Social Sciences
2026, Vol. 15, No. 4, pp. 223–233
https://doi.org/10.11648/j.ss.20261504.18

SciencePG
Science Publishing Group

Research Article

# Human-Coded Evaluation of Machine Translation and Large Language Model Agents for Chinese City Publicity Texts

**Xiongfei Wang***

Faculty of English Language and Culture, Guangdong University of Foreign Studies, Guangzhou, China

## Abstract

This study evaluates how web-based machine translation (MT) systems and large language model (LLM) translation agents perform in the English translation of Chinese city publicity texts. City publicity translation is a high-stakes form of institutional intercultural communication: it must be factually accurate, culturally legible, pragmatically appropriate, accessible to international readers, and capable of representing a city image without exaggeration or distortion. A corpus of 90 official Chinese source segments from Qingdao, Xi'an, and Hangzhou was translated under four conditions: DeepL web MT, Google Translate web MT, a GPT-5.5 translation agent, and a DeepSeek V4 Pro translation agent, yielding 360 English translations. Two trained coders independently evaluated all translations on five 1-5 dimensions: accuracy, cultural adequacy, pragmatic appropriateness, audience accessibility, and city image representation. Formal coding showed acceptable to strong reliability: Cohen's kappa for primary issue coding was 0.777, and quadratic weighted kappa values for the five rating dimensions ranged from 0.786 to 0.847. The strongest composite score was observed for DeepL web MT (M=4.683), followed by GPT-5.5 (M=4.599), DeepSeek V4 Pro (M=4.553), and Google Translate (M=4.261). Paired permutation tests showed that DeepL, GPT-5.5, and DeepSeek V4 Pro all significantly outperformed Google Translate on composite quality, while the differences between DeepL and the two LLM agents were not statistically significant. The findings therefore do not support a simple claim that LLM agents uniformly surpass MT. Instead, they suggest that LLM agents can reach a strong MT baseline and may offer pragmatic and audience-oriented affordances, while their value depends on the benchmark system, the target discourse function, and the evaluation dimension.

## Keywords

City Publicity Translation, Machine Translation, Large Language Models, Human Coding, Intercoder Reliability, Cultural Adequacy, Public Communication

## 1. Introduction

Chinese city publicity texts increasingly circulate in multilingual digital environments. Municipal profiles, tourism introductions, cultural heritage descriptions, and international communication materials do more than transmit information. They construct an outward-facing image of the city, mediate culturally dense local meanings, and address readers who may not share the historical, administrative, or symbolic knowledge assumed by the Chinese source text. English translation in this setting is therefore both a linguistic and an intercultural communication task. The translated text must remain faithful to the source, but it must also function as public-facing

*Correspondence: Xiongfei Wang (pencinus@qq.com)

**Received:** 26 June 2026; **Accepted:** 24 July 2026; **Published:** 13 August 2026

discourse for readers who may approach the city through tourism, investment, education, cultural exchange, or international news.

The rapid adoption of web machine translation (MT) and large language model (LLM)-based translation agents has changed how such texts are produced. Neural machine translation has become the dominant paradigm of contemporary MT, with sequence-to-sequence and Transformer architectures reshaping the field after the work of Bahdanau et al. [13], Wu et al. [14], and Vaswani et al. [15]. Commercial systems such as DeepL and Google Translate are accessible, fast, and widely used for draft translation. LLM agents, by contrast, can be prompted to consider audience, genre, public-facing tone, and cultural explanation. Studies of GPT-style models suggest that they can be competitive for high-resource translation directions, especially when prompt design and context are well controlled [9]. Yet the same literature also warns that LLM translation quality is uneven across languages, domains, and evaluation settings.

These differences make the city publicity genre a useful test case for artificial intelligence (AI)-mediated translation. If LLM agents have an advantage, it should appear not only in semantic accuracy but also in cultural adequacy, pragmatic appropriateness, accessibility, and city image representation. Such dimensions are not peripheral. A city publicity text may contain compressed institutional titles, scenic metaphors, policy slogans, cultural heritage references, or place-name histories. A literal translation can be intelligible yet unsuitable as external communication; an explanatory translation can be accessible yet over-expanded; and a fluent translation can still distort a historically sensitive term. These risks align with broader concerns in translation studies that quality cannot be reduced to lexical equivalence or surface fluency [1-3].

However, current discussion of LLM translation often moves faster than evidence. Claims about LLM superiority are frequently made without a strong human-coded baseline, without blinding, or without separating different quality dimensions. In public-sector international communication, such simplification is risky. A fluent output may still misrepresent institutional voice; a culturally explanatory output may introduce unsupported additions; and a literal but accurate output may still fail to serve international readers. This study therefore compares web MT and LLM-agent translation using a controlled corpus design and multidimensional human coding.

The contribution of this study is threefold. First, it provides a dataset-based comparison of four translation conditions for Chinese city publicity texts, a genre underrepresented in empirical MT and LLM evaluation. Second, it develops a human-coding framework that combines scalar quality dimensions with issue-code diagnosis, drawing on the logic of human evaluation and multidimensional quality assessment rather than relying on automatic metrics alone. Third, it shows that the relevant conclusion is not a categorical claim about "MT versus LLM", but a benchmark-sensitive finding: the two LLM agents significantly outperform Google Translate but do not significantly outperform DeepL.

## 2. Literature Review

### 2.1. From Neural Machine Translation to LLM Translation

Modern MT research has moved from phrase-based and statistical systems toward neural architectures. Bahdanau et al. [13] introduced attention-based neural machine translation, which addressed limitations in fixed-length sentence representation. Wu et al. [14] presented Google's neural machine translation system as a large-scale industrial approach to improving fluency and adequacy. Vaswani et al. [15] then introduced the Transformer architecture, which has become central to contemporary MT and LLM systems. Koehn [16] emphasizes that neural machine translation quality depends not only on model architecture but also on data, domain, language pair, and evaluation method.

LLM translation builds on this trajectory but differs from specialized MT systems in important ways. GPT-style models can be prompted for zero-shot or few-shot translation without being exclusively trained as translation engines. Hendy et al. [9] conducted a broad evaluation of GPT models for machine translation and found competitive performance in high-resource settings, while also noting limitations in low-resource and domain-shifted conditions. Kocmi and Federmann [10] further showed that LLMs can be powerful translation-quality evaluators under certain prompt settings, but their work is more directly about evaluation than production. Qian et al. [11] ask what information LLMs need for MT evaluation and report that reference translations, source information, and prompt design affect reliability. Qian et al. [12] similarly show that LLM-based quality estimation can be unstable in user-generated content, especially when outputs fail to follow the requested format.

These findings suggest caution in treating LLM translation as a uniform improvement over MT. LLMs may be more context-sensitive and flexible, but they also introduce new risks: prompt dependence, over-generation, refusals or format instability in evaluation settings, and plausible but unsupported elaboration. For city publicity translation, these risks matter because the target text must be persuasive but not invented, culturally accessible but not over-explained, and institutionally polished but not rhetorically inflated.

### 2.2. Human Evaluation and Multidimensional Quality Assessment

Human evaluation remains central to MT research because automatic metrics often fail to capture the full range of translation quality. Bilingual Evaluation Understudy (BLEU), introduced by Papineni et al. [17], has been influential because

it provides a low-cost corpus-level comparison against reference translations. Yet BLEU is limited for public-facing translation because it does not directly measure pragmatic appropriateness, cultural mediation, institutional tone, or city image representation. Freitag et al. [4] argue that human evaluation of high-quality MT is difficult and that inadequate evaluation procedures can lead to incorrect system rankings. Their MQM-based study highlights the importance of expert annotation, error typologies, and document context.

The broader Multidimensional Quality Metrics (MQM) tradition provides a useful model for analytic translation evaluation. Lommel et al. [5] describe MQM as a framework that combines error typology with scoring models, enabling fine-grained assessment beyond general fluency or adequacy. Recent work has continued to adapt MQM-like approaches to different discourse conditions. Li et al. [18] develop MQM-Chat for chat translation, emphasizing that stylized and dialogue-specific content requires tailored evaluation. Magdy et al. [19] propose a linguistically motivated multidimensional framework for dialect- and culture-sensitive MT evaluation, arguing that language-agnostic metrics may miss pragmatic and sociolinguistic errors.

This study does not implement full MQM span-level annotation, but it follows the same analytic principle: translation quality should be decomposed into interpretable dimensions and error categories. The five score dimensions in the present study correspond to the needs of city publicity translation rather than to a generic MT benchmark. Accuracy captures source fidelity; cultural adequacy captures culture-specific transfer; pragmatic appropriateness captures institutional and public-facing tone; audience accessibility captures target-reader legibility; and city image representation captures the public communication function of the translation.

Intercoder reliability is also essential. Cohen's kappa and weighted kappa are widely used to assess agreement beyond chance for categorical and ordinal judgments [6-8]. In translation evaluation, reliability is especially important because coders may differ in expectations about literalness, fluency, cultural explanation, and acceptable adaptation. Licht et al. [20] note that human evaluation across language pairs can suffer from variability, and they propose calibrated methods to improve consistency. The present study therefore reports reliability before interpreting condition-level results.

## 2.3. Cultural and Pragmatic Dimensions

Translation studies has long treated translation as intercultural mediation rather than word substitution. Katan [2] argues that translators mediate across cultural frames, while House [1] emphasizes that translation quality assessment must consider register, genre, and function. Munday et al. [3] similarly present translation as a field in which linguistic choices are bound to communicative purpose and social context. O'Brien [25] frames translation increasingly as human-computer interaction, which is relevant here because AI-assisted translation changes how human evaluators, tools, and institutional communication goals interact.

Culture-specific translation remains difficult for MT systems. Tang [21] shows that Chinese idiom translation is challenging even when datasets are constructed to support machine and human learning. Yao et al. [22] propose culturally aware MT benchmarking and argue that culture-specific items require pragmatic assessment, not only semantic matching. Wang et al. [23] find that Google Translate can struggle with Chinese cultural and historical expressions when compared with human expert translations. These studies support the need for evaluating Chinese city publicity translation through cultural and pragmatic dimensions.

Publicity-oriented translation adds another layer. Jiang and Zhang [24] examine hedges in translations of publicity-oriented documents and show that translated institutional discourse can shift in frequency and strategy across time and directionality. Although their focus differs from the present study, their work supports the view that publicity translation is a register-sensitive domain in which rhetorical and pragmatic choices matter. For Chinese city publicity texts, the translation must often negotiate between official discourse, tourism appeal, and international readability. A direct translation of a phrase such as "golden name card" may preserve the metaphoric form but fail as idiomatic public-facing English; a more adaptive translation may improve readability but risk over-explicitation or promotional exaggeration.

## 2.4. Research Gap

The literature supports three conclusions. First, MT and LLM translation systems must be evaluated against strong baselines and in specific domains rather than as broad technology categories. Second, human-coded multidimensional evaluation remains necessary when the target task includes cultural, pragmatic, or institutional functions. Third, Chinese publicity translation provides an important but underexamined test case because it combines factual reference, cultural representation, and public image construction.

This study addresses the gap by comparing DeepL, Google Translate, GPT-5.5, and DeepSeek V4 Pro on the same set of Chinese city publicity source texts. It asks whether LLM agents outperform web MT systems, but it does so in a way that allows the answer to be benchmark-sensitive rather than predetermined.

## 2.5. Research Questions

This study addresses three research questions (RQs). The questions separate system comparison from broad categorical claims about MT and LLMs because the category label itself is less informative than the performance of a specific system under a specific translation task.

RQ1. How do DeepL, Google Translate, GPT-5.5, and DeepSeek V4 Pro compare across accuracy, cultural adequacy,

pragmatic appropriateness, audience accessibility, and city image representation?

RQ2. Which translation problems are most salient across the four translation conditions?

RQ3. Do LLM translation agents outperform web MT systems, or do their advantages depend on the benchmark system and evaluation dimension?

# 3. Method

## 3.1. Corpus and Translation Conditions

The corpus contains 90 Chinese city publicity segments from Qingdao, Xi'an, and Hangzhou. The texts cover four broad text types: city profiles, tourism publicity, cultural heritage, and international communication. These texts were selected because they combine factual information with public-facing representation. Many contain local place names, administrative expressions, historical events, scenic metaphors, and culturally loaded descriptions that require more than literal semantic transfer.

Each source segment was translated under four conditions: DeepL web MT, Google Translate web MT, GPT-5.5 LLM agent, and DeepSeek V4 Pro LLM agent. The full design produced 360 translation records. The two web MT systems were used as mainstream accessible translation baselines. The two LLM-agent conditions were included to test whether prompt-able agents can provide better audience adaptation and cultural-pragmatic mediation.

The design is paired by source segment: every Chinese source has four English translations. This makes it possible to compare systems while holding source difficulty constant. It also matches the practical evaluation problem faced by institutions: given the same Chinese publicity text, which AI-assisted output is more suitable for public-facing English communication? The materials consisted of publicly available city publicity texts and system-generated translations; no personal data, sensitive private records, or participant-identifiable information were collected. AI tools were used as the evaluated translation conditions and as limited writing-support tools for language polishing and manuscript formatting, while all research design decisions, coding procedures, data interpretation, and final manuscript responsibility remained with the author.

## 3.2. Blinded Human Coding

Two coders independently evaluated all 360 translations. The coding workbooks were blinded: coders saw the Chinese source text, English output, city, text type, and a blinded item identifier, but not the translation condition. For each source segment, the four translations remained grouped but were randomly ordered within the group. The condition mapping was held in a separate file and used only after independent coding was complete.

Each translation was rated on five 1-5 dimensions: accuracy, cultural adequacy, pragmatic appropriateness, audience accessibility, and city image representation. Coders also assigned a primary issue code and a secondary issue code. The issue-code inventory included None, terminology error, omission, addition, literalism, cultural opacity, pragmatic mismatch, factual drift, over-explicitation, readability problem, encoding/format artifact, and Other.

The first pilot coding round was discarded after the coder pair was changed. A second pilot round with the new pair showed sufficient agreement to proceed. The present results are based only on the formal 360-record coding and do not use the superseded first-round pilot data.

## 3.3. Reliability and Statistical Analysis

Intercoder reliability was calculated before any reconciliation. Cohen's kappa was used for categorical issue-code agreement. Quadratic weighted kappa and intraclass correlation coefficient (ICC) were used for ordinal 1-5 score dimensions. For condition-level score analysis, the two coders' ratings were averaged for each translation record. A composite score was calculated as the mean of the five-dimension scores. Issue-code distributions were treated as diagnostic evidence pending final adjudication, while scalar ratings were used for the main condition comparison because their intercoder reliability was consistently high.

Because each condition translated the same 90 source segments, condition comparisons used paired source-level contrasts. Pairwise composite differences were summarized using mean paired difference, Cohen's dz, and two-tailed paired permutation tests based on random sign flipping. This procedure is appropriate for the within-source design and avoids relying on distributional assumptions that may be fragile with bounded 1-5 ratings.

## 3.4. Operationalization of Publicity-Translation Quality

The five rating dimensions were designed to reflect both translation studies theory and the communicative requirements of city publicity. Accuracy was defined as the degree to which propositional content, named entities, temporal relations, institutional roles, and causal relations were preserved. A translation could be fluent and still receive a lower accuracy score if it changed a historical relation, mistranslated an event name, or omitted a limiting condition from the Chinese source. Cultural adequacy was defined as the handling of culturally marked expressions, place-based symbolism, historical references, idioms, heritage terms, and metaphoric language. This dimension therefore covered not only whether a cultural item was translated, but also whether the target-language rendering made the cultural item interpretable without flattening its local specificity.

Pragmatic appropriateness was defined as the fit between

the English translation and the institutional-publicity function of the source text. City publicity texts often occupy a mixed register: they are informative, promotional, culturally representative, and institutionally cautious at the same time. A translation could lose points on this dimension if it sounded like a mechanical tourist note, an inflated advertising slogan, or an informal travel blog when the source required public-institutional style. Audience accessibility was defined from the perspective of international readers who do not necessarily know Chinese administrative categories, place-name histories, or culturally compressed expressions. This dimension rewarded clear, idiomatic, and reader-oriented English, while penalizing calques that preserved Chinese form at the expense of target-reader comprehension. City image representation was defined as the extent to which the translation supported a credible and proportionate representation of the city. This dimension was deliberately separated from pragmatic appropriateness because a translation may be grammatically and pragmatically acceptable while still weakening, exaggerating, or distorting the intended public image.

The primary-issue code complemented the five scores by requiring coders to identify the most salient problem in each translation. This design prevents the scalar ratings from becoming impressionistic general quality judgments. For example, two translations may receive similar composite scores for different reasons: one may contain a small factual drift in an otherwise elegant sentence, while another may be accurate but culturally opaque and awkward for international readers. The issue-code layer makes these differences analyzable. It also aligns the study with human-evaluation and MQM-inspired approaches in which error diagnosis and quality scoring are treated as related but not identical forms of evidence [4, 5].

This operationalization also responds to a limitation in many comparisons of LLMs and MT systems. If evaluation relies only on overall preference or a single adequacy-fluency score, the result may reward surface fluency and under-detect cultural or pragmatic risk. The present rubric therefore makes the evaluation task harder but more relevant to the target genre. A system that performs well under this framework must preserve source meaning, avoid unsupported elaboration, communicate to international readers, and maintain a credible institutional voice. These demands are precisely why city publicity translation is a useful test case for AI-mediated translation.

# 4. Results

## 4.1. Data Completion and Reliability

All 360 translation records were successfully matched across Coder A, Coder B, and the randomization mapping table. The final dataset contains 90 source segments, four translations per source, and 90 translations per condition.

Table 1 reports the core intercoder reliability results. Primary issue-code agreement was substantial (kappa=0.777). The five scalar dimensions also showed acceptable to strong reliability, with quadratic weighted kappa values ranging from 0.786 for accuracy to 0.847 for cultural adequacy. These results support the use of coder-averaged ratings for condition-level analysis.

*Table 1. Intercoder reliability summary.*

| Measure | Result |
|---|---|
| Primary issue code | Cohen's kappa=.777; N=360 |
| Accuracy | Quadratic weighted kappa=.786 |
| Cultural adequacy | Quadratic weighted kappa=.847 |
| Pragmatic appropriateness | Quadratic weighted kappa=.822 |
| Audience accessibility | Quadratic weighted kappa=.809 |
| City image representation | Quadratic weighted kappa=.834 |

The reliability pattern is important for the interpretation of the study. It indicates that the coders did not merely agree on obvious adequacy or fluency judgments; they also achieved consistent ratings on culturally and pragmatically oriented dimensions. This supports the methodological claim that city publicity translation can be evaluated through a multidimensional human-coding framework.

## 4.2. Condition-Level Quality Scores

Table 2 reports composite scores by translation condition. DeepL achieved the highest composite mean (M=4.683), followed by GPT-5.5 (M=4.599), DeepSeek V4 Pro (M=4.553), and Google Translate (M=4.261). The ranking qualifies a simple LLM-versus-MT narrative. The two LLM agents clearly exceeded Google Translate, but neither exceeded DeepL on composite quality.

*Table 2. Composite scores by condition.*

| Condition | Composite mean |
|---|---|
| DeepL web machine translation | 4.683 |
| Google Translate web machine translation | 4.261 |
| GPT-5.5 large language model agent | 4.599 |
| DeepSeek V4 Pro large language model agent | 4.553 |

The dimension-level scores further clarify the pattern. DeepL received the highest mean for accuracy (4.850), pragmatic appropriateness (4.633), and city image representation (4.683). DeepSeek V4 Pro was close to DeepL on cultural adequacy (4.506 vs. 4.567) and audience accessibility (4.611 vs.

4.683). GPT-5.5 performed strongly on accuracy (4.811), audience accessibility (4.594), and city image representation (4.617). Google Translate was consistently lowest across all five dimensions, especially cultural adequacy (4.122) and pragmatic appropriateness (4.183).

Dimension profile by condition:

DeepL web MT: Accuracy=4.850; Cultural adequacy=4.567; Pragmatic appropriateness=4.633; Audience accessibility=4.683; City image representation=4.683.

Google Translate web MT: Accuracy=4.544; Cultural adequacy=4.122; Pragmatic appropriateness=4.183; Audience accessibility=4.217; City image representation=4.239.

GPT-5.5 LLM agent: Accuracy=4.811; Cultural adequacy=4.428; Pragmatic appropriateness=4.544; Audience accessibility=4.594; City image representation=4.617.

DeepSeek V4 Pro LLM agent: Accuracy=4.750; Cultural adequacy=4.506; Pragmatic appropriateness=4.411; Audience accessibility=4.611; City image representation=4.489.

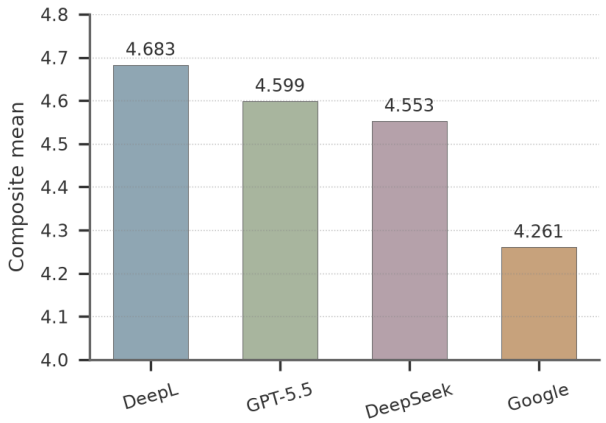


***Figure 1.*** *Composite Quality Scores Across Four Translation Conditions.*

## 4.3. Paired Composite Comparisons

Table 3 reports paired permutation tests for composite scores. DeepL significantly outperformed Google Translate (mean difference=0.422, dz=0.648, p<0.0001). GPT-5.5 and DeepSeek V4 Pro also significantly outperformed Google Translate. However, DeepL did not significantly differ from GPT-5.5 (p=0.216), and its difference from DeepSeek V4 Pro was only marginal (p=0.0646). GPT-5.5 and DeepSeek V4 Pro were not meaningfully different from each other (p=0.5325).

***Table 3.*** *Paired permutation tests for composite score.*

| Paired contrast | Result |
|---|---|
| DeepL vs. Google Translate | Difference=+0.422; dz=0.648; p<.0001 |
| DeepL vs. GPT-5.5 | Difference=+0.084; dz=0.133; p=.216 |
| DeepL vs. DeepSeek V4 Pro | Difference=+0.130; dz=0.201; p=.0646 |
| Google Translate vs. GPT-5.5 | Difference=-0.338; dz=-0.484; p<.0001 |
| Google Translate vs. DeepSeek V4 Pro | Difference=-0.292; dz=-0.401; p=.0004 |
| GPT-5.5 vs. DeepSeek V4 Pro | Difference=+0.046; dz=0.067; p=.5325 |

These results indicate that the main empirical contrast is not "LLM agents versus MT" as a broad category. Instead, the weaker condition is Google Translate, while DeepL and the two LLM-agent conditions form a stronger group with relatively small internal differences. This finding is important for applied translation workflows because it shows that the value of LLM translation should be judged against a strong MT baseline rather than against MT in general.

## 4.4. Qualitative Case Evidence

To make the score patterns interpretable, three coder-agreed cases were examined qualitatively. These examples are not used as isolated proof of system quality; rather, they illustrate the kinds of issues reflected in the aggregate scores.

***Table 4.*** *Representative qualitative cases.*

| Case | Coder-agreed interpretation |
|---|---|
| F069, Google Translate | Literalism; composite score=3.3; the metaphor was transferred too literally for public-facing English. |
| F020, DeepL | Factual drift; composite score=2.7; a historically dense item was rendered inaccurately. |
| F044, DeepSeek V4 Pro | Over-explicitation; composite score=3.6; the output added promotional elaboration beyond the source. |

The examples clarify why the quantitative findings should not be read as a simple fluency ranking. Google Translate's lower score is partly associated with literal calques and public-facing awkwardness. DeepL's strong average performance does not remove the need for fact checking. LLM-agent outputs can sound rhetorically polished, but this polish may involve additional evaluative language or over-expansion.

These qualitative observations support the paper's central interpretation: AI-mediated city publicity translation should be assessed as a balance among accuracy, cultural mediation, audience accessibility, and institutional voice.

### 4.5. Issue-Code Evidence

Issue-code reliability was substantial, but 48 issue-code disagreements remained. Two score disagreements of two points or more were also flagged. For this reason, the present revision treats issue-code results as diagnostic rather than final. The coder-level issue distribution already suggests a consistent qualitative pattern: Google Translate produced more visible problem codes, especially literalism, terminology error, readability problem, and factual drift. DeepL had the highest proportion of None primary codes. The LLM-agent outputs reduced many Google-type surface problems but still require close checking for additions, over-explicitation, and pragmatic shifts.

The next version should include an adjudicated issue-code table. At the current stage, the scalar ratings are sufficiently reliable for condition-level analysis, while issue-code interpretation should remain cautious until reconciliation is complete.

## 5. Discussion

### 5.1. Benchmark Choice Matters

The formal results refine the study's central claim. LLM agents did not uniformly outperform web MT. Rather, they performed substantially better than Google Translate and approximately at the level of DeepL. The qualitative cases further show why this pattern should be interpreted as a benchmark-sensitive result rather than a technology-category result. If Google Translate is the MT benchmark, the evidence supports a clear LLM-agent advantage. If DeepL is the benchmark, the evidence supports LLM-agent parity rather than LLM-agent superiority.

This distinction matters for both research and practice. In research, it warns against treating "MT" as a homogeneous category. In practice, it suggests that institutions should not evaluate LLM tools against weak baselines only. A city government or cultural institution deciding whether to use an LLM agent should compare it with the strongest available MT workflow, not simply with a generic web translation output.

### 5.2. Accuracy Alone Is Insufficient

The results also support the multidimensional evaluation framework. Accuracy scores were high across all four systems, which suggests that accuracy alone would obscure important differences. The qualitative cases also show that high fluency may coexist with literalism, factual drift, or over-explicitation. The largest practical separation appeared in cultural adequacy, pragmatic appropriateness, audience accessibility, and city image representation. These dimensions are central to international city communication because the task is not merely to avoid mistranslation but to make the city legible, credible, and institutionally appropriate for international readers.

This finding aligns with cultural and pragmatic approaches to translation quality. City publicity translation is not a domain in which literal transfer is automatically desirable. It requires judgment about when to preserve local specificity, when to explain it, and when to avoid adding unsupported promotional language. The "golden name card" example illustrates the cost of literal metaphor transfer. The "Sino-German War" example illustrates the risk of historically sensitive factual drift. The "legendary 72 Valleys" example illustrates the risk of rhetorical over-expansion. These are different problems, and a single automatic metric would struggle to represent their communicative significance.

### 5.3. Practical Workflow Implications

The findings have practical implications for municipal translation workflows. Google Translate may be useful for rough comprehension, but the present data do not support relying on it for polished city publicity translation. DeepL appears to be a strong baseline for English drafting. LLM agents are promising for controlled adaptation, especially when prompts specify audience, genre, and public-facing tone. Nevertheless, human review remains necessary because high average scores can coexist with localized risks such as factual drift, unsupported addition, or over-explicitation.

A defensible workflow would therefore be hybrid. For low-risk factual segments, DeepL or an LLM agent may generate an initial draft. For culturally dense, historically sensitive, or image-building passages, a human translator or editor should review terminology, factual relations, and public-facing tone. LLM agents may be especially useful in generating alternative phrasings or audience-adapted versions, but their suggestions should be checked against the Chinese source and, where available, official English references.

### 5.4. Methodological Contribution

The study also contributes methodologically. The coding framework performed reliably across two independent coders and 360 translation records. The high weighted kappa values suggest that dimensions such as cultural adequacy, audience accessibility, and city image representation can be operationalized for empirical evaluation. This is important because public-sector translation quality often remains discussed in broad normative terms. A transparent coding framework makes the evaluation more reproducible and allows translation technology claims to be tested against human judgments.

The study also shows the value of pairing quantitative and qualitative evidence. The condition means identify system-

level patterns, while the case evidence shows why those patterns matter. Together, they support a more precise conclusion than either method alone.

## 5.5. Implications for LLM-Agent Evaluation

The findings have implications for how LLM translation agents should be evaluated. First, an LLM-agent condition should not be treated as a monolithic technology. In practical use, the output depends on the model, prompt, context window, system instruction, temperature setting, and the human operator's ability to specify genre and audience. This means that future studies should report the agent configuration as part of the method, just as MT studies report system versions and language directions where possible. Without this information, the comparison becomes difficult to reproduce and the claim of superiority becomes too broad.

Second, the results suggest that the most relevant comparison is not between "traditional MT" and "LLMs" in the abstract, but between specific workflows. DeepL as a strong web MT baseline produced the highest mean score in this dataset, while the two LLM agents performed close to DeepL and substantially better than Google Translate. This pattern is consistent with recent work showing that LLMs can be strong translation systems and evaluators, but that their performance varies with task setting and evaluation design [9-12]. For a city-publicity workflow, the practical question is therefore whether an LLM agent adds value beyond a strong MT draft plus human editing. The current evidence suggests that the answer is conditional: LLM agents are useful when they are prompted for audience adaptation and alternative phrasing, but they do not remove the need for source-based verification.

Third, the study suggests that evaluation should include negative affordances as well as positive affordances. LLM agents are often praised for producing natural and rhetorically polished target texts. In city publicity translation, however, rhetorical polish can become a source of risk if it introduces scenic adjectives, evaluative intensifiers, or explanatory content not licensed by the source. This does not mean that LLM agents are unsuitable. It means that their strengths must be governed by a review protocol that distinguishes legitimate adaptation from unsupported addition. In this sense, the LLM agent is better understood as an adaptive drafting partner than as an autonomous replacement for human translation judgment.

Finally, the study contributes to a broader methodological debate about automatic metrics and human evaluation. Automatic metrics are useful for large-scale system development, but they are not sufficient for institutional-publicity translation because the target of evaluation includes reader uptake, cultural legibility, and public image. A multidimensional human-coded design is more expensive, but it produces evidence that is closer to the communicative risks faced by local governments, tourism bureaus, museums, and cultural institutions. This is especially important for Chinese-English translation, where culturally compressed source expressions can be formally short but pragmatically dense. The present results therefore support a mixed evaluation model: automatic or tool-based screening may be used for efficiency, but final claims about public-facing quality should rest on human evaluation with transparent dimensions and reliability reporting.

## 5.6. Implications for Corpus-Based Translation Research

The paired design also has implications for corpus-based translation research. Much discussion of AI translation relies on isolated examples: a striking mistranslation, a polished LLM rewrite, or a favorable comparison between one tool and one human-preferred output. Such examples are useful for illustrating mechanisms, but they cannot establish system-level tendencies. The present corpus design addresses this limitation by keeping the source segment constant across four translation conditions. This makes the comparison sensitive to source difficulty. A culture-loaded scenic metaphor, a historically dense sentence, or an institutional slogan is difficult for all systems; the paired design asks which system handles that same difficulty more successfully.

At the same time, the results show why corpus size and annotation depth must be balanced. A larger corpus would provide more statistical power and broader coverage, but detailed human coding becomes costly. A smaller corpus allows careful qualitative interpretation, but risks overgeneralization. The present dataset of 90 source segments and 360 translations is therefore best understood as a focused comparative corpus rather than a universal benchmark. Its value lies in the combination of paired design, blinded coding, reliability reporting, and dimension-specific analysis. Future studies could extend this design by adding more cities, more genres, and more language directions while preserving the same paired and blinded structure.

The findings also support a stronger distinction between "translation output" and "translation workflow." A corpus can compare final outputs, but public-sector translation is normally a workflow involving drafting, editing, verification, and approval. An LLM agent may underperform as an unreviewed final translator but still improve the workflow by generating paraphrase options, explaining cultural references, or helping editors identify audience-accessibility problems. Conversely, a strong MT system may produce the best first draft but provide less interactive support during revision. Future corpus studies should therefore consider adding post-editing time, editor preference, revision logs, and final publishability judgments. Such measures would connect product-oriented evaluation with process-oriented translation research.

Finally, the study suggests that culturally oriented evaluation should not be reduced to counting culture-specific terms. Cultural adequacy in city publicity translation is relational: it depends on the source expression, the presumed international reader, the communicative function, and the city's desired

public image. A term may need literal preservation in a heritage context, paraphrase in a tourism context, and institutional explanation in an investment or international-exchange context. This is one reason why the present rubric separates cultural adequacy, pragmatic appropriateness, audience accessibility, and city image representation. This study also treats MT use as part of translation literacy and workflow design [26-28]. These dimensions overlap in real reading but separating them analytically helps identify whether a translation fails because it is wrong, opaque, awkward, inflated, or image-damaging. This distinction is essential for building an evidence-based account of AI-mediated publicity translation.

## 6. Conclusion and Limitations

### 6.1. Conclusion

This study compared DeepL web MT, Google Translate web MT, GPT-5.5 LLM-agent translation, and DeepSeek V4 Pro LLM-agent translation for Chinese city publicity texts using 360 blinded human-coded translation records. The results show reliable human evaluation and meaningful condition-level differences. DeepL achieved the highest composite score, while GPT-5.5 and DeepSeek V4 Pro significantly outperformed Google Translate but did not significantly outperform DeepL. The most defensible conclusion is therefore not that LLM agents simply surpass MT, but that LLM agents can reach a strong MT baseline and provide useful affordances for culturally and pragmatically sensitive translation when combined with human review.

For international city communication, the relevant question is not whether AI can translate, but which AI-assisted workflow can preserve factual accuracy, cultural meaning, public-facing tone, audience accessibility, and credible city image at the same time. The evidence in this study supports a benchmark-sensitive, human-supervised approach to AI-mediated translation.

The broader implication is that translation technology evaluation should move from tool enthusiasm to task-specific accountability. For city publicity translation, the final product is not merely an English sentence but an institutional representation of place, memory, culture, and credibility. AI systems can accelerate drafting and support revision, but their outputs still require human judgment capable of reading both the Chinese source and the international communicative situation. A responsible workflow should therefore combine strong AI drafting, transparent evaluation criteria, documented human review, and post-editing decisions that can be justified against the source text and the intended public function.

### 6.2. Limitations

Several limitations remain. The most immediate limitation is that issue-code disagreements have not yet undergone final adjudication, so issue-code patterns are treated as diagnostic rather than definitive in this revision. The scalar score results are reliable enough for the main condition comparison, but the issue-code results should be finalized after reconciliation.

Second, the source corpus covers three Chinese cities and four text types. Broader geographic and genre coverage is needed before generalizing to all city publicity translation. Third, the rating scale produced generally high scores, creating a possible ceiling effect. Future work could add more fine-grained criteria, span-level error annotation, or post-editing effort measures. Fourth, LLM-agent performance depends on the prompt, model version, and task framing. Finally, official English reference texts were available only for part of the corpus, so the study relies mainly on source-based human evaluation rather than a full official-reference benchmark.

## Abbreviations

| | |
|---|---|
| AI | Artificial Intelligence |
| BLEU | Bilingual Evaluation Understudy |
| DOI | Digital Object Identifier |
| GPT | Generative Pre-trained Transformer |
| ICC | Intraclass Correlation Coefficient |
| LLM | Large Language Model |
| MQM | Multidimensional Quality Metrics |
| MT | Machine Translation |
| RQ | Research Question |

## Author Contributions

**Xiongfei Wang:** Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing

## Funding

This work was supported by the 2024 Postgraduate Research and Innovation Projects at Guangdong University of Foreign Studies under the project Metaphor Interpretation and Translation of Large Language Models (Grant No. 24GWCXXM-016).

## Data Availability Statement

The anonymized corpus coding tables, randomization mapping, codebook, and statistical analysis outputs generated for this study are publicly available on Zenodo at Digital Object Identifier (DOI): 10.5281/zenodo.21133206. The released materials retain the research data needed to reproduce the aggregate results.

## Conflicts of Interest

The author declares no conflict of interest.

## References


[1] House, J. (2015). Translation Quality Assessment: Past and Present. Routledge.

[2] Katan, D. (2004). Translating Cultures: An Introduction for Translators, Interpreters and Mediators (2nd ed.). St. Jerome/Routledge. https://doi.org/10.4324/9781315759692

[3] Munday, J., Ramos Pinto, S., & Blakesley, J. (2022). Introducing Translation Studies: Theories and Applications (5th ed.). Routledge. https://doi.org/10.4324/9780429352461

[4] Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., & Macherey, W. (2021). Experts, errors, and context: A large-scale study of human evaluation for machine translation. Transactions of the Association for Computational Linguistics, 9, 1460-1474. https://doi.org/10.1162/tacl_a_00437

[5] Lommel, A., Gladkoff, S., Melby, A., Wright, S. E., Strandvik, I., Gasova, K., Vaasa, A., Benzo, A., Marazzato Sparano, R., Foresi, M., Innis, J., Han, L., & Nenadic, G. (2024). The multi-range theory of translation quality measurement: MQM scoring models and statistical quality control. arXiv preprint arXiv:2405.16969. https://arxiv.org/abs/2405.16969

[6] Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37-46. https://doi.org/10.1177/001316446002000104

[7] Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin, 70(4), 213-220. https://doi.org/10.1037/h0026256

[8] Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159-174. https://doi.org/10.2307/2529310

[9] Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y. J., Afify, M., & Awadalla, H. H. (2023). How good are GPT models at machine translation? A comprehensive evaluation. arXiv preprint arXiv:2302.09210. https://arxiv.org/abs/2302.09210

[10] Kocmi, T., & Federmann, C. (2023). Large language models are state-of-the-art evaluators of translation quality. arXiv preprint arXiv:2302.14520. https://arxiv.org/abs/2302.14520

[11] Qian, S., Orasan, C., Kanojia, D., & do Carmo, F. (2024a). Are large language models state-of-the-art quality estimators for machine translation of user-generated content? arXiv preprint arXiv:2410.06338. https://arxiv.org/abs/2410.06338

[12] Qian, S., Sindhujan, A., Kabra, M., Kanojia, D., Orasan, C., Ranasinghe, T., & Blain, F. (2024b). What do large language models need for machine translation evaluation? arXiv preprint arXiv:2410.03278. https://arxiv.org/abs/2410.03278

[13] Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. International Conference on Learning Representations. https://arxiv.org/abs/1409.0473

[14] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., & Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144. https://arxiv.org/abs/1609.08144

[15] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30. https://papers.nips.cc/paper/7181-attention-is-all-you-need

[16] Koehn, P. (2020). Neural Machine Translation. Cambridge University Press.

[17] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: A method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 311-318. https://doi.org/10.3115/1073083.1073135

[18] Li, Y., Suzuki, J., Morishita, M., Abe, K., & Inui, K. (2024). MQM-Chat: Multidimensional Quality Metrics for Chat Translation. arXiv preprint arXiv:2408.16390. https://arxiv.org/abs/2408.16390

[19] Magdy, S. M., Alwajih, F., El Mekki, A., El-Sayed, W., & Abdul-Mageed, M. (2026). LQM: Linguistically motivated multidimensional quality metrics for machine translation. arXiv preprint arXiv:2604.18490. https://arxiv.org/abs/2604.18490

[20] Licht, D., Gao, C., Lam, J., Guzman, F., Diab, M., & Koehn, P. (2022). Consistent human evaluation of machine translation across language pairs. arXiv preprint arXiv:2205.08533. https://arxiv.org/abs/2205.08533

[21] Tang, K. (2022). PETCI: A parallel English translation dataset of Chinese idioms. arXiv preprint arXiv:2202.09509. https://arxiv.org/abs/2202.09509

[22] Yao, B., Jiang, M., Bobinac, T., Yang, D., & Hu, J. (2023). Benchmarking machine translation with cultural awareness. arXiv preprint arXiv:2305.14328. https://arxiv.org/abs/2305.14328

[23] Wang, X., Beard, R., & Chandra, R. (2024). Evaluation of Google Translate for Mandarin Chinese translation using sentiment and semantic analysis. arXiv preprint arXiv:2409.04964. https://arxiv.org/abs/2409.04964

[24] Jiang, Z., & Zhang, Z. (2023). Hedges in bidirectional translations of publicity-oriented documents. arXiv preprint arXiv:2305.12146. https://arxiv.org/abs/2305.12146

[25] O'Brien, S. (2012). Translation as human-computer interaction. Translation Spaces, 1(1), 101-122. https://doi.org/10.1075/ts.1.05obr

[26] Bowker, L. (2020). Machine translation literacy instruction for international business students and business English instructors. Journal of Business & Finance Librarianship, 25(1-2), 25-43. https://doi.org/10.1080/08963568.2020.1794739

[27] Zouhar, V., Tamchyna, A., Popel, M., & Bojar, O. (2021). Neural machine translation quality and post-editing performance. arXiv preprint arXiv:2109.05016. https://arxiv.org/abs/2109.05016

[28] Ye, Y., & Toral, A. (2020). Fine-grained human evaluation of Transformer and recurrent approaches to neural machine translation for English-to-Chinese. arXiv preprint arXiv:2006.08297. https://arxiv.org/abs/2006.08297