A Standardized Framework for Evaluating Language Quality of Chinese Large Language Models

Aijia Zhong[1] and Lei Li[2,*]

[1]*School of International College of Languages and Cultures, Yunnan University of Finance and Economics, Kunming 650000, Yunnan, China*
[2]*Jiangxi institute of Applied Science and Technology, Nanchang 330001, Jiangxi, China*
E-mail: 15887006637@163.com; 17870569828@163.com
[*]*Corresponding Author*

Received 07 April 2026; Accepted 24 May 2026

Abstract

Chinese large language models (CLLMs) are rapidly transitioning from research to deployed infrastructure across multiple sectors. Yet current evaluation practice remains fragmented and benchmark-centric, conflating language quality with general capability and weakening comparability across models. This is especially critical in standards-oriented contexts, where language quality must be treated as a multidimensional construct encompassing linguistic correctness, semantic adequacy, discourse coherence, style appropriateness, factual grounding, safety compliance, and robustness. In this study, we propose a standardized framework for evaluating CLLM language quality as a pre-standardization, which integrates five core elements, specifically a layered quality model, scenario-driven test specification, multi-source evidence collection (automated metrics, expert review, calibrated LLM-as-a-judge), transparent scoring and conformity schemes, and governance structures aligned with AI standards practice. Unlike conventional

Journal of ICT Standardization, Vol. 14_3, 421–460.
doi: 10.13052/jicts2245-800X.1436

approaches, the framework explicitly separates language quality from broader capability, introduces a hierarchical indicator system tailored to Chinese linguistic phenomena, and defines a reproducible evaluation workflow with quality assurance and version control. To operationalize the proposal, we also develop a reference architecture for quality dimensions, indicators, scoring logic, reporting structure, and pre-standard deliverables. It further demonstrates how the framework can support both research benchmarking and practical deployment scenarios, including enterprise acceptance testing and sector-specific profile extension. The resulting framework provides a technically grounded and standards-oriented blueprint for CLLM language-quality evaluation, with potential value as both a de facto industrial evaluation specification and a foundation for future formal standardization in the information and communication technology (ICT) and artificial intelligence (AI) quality ecosystem.

Keywords: Chinese large language model, language quality evaluation, standardization, benchmark, LLM-as-a-judge, AI quality model, conformity assessment, pre-standardization.

1 Introduction

Large language models (LLMs) have rapidly become a core infrastructure layer for search, office productivity, customer service, education, software engineering, and intelligent content generation. Their rise follows the broader transition from task-specific models to foundation models trained at scale and adapted across downstream scenarios [1–3]. As these systems move from laboratory demonstrations to real deployment, evaluation has become a central technical and governance problem rather than a purely academic exercise. Benchmark scores now influence model selection, procurement, product safety decisions, and public claims about capability and quality [4, 5]. Existing LLM evaluation practice has expanded quickly, but it remains fragmented. General benchmarks such as massive multitask language understanding (MMLU) and Beyond the Imitation Game benchmark (BIG-bench) measure broad knowledge and reasoning, while newer conversational and preference-oriented benchmarks such as MT-Bench, a multi-turn benchmark for evaluating chat-oriented model responses, and AlpacaEval, an automated instruction-following evaluation benchmark based on model preference judgments, emphasize interactive assistant behavior [6–9]. At the same time, recent surveys have shown that LLM evaluation pipelines often vary in data

construction, prompting settings, decoding controls, judging procedures, and aggregation methods, which undermines reproducibility and makes cross-study comparison difficult [5, 10, 11]. For Chinese large language models (CLLMs), the challenge is even sharper. Chinese-language evaluation has grown substantially in the last two years, with influential resources such as C-Eval, CMMLU, and AGIEval providing rigorous measurement of knowledge and reasoning under Chinese or bilingual settings [12–14]. More recent work has extended this landscape to Chinese factuality and hallucination assessment, including Chinese SimpleQA, HalluQA, UHGEval, and Chinese SafetyQA [15–18]. These resources have materially improved the empirical basis for comparing CLLMs, yet most were designed as benchmarks for capability comparison, not as standardized frameworks for language-quality assurance.

This distinction matters because language quality is not equivalent to general capability. A model may perform strongly on exam-style multiple-choice questions while still failing on discourse coherence, register control, terminology consistency, idiomatic naturalness, punctuation norms, pragmatic appropriateness, factual precision in long-form generation, or robustness across prompt variants. Prior work on natural language generation has long shown that surface-overlap metrics such as BLEU and ROUGE are insufficient proxies for human judgments of text quality, while learned and LLM-based evaluators, although often stronger, still introduce their own biases and stability concerns [19–26]. The problem is particularly important for Chinese because Chinese language quality involves phenomena that are often underrepresented in generic multilingual benchmarks: script variation, punctuation conventions, idiom and chengyu use, classical and modern register interaction, culturally grounded implicature, domain-sensitive honorific or institutional language, and code-switching across Chinese-English technical discourse. A useful evaluation framework therefore needs to go beyond correctness on fixed-answer tasks and capture how well a model produces acceptable, context-appropriate, and trustworthy Chinese in realistic deployment scenarios [4, 12–18]. Recent work also suggests that evaluation itself should be treated as a design problem. The holistic evaluation of language models (HELM) argues for broad scenario-and-metric coverage rather than single-score reporting [4]; GEM and Dynabench emphasize the importance of benchmark evolution, documentation, and robustness to distributional drift [10, 27]; and judge-benchmarking studies, together with other work on LLM-as-a-judge, show that automatic judging can be valuable but must be validated carefully against human criteria and bias controls [8, 9, 11, 19, 26, 28, 29].

These findings point toward the need for evaluation of artifacts that are transparent, modular, and interoperable across institutions.

Motivated by this gap, this paper studies a standardized framework for evaluating the language quality of CLLMs. Instead of proposing yet another benchmark, it focuses on the intermediate layer between isolated benchmark design and formal standardization: terminology, quality dimensions, test-unit design, evidence channels, scoring logic, reporting templates, and conformance-oriented workflows. The contributions of the paper are fourfold. First, it defines CLLM language quality as a structured and multi-dimensional construct that is related to, but distinct from, general capability. Second, it synthesizes benchmark research, factuality assessment, human evaluation, and LLM-based judging into a unified evaluation architecture. Third, it proposes a repeatable workflow for corpus construction, scoring, human review, automatic judging, and reporting. Fourth, it frames the result as a pre-standardization artifact that can support future industry profiles, consortium agreements, and eventually formal standard-setting efforts.

In this paper, the term "pre-standardization framework" refers to an intermediate technical artifact that is more structured than an individual benchmark but not yet a formal normative standard. It provides common terminology, quality dimensions, evidence categories, scoring principles, and reporting expectations that can be reused across evaluation settings. Unless otherwise stated, the framework elements are intended as recommended design components rather than mandatory certification requirements. Mandatory requirements may be introduced later through specific evaluation profiles, sector-specific implementation rules, or formal standard-setting processes. To further ensure consistent terminology, this paper distinguishes among the framework, the quality model, scenario profiles, and evaluation profiles. The framework refers to the overall pre-standardization architecture that connects evaluation objectives, quality dimensions, evidence channels, scoring logic, reporting artifacts, and conformity-oriented workflows. The quality model refers more specifically to the hierarchical structure of language-quality dimensions, sub-dimensions, and measurable indicators. A scenario profile refers to a task- or deployment-oriented evaluation setting, such as formal document drafting, factual question answering, customer-service dialogue, educational explanation, or constrained creative writing. An evaluation profile refers to the complete operational configuration used in a particular assessment, including the selected scenario profile, activated quality dimensions, indicator weights, evidence channels, gating conditions, quality-assurance procedures, and reporting requirements. This distinction

helps maintain consistency between the conceptual framework, the measurable quality structure, and the practical evaluation configuration used in implementation.

2 Related Work

2.1 General LLM Benchmarking Foundations

The modern LLM evaluation literature began from broad capability measurement. Foundational papers such as GPT-3, MMLU, and BIG-bench helped establish the now-standard practice of assessing large models across many tasks, domains, and levels of difficulty [1, 6, 7]. HELM subsequently argued that this approach remained incomplete because benchmarking had become too dependent on isolated datasets and single leaderboards; it proposed a scenario-based evaluation philosophy in which models are assessed across multiple use contexts and multiple desiderata, including accuracy, calibration, robustness, fairness, efficiency, and toxicity [4]. This shift from "one benchmark, one score" toward broader coverage is directly relevant to any standardization effort for LLM language quality.

At the same time, benchmark research has repeatedly documented the limits of static test sets. Dynabench showed that benchmark saturation can hide brittle failure modes and that more dynamic human-and-model-in-the-loop evaluation is needed [10]. GEM made a similar point for natural language generation by linking datasets, metrics, human evaluation, and documentation practices in a living benchmark ecosystem [27]. Recent surveys of LLM evaluation likewise stress that reproducibility depends not only on the benchmark itself but also on prompt templates, decoding settings, judge instructions, aggregation rules, and release transparency [5]. These studies provide an important methodological foundation for treating evaluation as a system rather than as a single number.

2.2 Chinese-centric Benchmark Development

Chinese-specific benchmark construction has progressed rapidly. C-Eval introduced a multi-level, multi-discipline Chinese evaluation suite spanning 52 subjects and four difficulty levels, becoming a landmark resource for Chinese knowledge and reasoning evaluation [12]. CMMLU expanded this line of work with a broader Chinese multitask benchmark covering natural sciences, social sciences, engineering, and the humanities [13]. AGIEval complemented both by using standardized human examinations, including

Chinese and English tests, thereby connecting model evaluation more directly to human-centered assessment settings [14]. Together, these benchmarks established the empirical backbone of current CLLM comparison.

However, benchmark expansion in Chinese has also revealed a conceptual gap between capability measurement and language-quality evaluation. Many influential Chinese benchmarks still emphasize objective correctness, exam performance, or short-answer knowledge tasks [12–14]. By contrast, practical CLLM deployment often depends on qualities that are harder to reduce to answer accuracy alone, such as fluency under domain constraints, style stability, discourse planning, linguistic naturalness, and pragmatic appropriateness. This gap is increasingly visible in Chinese factuality and hallucination research.

Several recent datasets focus directly on reliability in Chinese generation. HalluQA evaluates hallucinations in CLLM question answering through adversarially designed prompts [16]. UHGEval moves closer to real deployment by benchmarking hallucination under unconstrained Chinese generation rather than tightly constrained answer formats [17]. Chinese SimpleQA provides a carefully curated factual short-answer benchmark for CLLMs [15], while Chinese SafetyQA extends short-form factuality toward safety-sensitive content [18]. These works demonstrate that Chinese evaluation is moving beyond exam-style knowledge tests, but they still remain largely benchmark-centric and do not yet provide a general, standardized language-quality framework.

2.3 Language Quality Metrics, Factuality, and Reliability

A second major thread of related work comes from natural language generation metrics and factuality assessment. Classic overlap-based metrics such as BLEU and ROUGE were influential because they are easy to compute and compare, but their limitations for open-ended generation are well known, especially when multiple valid outputs exist or when discourse-level quality matters more than lexical overlap [22, 23]. Embedding-based and learned metrics such as BERTScore and BLEURT improved correlation with human judgments in many settings [24, 25], yet they still inherit assumptions from reference-based evaluation and are not sufficient on their own for assessing long-form Chinese assistant outputs.

More recent factuality-oriented work is especially relevant to language quality because high-quality language is not useful if it is not trustworthy. TruthfulQA exposed the extent to which language models can reproduce

common misconceptions rather than truthful answers [30]. HaluEval and FActScore contributed two complementary directions: the former provides a large-scale hallucination benchmark, while the latter offers a fine-grained atomic scoring approach for long-form factual precision [20, 21]. Surveys on hallucination and factuality have since synthesized a broad set of failure modes, metrics, and mitigation strategies, reinforcing that reliability must be treated as a first-class evaluation dimension rather than a side constraint [31, 32]. For a Chinese language-quality framework, these studies imply that fluency, coherence, and style cannot be separated from factual grounding and evidential reliability.

2.4 Judge-based Evaluation and Research Gap

Another closely related line of research concerns LLM-based evaluation itself. Chiang and Lee showed early that strong LLMs can approximate expert human judgments in some text-evaluation settings [28]. G-Eval demonstrated that carefully structured prompts and form-filling rubrics can improve agreement with humans for summarization and dialogue evaluation [19]. Zheng et al. pushed this further with MT-Bench and Chatbot Arena, showing that strong LLM judges can align well with human preference data at scale, but also identifying biases such as position bias, verbosity bias, and self-enhancement bias [8]. Later work, including AlpacaEval, large-scale empirical studies of LLM judges, and more recent meta-evaluations, confirmed both the value and the fragility of judge-based pipelines [9, 11, 26].

For a standards-oriented framework, the lesson is clear: LLM-as-a-judge is useful, but it cannot be treated as a self-authenticating oracle. Instead, it should be one evidence channel within a controlled evaluation design that includes rubric specification, human anchoring, bias analysis, disagreement tracking, and versioned reporting. This point is especially important in Chinese evaluation, where small prompt wording changes, script choices, and cultural framing may influence both candidate outputs and judge behavior.

The literature therefore leaves three unresolved problems. First, existing Chinese benchmarks provide strong task-level evidence, but they do not define a shared terminology or hierarchical quality model for CLLM language quality [12–18]. Second, existing metric research offers many useful tools, but no consensus architecture for combining reference-based metrics, factuality checks, human review, and LLM-based judging into a single reproducible workflow [11, 19–26]. Third, benchmark papers and surveys increasingly call for more transparent and reproducible evaluation practice,

yet most current studies still optimize for model comparison rather than for standardizable reporting and conformance assessment [4, 5, 10, 27].

Accordingly, the research gap is not merely the absence of another benchmark. The missing contribution is an implementable pre-standard framework that translates scattered benchmark evidence into a coherent evaluation specification for CLLM language quality. The remainder of this paper addresses that gap by proposing such a framework.

3 Limitations of Current Practice

Current CLLM evaluation practice shows some recurring limitations. First, language quality and task capability are frequently conflated. Multiple-choice accuracy, benchmark rank, or exam success is often treated as evidence of overall language quality, even though these signals do not directly measure discourse organization, style control, rhetorical appropriateness, or terminology stability. Second, evaluation dimensions are incomplete. Many industrial evaluations focus on fluency and factuality but ignore critical properties such as instruction adherence in Chinese, formal-vs-colloquial register control, structural consistency in long documents, robustness to prompt paraphrase, and the interaction between safety alignment and naturalness. Third, Chinese-specific linguistic phenomena are underrepresented. Common evaluation sets do not sufficiently cover punctuation conventions, idioms and chengyu, sentence-final particles, abbreviation expansion, mixed Chinese-English terminology, official document style, classical references, or regionally sensitive lexical usage. These issues can lead to concrete quality failures in practical Chinese-language deployment. For example, a model may generate semantically correct content but use punctuation patterns that are inappropriate for formal Chinese writing, mix simplified and traditional Chinese forms without contextual justification, or introduce colloquial particles into an institutional notice where a formal register is expected. In technical or professional settings, the model may translate the same Chinese-English term inconsistently across a document, weakening terminology stability and user trust. In customer-service dialogue, a response may be factually adequate but pragmatically unsuitable if its tone is too abrupt, overly casual, or inconsistent with expected politeness norms. These examples show why CLLM language-quality evaluation must assess not only answer correctness, but also script consistency, punctuation conventions, register control, terminology stability, and pragmatic appropriateness. Fourth, open-ended evaluation lacks procedural standardization. Teams differ in prompts, reviewer qualifications,

number of raters, dispute resolution, and scoring normalization. As a result, results are difficult to reproduce or compare. Fifth, evidence sources are often improperly weighted. Purely automatic metrics can miss pragmatic and discourse failures; purely human review is costly and inconsistent; purely LLM-based judging can inherit model-specific biases. Without a principled multi-evidence design, evaluation outcomes remain fragile. Lastly, reporting is weak. Many reports provide only top-level averages with no scenario breakdown, confidence intervals, error taxonomy, or metadata on model version, prompt template, decoding settings, and test-set provenance. This prevents auditability and reduces usefulness for regulators, buyers, integrators, and standards bodies. These limitations motivate a framework that is dimension-complete, Chinese-aware, reproducible, and explicitly aligned with standardization logic.

4 Proposed Standardized Framework

4.1 Overall Architecture and Functional Logic of the Standardized Framework

This section describes the proposed pre-standardization framework and its embedded quality model. The framework provides the overall evaluation architecture, while the quality model defines the hierarchical relationship among quality dimensions, sub-dimensions, indicators, evidence channels, and reporting outputs. Figure 1 presents this structure as a five-layer architecture that organizes CLLM language-quality evaluation from abstract quality objectives to operational outputs.

Figure 1 presents the proposed standardized framework as a five-layer architecture that organizes CLLM language-quality evaluation from abstract quality objectives to operational outputs. The framework is designed as a top-down and traceable chain: it begins by defining what should be evaluated, then specifies how relevant evidence is elicited and interpreted, and finally translates evaluation results into standardized reporting artifacts. This structure is intended to support not only academic comparison, but also pre-standardization work and future conformity-oriented assessment.

At the first layer, the framework defines the core quality dimensions that constitute language quality in principle, including correctness, adequacy, coherence, style, factuality, safety, robustness, and interaction quality. These dimensions serve as the normative foundation of the framework. Their role is to establish a stable conceptual vocabulary for discussing language

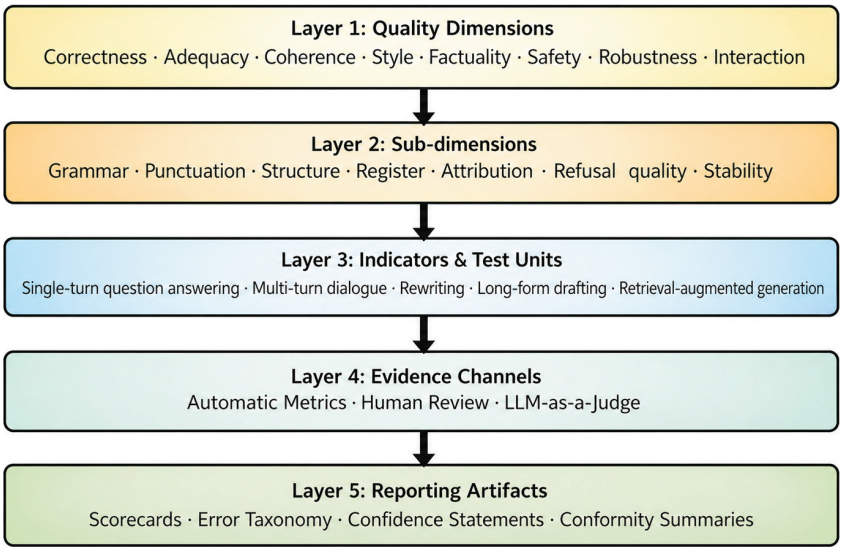

Figure 1 Layered architecture of the proposed standardized framework.

quality across tasks, products, and institutional settings, and to prevent evaluation from being reduced to narrow benchmark optimization or isolated performance indicators.

The second layer refines these broad dimensions into operational sub-dimensions such as grammar, punctuation, structure, register, attribution, refusal quality, and stability. This decomposition is necessary because high-level categories alone are too coarse to support reliable evaluation or actionable reporting. A model may appear fluent at a global level while still showing weaknesses in discourse organization, punctuation conventions, factual attribution, or behavioral consistency. By disaggregating broad quality objectives into more specific sub-dimensions, the framework improves interpretability, facilitates metric assignment, and supports more precise requirement setting in a standardization context.

The third layer translates sub-dimensions into indicators and test units, thereby connecting the conceptual model to executable evaluation practice. Representative task forms include single-turn question answering, multi-turn dialogue, rewriting, long-form drafting, and retrieval-augmented generation. In the proposed framework, these tasks are not treated merely as benchmark categories, but as evidentiary environments through which specific aspects of language quality can be observed. For example, single-turn question

answering may reveal correctness and factual adequacy, while multi-turn dialogue is more suitable for evaluating coherence, interaction consistency, and contextual stability. Long-form drafting can expose discourse planning, style control, and citation behavior, whereas rewriting tasks can isolate instruction fidelity, register transfer, and structural preservation. This layer therefore ensures that task selection is justified by evaluative relevance rather than by dataset convenience.

The fourth layer defines the evidence channels through which quality judgments are formed. The framework integrates automatic metrics, human review, and LLM-as-a-judge as complementary sources of evidence. Each channel contributes different strengths: automatic metrics offer scalability and repeatability, human review provides semantic, pragmatic, and culturally grounded sensitivity, and LLM-based judging can improve efficiency when used under controlled conditions. Rather than treating these methods as interchangeable alternatives, the framework positions them within a common evidential layer so that evaluation is based on triangulation rather than a single measurement source. This design is especially important for CLLM language-quality assessment, where nuanced linguistic and pragmatic phenomena often cannot be captured adequately by one method alone.

The fifth layer converts evaluation outcomes into standardized reporting artifacts, including scorecards, error taxonomies, confidence statements, and conformity summaries. This final layer gives the framework practical value beyond the laboratory, because evaluation becomes institutionally useful only when results can be documented, compared, audited, and communicated in a repeatable form. Structured reporting enables the framework to support procurement, governance, product documentation, profile specification, and future conformity assessment.

Taken together, the five-layer architecture connects quality definition, indicator design, evidence collection, scoring, and reporting within a single evaluation chain. This structure allows language-quality claims to be traced from high-level dimensions to observable evidence and documented outputs. By organizing evaluation in this way, the framework moves beyond benchmark-centered comparison and provides a more suitable basis for reproducible, standards-oriented CLLM language-quality assessment.

4.2 Mapping Quality Dimensions to Measurable Indicators

A central requirement of the proposed framework is that each quality dimension shown in Figure 1 must be traceable to measurable indicators rather

than remaining a purely conceptual label. This traceability is necessary if the framework is to function as a pre-standard artifact rather than a descriptive taxonomy. In practical evaluation settings, broad dimensions such as correctness, coherence, style, or robustness are not directly measurable unless they are translated into observable textual or behavioral features. The framework therefore adopts a mapping principle in which each high-level dimension is associated with one or more sub-dimensions, and each sub-dimension is further linked to explicit indicators that can be instantiated through test units. This layered mapping preserves conceptual clarity while enabling operational execution.

Correctness is interpreted in the framework as the extent to which the semantic content of a model response satisfies the task requirements and avoids substantive error. Its indicators may include answer accuracy, instruction fulfillment, semantic consistency with the input, and the absence of contradiction or unsupported claims. Adequacy, although related to correctness, is treated as a distinct dimension because a response may be factually correct yet incomplete, poorly scoped, or insufficiently responsive to the communicative intent of the prompt. Adequacy is therefore mapped to indicators such as relevance, completeness, coverage of requested constraints, and alignment with the intended use context. This distinction is especially important in Chinese-language applications involving formal, institutional, or customer-facing communication, where partial fulfillment may still constitute an unsatisfactory output.

Coherence is mapped to indicators that capture both local and global discourse quality. At the local level, the framework considers sentence-to-sentence continuity, referential clarity, connective appropriateness, and avoidance of abrupt topic drift. At the global level, it examines structural organization, progression of ideas, consistency of argumentative or explanatory flow, and the maintenance of thematic unity across long-form outputs. This distinction allows the framework to evaluate not only whether a response is readable at the sentence level, but also whether it remains organized and intelligible as a complete discourse unit. In Chinese long-form drafting and rewriting tasks, these indicators are particularly important because grammaticality alone does not guarantee acceptable document-level quality.

Style is treated in the framework as the controlled expression of linguistic form relative to contextual expectations. Its indicators include fluency, naturalness, register appropriateness, terminology consistency, rhetorical fit, and conformity with genre-specific conventions. In Chinese evaluation, this dimension carries additional weight because style often interacts with script

choice, punctuation conventions, idiomatic density, institutional phrasing, and culturally grounded norms of politeness or indirectness. A model may therefore be semantically correct while still failing stylistically if it produces language that is awkward, inconsistent in tone, or inappropriate for the intended audience. The framework accordingly treats style not as cosmetic variation but as a substantive component of language quality.

Factuality is mapped to indicators related to attribution, evidence consistency, source-groundedness, and resistance to hallucination. This dimension extends beyond the binary question of whether an isolated statement is true. It also includes whether the response distinguishes fact from conjecture, whether claims are traceable to provided evidence in retrieval-augmented settings, and whether the response maintains epistemic discipline when uncertainty is present. This treatment is necessary because high-quality language in deployment scenarios must be not only fluent but also trustworthy. In the framework, factuality therefore interacts closely with correctness, but it remains analytically distinct because it concerns the evidential grounding and reliability of generated content rather than only task-level answer success.

Safety is mapped to indicators that reflect harmful content avoidance, refusal appropriateness, de-escalation quality, and compliance with scenario-specific constraints. Importantly, the framework does not reduce safety to mere refusal frequency. Instead, it evaluates whether the model refuses when refusal is warranted, provides safe alternatives where appropriate, avoids over-refusal in benign contexts, and maintains a linguistically coherent and context-sensitive tone during safety responses. This is especially relevant in Chinese-language deployment, where acceptable refusal behavior may depend not only on policy consistency but also on pragmatic framing and user expectations.

Robustness is defined as the stability of quality under perturbation, reformulation, and contextual variation. Its indicators include consistency across paraphrased prompts, resistance to minor noise or ambiguity, preservation of output quality under multi-turn interaction, and tolerance to domain or formatting changes. This dimension is essential because evaluation results that depend heavily on a single prompt template or narrow benchmark phrasing are difficult to generalize or standardize. By explicitly mapping robustness to perturbation-sensitive indicators, the framework reinforces the idea that language quality should be assessed as a stable behavioral property rather than a one-off benchmark score.

Interaction quality is mapped to indicators such as conversational continuity, turn-level responsiveness, memory consistency within dialogue,

clarification behavior, and adaptation to user intent over multiple exchanges. This dimension reflects the practical reality that many CLLM applications are interactive rather than single-shot. A model that performs well on isolated prompts may still fail to sustain a coherent and user-aligned dialogue. Figure 1 therefore places interaction within the highest layer of the architecture to make clear that conversational behavior is not an optional add-on, but one of the major quality dimensions that standardized evaluation should cover.

Overall, the mapping process serves two functions. First, it decomposes abstract quality dimensions into operationally meaningful components. Second, it creates a traceable path from conceptual requirements to measurable evidence. This traceability allows evaluators to justify why a given task, metric, or review criterion is relevant to a specific language-quality claim.

4.3 Evidence Integration and Scoring Logic

While Figure 1 distinguishes the evidence layer from the reporting layer, the value of the framework depends on how evidence is integrated into coherent evaluation outcomes. The proposed approach does not assume that a single metric, judge, or benchmark can adequately represent CLLM language quality. Instead, it adopts an evidence integration model in which multiple channels contribute complementary information under a controlled scoring logic. This design reflects the fact that language quality is inherently multidimensional and that different dimensions are more appropriately assessed through different forms of evidence.

Automatic metrics provide the first evidence channel because they enable large-scale, repeatable, and low-cost assessment. Their utility is strongest in scenarios where the evaluation objective can be approximated through lexical, embedding-based, or structured factual comparisons. For example, overlap-based or semantic similarity metrics may offer useful signals for constrained rewriting or summarization tasks, while factuality-specific evaluators may support the assessment of grounded long-form generation. However, the framework treats automatic metrics as partial indicators rather than definitive judgments. Their outputs must be interpreted relative to task type, metric validity, and language-specific limitations. This is particularly important for Chinese, where acceptable outputs may vary significantly in lexical realization, discourse organization, or stylistic convention, thereby limiting the reliability of any single automatic metric.

Human review constitutes the second evidence channel and serves as the principal anchor for nuanced quality interpretation. Human evaluators are

necessary for judging pragmatic appropriateness, discourse quality, stylistic naturalness, refusal quality, and other aspects of language behavior that remain difficult to formalize computationally. Within the framework, human review is not treated as an unstructured fallback, but as a controlled procedure guided by explicit rubrics linked to the quality dimensions and sub-dimensions in Figure 1. This linkage is important because it reduces evaluator drift and increases comparability across review sessions. It also allows human judgments to function as calibration references for other evidence channels, especially when the framework is used to assess domain-specific Chinese outputs involving institutional, technical, or culturally sensitive language.

The third evidence channel is LLM-as-a-judge. The framework includes this channel because recent research has shown that strong language models can provide scalable evaluative signals under carefully designed prompting and rubric conditions. Nevertheless, the framework explicitly avoids treating model-based judging as a self-validating substitute for human assessment. Instead, LLM judges are positioned as structured auxiliary evaluators whose outputs must be examined for consistency, positional bias, verbosity bias, prompt sensitivity, and domain transfer limitations. In standardized evaluation settings, their role is therefore conditional and controlled. They may be used to expand coverage, identify candidate discrepancies, or provide intermediate scoring where human review is resource-constrained, but their judgments should remain anchored to predefined criteria and periodically checked against human evaluation.

In operational use, human review should serve as the primary calibration anchor for open-ended language-quality judgments. LLM-as-a-judge may be used after its outputs have been compared with a human-labeled development subset under the same rubric. When systematic disagreement is observed, the report should identify the affected dimensions, such as style, factuality, refusal quality, or discourse coherence, rather than merely averaging the scores. For high-impact scenarios, human adjudication should override LLM-judge outputs when the disagreement affects a gating condition or conformity conclusion. For lower-risk or large-scale monitoring scenarios, LLM-judge outputs may be used as screening signals, provided that periodic human checks are maintained.

The scoring logic of the framework is correspondingly hierarchical. Evaluation begins at the indicator level, where each test unit generates raw observations from one or more evidence channels. These observations are then normalized into sub-dimension scores according to task-specific and channel-specific scoring rules. Sub-dimension scores are subsequently

aggregated into dimension-level results, provided that minimum evidential sufficiency conditions are met. This means that a dimension score should not be interpreted as valid unless the relevant evidence channels and test coverage requirements have been satisfied. Such a design discourages superficial reporting and helps distinguish between true quality evidence and incomplete measurement.

A key principle of the scoring logic is that aggregation should preserve interpretability. The framework therefore discourages premature collapse into a single overall score unless the reporting context explicitly requires it. Even when an aggregate index is reported, the underlying scenario, dimension, and sub-dimension results should remain visible so that stakeholders can identify strengths, weaknesses, and areas of uncertainty.

The framework also treats uncertainty and disagreement as part of the evaluation outcome. Divergence between automatic metrics, human review, and LLM-based judging may indicate ambiguity, evaluator limitations, task underspecification, or instability in model behavior. Therefore, disagreement should be documented rather than hidden through simple averaging, especially when it affects dimension-level scores or gating decisions.

Finally, the evidence integration model supports extensibility. Because Figure 1 organizes the framework into layers, new metrics, judge models, or task types can be added without altering the foundational logic of the architecture. What must remain stable is not the exact implementation of each evidence channel, but the traceable relationship between dimensions, indicators, evidence sources, and reporting artifacts. This balance between structural stability and methodological extensibility is one of the reasons the framework is suitable for emerging-technology standardization. It allows the evaluation system to evolve with CLLM capabilities while preserving the consistency needed for comparable, auditable, and eventually standardizable language-quality assessment.

4.4 Operational Summary of the Framework

Tables 1 and 2, and Figures 2 and 3, provide the operational context of the proposed framework. While Figure 1 defines the layered architecture at a conceptual level, these additional elements clarify why the framework is needed, how it is executed, and what dimensions it ultimately evaluates. Their purpose is to translate the framework from a conceptual proposal into an implementable and auditable evaluation model.

Table 1 Representative differences between prevailing LLM evaluation practice and the proposed standards-oriented framework

Aspect	Prevailing Evaluation Practice	Limitation for Standardized Assessment	Proposed Framework Response
Evaluation target	General capability, benchmark ranking, or task accuracy	Language quality is not isolated as a first-class assessment object	Defines language quality as an explicit, multi-dimensional evaluation target
Scenario coverage	One benchmark, one dataset, or one prompt family	Limited representativeness across real deployment contexts	Uses scenario-family profiles spanning question answering, dialogue, rewriting, long-form drafting, and retrieval-augmented generation
Chinese-language specificity	Often inherited from multilingual or translated benchmarks	Chinese-specific linguistic and pragmatic phenomena are underrepresented	Introduces Chinese-specific quality profiles covering script, punctuation, register, idiomaticity, attribution, and interaction norms
Evidence model	Single metric, single benchmark score, or single-judge pipeline	Weak validity, weak reproducibility, and poor cross-setting comparability	Combines automatic metrics, human review, and LLM-based judging under explicit evidence-fusion rules
Robustness treatment	Often evaluated only once per prompt or setting	Prompt sensitivity and run variance remain hidden	Requires perturbation testing, rerun analysis, and stability reporting
Reporting form	Leaderboards, mean scores, or informal summaries	Limited auditability and weak suitability for governance or procurement	Produces structured scorecards, error taxonomies, confidence statements, and conformity-style summaries
Reproducibility control	Partial disclosure of prompts, settings, or judge conditions	Results are difficult to replicate across institutions	Requires protocol freezing, execution logging, calibration records, and versioned reporting artifacts
Standardization readiness	Benchmark-centric and research-oriented	Difficult to translate into pre-standard or formal standard documents	Provides layered terminology, evaluation workflow, and reporting structure suitable for pre-standardization

Table 2 Top-level quality dimensions, representative sub-dimensions, and example indicators in the proposed framework

Dimension	Representative Sub-Dimensions	Example Indicators
Q1 Linguistic correctness	Grammar, punctuation, orthography, lexical choice, terminology consistency	Grammatical error rate, punctuation compliance, orthographic accuracy, term consistency score
Q2 Semantic adequacy	Instruction fulfillment, completeness, relevance, faithfulness to source/input	Task completion score, omission rate, constraint satisfaction rate, semantic alignment rating
Q3 Discourse coherence	Organization, topic continuity, referential consistency, long-context stability	Discourse structure score, contradiction flags, entity continuity score, context retention score
Q4 Stylistic appropriateness	Register control, tone alignment, genre fit, audience adaptation	Style-match rating, tone appropriateness score, genre compliance, audience suitability assessment
Q5 Factual grounding	Accuracy, attribution, evidential support, uncertainty handling	Unsupported-claim rate, citation-support rate, fact precision score, calibrated uncertainty expression
Q6 Safety-compliant expression	Refusal quality, non-toxicity, privacy-aware wording, harm avoidance	Policy adherence score, refusal helpfulness rating, toxicity score, privacy-risk flags
Q7 Robustness and stability	Prompt sensitivity, paraphrase consistency, rerun variance, decoding stability	Variance across runs, paraphrase agreement, robustness under perturbation, stability index
Q8 Interactive quality	Clarification behavior, repair capability, dialogue continuity, memory consistency across turns	Clarification appropriateness, recovery success rate, turn-to-turn coherence score, intent retention score

Table 1 provides the comparative rationale for the framework by contrasting it with prevailing LLM evaluation practice. The comparison highlights that much current evaluation remains centered on benchmark ranking, task accuracy, or isolated model comparison, whereas standardized language-quality assessment requires explicit treatment of evaluation target definition, scenario coverage, Chinese-language specificity, evidence integration, robustness testing, reproducibility control, and reporting discipline. In this

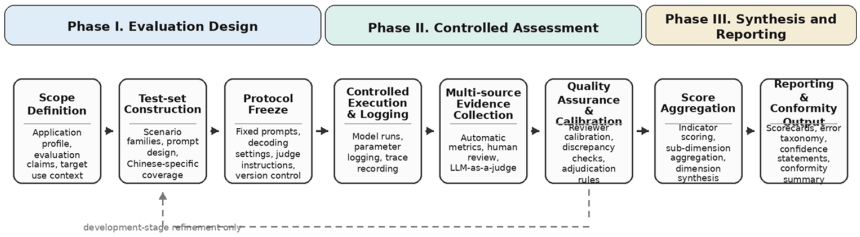

Figure 2 Standardized evaluation workflow for Chinese large language model language-quality assessment.

way, Table 1 does more than summarize differences in practice; it establishes why CLLM language-quality evaluation should be approached as a pre-standardization problem rather than as a simple extension of existing benchmark culture.

Table 2 complements this rationale by translating the framework into a measurable quality model. It summarizes the top-level dimensions, representative sub-dimensions, and example indicators that anchor the evaluation process. Its central function is to preserve traceability between abstract quality claims and concrete observations. By showing how each dimension can be decomposed into observable traits and corresponding indicators, the table provides the operational bridge between the architecture in Figure 1 and the evidence and scoring logic discussed in Sections 4.2 and 4.3.

Figure 2 presents the evaluation workflow as a governed sequence of stages, beginning with scope definition and ending with reporting and conformity summary. This workflow is important because it frames evaluation not as an isolated experiment, but as a controlled process. Scope definition establishes the assessment target and intended claims; test-set construction and protocol freeze create methodological stability; controlled execution and evidence collection support reproducibility; and aggregation and reporting convert raw findings into structured outputs suitable for comparison, audit, and decision-making. Figure 2 therefore captures the procedural discipline required for standards-oriented evaluation.

Figure 3 provides a concise interpretive view of the top-level quality dimensions and their primary assessment focus. Unlike Table 2, which emphasizes operational mapping, Figure 3 emphasizes conceptual readability. It offers a compact summary of the multidimensional nature of language quality and makes visible the fact that the framework is not limited to correctness or benchmark accuracy. Instead, it spans linguistic form, semantic adequacy, discourse behavior, stylistic appropriateness, factual grounding,

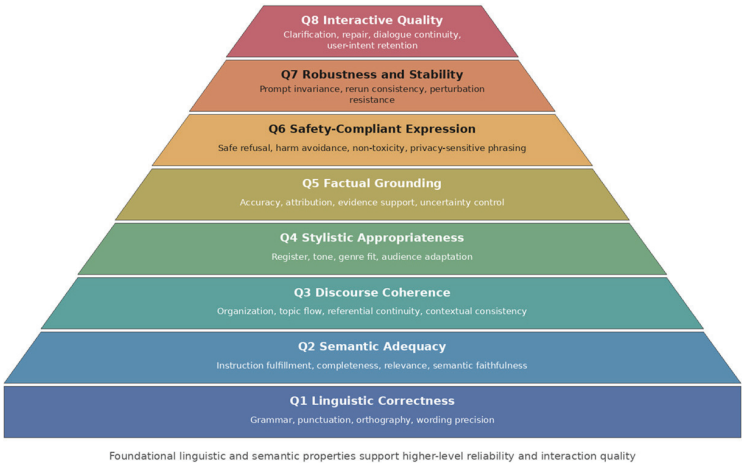

Figure 3 Pyramid of top-level quality dimensions in the proposed framework.

safety-compliant expression, robustness, and interaction quality. This visual summary is particularly useful for communicating the structure of the framework to stakeholders who need a clear overview before engaging with detailed indicators and scoring rules.

Taken together, Tables 1 and 2, and Figures 2 and 3, serve complementary roles within the framework. Table 1 explains the comparative need for a standards-oriented approach, Table 2 defines the measurable quality model, Figure 2 formalizes the evaluation workflow, and Figure 3 summarizes the multidimensional evaluation scope. Their combined function is to show that the proposed framework is not only conceptually structured, but also operationally organized, procedurally governed, and suitable for repeatable application across research and deployment settings.

Table 2 then translates the high-level architecture into a more operational quality model by summarizing the top-level dimensions, representative sub-dimensions, and example indicators. Its function is to make the framework measurable. Table 2 demonstrates that each quality dimension can be decomposed into observable traits and corresponding indicators, thereby supporting traceability between abstract quality claims and concrete evaluation evidence. This traceability is central to the framework because standardized assessment requires not only a conceptual taxonomy, but also a mechanism through which evaluators can justify how a score or judgment was produced. In this sense, Table 2 serves as the bridge between the layered logic of Figure 1 and the actual scoring and reporting procedures described later in the paper.

Figure 2 complements the dimensional model by presenting the evaluation workflow as a controlled sequence of stages, beginning with scope definition and ending with reporting and conformity summary. The significance of this workflow is that it treats evaluation as a governed process rather than an isolated experiment. Scope definition ensures that the assessment target, deployment context, and intended claims are explicit. Test-set construction and protocol freeze create methodological stability before execution begins. Controlled execution, evidence collection, and calibration ensure that results are not merely obtained, but generated under reproducible conditions. Aggregation and reporting then translate raw evidence into structured outputs that can support comparison, auditing, and decision-making. Figure 2 therefore expresses the procedural discipline required for any framework intended to support pre-standardization or future formal standardization.

Figure 3 provides a concise visual summary of the top-level quality dimensions and their primary assessment focus. Its role is interpretive rather than procedural. Where Table 2 gives a more detailed mapping of dimensions to sub-dimensions and indicators, Figure 3 offers a compact overview of the conceptual coverage of the framework. This is useful because the proposed approach is broader than many existing benchmark-centered methods. It makes explicit that language quality in CLLMs includes not only linguistic form and semantic adequacy, but also discourse behavior, stylistic control, factual grounding, safety-compliant expression, robustness, and interactive quality. By presenting these dimensions together, Figure 3 reinforces the argument that language quality is inherently multi-dimensional and cannot be adequately represented by a single benchmark score or a single task family.

Taken together, Tables 1 and 2, and Figures 2 and 3, strengthen the paper in three ways. First, they improve conceptual clarity by separating comparative motivation, operational dimensions, procedural workflow, and top-level scope summary into distinct but connected artifacts. Second, they improve technical rigor by showing that the framework is not limited to general principles, but includes explicit mappings, staged procedures, and structured outputs. Third, they improve standardization relevance by presenting the framework in forms that are compatible with specification writing, evaluation guideline development, and future conformity-oriented reporting. For these reasons, Tables 1 and 2, and Figures 2 and 3, should be understood not as supplementary illustrations, but as integral parts of the proposed standardized evaluation framework.

5 Metric System and Scoring Model

5.1 Indicator Design

Evaluation indicators should satisfy six core properties: relevance, interpretability, repeatability, sensitivity, robustness, and practical cost-effectiveness. The framework explicitly discourages dependence on any single type of indicator, since no individual measure can adequately capture the full range of language-quality phenomena. Instead, it advocates a mixed-indicator design that combines complementary evidence channels.

Automatic indicators may include grammar and typographical error detection, reference-based semantic similarity when gold outputs are available, terminology-consistency checks, structural compliance checks, retrieval-grounded attribution measures, and stability statistics computed across repeated runs. These indicators are valuable for scalability, repeatability, and regression monitoring.

Human-review indicators should be used to assess qualities such as clarity, coherence, stylistic appropriateness, faithfulness, and practical usefulness. Human assessment should rely on anchored scoring rubrics with explicit criteria and examples, rather than unstructured Likert-style judgments alone, in order to improve consistency and auditability.

Judge-model indicators can support pairwise comparison, rubric-based scoring, and explanation extraction. However, such indicators should be treated as supplementary evidence and used only after calibration against human-annotated data. Their deployment should also include bias-mitigation procedures, such as prompt-order randomization, judge-ensemble comparison, and periodic consistency checks.

5.2 Recommended Scoring Architecture

Each test unit is assigned a set of normalized indicator scores on a common scale from 0 to 100. Aggregation is then performed hierarchically. At the sub-dimension level, the score is computed as:

$$S_{\text{sub}} = \sum_i w_i x_i, \sum_i w_i = 1$$

where x_i denotes the normalized indicator score and w_i its corresponding weight.

Weight selection should be profile-specific and explicitly documented. In a research benchmark profile, weights may be assigned to maintain balanced

coverage across dimensions. In an enterprise acceptance profile, weights may reflect deployment risk, user impact, and the relative importance of different quality failures. In a sector-specific profile, weights may be determined through expert consensus, stakeholder requirements, or pilot validation. The framework does not prescribe one universal weighting scheme because language-quality priorities differ across scenarios. However, it requires that all weights be declared before evaluation, justified in the evaluation protocol, and preserved in the report so that results remain interpretable and reproducible. At the dimension level, the score is defined as:

$$S_{\mathrm{dim}} = \sum_{j} \alpha_j S_{\mathrm{sub},j}$$

where $S_{\mathrm{sub},j}$ represents the relevant sub-dimension scores and α_j the aggregation weights. At the scenario level, the score is computed as:

$$S_{\mathrm{scn}} = \sum_{k} \beta_k S_{\mathrm{dim},k}$$

where only the dimensions relevant to the scenario are activated. At the overall profile level, the aggregate score is:

$$S_{\mathrm{total}} = \sum_{m} \gamma_m S_{\mathrm{scn},m}$$

This layered structure preserves interpretability by allowing performance to be inspected not only through a single overall score, but also through scenario-level, dimension-level, and sub-dimension-level diagnostics.

5.3 Gating and Minimum Conditions

Gating rules differ from ordinary weights because they define minimum acceptability conditions rather than relative contribution to an aggregate score. A low-weighted dimension may still be subject to a gate if failure in that dimension creates unacceptable risk for the declared profile. For example, a factual question-answering (QA) or retrieval-augmented generation (RAG) profile may assign substantial weight to semantic adequacy and factual grounding, but it may also require a minimum grounding score before any overall pass decision is possible. This separation between weighted performance and mandatory thresholds prevents critical failures from being hidden by strong performance in less critical dimensions.

A robust evaluation standard should prevent serious deficiencies from being obscured by high average performance in other areas. For this reason, the framework incorporates gating rules and minimum acceptance conditions in addition to weighted aggregation. For example, factual grounding below a specified threshold should prevent a system from receiving a qualified rating in information-intensive scenarios. Similarly, failure to meet safety-expression requirements should block deployment in public-service or high-risk use settings. In the same way, robustness below a minimum threshold may trigger only a conditional pass, even when average quality scores appear strong. Such gating mechanisms are essential for real-world acceptance testing because they ensure that critical risks are handled explicitly rather than diluted through averaging.

5.4 Uncertainty and Agreement

Whenever human review or judge-model assessment is involved, the evaluation report should document uncertainty and agreement. This includes inter-rater agreement statistics, procedures for resolving annotation disagreements, and confidence intervals or bootstrap ranges wherever feasible. For tasks involving repeated generation, variability across multiple runs should also be reported explicitly. Run-to-run variance is not merely noise; it is an important indicator of stability and reliability, and should therefore be treated as part of the evaluation outcome rather than omitted from reporting.

5.5 Error Taxonomy

Overall scores by themselves do not provide enough information for meaningful diagnostic analysis. For this reason, the framework requires an explicit error taxonomy that classifies observed failures in a consistent and interpretable manner. At a minimum, the taxonomy should include the following top-level categories: linguistic errors, semantic omissions or distortions, factual errors or unsupported claims, discourse inconsistencies, style or register mismatches, unsafe or non-compliant expressions, interaction failures, and robustness failures.

The use of a structured taxonomy enables evaluators to move beyond summary scores and identify the specific sources of model weakness. It also strengthens comparability across systems, supports more reliable benchmark maintenance, and provides a clearer foundation for future refinement and extension of the framework.

5.6 Reference Implementation Guidance

For practical deployment, the framework should be accompanied by a reference implementation that specifies a minimum set of evaluation requirements. As a baseline, the core evaluation profile should cover at least three scenario families in order to avoid overly narrow conclusions about system performance. Each scenario family should include no fewer than 100 test units for a basic release-level assessment, so that reported results have an adequate empirical basis.

For quality assurance, critical subsets should be subjected to dual human review, especially in cases where errors may have significant downstream consequences. When judge models are used, they should first be calibrated on a labeled development subset to ensure reasonable alignment with human judgment. In addition, prompt templates should be version controlled, and all test units should be associated with immutable identifiers so that evaluation results remain reproducible over time.

To support transparency and external scrutiny, the reference implementation should also disclose the scoring formulas used in aggregation and provide at least a partial set of prompt exemplars. Taken together, these practices establish a realistic and operational evaluation baseline that can be adopted by both academic researchers and industrial benchmarking programs.

In this paper, a scenario profile denotes a task- or deployment-oriented evaluation configuration that activates different quality dimensions and weighting priorities according to the intended use context.

Table 3 Evidence channels and recommended usage

Evidence Channel	Strengths	Weaknesses	Recommended Role
Automatic metrics	Scalable, objective, and repeatable	Often insensitive to pragmatics, discourse quality, and contextual appropriateness	Baseline measurement, regression tracking, and large-scale monitoring
Human review	High face validity and strong sensitivity to nuanced quality differences	Expensive and potentially inconsistent without calibration	Primary evidence source for open-ended language quality
LLM-as-a-judge	Scalable for comparative and rubric-based assessment of open-ended outputs	Vulnerable to bias, instability, and reproducibility concerns	Secondary evidence channel after calibration and bias control

Table 4 Example scenario profiles and weight emphasis

Scenario Family	Dominant Dimensions	Typical Outputs	Special Notes
Formal document drafting	Q1, Q2, Q3, Q4, Q5	Notice, email, report, memo	High weighting should be assigned to structure, register control, and formatting consistency
Factual QA / RAG	Q2, Q5, Q6, Q7	Concise answers, grounded responses	Grounding quality and hallucination-related gating are essential
Customer-service dialogue	Q2, Q4, Q6, Q8	Multi-turn assistance dialogue	Repair strategies and safe redirection should receive explicit attention
Educational explanation	Q2, Q3, Q4, Q5	Step-by-step explanation	Clarity and level-appropriate presentation are especially important
Creative constrained writing	Q3, Q4, Q6, Q7	Story, rewrite, title, script	Naturalness should be balanced against task constraints and controllability

As an illustrative application, consider a formal document drafting profile for enterprise Chinese writing. The scope declaration may specify that the system is being evaluated for drafting notices, internal reports, and customer-facing emails. The activated dimensions may include linguistic correctness, semantic adequacy, discourse coherence, stylistic appropriateness, and factual grounding, while safety and robustness may be treated as gating or supporting dimensions depending on the deployment context. Test units would include prompts requiring formal register, consistent terminology, structured paragraphs, and faithful use of supplied source information. Automatic metrics may be used to detect terminology consistency, format compliance, and factual overlap with provided materials. Human reviewers would assess register appropriateness, coherence, completeness, and practical usefulness using anchored rubrics. A calibrated LLM judge may provide secondary scoring or flag candidate disagreements for human review. The final report would present dimension-level scores, major error categories, disagreement notes, and a conformity conclusion such as pass, conditional pass, or fail under the declared profile. This example illustrates how the framework connects scope definition, scenario selection, evidence collection, scoring, and reporting without requiring all dimensions to be weighted equally in every use case.

6 Reporting, Conformity, and Pre-Standard Deliverables

Reporting is treated as a formal component of the proposed framework rather than as a secondary presentation layer. In standards-oriented assessment, a score is useful only when the protocol, evidence, scoring rules, uncertainty, and limitations are documented in a reproducible form. For this reason, the framework requires a minimum reporting package that includes system identification, profile declaration, test-set provenance, protocol specification, score summary, uncertainty and agreement statistics, error analysis, conformity statement, change log, and limitations. These elements ensure that evaluation outcomes can be interpreted, compared, audited, and maintained across model or dataset versions.

Figure 4 summarizes the reporting workflow from system identification and profile declaration to scoring, uncertainty analysis, error analysis, and conformity statement. The workflow shows how evaluation evidence is organized into a structured report that can support comparison, audit, and decision-making. Figure 5 illustrates the evidence traceability chain. Raw outputs, prompts, source passages, execution records, and annotations are first converted into indicators and review evidence. These are then aggregated into quality results and finally linked to a conformity statement. This

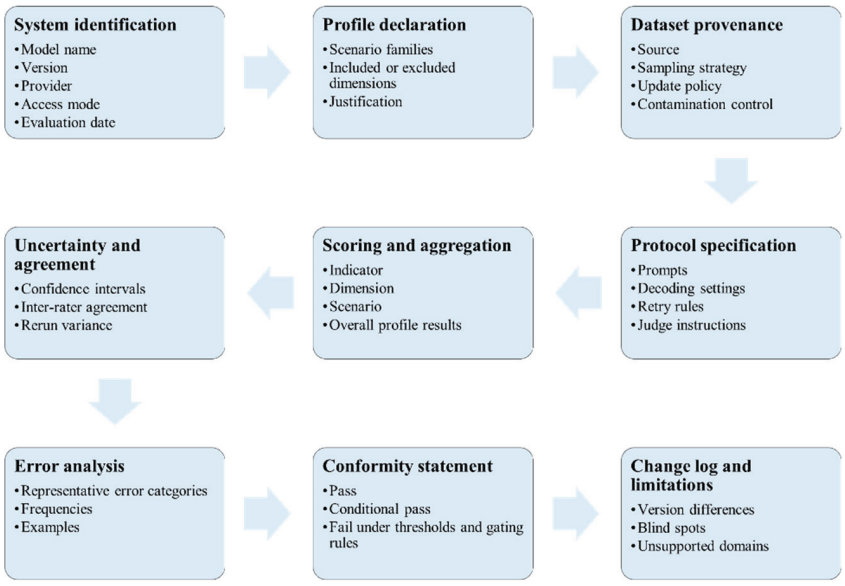

Figure 4 Standardized reporting workflow for language-quality evaluation.

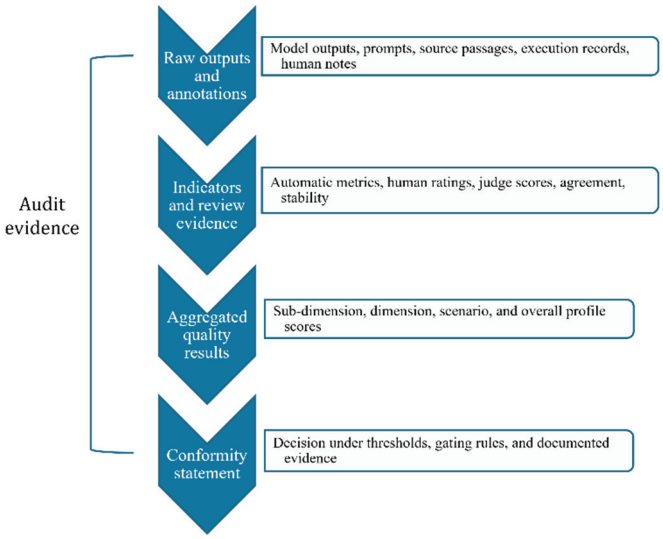

Figure 5 Evidence traceability chain from raw observations to conformity statement.

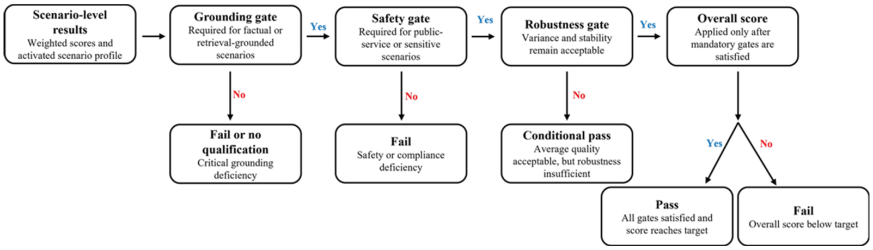

Figure 6 Example conformity-decision logic with gating conditions.

chain helps ensure that final conclusions remain explainable in terms of the observations and scoring rules that produced them.

Figure 6 shows how conformity decisions combine gating conditions with overall acceptance criteria. Critical dimensions such as factual grounding, safety, and robustness may operate as mandatory requirements before the overall weighted score is considered. This prevents strong performance in less critical areas from masking failures that are unacceptable for the declared evaluation profile.

Figure 7 further shows that a standardized report should be understood as a multi-perspective artifact rather than a simple score sheet. The report architecture includes technical, data, measurement, and decision perspectives. The technical perspective records model versioning, prompts, decoding

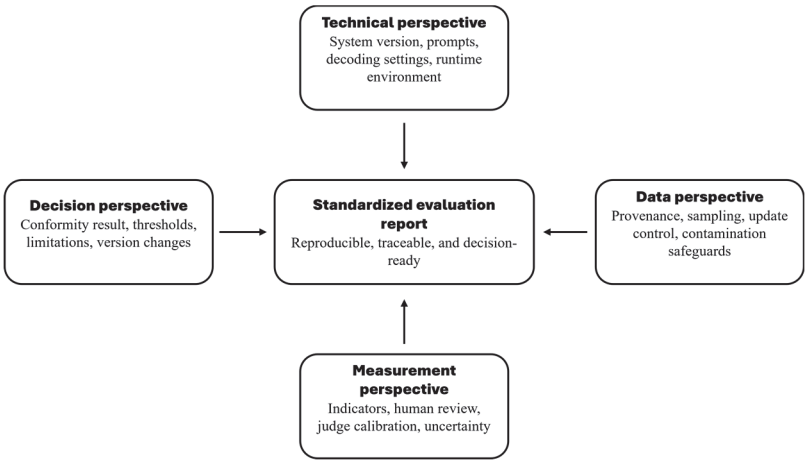

Figure 7 Multi-perspective architecture of a standardized evaluation report.

settings, and runtime conditions. The data perspective records provenance, sampling, update control, and contamination safeguards. The measurement perspective records indicators, review procedures, calibration, and uncertainty. The decision perspective records thresholds, conformity conclusions, limitations, and version changes. Organizing the report in this way improves interpretability for different stakeholders, including developers, evaluators, procurement teams, auditors, and standards bodies.

Figures 8 and 9 extend this logic from one-time reporting to lifecycle management and multidimensional interpretation. Figure 8 situates evaluation within a controlled release and maintenance cycle, showing that standardized assessment must support not only initial evaluation but also re-evaluation, benchmark refresh, and version tracking over time. Figure 9 illustrates how multidimensional score profiles can communicate strengths and weaknesses across dimensions more effectively than a single aggregate score. Such profiling is especially useful in deployment settings, where release readiness often depends on the pattern of quality performance rather than on one summary number alone.

From a pre-standardization perspective, the outputs of the framework can be organized into four candidate normative artifacts. The first is a terminology and definitions artifact covering scenario families, quality dimensions, evidence channels, and conformity outcomes. The second is a quality model and indicator catalog specifying the structure of the framework and the recommended indicator sets associated with each dimension. The third is an

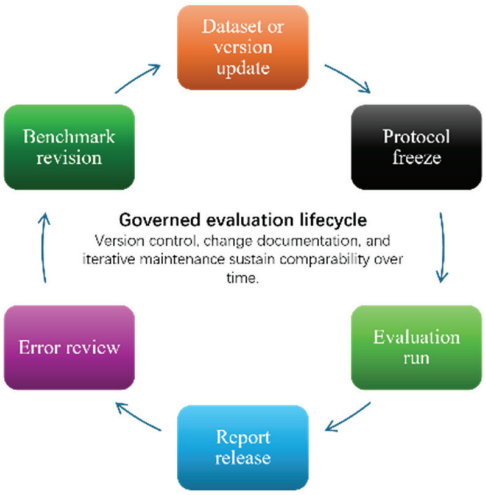

Figure 8 Evaluation-release lifecycle and benchmark maintenance.

Figure 9 Example multidimensional scorecard profile for a model under one scenario set.

Table 5 Minimum metadata required for a standardized report

Category	Required Items
System identification	Model name, version, provider, application programming interface (API) or local deployment, evaluation date
Protocol	System prompt, user template, decoding parameters, retry rules
Dataset provenance	Source, authoring policy, contamination controls, version ID
Execution log	Run ID, timestamp, hardware/service environment, failures or retries
Quality assurance	Rater training, calibration method, agreement statistics, adjudication process
Results	Indicator scores, dimension scores, scenario scores, gating decisions
Limitations	Known blind spots, unsupported domains, caveats for interpretation

evaluation procedure artifact defining execution rules, aggregation methods, uncertainty treatment, and gating logic. The fourth is a reporting template artifact standardizing how evidence, scores, and conformity conclusions are documented. Together, these artifacts provide a practical pathway through which the framework may evolve into consortium guidance, a technical specification, or a future formal standards proposal. In summary, the reporting and conformity components convert evaluation results into decision-ready artifacts. They make the framework usable not only for benchmarking, but also for procurement, release assessment, governance, and future standardization activities.

7 Implementation Pathways and Standardization Value

The principal contribution of the proposed framework is not the creation of another standalone benchmark, but the design of an integrated evaluation architecture for CLLM language quality. Existing studies often focus on one layer at a time, such as dataset construction, metric design, human review, or leaderboard comparison. By contrast, the present framework links quality modeling, test-unit design, evidence integration, conformity logic, and reporting structure within a single system. This integration is what makes the framework suitable as a bridge between research evaluation and standardization-oriented assessment.

A major strength of the framework is its profile-based implementation structure. Because the common quality model is separated from scenario-specific weighting, gating, and reporting choices, the framework can support

multiple deployment modes without sacrificing comparability. One deployment mode is an open benchmark profile for public research use, with transparent task descriptions, disclosed scoring logic, and partially open evaluation materials. A second is an enterprise acceptance profile for model onboarding, procurement evaluation, release gating, and regression monitoring in internal environments. A third is a sector-specific profile in which the common framework backbone is retained but scenario families, indicators, thresholds, and conformity criteria are adapted to domains such as education, finance, healthcare, public administration, or media. This profile-based logic enables reuse of the same framework across distinct settings while preserving operational relevance.

Different stakeholders may emphasize different parts of the framework. Researchers may use the common quality model and scenario-family profiles to improve comparability across studies. Benchmark developers may use the indicator catalog, protocol-freezing rules, and error taxonomy to design more transparent evaluation sets. Enterprises may focus on acceptance profiles, gating rules, regression monitoring, and release-oriented scorecards. Procurement teams and auditors may rely on the reporting template, evidence traceability, and conformity statements. Standards organizations may use the terminology, workflow, and profile-based structure as inputs to future technical specifications.

The framework is also valuable because it places Chinese-specific linguistic and communicative properties at the center of evaluation design. Rather than treating Chinese as a downstream adaptation of English-oriented evaluation practice, it recognizes script choice, punctuation conventions, idiomaticity, register control, mixed Chinese-English terminology, and context-sensitive pragmatic norms as substantive parts of language quality. This emphasis improves the suitability of the framework for real Chinese deployment contexts and strengthens its relevance for future Chinese-language standardization work.

At the same time, practical implementation remains challenging. High-quality human review is costly and difficult to scale across multiple scenario families. LLM-as-a-judge can reduce cost, but only when calibration, bias monitoring, and agreement checks are maintained over time. Factuality-oriented evaluation is also unstable because external knowledge changes, making benchmark refresh and version control necessary. In addition, broader disclosure of prompts and materials may improve transparency while simultaneously increasing the risk of benchmark contamination. These tensions do not undermine the framework, but they do show why governed workflows,

profile control, and reporting discipline are necessary parts of any standards-oriented evaluation system.

A further implementation issue is that Chinese should not be treated as a uniform linguistic target. Standard written Chinese in Mainland usage, regional variation, domain-specific institutional language, mixed Chinese-English professional discourse, and cross-script or cross-register settings all introduce meaningful complexity. The proposed framework addresses this in part through profile design, dimension weighting, provenance control, and limitation reporting. However, additional work is needed for region-sensitive, multilingual-Chinese, and possibly dialect-aware extensions. Here again, the profile-based structure is useful because it allows controlled specialization without abandoning a shared evaluation backbone.

The proposed framework also has limitations. Conceptually, it defines a general evaluation architecture but does not prescribe a single universal weighting scheme or threshold set, because these must be adapted to scenario profiles and deployment risks. Operationally, high-quality human review remains costly, and LLM-as-a-judge requires continuous calibration and bias monitoring. Empirically, the framework still requires validation through shared evaluation profiles, cross-organization studies, and reliability analysis of indicators and scoring rules. Institutionally, adoption will depend on whether different organizations are willing to align terminology, reporting formats, and evidence requirements. These limitations indicate that the framework should be understood as a structured foundation for standardization-oriented evaluation rather than as a completed formal standard.

From a standardization perspective, the framework has value precisely because of its intermediate status. It is more structured and auditable than a conventional benchmark design, but less rigid than a finalized normative standard. That position makes it suitable for pilot adoption, inter-organizational comparison, consortium experimentation, and profile refinement before formal standard-setting. In practical terms, it can support benchmark development, internal model governance, external procurement evaluation, and the gradual harmonization of reporting practice across institutions.

Future work should therefore proceed along both empirical and institutional directions. On the empirical side, more studies are needed to validate indicator reliability, judge-model calibration, robustness-testing procedures, and Chinese-specific scenario coverage. On the institutional side, the framework should be tested through shared profiles, cross-organization reporting exercises, and prototype conformity procedures that can inform eventual

technical specifications or standards proposals. In this way, the framework can serve not only as a research contribution, but also as a practical foundation for future CLLM language-quality standardization.

8 Conclusion

This paper presents a standardized framework for the evaluation of Chinese large language model language quality. The framework is motivated by a clear gap between capability-oriented benchmarking and the practical requirements of standardization, procurement, governance, auditability, and reproducible product evaluation. In response to this gap, the paper has proposed an integrated architecture comprising a layered quality model, Chinese-specific evaluation dimensions and indicators, scenario-based assessment procedures, multi-source evidence mechanisms, gating and conformity logic, and standardized reporting artifacts. Taken together, these elements provide a technically grounded and standards-oriented blueprint for systematic Chinese language quality evaluation.

The significance of this work extends beyond the introduction of a new evaluation framework. It argues that language-quality assessment should be understood not merely as a benchmarking exercise, but as an engineering, governance, and standardization problem. A model is not adequately evaluated simply because it performs well on a public leaderboard; rather, it must be assessed through transparent procedures, interpretable quality dimensions, traceable evidence, and decision-ready reporting. By placing these requirements within a unified framework, the present study provides a practical foundation for more reliable and comparable evaluation practice.

More broadly, the framework offers a pathway by which fragmented research benchmarking can evolve toward a more formal evaluation infrastructure for Chinese language AI systems. If adopted and iteratively refined by researchers, industry stakeholders, and standards communities, it could serve initially as a de facto evaluation specification for research and industrial practice, and later as the basis for pre-standardization guidance or a formal standards contribution within the wider ICT and AI quality ecosystem. In this sense, the contribution of the paper lies not only in what it evaluates, but also in how it redefines the role of language-quality evaluation in the lifecycle of trustworthy Chinese large language models.

References

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., "Language Models are Few-Shot Learners," *Advances in Neural Information Processing Systems 33*, 2020.

[2] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al., "On the Opportunities and Risks of Foundation Models," *arXiv preprint* arXiv:2108.07258, 2021.

[3] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623, 2021.

[4] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, Y. Yao, C. Zhang, D. Narayanan, B. Wu, A. Kumar, et al., "Holistic Evaluation of Language Models," *Transactions on Machine Learning Research*, 2023.

[5] M. T. R. Laskar, S. Alqahtani, M. S. Bari, M. Rahman, M. A. M. Khan, H. Khan, I. Jahan, A. Bhuiyan, C. W. Tan, M. R. Parvez, E. Hoque, S. Joty, and J. Huang, "A Systematic Survey and Critical Review on Evaluating Large Language Models: Challenges, Limitations, and Recommendations," *arXiv preprint* arXiv:2407.04069, 2024.

[6] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, "Measuring Massive Multitask Language Understanding," in *Proceedings of the International Conference on Learning Representations*, 2021.

[7] A. Srivastava, A. Rastogi, A. Rao, A. W. M. Shoeb, A. Abid, A. Fisch, A. R. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso, et al., "Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models," *Transactions on Machine Learning Research*, 2023.

[8] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena," *Advances in Neural Information Processing Systems 36*, 2023.

[9] Y. Dubois, B. Galambosi, P. Liang, and T. B. Hashimoto, "Length-Controlled AlpacaEval: A Simple Way to Debias Automatic Evaluators," *arXiv preprint* arXiv:2404.04475, 2024.

[10] D. Kiela, M. Bartolo, Y. Nie, D. Kaushik, A. Geiger, A. H. O. Bassem, E. Dinan, J. Urbanek, A. Szlam, Z. H. Kalyan, et al., "Dynabench: Rethinking Benchmarking in NLP," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4110–4124, 2021.

[11] A. Bavaresco, M. Bernhard, M. Cettolo, F. Dell'Orletta, M. Federico, A. Guerberof, J. H. M. González, K. He, F. Heinemann, G. Houbrechts, et al., "LLMs instead of Human Judges? A Large Scale Empirical Study across 20 NLP Evaluation Tasks," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024.

[12] Y. Huang, Y. Bai, Z. Zhu, J. Zhang, J. Zhang, T. Su, J. Liu, C. Lv, Y. Cui, Y. Yuan, et al., "C-Eval: A Multi-Level Multi-Discipline Chinese Evaluation Suite for Foundation Models," in *Advances in Neural Information Processing Systems 36*, 2023.

[13] H. Li, Y. Zhang, F. Koto, Y. Yang, H. Zhao, Y. Gong, N. Duan, and T. Baldwin, "CMMLU: Measuring Massive Multitask Language Understanding in Chinese," in *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 9853–9879, 2024.

[14] W. Zhong, R. Cui, Y. Guo, Y. Liang, S. Lu, Y. Wang, A. Saied, W. Chen, and N. Duan, "AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models," in *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8794–8823, 2024.

[15] Y. He, W. Sun, Y. Chen, X. Wang, S. Zhou, Z. Wang, S. Jin, Y. Wang, X. Ren, J. Zhao, et al., "A Chinese Factuality Evaluation for Large Language Models," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 2025.

[16] Q. Cheng, T. Sun, W. Zhang, S. Wang, X. Liu, M. Zhang, J. He, M. Huang, Z. Yin, K. Chen, and X. Qiu, "Evaluating Hallucinations in Chinese Large Language Models," *arXiv preprint* arXiv:2310.03368, 2023.

[17] X. Liang, S. Song, S. Niu, Z. Li, F. Xiong, B. Tang, Y. Wang, D. He, P. Cheng, Z. Wang, and H. Deng, "UHGEval: Benchmarking the Hallucination of Chinese Large Language Models via Unconstrained Generation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

[18] Y. Tan, B. Zheng, B. Zheng, K. Cao, H. Jing, J. Wei, J. Liu, Y. He, W. Su, X. Zhu, B. Zheng, and K. Zhang, "Chinese SafetyQA: A

Safety Short-form Factuality Benchmark for Large Language Models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

[19] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, "G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, 2023.

[20] J. Li, X. Cheng, X. Zhao, J.-Y. Nie, and J.-R. Wen, "HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6449–6464, 2023.

[21] S. Min, K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi, "FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, 2023.

[22] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

[23] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Text Summarization Branches Out*, pp. 74–81, 2004.

[24] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT," in *Proceedings of the International Conference on Learning Representations*, 2020.

[25] T. Sellam, D. Das, and A. P. Parikh, "BLEURT: Learning Robust Metrics for Text Generation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7881–7892, 2020.

[26] Z. Li, X. Xu, T. Shen, C. Xu, J.-C. Gu, Y. Lai, C. Tao, and S. Ma, "Leveraging Large Language Models for NLG Evaluation: Advances and Challenges," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024.

[27] S. Gehrmann, T. Adewumi, K. Aggarwal, P. A. Ammanamanchi, A. Anadkat, K. Anubhai, J. A. Bale, C. Bandarkar, S. B. Das, B. Gautam, et al., "The GEM Benchmark: Natural Language Generation, its Evaluation and Metrics," in *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pp. 96–120, 2021.

[28] C.-H. Chiang and H.-Y. Lee, "Can Large Language Models Be an Alternative to Human Evaluators?" in *Proceedings of the 61st Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 15607–15631, 2023.

[29] C. M. Chan, W. Chen, Y. Su, J. Yu, W. Xue, S. Zhang, J. Fu, and Z. Liu, "ChatEval: Towards better LLM-based Evaluators through Multi-Agent Debate," *arXiv preprint* arXiv:2308.07201, 2023.

[30] S. Lin, J. Hilton, and O. Evans, "TruthfulQA: Measuring How Models Mimic Human Falsehoods," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, 2022.

[31] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu, "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions," *arXiv preprint* arXiv:2311.05232, 2023.

[32] C. Wang, X. Liu, Y. Yue, X. Tang, T. Zhang, C. Jiayang, Y. Yao, W. Gao, X. Hu, Z. Qi, and Y. Wang, "Survey on Factuality in Large Language Models: Knowledge, Retrieval and Domain-Specificity," *arXiv preprint arXiv:2310.07521*, 2023.

Biographies

Aijia Zhong is currently a junior majoring in Chinese Language and Literature at the International College of Languages and Cultures, Yunnan University of Finance and Economics, China, with research interests including linguistics, language symbol databases, and the application of artificial intelligence in language technology.

Lei Li is a lecturer at Jiangxi institute of Applied Science and Technology, China, with a master's degree, and his research area is computer science. He has presided over and participated in three provincial and university-level teaching reform projects, participated in one national-level cross-project, published three academic papers (one in SCI and one in EI), obtained one utility model patent, guided students to win the second prize of the national competition, three times at the third level, one first prize and one third prize in the provincial competition, and received the title of National Excellent Instructor.