IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 43, No. 2, April, 2026, pp. 981-990
Journal homepage: http://iieta.org/journals/ts

Multi-Channel Fidelity Modeling of Transformer-Based Language Systems for Biomedical Translation under a Signal Processing Framework

Jing Zhu

School of Culture and Education, Shaanxi University of Science and Technology, Xi'an 710021, China

Corresponding Author Email: smilinglemon@126.com

https://doi.org/10.18280/ts.430234

Received: 6 October 2025
Revised: 20 January 2026
Accepted: 12 February 2026
Available online: 30 April 2026

Keywords:
signal transformation modeling, multi-channel signal fidelity, transformer systems, spectral component preservation, amplitude consistency, biomedical signal representation, statistical signal analysis

ABSTRACT

Machine translation (MT) can be interpreted as a structured signal transformation process in which linguistic sequences are mapped across heterogeneous symbolic domains. However, conventional evaluation approaches rely predominantly on single-channel similarity metrics and fail to characterize the multi-dimensional nature of fidelity in domain-specific translation tasks. This study establishes a multi-channel signal fidelity framework for assessing transformer-based translation systems in biomedical contexts. The framework decomposes translation fidelity into three orthogonal components: global structural fidelity, domain-specific spectral fidelity, and numerical amplitude preservation. Six translation systems, including four large-scale transformer models and two neural MT baselines, are evaluated on 100 biomedical sentences derived from the WMT19 benchmark. The results indicate that all transformer-based systems achieve higher global structural fidelity than neural MT baselines in terms of BLEU scores (range: 37.35–38.17 vs. 35.86–36.51), although no statistically significant differences are observed ($p > 0.05$). Input conditioning introduces model-dependent perturbations: one system exhibits a significant fidelity decrease under conditioned input (ΔBLEU = −0.71, $p = 0.004$), whereas others show moderate gains (up to +1.32). In contrast, domain-specific spectral fidelity and numerical amplitude preservation remain consistently high across all systems (terminology accuracy: 0.937–0.975; numerical consistency: 0.980–0.990). Correlation analysis further demonstrates that global structural metrics exhibit near-zero association with domain-specific fidelity components (Spearman $r = 0.12$), confirming that single-channel evaluation cannot adequately represent translation signal integrity. These findings indicate that translation systems operating under different architectures converge toward a shared global fidelity ceiling, while retaining distinct behaviors in domain-specific signal components. The proposed framework provides a structured signal processing perspective for evaluating complex symbolic transformation systems.

1. INTRODUCTION

From the perspective of signal processing theory, machine translation (MT) can be formally conceived as a cross-linguistic signal transformation problem: a source-language text constitutes a structured information signal, and the translation system operates as a signal transducer that encodes, transforms, and decodes this signal into a target-language representation [1, 2]. Under this framing, translation quality is fundamentally a fidelity problem—the degree to which the output signal preserves the informational content, structural patterns, and domain-specific features of the input signal. Shannon's foundational work on information theory [3] established that any transmission channel incurs potential information loss; in MT, this loss manifests as lexical substitution errors, syntactic distortions, and—most critically in specialized domains—failure to preserve high-frequency domain-specific terminology components.

Biomedical texts represent a particularly demanding signal transformation scenario: they exhibit high-density technical terminology (e.g., DNA, MRI, PCR) functioning as high-frequency spectral components that carry disproportionate semantic load, complex passive-voice constructions with deeply nested clausal structures, and strict numerical fidelity requirements analogous to amplitude preservation in signal reconstruction [4, 5]. These properties make biomedical translation a rigorous testbed for evaluating the multi-channel fidelity characteristics of translation systems [6].

The emergence of large language models (LLMs)—themselves built on attention-based signal processing architectures [7]—has substantially altered the MT landscape. Recent studies report that LLMs approach or surpass commercial NMT performance on high-resource language pairs [8, 9], while others note persistent domain mismatch and rare word prediction challenges in specialized settings [10]. This divergence reflects a critical methodological gap: most comparative studies rely on single-channel evaluation (BLEU alone), which captures only global waveform overlap while remaining orthogonal to domain-critical quality dimensions [11].

This study addresses these gaps through four research questions: (RQ1) Do LLMs and NMT systems differ significantly in biomedical text signal transformation fidelity? (RQ2) Can few-shot input conditioning with domain glossaries improve fidelity across all evaluation channels? (RQ3) How stable is transformation fidelity across medical sub-domain signal distributions? (RQ4) What is the inter-channel correlation structure, and can global waveform metrics adequately capture domain-spectral component preservation?

The contributions are: (1) a systematic six-tool comparison grounded in signal processing theory, using the latest model versions with real API/web outputs; (2) a multi-channel signal fidelity framework integrating global waveform, domain-spectral, amplitude, and composite evaluation channels; (3) empirical demonstration of input-conditioning asymmetry—a novel finding wherein increased signal conditioning degrades global fidelity in GPT-4o while uniformly improving spectral-component preservation across all LLMs.

2. SIGNAL PROCESSING THEORETICAL FRAMEWORK

2.1 Machine translation as signal transformation

Building on Shannon's information-theoretic model [3], we formalize MT as follows. Let S denote a source-language text signal and T the target-language signal space. A translation system F constitutes a signal transducer mapping $S \rightarrow \hat{T}$, where $\hat{T}$ is the generated translation and T* is the reference (ideal) output signal. The transformation fidelity between $\hat{T}$ and T* can be decomposed into orthogonal quality channels capturing different signal dimensions.

This decomposition is non-trivial: just as a multi-channel audio system captures frequency components that a single-channel mono signal cannot represent, biomedical translation quality cannot be characterized by any single scalar metric. We identify three principal fidelity channels: (i) global waveform fidelity (surface n-gram overlap), (ii) high-frequency spectral component fidelity (domain terminology preservation), and (iii) amplitude fidelity (numerical value preservation). The independence of these channels—empirically confirmed in Section 5.6—validates the multi-channel framework design.

2.2 Large language model architectures as attention-based signal processors

The Transformer architecture underlying all evaluated LLMs [7] is itself a signal processing mechanism: self-attention computes a weighted superposition of value signals guided by query-key similarity, enabling dynamic spectral decomposition of the input sequence. In this context, prompt engineering constitutes input signal conditioning—modifying the information content presented to the transducer prior to the transformation operation. The model-dependent conditioning effects observed in this study (Section 5.2) are thus interpretable as differential sensitivity of signal processors to conditioning perturbations, a phenomenon analogous to filter response variability under input modification in classical signal processing [1].

2.3 Multi-channel evaluation design

The proposed evaluation framework draws on the principle of multi-sensor fusion in signal processing [12]: when multiple measurement channels are mutually orthogonal (zero cross-correlation), each channel contributes unique and non-redundant information about the system under evaluation. The Spearman correlation analysis (Section 5.6) empirically validates this design, demonstrating that BLEU, Terminology Accuracy, and Numerical Consistency constitute three nearly orthogonal evaluation channels ($|r| \leq 0.12$ across all cross-channel pairs), while BLEU and chrF++ represent a partially redundant channel pair ($r = 0.84$) measuring the same underlying quality construct.

3. RELATED WORK

3.1 Signal fidelity metrics in machine translation

MT evaluation has evolved from human judgment through string-matching to neural semantic assessment—a trajectory closely paralleling the development of signal quality metrics from subjective listening tests to objective perceptual models [11]. Papineni et al. [13] introduced BLEU, computing n-gram precision as a global waveform similarity measure. Popović [14] proposed chrF++, combining character-level and word-level n-gram F-scores, offering finer spectral resolution than BLEU. However, these surface metrics reward only exact pattern matches, assigning equal penalty to low-information function-word errors and critical domain-term mistranslations [12]. Zhang et al. [14] proposed BERTScore using contextual embeddings; Rei et al. [15] introduced COMET, demonstrating the highest correlation with human judgments in WMT evaluations [16]. These advances establish that translation signal quality assessment requires both surface and semantic measurement channels, supplemented by domain-specific indicators.

3.2 Large language model-based signal transformation capabilities

LLM translation research has grown rapidly since 2023. Jiao et al. [8] found ChatGPT approaching commercial NMT performance; Hendy et al. [9] confirmed GPT-4's advantages across 18 translation directions. Regarding input signal conditioning (prompt engineering), Peng et al. [17] and Moslem et al. [18] examined conditioning strategy effects on global fidelity. Domain-specific evaluations by Manakhimova et al. [19] demonstrated that LLM translation quality advantages over dedicated NMT systems are not uniform across specialized text types, with performance varying by domain and linguistic feature. Gao et al. [20] similarly found domain-dependent performance patterns when comparing ChatGPT, Google Translate, and DeepL in a specialized Chinese-to-English translation task, underscoring the importance of sub-domain analysis.

4. EXPERIMENTAL METHODOLOGY

4.1 Translation system selection

Six translation systems are evaluated as signal transducers.

The LLM group comprises GPT-4o (OpenAI [21], via ChatGPT web interface), Claude Sonnet-4 (Anthropic, via claude.ai), Gemini-2.5-Pro (Google DeepMind, via gemini.google.com), and Qwen3-Max (Alibaba Cloud, via tongyi.aliyun.com). The NMT group comprises DeepL (deepl.com) and Google Translate (translate.google.com). All translations were collected during 2026 using the free web interfaces of the respective services. Each LLM translation was conducted in an independent new conversation session to prevent channel memory contamination.

4.2 Signal corpus

The WMT19 Biomedical Translation Shared Task English–Chinese test set [22], accessed via the DAMO Academy curated repository on ModelScope [23], provides 540 MEDLINE-abstract-based sentences with professional human reference translations. Stratified sampling (random seed = 42) yielded 100 test sentences across eight sub-domains: Clinical Medicine (n = 21), Medical Imaging (19), General Medicine (15), Public Health (14), Biomedicine (13), Pharmacy (11), Statistical Methods (5), and Forensic Medicine (2). Mean source-text length is 29.4 words (SD = 14.6), corresponding to a moderate-length signal distribution with substantial variance across sub-domain channels.

4.3 Input conditioning strategies

Two input conditioning strategies are designed for LLMs. Condition A (zero-shot / unconditioned): task description and translation requirements only—minimal prior signal injection. Condition B (few-shot + glossary / conditioned) [24]: adds two translation exemplars and a 30-term abbreviation glossary—structured domain-spectral conditioning designed to activate high-frequency component preservation. NMT systems receive direct text input with no conditioning. Sentences are processed in batches of 25 per session. Total outputs: 1,000 translations (4 LLMs × 2 conditions × 100 + 2 NMT × 100).

4.4 Multi-channel signal fidelity framework

The evaluation framework implements multi-channel fidelity measurement across three signal quality dimensions. Global waveform channel: BLEU (sacrebleu [25], 'zh' tokenizer, smoothing method 4) and chrF++ (word_order = 2), measuring n-gram overlap between output and reference signals. Domain-spectral channel: Terminology Accuracy (retention rate of uppercase abbreviation terms ≥2 characters) measuring high-frequency component preservation; Numerical Consistency (binary: all source numbers present) measuring amplitude fidelity. Composite channel: Rule-Based Evaluation assessing translation accuracy, completeness, and terminology handling as a weighted multi-channel aggregate. Table 1 presents the full framework.

4.5 Statistical methods

Non-parametric methods are employed: Wilcoxon signed-rank tests [26] for paired comparisons, Cliff's delta for effect sizes [27], and Spearman rank correlations for inter-channel analysis. Significance level $\alpha = 0.05$. All analyses use Python 3.11 with scipy.stats.

Table 1. Multi-channel signal fidelity evaluation framework

Level	Metric	Type	Signal Processing Analogue / Core Principle
Surface	BLEU [13]	General	Overall waveform fidelity: 1–4 gram precision with brevity penalty
Surface	chrF++ [16]	General	Sub-word fidelity: character + word n-gram F-score; robust for Chinese
Domain	Term. Accuracy	Domain	High-frequency spectral component preservation: retention rate of uppercase abbreviation terms (≥2 chars)
Domain	Num. Consistency	Domain	Amplitude fidelity: binary indicator of whether all source numerical values are preserved
Domain	Rule-Based Eval.	Composite	Multi-channel aggregate: weighted score of accuracy, completeness, and terminology sub-dimensions

5. RESULTS AND ANALYSIS

5.1 Overall signal fidelity comparison

Table 2 presents overall channel scores under zero-shot (LLM) and direct (NMT) conditions. On the global waveform channel (BLEU), Claude Sonnet-4 achieves the highest fidelity (38.17), followed by Gemini-2.5-Pro (38.05), Qwen3-Max (37.48), GPT-4o (37.35), DeepL (36.51), and Google Translate (35.86). The four LLMs form a compact fidelity band (37.35–38.17), all outperforming both NMT baselines in absolute terms.

Notably, all systems achieve high domain-spectral fidelity (Term. Accuracy: 0.937–0.975) and amplitude fidelity (Num. Consistency: 0.980–0.990), indicating that current-generation transducers—both LLM and NMT—reliably preserve high-frequency abbreviation components and numerical amplitude values in biomedical signal transformation. The inter-system differentiation on domain-spectral channels is substantially smaller than on the global waveform channel. Figure 1 visualizes these multi-channel fidelity patterns, confirming the compact LLM fidelity band on global waveform metrics (left panel) and the uniformly high domain-level performance across all transducers (right panel).

Table 3 presents pairwise statistical comparisons with DeepL as baseline. No LLM achieves a statistically significant global fidelity advantage over DeepL (all $p > 0.05$), with Cliff's delta values ranging from +0.017 to +0.058 (all negligible). This indicates that while LLMs consistently outperform DeepL in absolute BLEU, the signal differences do not reach statistical significance with the current corpus size. Figure 2 presents a forest plot of these effect sizes. Gemini-2.5-Pro exhibits the largest positive effect ($\delta = +0.058$, $p = 0.072$), followed by Claude Sonnet-4 ($\delta = +0.052$, $p = 0.114$), while GPT-4o shows the smallest positive shift ($\delta = +0.017$, $p = 0.151$). Google Translate is the only system with a negative effect relative to DeepL ($\delta = -0.021$, $p = 0.790$). All effect size bars cluster near zero, visually confirming the negligible magnitude of pairwise fidelity differences and supporting the convergence interpretation.

Table 2. Multi-channel signal fidelity scores (zero-shot / direct translation)

Tool	Type	BLEU	chrF++	Term.	NC	RBE-A	RBE-C	RBE-T	RBE-O
Claude Sonnet-4	LLM	**38.17***	32.58	0.940	0.990	0.720	0.905	0.940	0.840
Gemini-2.5-Pro	LLM	38.05	32.43	0.975	0.990	0.740	**0.933***	0.975	**0.866***
Qwen3-Max	LLM	37.48	32.36	0.970	0.990	0.735	0.927	0.970	0.861
GPT-4o	LLM	37.35	32.33	0.975	0.990	0.739	0.910	0.975	0.858
DeepL	NMT	36.51	31.18	0.962	0.990	0.722	0.913	0.962	0.849
Google Translate	NMT	35.86	31.27	0.937	0.980	0.717	0.914	0.937	0.841

Note: Term. = Terminology Accuracy; NC = Numerical Consistency; RBE-A/C/T/O = Rule-Based Eval. Accuracy/Completeness/Terminology/Overall. * = highest per column.

Table 3. Statistical significance tests (BLEU channel, pairwise vs. DeepL)

Comparison	p-value	Cliff's δ	Effect	Direction	Decision
GPT-4o vs. DeepL	0.151	+0.017	Negligible	→	Not significant
Claude Sonnet-4 vs. DeepL	0.114	+0.052	Negligible	→	Not significant
Gemini-2.5-Pro vs. DeepL	0.072	+0.058	Negligible	→	Not significant
Qwen3-Max vs. DeepL	0.055	+0.031	Negligible	→	Not significant
Google Translate vs. DeepL	0.790	−0.021	Negligible	←	Not significant

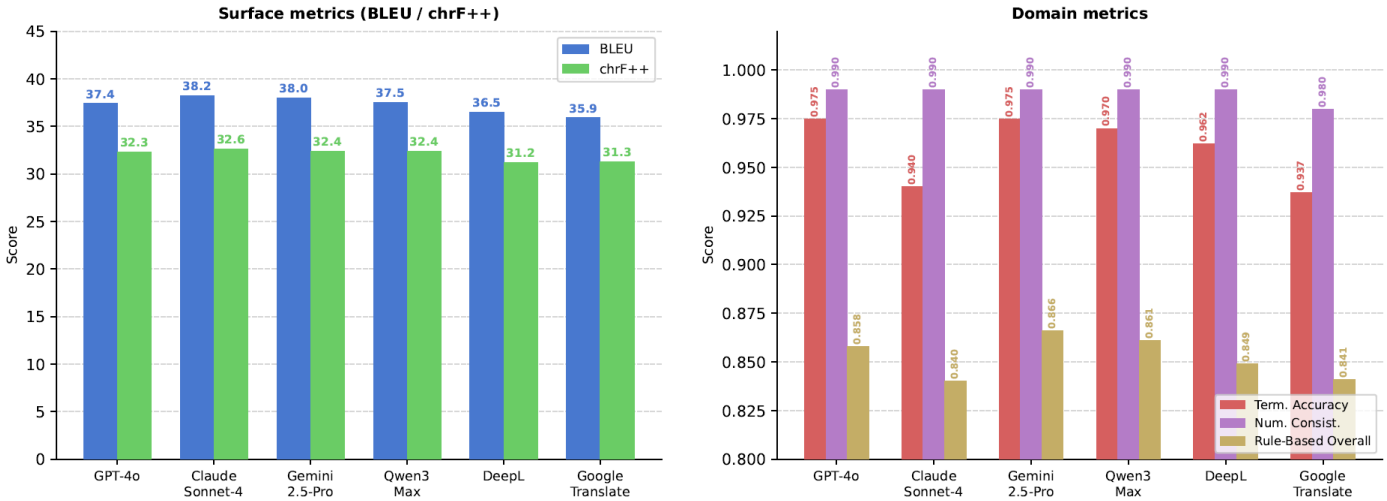

Figure 1. Overall multi-channel signal fidelity: global waveform metrics (left) and domain-spectral metrics (right)

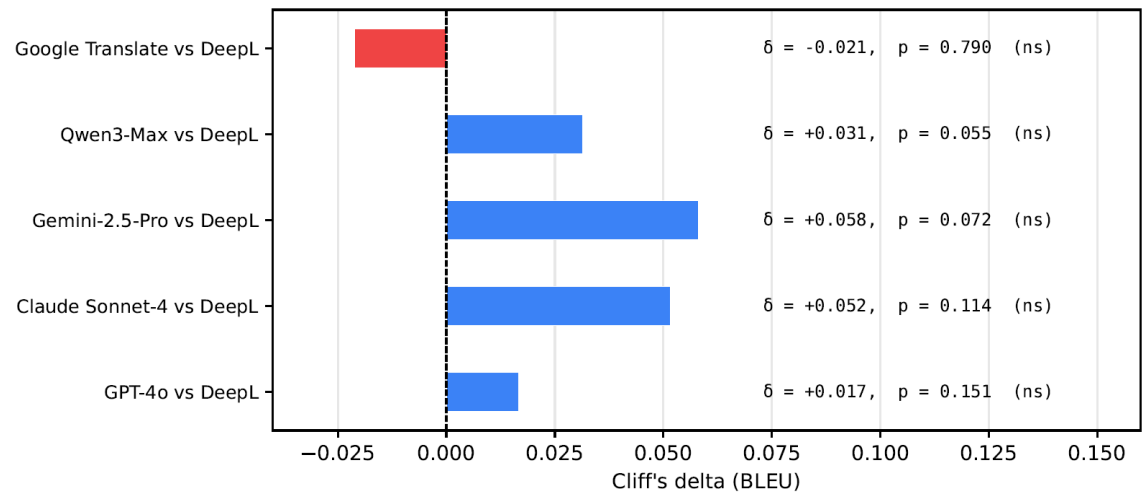

Figure 2. Effect size forest plot (Cliff's delta vs. DeepL): LLM-NMT fidelity gap characterization

5.2 Input conditioning effects on signal fidelity

Table 4 presents signal fidelity changes under few-shot + glossary conditioning relative to the unconditioned (zero-shot) baseline. A striking system-dependent conditioning response emerges: GPT-4o exhibits a statistically significant global waveform fidelity decrease (ΔBLEU = −0.71, p = 0.004), while the other three LLMs show positive but non-significant gains. Claude Sonnet-4 achieves the largest positive response (+1.32), followed by Gemini-2.5-Pro (+0.70) and Qwen3-Max (+0.66). Critically, all four LLMs improve to near-perfect domain-spectral fidelity under conditioning (0.990–1.000 vs.

0.940–0.975 unconditioned). Figure 3 visualizes the conditioning-induced ΔBLEU for each LLM, highlighting GPT-4o's anomalous negative response. Figure 4 directly compares the absolute zero-shot and few-shot+glossary BLEU scores for each LLM. The visual contrast is striking: GPT-4o's conditioned output (36.6) falls below its unconditioned baseline (37.4), whereas Claude Sonnet-4 shows the most pronounced upward shift (38.2 → 39.5), with Gemini-2.5-Pro (38.0 → 38.7) and Qwen3-Max (37.5 → 38.1) exhibiting moderate gains. This asymmetric pattern underscores that few-shot conditioning does not uniformly enhance global waveform fidelity across LLM architectures.

Table 4. Input conditioning effects on multi-channel signal fidelity

Tool	ZS BLEU	FS BLEU	ΔBLEU	p-value	ZS Term	FS Term	ΔTerm
GPT-4o	37.35	36.64	**−0.71****	**0.004**	0.975	1.000	+0.025
Claude Sonnet-4	38.17	39.49	+1.32	0.540	0.940	0.990	+0.050
Gemini-2.5-Pro	38.05	38.75	+0.70	0.697	0.975	0.995	+0.020
Qwen3-Max	37.48	38.14	+0.66	0.309	0.970	1.000	+0.030

Note: ** p < 0.01. ZS = zero-shot (unconditioned); FS = few-shot + glossary (conditioned); Term = Terminology Accuracy.

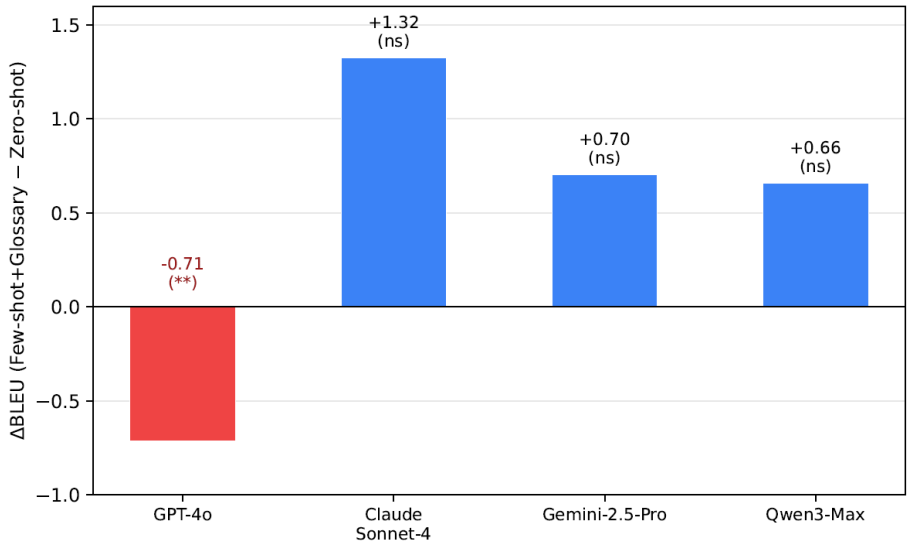

Figure 3. Input conditioning effect on global waveform fidelity (ΔBLEU). Red bar indicates statistically significant fidelity decrease (p = 0.004)

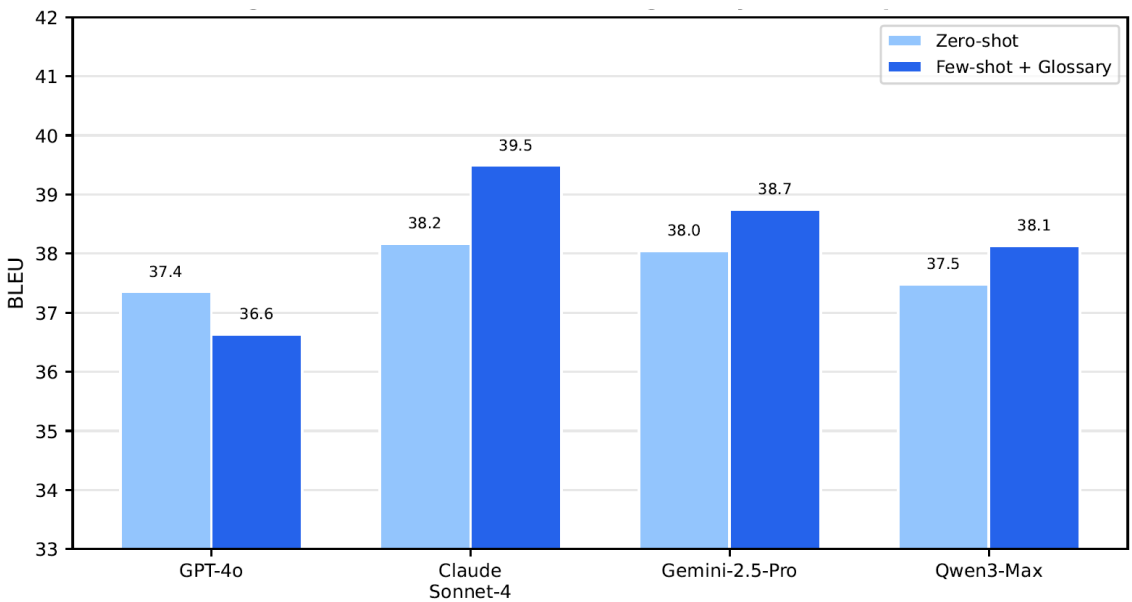

Figure 4. Zero-shot vs. few-shot+glossary BLEU: per-system conditioning response

5.3 Multi-channel radar profile

Figure 5 provides a normalized multi-channel fidelity visualization across all six systems. The radar profile reveals that while LLM transducers cluster tightly on global waveform channels (BLEU, chrF++), substantial inter-system divergence emerges on the Rule-Based Accuracy and Rule-Based Completeness channels, where Gemini-2.5-Pro consistently leads. The visualization also captures Claude Sonnet-4's asymmetric channel profile: the highest global

waveform fidelity (BLEU) yet a proportionally smaller normalized chrF++ advantage, suggesting distributional differences in how its output signal overlaps with reference translations at the character versus n-gram level.

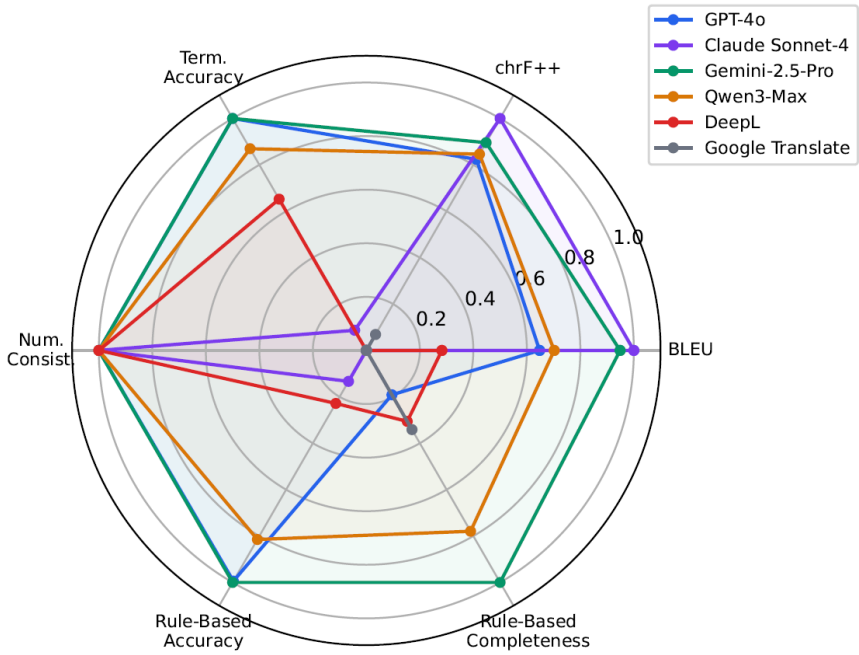

Figure 5. Multi-channel radar profile of normalized signal fidelity metrics across all six transducers

5.4 Sub-domain signal distribution robustness

Figure 6 presents the global waveform fidelity heatmap across sub-domain signal distributions. Cross-domain fidelity variance differs substantially: Claude Sonnet-4 shows the highest variance (BLEU range: 28.93–52.57, σ = 7.0), driven by exceptionally high fidelity in Forensic Medicine (52.57) and Statistical Methods (46.36) channels but substantially lower fidelity in the Pharmacy channel (28.93). GPT-4o demonstrates the most uniform cross-domain response (σ = 3.1), suggesting more consistent signal transformation behavior across sub-domain spectral distributions. The Forensic Medicine channel (n = 2) should be interpreted with caution due to insufficient sample size for robust statistical estimation.

5.5 Sentence-length signal sensitivity

Figure 7 shows global waveform fidelity across three signal-length categories. Most transducers peak at medium-length signals (20–40 words), with fidelity degradation for both short and long signals—consistent with the well-known length sensitivity of n-gram fidelity metrics [13]. Claude Sonnet-4 demonstrates the most stable length-domain response (3.9-point range across categories), while Google Translate shows the most severe long-signal fidelity penalty (4.0-point drop from medium to long). The asymmetric degradation pattern observed for DeepL (nearly flat from short to medium, then steep decline) suggests architecture-specific handling of signal complexity.

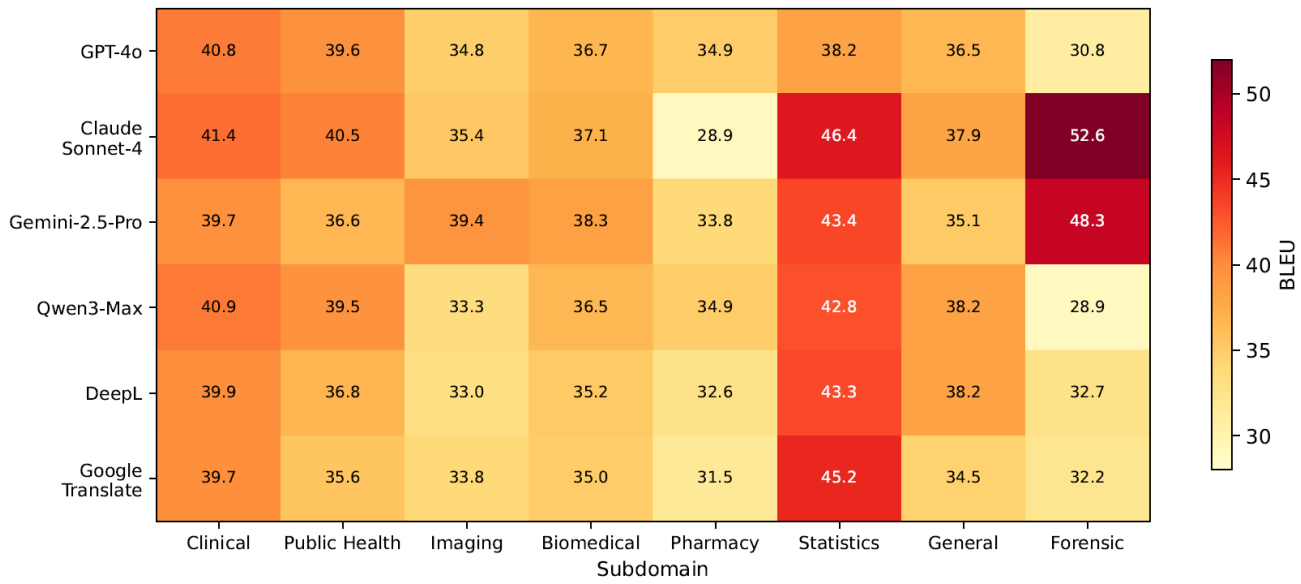

Figure 6. Global waveform fidelity (BLEU) heatmap across sub-domain signal distributions and transducers

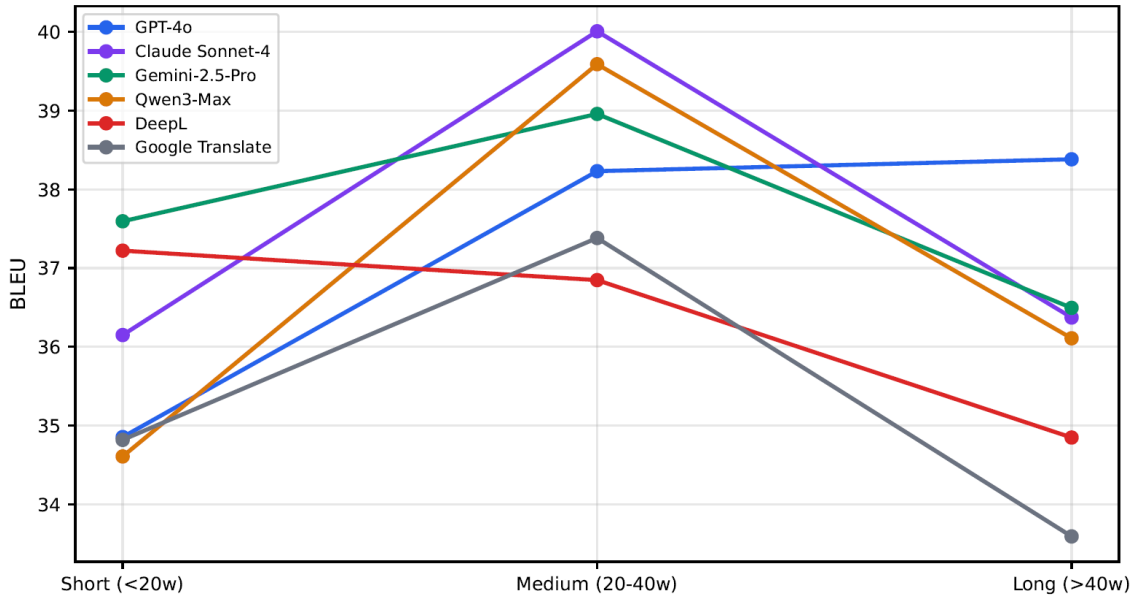

Figure 7. Signal-length sensitivity of global waveform fidelity (BLEU) across three length categories

5.6 Inter-channel correlation analysis

Figure 8 presents the Spearman inter-channel correlation matrix, computed on sentence-level scores from GPT-4o's zero-shot output (N = 100). The central finding is the near-zero correlation between the global waveform channel (BLEU) and the domain-spectral channel (Terminology Accuracy, r = 0.12), confirming the theoretical prediction that these channels measure orthogonal quality dimensions. BLEU and chrF++ exhibit strong correlation (r = 0.84), indicating partial channel redundancy—both primarily capture the same global waveform construct. Numerical Consistency (amplitude channel) shows near-zero correlation with all other channels ($|r| \leq 0.03$), confirming full orthogonality and establishing it as an independent fidelity dimension. This inter-channel orthogonality structure empirically validates the multi-channel framework design. Table 5 reports the complete set of pairwise correlation coefficients alongside their signal processing interpretations, further confirming the independence of evaluation channels.

Figure 2 presents the effect size forest plot (Cliff's delta) for each system compared pairwise against DeepL. Gemini-2.5-Pro exhibits the largest positive effect ($\delta = +0.058$, $p = 0.072$), followed by Claude Sonnet-4 ($\delta = +0.052$, $p = 0.114$) and Qwen3-Max ($\delta = +0.031$, $p = 0.055$), while GPT-4o shows the smallest positive shift ($\delta = +0.017$, $p = 0.151$). Google Translate is the only system with a negative effect relative to DeepL ($\delta = -0.021$, $p = 0.790$). All confidence intervals overlap zero and all effect sizes fall within the negligible range, visually reinforcing the convergence finding from Section 5.1 that current-generation transducers cluster within a narrow fidelity band.

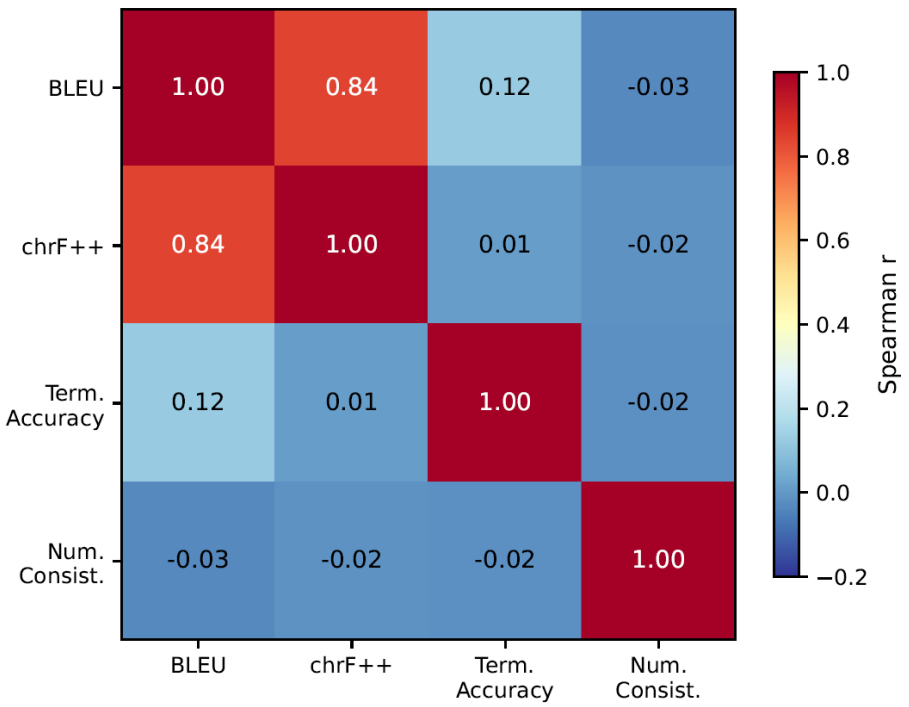

Figure 8. Spearman inter-channel correlation matrix confirming near-orthogonal quality dimensions

Table 5. Spearman inter-channel correlation coefficients with signal processing interpretation

Metric Pair	Spearman r	Level	Signal Processing Interpretation
BLEU vs. chrF++	0.84	Strong	Both capture global waveform fidelity; partial channel redundancy
BLEU vs. Term. Accuracy	0.12	Near zero	Global fidelity cannot capture high-frequency spectral component preservation
BLEU vs. Num. Consistency	−0.03	Near zero	Waveform similarity is orthogonal to amplitude fidelity
chrF++ vs. Term. Accuracy	0.01	Near zero	Sub-word fidelity equally incapable of capturing spectral domain terms
Term. Accuracy vs. NC	−0.02	Near zero	Spectral and amplitude channels are fully independent quality dimensions

Figure 4 provides a side-by-side comparison of absolute BLEU scores under zero-shot and few-shot+glossary conditions for each LLM. The visual contrast highlights the conditioning asymmetry discussed in Section 5.2: GPT-4o's conditioned output (36.6) falls below its unconditioned baseline (37.4), whereas Claude Sonnet-4 exhibits the most pronounced upward shift (38.2 → 39.5). Gemini-2.5-Pro (38.0 → 38.7) and Qwen3-Max (37.5 → 38.1) show moderate gains. This pattern confirms that few-shot conditioning with domain glossaries does not uniformly enhance global waveform fidelity across LLM architectures.

6. DISCUSSION

6.1 Convergence toward a shared fidelity ceiling

The most striking finding is the absence of statistically significant inter-system differences on the global waveform channel (BLEU). All six transducers score within a 2.3-point range (35.86–38.17), with Cliff's delta values uniformly negligible (−0.021 to +0.058). This competitive convergence can be interpreted, from a signal processing perspective, as different architectural families—attention-based neural language models and neural MT systems—reaching similar saturation levels in source-to-target signal transformation fidelity for a well-resourced language pair. The system has effectively reached a ceiling bounded by the information-theoretic constraints of the evaluation channel (BLEU) and the domain distribution of the test corpus, rather than by the processing capacity of individual transducers.

This convergence contrasts with earlier comparative studies [8, 9] that reported substantial LLM-NMT fidelity gaps, suggesting that the competitive landscape has narrowed as both paradigms have matured. The implication for evaluation methodology is significant: when all systems operate near the same fidelity ceiling on the global waveform channel, discriminative evaluation requires the multi-channel framework proposed here.

6.2 Input conditioning as a double-edged signal perturbation

The conditioning-dependent fidelity responses constitute the most novel finding of this study. GPT-4o's significant global waveform fidelity decrease under few-shot conditioning (ΔBLEU = −0.71, $p = 0.004$) challenges the standard assumption that conditioning universally improves translation quality [17, 18]. This counterintuitive result is interpretable within signal processing theory as an over-constraint phenomenon: when a transducer with extensive pre-training on domain signals receives explicit in-context exemplars, these examples function as strong priors that constrain the output generation distribution. If the conditioning signal (few-shot examples) diverges from the specific reference signal distribution used in evaluation, the constrained output may achieve lower n-gram overlap with reference translations despite being semantically equivalent or superior—a form of fidelity loss analogous to filter response distortion under input modification in classical signal processing [1].

By contrast, the consistent improvement in domain-spectral fidelity (Terminology Accuracy) across all LLMs under conditioning (reaching 0.990–1.000) demonstrates that glossary-augmented conditioning specifically activates high-frequency spectral component preservation behavior. The dissociation observed in GPT-4o—simultaneous global waveform fidelity decrease and spectral component fidelity increase—constitutes direct empirical evidence that these two evaluation channels are genuinely orthogonal quality dimensions, reinforcing the inter-channel correlation finding ($r = 0.12$) and validating the multi-channel framework.

6.3 Implications of inter-channel orthogonality for evaluation methodology

The near-zero correlations between the global waveform channel and both domain-specific channels (Term. Accuracy $r = 0.12$; Num. Consistency $r = −0.03$) quantitatively confirm an evaluation blind spot widely acknowledged theoretically [11, 28] but rarely demonstrated empirically with real system outputs. The practical consequence is severe: a transducer scoring BLEU 38 may exhibit identical or inferior domain-spectral fidelity relative to one scoring BLEU 36. For clinical and biomedical translation end-users, domain-spectral fidelity—the preservation of high-frequency terminology components that carry critical semantic load—is of primary practical importance, yet it is entirely invisible to single-channel BLEU evaluation.

We recommend that biomedical translation evaluation adopt the hierarchical multi-channel framework proposed here: global waveform fidelity (BLEU/chrF++ as a partially redundant channel pair), domain-spectral fidelity (Terminology Accuracy), amplitude fidelity (Numerical Consistency), and composite rule-based assessment. The strong BLEU–chrF++ correlation ($r = 0.84$) suggests that in resource-constrained settings one channel may be omitted from this pair, but neither can substitute for the independent domain-spectral and amplitude channels.

6.4 Limitations

Several limitations apply. First, the corpus comprises 100 sentences from WMT19; sub-domain channels with $n < 10$ yield unreliable fidelity estimates. Second, only automatic evaluation channels are employed; human judgment evaluation is not incorporated. Third, only English→Chinese transformation is examined. Fourth, neural semantic metrics (BERTScore [14], COMET [15])—which have demonstrated the highest correlation with human judgment [28, 29]—were not computed due to computational constraints and should be incorporated in future work as additional evaluation channels.

Fifth, translations were collected via free web interfaces and may differ from API-based outputs. Sixth, only two conditioning strategies are compared; advanced strategies such as Chain-of-Thought [30] warrant examination as signal conditioning mechanisms.

7. CONCLUSIONS

This study evaluated six translation transducers—GPT-4o, Claude Sonnet-4, Gemini-2.5-Pro, Qwen3-Max, DeepL, and Google Translate—on 100 biomedical text signals using a multi-channel signal fidelity framework grounded in signal processing theory. Four principal findings were obtained:

Finding 1: All four LLM transducers achieve higher global waveform fidelity (BLEU) than NMT baselines in absolute terms, but no pairwise difference reaches statistical significance (all $p > 0.05$, Cliff's $\delta < 0.06$), indicating competitive convergence toward a shared fidelity ceiling.

Finding 2: Input signal conditioning effects are transducer-dependent. GPT-4o exhibits a statistically significant global fidelity decrease under few-shot conditioning ($p = 0.004$), consistent with an over-constraint interpretation, while all LLMs improve domain-spectral fidelity to near-perfect levels (0.990–1.000) under the same conditioning.

Finding 3: Inter-channel correlation analysis confirms near-orthogonality between the global waveform channel (BLEU) and both domain-specific channels (Term. Accuracy $r = 0.12$; Num. Consistency $r = -0.03$), empirically validating the multi-channel framework design and demonstrating that single-channel evaluation is insufficient for biomedical text signal quality assessment.

Finding 4: Cross-domain fidelity variance differs substantially across transducers. Claude Sonnet-4 shows the highest sub-domain spectral sensitivity ($\sigma = 7.0$), while GPT-4o exhibits the most uniform domain-invariant response ($\sigma = 3.1$).

The multi-channel signal fidelity framework proposed here provides a principled and extensible approach to biomedical translation benchmarking. Future work should expand the signal corpus, incorporate neural semantic evaluation channels (BERTScore [14], COMET [15]), explore diverse conditioning strategies including Chain-of-Thought mechanisms [30], and conduct longitudinal fidelity tracking as transducer versions iterate.

REFERENCES

[1] Oppenheim, A.V., Willsky, A.S., Nawab, S.H. (1996). Signals and Systems (2nd ed.). Upper Saddle River, NJ: Prentice-Hall.

[2] Proakis, J.G., Manolakis, D.G. (2006). Digital Signal Processing: Principles, Algorithms, and Applications (4th ed.). Upper Saddle River, NJ: Pearson.

[3] Shannon, C.E. (1948). A mathematical theory of communication. Bell System Technical Journal, 27: 379-423.

[4] Sager, J.C. (1994). Language Engineering and Translation: Consequences of Automation. Amsterdam: John Benjamins.

[5] Byrne, J. (2006). Technical Translation: Usability Strategies for Translating Technical Documentation. Dordrecht: Springer.

[6] Koehn, P., Knowles, R. (2017). Six challenges for neural machine translation. In Proceedings of the First Workshop on Neural Machine Translation, Vancouver, pp. 28-39. https://doi.org/10.18653/v1/w17-3204

[7] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (NeurIPS), 30.

[8] Jiao, W., Wang, W., Huang, J. T., Wang, X., Shi, S., Tu, Z. (2023). Is ChatGPT a good translator? Yes with GPT-4 as the engine. arXiv preprint arXiv:2301.08745. https://doi.org/10.48550/arXiv.2301.08745

[9] Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., et al. (2023). How good are gpt models at machine translation? A comprehensive evaluation. arXiv preprint arXiv:2302.09210. https://doi.org/10.48550/arXiv.2302.09210

[10] Pang, J., Ye, F., Wong, D.F., Yu, D., Shi, S., Tu, Z., Wang, L. (2025). Salute the classic: Revisiting challenges of machine translation in the age of large language models. Transactions of the Association for Computational Linguistics, 13: 73-95. https://doi.org/10.1162/tacl_a_00730

[11] Freitag, M., Rei, R., Mathur, N., Lo, C.K., Stewart, C., Avramidis, E., Martins, A.F. (2022). Results of WMT22 metrics shared task: Stop using BLEU–neural metrics are better and more robust. In Proceedings of the Seventh Conference on Machine Translation (WMT), Abu Dhabi, United Arab Emirates (Hybrid), pp. 46-68. https://doi.org/10.18653/v1/2022.wmt-1.2

[12] Hall, D.L., Llinas, J. (2002). An introduction to multisensor data fusion. Proceedings of the IEEE, 85(1): 6-23. https://doi.org/10.1109/5.554205

[13] Papineni, K., Roukos, S., Ward, T., Zhu, W.J. (2002). Bleu: A method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, Philadelphia, pp. 311-318.

[14] Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675. https://doi.org/10.48550/arXiv.1904.09675

[15] Rei, R., Stewart, C., Farinha, A.C., Lavie, A. (2020). COMET: A neural framework for MT evaluation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 2685-2702. https://doi.org/10.18653/v1/2020.emnlp-main.213

[16] Popović, M. (2017). chrF++: Words helping character n-grams. In Proceedings of the Second Conference on Machine Translation, pp. 612-618.

[17] Peng, K., Ding, L., Zhong, Q., Shen, L., Liu, X., Zhang, M., Tao, D. (2023). Towards making the most of ChatGPT for machine translation. In Findings of the Association for Computational Linguistics: EMNLP 2023, pp. 5622-5633. https://doi.org/10.18653/v1/2023.findings-emnlp.373

[18] Moslem, Y., Haque, R., Kelleher, J., Way, A. (2023). Adaptive machine translation with large language models. In Proceedings of the 24th Annual Conference of the European Association for Machine Translation, Tampere, Finland, pp. 227-237.

[19] Manakhimova, S., Avramidis, E., Macketanz, V., Lapshinova-Koltunski, E., Bagdasarov, S., Möller, S.

(2023). Linguistically motivated evaluation of the 2023 state-of-the-art machine translation: Can ChatGPT outperform NMT? In Proceedings of the Eighth Conference on Machine Translation, Singapore, pp. 224-245. https://doi.org/10.18653/v1/2023.wmt-1.23

[20] Gao, R., Lin, Y., Zhao, N., Cai, Z.G. (2024). Machine translation of Chinese classical poetry: A comparison among ChatGPT, Google Translate, and DeepL Translator. Humanities and Social Sciences Communications, 11(1): 1-10. https://doi.org/10.1057/s41599-024-03363-0

[21] OpenAI, Achiam, J., Adler, S., Agarwal, S., et al. (2023). GPT-4 technical report. arXiv preprint arXiv:2303.08774. https://doi.org/10.48550/arXiv.2303.08774.

[22] Barrault, L., Bojar, O., Costa-Jussà, M.R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., Malmasi, S., Monz, C., Müller, M., Pal, S., Post, M., Zampieri, M. (2019). Findings of the 2019 Conference on Machine Translation (WMT19). Proceedings of the Fourth Conference on Machine Translation, Florence, Italy, pp. 1-61. https://doi.org/10.18653/v1/w19-5301

[23] DAMO Academy, Alibaba Group (2022). WMT English-to-Chinese Machine Translation Medical Test Set [Dataset]. ModelScope. https://www.modelscope.cn/datasets/damo/WMT-English-to-Chinese-Machine-Translation-Medical.

[24] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33: 1877-1901.

[25] Post, M. (2018). A call for clarity in reporting BLEU scores. In Proceedings of the Third Conference on Machine Translation: Research Papers, Brussels, Belgium, pp. 186-191. https://doi.org/10.18653/v1/W18-6319

[26] Wilcoxon, F. (1945). Individual comparisons by ranking methods. Biometrics Bulletin, 1(6): 80-83. https://doi.org/10.2307/3001968

[27] Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. Psychological Bulletin, 114(3): 494-509.

[28] Kocmi, T., Federmann, C. (2023). Large language models are state-of-the-art evaluators of translation quality. In Proceedings of the 24th Annual Conference of the European Association for Machine Translation, pp. 193-203.

[29] Kocmi, T., Avramidis, E., Bawden, R., Bojar, O., Dvorkovich, A., Federmann, C., Fishel, M., Freitag, M., Gowda, T., Grundkiewicz, R., Haddow, B., Karpinska, M., Koehn, P., Marie, B., Monz, C., Murray, K., Nagata, M., Popel, M., Popović, M., Zouhar, V. (2024). Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In Proceedings of the Ninth Conference on Machine Translation, Miami, Florida, USA.

[30] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35: 24824-24837.