0.45	0.121	0.0545	0.039 (<0.1)
	C22	0.35	0.121	0.0424	
	C23	0.20	0.121	0.0242	
B3 (0.078)	C31	0.30	0.078	0.0234	0.045 (<0.1)
	C32	0.45	0.078	0.0351	
	C33	0.25	0.078	0.0195	
B4 (0.186)	C41	0.48	0.186	0.0893	0.037 (<0.1)
	C42	0.32	0.186	0.0595	
	C43	0.20	0.186	0.0372	
B5 (0.186)	C51	0.50	0.186	0.0930	0.041 (<0.1)
	C52	0.30	0.186	0.0558	
	C53	0.20	0.186	0.0372	
B6 (0.144)	C61	0.45	0.144	0.0648	0.036 (<0.1)
	C62	0.35	0.144	0.0504	
	C63	0.20	0.144	0.0288	

B. CONSTRUCTION OF THE EVALUATION MODEL BASED ON THE FUZZY COMPREHENSIVE EVALUATION METHOD

1) Determination of Evaluation Grades and Membership Functions

The service quality evaluation grades were divided into 4 levels: Excellent (V1, 90-100 points), Good (V2, 80- 89 points), Qualified (V3, 70-79 points), and Unqualified (V4, <70 points).

A trapezoidal membership function was used to calculate the membership degree of each indicator to the evaluation grades. The threshold values of the membership functions are determined based on the following basis: (1) Referring to industry standards such as GB/T 19363.1-2003 "Translation Service Specifications" and relevant technical indicators of mainstream multimodal language service platforms; (2) Conducting a pre-survey of 50 professionals and language service field to collect their opinions on reasonable threshold ranges; (3) Verifying through statistical analysis of historical service data of the empirical platform (one year of operation data). Taking "Text Translation Accuracy Rate (C11)" as an example, the membership functions are as follows:

- Excellent (V1): $\mu_1(x) = 1$, $x \leq 3\%$; $\mu_1(x) = (5\% - x)/(5\% - 3\%)$, $3\% < x \leq 5\%$; $\mu_1(x) = 0$, $x > 5\%$.
- Good (V2): $\mu_2(x) = (x - 3\%)/(5\% - 3\%)$, $3\% < x \leq 5\%$; $\mu_2(x) = (8\% - x)/(8\% - 5\%)$, $5\% < x \leq 8\%$; $\mu_2(x) = 0$, $x \leq 3\%$ or $x > 8\%$.
- Qualified (V3): $\mu_3(x) = (x - 5\%)/(8\% - 5\%)$, $5\% <$

$x \leq 8\%$; $\mu_3(x) = 0$, $x \leq 5\%$ or $x > 8\%$.

- Unqualified (V4): $\mu_4(x) = 1$, $x > 8\%$; $\mu_4(x) = 0$, $x \leq 8\%$.

2) Establishment of the Fuzzy Evaluation Matrix

Based on the measured indicator data of the empirical object, the membership degree of each indicator to the evaluation grades was calculated to form the fuzzy evaluation matrix R. For example, if the membership degree corresponding to the measured value of an indicator is [0.6, 0.3, 0.1, 0], it means 60% of the indicator belongs to "Excellent", 30% to "Good", 10% to "Qualified", and 0% to "Unqualified".

3) Calculation of Fuzzy Comprehensive Evaluation

The synthetic operation of "weight vector × fuzzy evaluation matrix" (adopting the "weighted average type" operator M(·,⊕)) was used to calculate the comprehensive evaluation results of the Criterion Level and Target Level. The formulas are as follows:

- Comprehensive evaluation of the Criterion Level: $B = W \times R = [\sum(W_i \times R_{ij})]$ $(i = 1, 2, \ldots, n;\ j = 1, 2, 3, 4)$
- Comprehensive evaluation of the Target Level: $A = W_B \times B = [\sum(W_{Bk} \times B_{kj})]$ $(k = 1, 2, \ldots, 6;\ j = 1, 2, 3, 4)$

Finally, the service quality grade was determined using the "maximum membership degree principle" or "weighted scoring method". To address the non-linear characteristics of user satisfaction, the weighted scoring method was optimized by introducing a non-linear correction factor. The corrected formula for the comprehensive score is:

$$\begin{aligned} &\text{Comprehensive Score} \\ &= B_1 \times 95 \times 1.05 + B_2 \times 85 \times 1.00 \\ &+ B_3 \times 75 \times 0.95 + B_4 \times 65 \times 0.90. \end{aligned}$$

(Note: The correction factors are determined based on the non-linear correlation analysis between user satisfaction scores and objective indicators, with higher weights assigned to higher grades to reflect the marginal utility of high-quality services.)

IV. EMPIRICAL STUDY

A. SELECTION OF EMPIRICAL OBJECT

This study selects the "Zhilian Multimodal Language Service Platform" launched by a leading domestic AI enterprise as the empirical object. This platform integrates three core modalities: text translation, voice interaction, and image text recognition. Its technical architecture and service scale align with the typical characteristics of AI-driven multimodal language services, making it of high research value.

B. DATA COLLECTION AND PROCESSING

1) Data Collection Scheme

In line with the data source requirements of the evaluation index system, a "multi-channel collaborative collection" strategy was adopted, with a period from October 1 to December 30, 2024 (90 days in total). The specific scheme is as follows:

(1) Platform Log Data:

Quantitative indicator data were extracted from the platform's technical background, including text translation error rate, speech recognition accuracy rate, response delay, and data encryption level. A total of 1,200,000 valid log records were collected.

(2) Experimental Test Data:

A 15-person test team (expanded from the original 5-person team) was formed, including 5 professionals with technical backgrounds in advanced engineering fields. Standardized test samples were designed, covering 10 typical application scenarios in engineering and technical service contexts (such as technical specification translation, product manual conversion, etc.), to simulate real user scenarios to initiate tasks, and record indicators such as multimodal processing time and image text recognition accuracy rate. A total of 3,600 sets of test data were obtained.

(3) User Questionnaire Survey:

An online stratified sampling method was used to distribute questionnaires to users in the platform's three major service fields. A total of 1,500 questionnaires were distributed, and 1,386 valid questionnaires were recovered (recovery rate of 92.4%). The Cronbach's α coefficient for questionnaire reliability test was 0.893, indicating good data reliability.

(4) Expert Evaluation Data:

Eight experts were invited to form an evaluation team to score indicators related to multimodal fusion degree.

2) Data Preprocessing

The following preprocessing operations were performed on the collected raw data:

(1) Outlier Handling: The "3σ principle" was used to eliminate extreme outliers (a total of 286 data points were removed, accounting for 0.02%).

(2) Missing Value Imputation: For a small number of missing items in user questionnaires (missing rate < 3%), the "domain mean imputation method" was used for supplementation.

(3) Data Standardization: Indicator data with different dimensions were converted into standardized values in the [0, 1] range to ensure comparability between indicators.

C. IMPLEMENTATION OF THE EVALUATION PROCESS

1) Calculation of Measured Indicator Values

Based on the preprocessed data, the measured values of 20 indicators at the Indicator Level were calculated.

The measured results of some core indicators are shown in Table 4:

TABLE 4. Measured Values and Evaluation Grades of Core Platform Indicators

Criterion Level	Indicator Level	Measured Value	Evaluation Grade	Membership Vector
B1	C11	96.8% (Error Rate: 3.2%)	Good	[0.4, 0.6, 0, 0]
	C12	93.5%	Good	[0.2, 0.8, 0, 0]
	C13	91.2%	Good	[0.3, 0.7, 0, 0]
B2	C21	0.8s	Good	[0, 0.6, 0.4, 0]
	C22	4.2s	Good	[0, 0.4, 0.6, 0]
	C23	2.1s	Qualified	[0, 0, 0.8, 0.2]
B3	C31	Text/Voice/Image	Good	[0, 1.0, 0, 0]
	C32	Occasional freezes (0.8 times/10 minutes)	Good	[0.1, 0.9, 0, 0]
	C33	90.3%	Good	[0, 0.7, 0.3, 0]
B4	C41	1.2 deviations	Excellent	[0.8, 0.2, 0, 0]
	C42	97.5%	Excellent	[0.9, 0.1, 0, 0]
	C43	Basically natural	Good	[0, 0.8, 0.2, 0]
B5	C51	4.2 points	Good	[0.3, 0.7, 0, 0]
	C52	92.8%	Good	[0.2, 0.8, 0, 0]
	C53	87.5%	Good	[0, 0.9, 0.1, 0]
B6	C61	AES-128 + RSA-1024	Good	[0, 1.0, 0, 0]
	C62	0.008%	Excellent	[0.9, 0.1, 0, 0]
	C63	96.5%	Good	[0, 0.7, 0.3, 0]

2) Fuzzy Comprehensive Evaluation of the Criterion Level

Based on the indicator weights in Table 3 and the membership vectors in Table 4, the "weighted average type" operator was used to calculate the comprehensive evaluation results of the Criterion Level (Table 5).

Taking "B1 Accuracy" as an example, its indicator weight vector is [0.42, 0.35, 0.23], and the membership matrix is as follows:

$$R_{B1} = \begin{bmatrix} 0.4 & 0.6 & 0 & 0 \\ 0.2 & 0.8 & 0 & 0 \\ 0.3 & 0.7 & 0 & 0 \end{bmatrix} \tag{1}$$

Then, the comprehensive evaluation vector of B1 is:

$$\begin{aligned} \mathbf{B}_1 &= [0.42 \times 0.4 + 0.35 \times 0.2 + 0.23 \times 0.3, \\ &\quad 0.42 \times 0.6 + 0.35 \times 0.8 + 0.23 \times 0.7,\ 0,\ 0] \\ &= [0.319,\ 0.681,\ 0,\ 0] \end{aligned} \tag{2}$$

TABLE 5. Fuzzy Comprehensive Evaluation Results of the Criterion Level

Criterion Level	Weight (W_B)	Comprehensive Evaluation Vector ([Excellent, Good, Qualified, Unqualified])	Weighted Score (Points)	Evaluation Grade
B1	0.285	[0.319, 0.681, 0, 0]	88.19	Good
B2	0.121	[0.000, 0.380, 0.540, 0.080]	79.52	Qualified
B3	0.078	[0.037, 0.853, 0.110, 0.000]	84.27	Good
B4	0.186	[0.710, 0.290, 0.000, 0.000]	92.90	Excellent
B5	0.186	[0.175, 0.785, 0.040, 0.000]	86.35	Good
B6	0.144	[0.640, 0.290, 0.070, 0.000]	90.70	Excellent

3) Comprehensive Evaluation of the Target Level

Taking the comprehensive evaluation vectors of the Criterion Level as input and combining them with the Criterion Level weight vector [0.285, 0.121, 0.078, 0.186, 0.186, 0.144], the comprehensive evaluation result of the Target Level was calculated:

$$\begin{aligned} \mathbf{A} &= [0.285 \times 0.319 + 0.121 \times 0.000 + 0.078 \times 0.037 \\ &\quad + 0.186 \times 0.710 + 0.186 \times 0.175 \\ &\quad + 0.144 \times 0.640,\ \ldots] \\ &= [0.368,\ 0.523,\ 0.101,\ 0.008] \end{aligned} \tag{3}$$

The comprehensive score was calculated using the weighted scoring method:

$$\begin{aligned} \text{Comprehensive Score} &= 0.368 \times 95 \\ &\quad + 0.523 \times 85 + 0.101 \times 75 \\ &\quad + 0.008 \times 65 = 87.26 \end{aligned} \tag{4}$$

According to the evaluation grade division standard (80-89 points = Good), the comprehensive evaluation grade of the multimodal language service quality of the Zhilian Platform is Good.

4) Analysis of Evaluation Results

(1) Advantageous Dimensions:

The Multimodal Fusion Degree (B4, 92.90 points) and Security (B6, 90.70 points) performed excellently, indicating that the Zhilian Platform has core competitiveness in the collaborative processing of multimodal information and user privacy protection.

(2) Shortcoming Dimension:

Timeliness (B2, 79.52 points) was the only dimension rated as Qualified, with the main issues focusing on Peak Response Speed (C23, 2.1s) and Multimodal Processing Time (C22, 4.2s). Analysis of the causes shows that when the platform's concurrency exceeds the peak of 300,000 times per hour, there is a delay in server resource scheduling.

(3) Verification of Core Influencing Factors:

The weight ratio of Accuracy (B1, weight 0.285) and User Satisfaction (B5, weight 0.186) reaches 47.1%, which effectively indicates that Accuracy and User Satisfaction are the core factors affecting service quality. This conclusion is supported by the expanded sample size and industry-specific expert scoring, enhancing its reliability and generalizability.

V. RESEARCH CONCLUSIONS

A. CONSTRUCTION OF A MULTIDIMENSIONAL EVALUATION INDEX SYSTEM

Six Criterion Level dimensions—Accuracy, Timeliness, Interactivity, Multimodal Fusion Degree, User Satisfaction, and Security—and 20 Indicator Level indicators were accurately identified, covering three aspects: technology, service, and users.

B. ESTABLISHMENT OF A SCIENTIFIC EVALUATION MODEL

By integrating the Analytic Hierarchy Process (AHP) and the Fuzzy Comprehensive Evaluation method, the model realizes the objective assignment of indicator weights, addresses the fuzziness of service quality, and achieves the transformation from qualitative description to quantitative evaluation. The introduction of non-linear correction factors in the scoring method further improves the scientificity of the model, offering a robust framework for assessing the multifaceted quality factors inherent in manufacturing and digital representation services.

C. EMPIRICAL VERIFICATION OF MODEL EFFECTIVENESS

Taking the Intelligent Language Service Platform as the empirical object, through multi-channel data collection (expanded sample size and extended data collection period) and comprehensive evaluation, it was concluded that its service quality grade is "Good" (87.26 points), and Timeliness was identified as the main shortcoming.

D. IDENTIFICATION OF CORE INFLUENCING FACTORS

The study found that Accuracy (weight 0.285) and User Satisfaction (weight 0.186) are the core factors affecting the quality of multimodal language services.

AUTHOR CONTRIBUTIONS

Yongjun Zhang designed, collected and analyzed the data, and drafted the manuscript. Yongjun Zhang conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Yongjun Zhang participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research received no external funding.

HUMAN RESEARCH SUBJECTS

This study was approved by the Ethics Committee of Henan College of Transportation, and was performed in accordance with the principles of the Declaration of Helsinki. All eligible participants signed an informed consent form.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] J. Holmqvist, V. Y. Van, and C. Grönroos, "Language use in services: Recent advances and directions for future research," Journal of Business Research, vol. (72), pp. 114-118, 2017, doi: 10.1016/j.jbusres.2016.10.005.
[2] S. B. Qu, "Observations and Reflections on Language Service Issues in The Report on the State of China's Language Life," Journal of Yunnan Normal University (Philosophy and Social Sciences Edition), vol. 42, no. 5, pp. 6, 2010, doi: 10.3969/j.issn.1000-5110.2010.05.004.
[3] N. Markl, "Defining language and managing its use: Language technology as language management," Language & Communication, vol. 105, pp. 245-258, 2025, doi: 10.1016/j.langcom.2025.10.004.
[4] W. Q. Wang, "Research on the Application Status and Prospects of Artificial Intelligence in the Language Service Industry," Southern Forum, vol. (5), pp. 2, 2018, doi: 10.3969/j.issn.1004-1133.2018.05.007.
[5] Y. Z. Liu, "A Brief Analysis of the Application and Development of Artificial Intelligence in the Language Service Industry from the Emergence of AI Translation," China Strategic Emerging Industries, vol. (04), pp. 153-154, 2018, doi: 10.19474/j.cnki.10-1156/f.003242.
[6] N. Manić, "Translation as a form of language activism: an ecolinguistic perspective," Language Sciences, vol. (113), Art. no. 101770, 2026, doi: 10.1016/j.langsci.2025.101770.
[7] Q. L. Cui, "Research on the Application of Artificial Intelligence in Language Service Enterprises," Foreign Language and Literature (Journal of Sichuan International Studies University), vol. 37, no. 1, pp. 26-32,73, 2021, doi: 10.3969/j.issn.1674-6414.2021.01.004.
[8] Y. Zou, W. T. Zheng, and Z. Y. Huang, "Construction of a Teaching and Training Platform Based on the "Translation Workshop" of Multimodal Sports Corpora," Journal of Central China Normal University (Natural Science Edition), vol. 52, no. 6, pp. 9, 2018, doi: 10.19603/j.cnki.1000-1190.2018.06.016.
[9] M. E. Foster and J. Oberlander, "Corpus-based Generation of Head and Eyebrow Motion for an Embodied Conversational Agent," Language Resources & Evaluation, vol. 41, no. 3-4, pp. 305-323, 2007, doi: 10.1007/s10579-007-9055-3.
[10] Z. X. Cai and G. H. Xu, "Artificial Intelligence and Its Applications (2nd ed.)," Beijing, China: Tsinghua University; 1996.
[11] S. M. Liang, "Improvement of Saaty's Analytic Hierarchy Process," College Mathematics, vol. (2), pp. 140-141, 1995.
[12] Q. Liu, "Identifying and correcting the defects of the Saaty analytic hierarchy/network process: A comparative study of the Saaty analytic hierarchy/network process and the Markov chain-based analytic network process," Operations Research Perspectives, vol. (9), Art. no. 100244, 2022, doi: 10.1016/j.orp.2022.100244.

[13] Y. Xu, “Construction of Marketing Management Performance Evaluation Model Based on Analytic Hierarchy Process,” International Journal of Information System Modeling and Design, vol. 16, no. 1, pp. 27, 2025, doi: 10.4018/IJISMD.386136.
[14] C. Y. Wu and M. Ma, “Fundamentals of Fuzzy Analysis,” Beijing, China: Defense Industry Press; 1991.
[15] C. Wenham and L. Stout, “A legal mapping of 48 WHO member states’ inclusion of public health emergency of international concern, pandemic, and health emergency terminology within national emergency legislation in responding to health emergencies,” The Lancet, vol. 403, no. 10435, pp. 1504-1512, 2024, doi: 10.1016/S0140-6736(24)00156-9.
[16] S. Wang and Y. Shao, “Integrative innovation of large language models in industries: technologies, applications, and challenges,” Data Science and Management, 2025, doi: 10.1016/j.dsm.2025.06.005.
[17] T. Jia, “Design and Application System of Machine Translation Language Model Based on Cloud LM Algorithm,” Procedia Computer Science, vol. (259), pp. 897-905, 2025, doi: 10.1016/j.procs.2025.04.042.
[18] Y. F. Pan, “Language Service Capability Reflects "Soft Power",” Overseas Chinese Education Trends, vol. (10), pp. 1, 2017.
[19] L. F. Wang, Q. L. Cui, and Y. Y. Meng, “Blue Book on Language Services for Chinese Enterprises "Going Global",” Beijing, China: University of International Business and Economics Press; 2016.
[20] Z. M. Z. Shao, H. Q. Sun, and M. J. Chai, “Analysis of the Current Situation of Language Services from the Perspective of Foreign Cultural Trade,” Chinese Translators Journal, vol. 38, no. 3, pp. 7, 2017.
[21] L. Mu, H. Z. Shen, and B. Zou, “Research on the Competence Characteristics of Translators for the International Language Service Industry—Based on a Survey of the Top 100 Global Language Service Providers,” Shanghai Journal of Translators, vol. (1), pp. 9, 2017.
[22] X. Z. Si and Y. Z. Yao, “Research on China’s Translation Industry: From the Perspective of Industrial Economics,” Chinese Translators Journal, vol. 35, no. 5, pp. 5, 2014.
[23] D. L. Olson, “The Analytic Hierarchy Process,” Decision Aids for Selection Problems. NY, New York, USA: Springer; 1996. p.49-68, doi: 10.1007/978-1-4612-3982-6_5.
[24] L. F. Wang and S. B. Xu, “Introduction to the Analytic Hierarchy Process,” 1990.
[25] X. X. Chen, J. Qiu, and G. J. Liu, “Research on the Selection Method of Test Equipment Based on Analytic Hierarchy Process and Fuzzy Comprehensive Evaluation,” Acta Armamentarii, vol. (1), pp. 6, 2010.
[26] X. Y. Yang, “Application of Analytic Hierarchy Process and Fuzzy Comprehensive Evaluation in Engineering Bidding Evaluation,” Science & Technology Progress and Policy, vol. 19, no. 12, pp. 3, 2002, doi: 10.3969/j.issn.1001-7348.2002.12.053.
[27] D. L. Wang, “Application of Analytic Hierarchy Process and Fuzzy Comprehensive Evaluation in University Website Evaluation,” Journal of Shandong Agricultural University (Natural Science Edition), vol. 36, no. 2, pp. 4, 2005, doi: 10.3969/j.issn.1000-2324.2005.02.030.
[28] J. L. Ding and J. H. Ye, “Application of Analytic Hierarchy Process and Fuzzy Comprehensive Evaluation in Evaluating Teachers’ Classroom Teaching Quality,” Wuhan University Journal (Philosophy and Social Sciences Edition), vol. 56, no. 2, pp. 5, 2003, doi: 10.3969/j.issn.1672-7320.2003.02.019.