Indian Journal of Computer Science and Technology
https://www.doi.org/10.59256/indjcst.20260503091
Volume 5, Issue3 (September-December 2026), PP: 769-779.
www.indjcst.com
ISSN No: 2583-5300

Machine Translation: Evolution, Multilingual Models, Evaluation, and Challenges

Sivakumar Eswaran[*1], R. Anitha[2]
[*1]*Department of Information Technology, Sri Venkateswara College of Engineering, Sripermbudur, Tamilnadu, India.*
[2]*Department of Computer Science and Engineering, Sri Venkateswara College of Engineering, Sripermbudur, Tamilnadu, India.*
[*]*Corresponding Author.*

To Cite this Article: *Sivakumar Eswaran[*1], R. Anitha[2], "Machine Translation: Evolution, Multilingual Models, Evaluation, and Challenges", Indian Journal of Computer Science and Technology, Volume 05, Issue 03 (September-December 2026), PP: 769-779.*

Copyright: ©2026 This is an open access journal, and articles are distributed under the terms of the Creative Commons Attribution License; Which Permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract: *Machine Translation (MT) is one of the most influential research areas in Natural Language Processing (NLP). The automated process of translating text from one language to another, has witnessed significant advancements in recent years. Over roughly seven decades it has moved from rule-based systems (RBMT) to statistical (SMT), neural (NMT) and Transformer-based models, and more recently to massively multilingual systems and Large Language Models (LLMs). This survey is written as an entry point for researchers new to the field. It reviews the main paradigms and their architectural milestones, massively multilingual systems, LLM-based translation, speech and image-grounded translation, and a dedicated review of machine translation for Indian languages covering early systems, parallel corpora, transliteration resources and Indic-specific evaluation. It also surveys automatic and human evaluation methods, and discusses open challenges including low-resource languages, hallucination, bias, document-level translation, evaluation validity and privacy. The narrative review includes selected literature available up to November 2026 and closes with future research directions.*

Key Word: *Machine Translation; Neural Machine Translation; Multilingual Models; Large Language Models; Indian Languages; Evaluation Metrics.*

I. INTRODUCTION

Language plays a fundamental role in human communication, cultural exchange and knowledge dissemination. The rapid growth of digital content and international collaboration has created increasing demand for technologies that support multilingual communication. Machine Translation (MT) addresses this need by automatically translating text, and increasingly speech, from one language to another. Modern MT is usually traced to Warren Weaver's 1949 memorandum and the 1954 Georgetown–IBM demonstration1. The earliest systems relied on manually constructed linguistic rules and bilingual dictionaries. These Rule-Based Machine Translation (RBMT) systems modelled morphological, syntactic and semantic structure explicitly. One of the earliest Indian initiatives was AnglaBharti, which used an interlingua-style framework for translating English into Indian languages2.

The availability of large bilingual corpora from the late 1980s and 1990s led to Statistical Machine Translation (SMT), which learned translation patterns directly from parallel text. Brown et al. introduced probabilistic translation models that became the basis of statistical MT3, and phrase-based models later improved quality by translating groups of words rather than single words4. The next breakthrough was Neural Machine Translation (NMT): sequence-to-sequence learning enabled end-to-end translation with recurrent networks5, and the attention mechanism let the decoder focus on relevant source words6. The Transformer architecture, which relies entirely on self-attention, then became the dominant framework7. More recently, massively multilingual models such as No Language Left Behind (NLLB), which covers about 200 languages8, and IndicTrans2, which covers all 22 scheduled Indian languages9, extended high-quality translation to many more languages. In parallel, LLMs have introduced zero-shot, few-shot and document-level translation through prompting and in-context learning10.

The objective of this survey is to give a structured, beginner-oriented review of MT paradigms, recent advances, Indian-language translation research, evaluation methodology and open challenges. Its main contributions are: (i) a chronological and conceptual account of MT from rule-based to LLM-based systems, with a timeline (Fig. 1) and a paradigm comparison (ii) a comparison of representative multilingual and LLM-based systems (iii) an extended review of Indian-language MT, including datasets, models and benchmarks (iv) a review of automatic and human evaluation, with a metric comparison and (v) a discussio n of open challenges and research directions.

Scope and Review Methodology

This is a narrative survey. Literature was gathered from the ACL Anthology, arXiv, IEEE Xplore, journal publishers (including Nature) and Google Scholar using combinations of the terms "machine translation", "statistical", "neural", "multilingual", "large language model", "evaluation", "Indian languages" and "Indic". Priority was given to (a) foundational papers that introduced a paradigm or technique, (b) peer-reviewed work from major venues (ACL, EMNLP, NAACL, TACL, NeurIPS, ICLR, EAMT, WMT and Nature), (c) findings of the WMT shared tasks, and (d) openly released resources for Indian languages. Coverage runs from the 1940s to November 2026. Widely adopted preprints are included where no archival version exists, and commercial systems without a published technical description are excluded. Because LLM-based translation is

changing quickly, the discussion of that area should be read as a snapshot, and the quantitative results quoted throughout are those reported by the original authors on their own test sets; they are not directly comparable across papers. This is a narrative overview rather than a reproducible systematic review; no claim of exhaustive coverage is made.

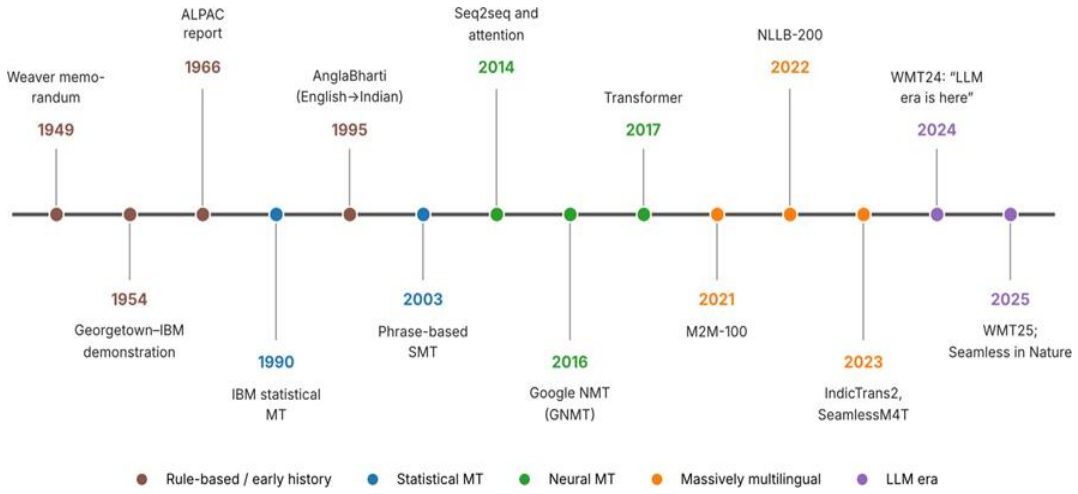


Fig. 1. Timeline of selected key milestones in machine translation (events are equally spaced and not drawn to scale; the diagram ends at 2025, while selected 2026 developments are discussed in the text).

II. EARLY HISTORY AND RULE-BASED MACHINE TRANSLATION

The 1954 Georgetown–IBM experiment generated considerable optimism about fully automatic translation1. That optimism was tempered in 1966 when the ALPAC report judged progress to be disappointing and recommended shifting support toward basic research and translation aids, after which US funding for MT fell sharply11. Research continued, however, along linguistically motivated lines.

Rule-Based Machine Translation depends on handcrafted grammatical rules, bilingual lexicons, morphological analysers and syntactic transfer mechanisms. RBMT approaches generally fall into three categories: (1) direct translation, which maps source words to target words with limited analysis; (2) transfer-based translation, which analyses the source sentence, transfers the structure to the target language and then generates the output; and (3) interlingua-based translation, which maps the source sentence to a language-independent meaning representation from which any target language can be generated.

AnglaBharti is a notable Indian example: it used an intermediate pseudo-target representation, often described as pseudo-interlingua-like, rather than a fully language-independent semantic interlingua, to support translation from English into multiple Indian languages2. Rule-based approaches have also continued to be explored for morphologically rich languages; for example, Hurskainen and Tiedemann describe a rule-based English–Finnish system built around detailed morphological and syntactic analysis12. The primary advantage of RBMT is its linguistic interpretability and the control it gives developers over output. However, the manual effort required to build and maintain rules limits scalability, and RBMT systems tend to struggle with ambiguity, idiomatic expressions and linguistic variation that the rules do not explicitly cover.

III. STATISTICAL MACHINE TRANSLATION

Statistical Machine Translation emerged as a data-driven alternative to rule-based approaches. The seminal work of Brown et al. proposed probabilistic models that estimate the likelihood of a target sentence given a source sentence3, and the follow-up IBM models formalised word alignment and parameter estimation for such models13. Och and Ney's alignment-template approach modelled correspondences between groups of words and combined multiple knowledge sources in a log-linear framework14. Koehn et al. then proposed a phrase-based translation model and decoder and compared methods for extracting phrase pairs4; phrase-based SMT became the dominant approach and was made widely accessible through the open-source Moses toolkit15. Further advances included factored translation models, which add linguistic annotations such as lemmas and part-of-speech tags to the translation process16, and mixture-model approaches to domain adaptation17. Although SMT clearly outperformed earlier rule-based systems on many language pairs, it exhibited several limitations:

1. difficulty handling long-range dependencies and reordering;
2. limited semantic understanding;
3. sensitivity to the quality and domain of the training data;
4. a pipeline of separately trained and tuned components; and
5. Poor generalisation for low-resource languages.
6. These limitations motivated the transition toward neural methods.

IV. NEURAL MACHINE TRANSLATION

Neural Machine Translation models the translation process with neural networks trained on parallel data. Early neural translation research included the recurrent continuous translation models of Kalchbrenner and Blunsom18. Cho et al.19 introduced an RNN encoder–decoder for learning phrase representations and used neural phrase-pair probabilities as an additional feature in

a conventional phrase-based SMT system; it should therefore be viewed as an important neural contribution to SMT, rather than as a fully end-to-end NMT system. The sequence-to-sequence architecture of Sutskever et al. then demonstrated that an encoder network could compress a source sentence into a vector from which a decoder generates the translation5. Compressing a whole sentence into one fixed-length vector creates an information bottleneck, which Bahdanau et al. addressed with an attention mechanism that lets the decoder weight source positions dynamically6. Luong et al. introduced global and local attention variants20. Google's NMT system (GNMT) applied deep attentional networks in production and reported substantial gains over Google's phrase-based system in human evaluation21. Two practical advances were important for making NMT work at scale: subword segmentation with byte-pair encoding, which lets models handle rare and unseen words22, and back-translation, which exploits monolingual target-language data23. NMT achieved substantial improvements in:

1. translation fluency
2. context modelling within a sentence
3. handling of long-distance dependencies and
4. adaptability to multiple language pairs.

However, recurrent NMT systems were computationally demanding, trained sequentially and remained difficult to scale.

V. TRANSFORMER-BASED MACHINE TRANSLATION

The Transformer introduced by Vaswani et al. replaced recurrence with self-attention7. This design offers parallelisable training, improved scalability, better modelling of long-range dependencies and reduced training time, and it rapidly became the dominant framework for MT. The same architecture also underpins pretrained language models. BERT is an encoder-only model pretrained on English text and is not itself a translation system24, but it popularised the pretrain-then-fine-tune paradigm. Multilingual counterparts followed: XLM-R, a multilingual encoder25; mT5, a multilingual text-to-text model26; and mBART, a multilingual sequence-to-sequence model pretrained with denoising objectives and designed for translation27. Table no 1 summarises the main paradigms.

Paradigm	Period	Core idea	Strengths	Limitations	Key works
Rule-based (RBMT)	1950s–1990s	Hand-written rules and lexicons; direct, transfer or interlingua architectures	Interpretable; controllable output	Costly rule authoring; weak on ambiguity and idioms	2, 12
Paradigm	Period	Core idea	Strengths	Limitations	Key works
Statistical (SMT)	c. 1990–2015	Probabilistic models and phrase tables learned from parallel corpora	Learns from data; good for well-resourced pairs	Weak long-range dependencies;many Separately tuned components	3, 4, 14, 4
Recurrent NMT	c. 2014–2017	Encoder–decoder with attention, trained end-to-end	More fluent output; joint learning	Sequential computation; ostly to train	5, 6, 20, 21
Transformer NMT	2017 onward	Self-attention encoder–decoder	Parallel traini ng; strong quality	ttention cost grows quadratically with length; data-hungry	7
Massively multilingual	c. 2017 onward	One model shared across many languages	Cross-lingual transfer; a single model to deploy	Capacity shared between anguages; uneven data across languages	8, 28, 29
LLM-based	c. 2022 onward	General-purpose LLMs prompted or fine-tuned for translation	Zero/few-shot u se; flexible, context-aware	Computational cost; hallucination; weaker for low-resource languages	10, 30, 31

Table no 1: Comparison of machine translation paradigms (periods are approximate).

VI. MASSIVE MULTILINGUAL MACHINE TRANSLATION

Rather than training one model per language pair, multilingual NMT shares parameters across languages. Johnson et al. showed that a single model with a target-language token could translate between many languages, including zero-shot directions not seen in training28. M2M-100 moved beyond English-centric training by mining and training on data for direct translation between many language pairs across 100 languages29. Progress on low-resource languages also depends on evaluation data: the FLORES-101 benchmark provided a many-to-many evaluation set covering 101 languages32.

The No Language Left Behind (NLLB) project, developed by the NLLB Team, scaled multilingual MT to 200 languages (the released NLLB-200 model supports 202) and explicitly targets the imbalance between high-resource and low-resource languages8,33. The project introduced FLORES-200, multilingual datasets and translation-quality and toxicity evaluation frameworks. The largest NLLB model uses a sparsely gated mixture-of-experts architecture; this should not be taken to mean that every released NLLB variant is an MoE model. The authors report an average improvement of about 44% in BLEU relative to the previous state of the art, evaluated on FLORES-2008. This is a relative BLEU gain on the authors' benchmark, not a general 44%

improvement in translation quality. FLORES-200 contains evaluation data for 204 languages, which is distinct from the 202-language coverage of the released NLLB-200 model. Table no 2 compares representative multilingual and LLM-based systems.

A major 2026 development is Omnilingual MT, a family of systems reported to cover more than 1,600 languages34. The work combines multilingual data collection, monolingual-data strategies, back-translation and specialised model architectures, and introduces expanded evaluation resources including BOUQuET and a reference-free quality-estimation metric. The authors report human-judged passable translation for more than 440 languages and zero-shot translation into approximately 1,200 languages. These figures describe the authors' evaluation setup and quality threshold; broad language coverage should not be interpreted as uniformly high-quality translation for every language or domain34.

System	Year	Language coverage	Notable features	Source
mBART	2020	25 languages (pretraining)	Denoising sequence-to-sequence pretraining for translation	27
M2M-100	2021	100 languages	Direct many-to-many translation without pivoting through English	29
mT5	2021	101 languages	Multilingual text-to-text pretraining; not translation-specific	26
NLLB-200	2022 (preprint); 2024 (journal)	200 languages (released model: 202)	FLORES-200 evaluation data cover 204 languages; largest model uses MoE; reported ~44% relative BLEU gain on the authors' benchmark	33, 8
IndicTrans2	2023	All 22 scheduled Indian languages	Trained on BPCC (230M bitext pairs); evaluated on the IN22 benchmark	9
System	Year	Language coverage	Notable features	Source
SeamlessM4T	2023 (preprint); 2025 (journal)	2023 version: ~100 speech-input, 95 text, 35 speech-output languages; 2025 journal version: 101, 96 and 36, respectively	Speech-to-speech, speech-to-text, text-to-speech and text-to-text translation, plus ASR	39, 40
GPT-4 (LLM)	2023	General-purpose; no translation-specific language list	Evaluated on translation by independent studies	10,41
ALMA (LLM recipe)	2024	Fine-tuned moderate-size LLMs	Two-stage fine-tuning: monolingual data, then small high-quality parallel data	30
Omnilingual MT	2026	More than 1,600 languages in the system family	Specialised 1B–8B models; expanded multilingual data and evaluation; human-judged passable quality reported for 440+ languages	34

Table no 2: Representative multilingual and LLM-based translation systems (figures as reported by the original authors; not directly comparable).

VII. LARGE LANGUAGE MODELS FOR TRANSLATION

Large Language Models acquire multilingual abilities during large-scale self-supervised pretraining rather than being optimised specifically for translation, and can be steered through prompts and in-context examples. Early evaluations of GPT models found translation quality that was competitive for high-resource languages but limited for low-resource ones42. Zhu et al. evaluated eight popular LLMs, including ChatGPT and GPT-4, and reported that GPT-4 outperformed the supervised NLLB baseline in 40.91% of the translation directions tested; the same study also found that GPT-4 still faces a large gap to commercial systems such as Google Translate, especially for low-resource languages10. Both parts of this finding matter: LLMs are strong in many directions, yet remain uneven across languages.

Research on LLM-based MT has developed along several lines. Fine-tuning recipes such as ALMA, which continues training on monolingual data and then fine-tunes on a small amount of high-quality parallel data, show that moderate-size LLMs can be adapted to translation30. Position papers argue that the future of MT lies with LLMs43, while the WMT24 findings summarise the state of the art31. Pang et al. revisit classic MT challenges in the LLM setting44. Reported advantages of LLMs include adaptability through prompting and cross-domain generalisation; document-level, interactive and style-preserving translation are active research directions rather than guaranteed capabilities across all models and languages. Challenges include hallucination, inconsistency, computational cost and privacy. A 2026 study of 15 LLMs across 22 language pairs found that non-English-centric directions generally received lower COMET scores and linked weaker performance to lower activation of

language-specific vocabulary35. This provides one possible explanation for persistent low-resource disparities, not a universal causal account.

VIII. SPEECH AND MULTIMODAL MACHINE TRANSLATION

Traditional MT operates on text, whereas human communication also involves speech and visual context. Two directions are usually grouped under "multimodal" translation, and they are distinct.

Speech Translation

SeamlessM4T is a single model that supports speech-to-speech, speech-to-text, text-to-speech and text-to-text translation as well as automatic speech recognition39. The original 2023 release reports approximately 100 languages for speech input, 95 for text and 35 for speech output39. The 2025 journal version reports 101 speech-input languages, 96 text-input/output languages and 36 speech-output languages40. Coverage claims should therefore state the model version and modality rather than relying on the headline "about 100 languages".

Image-Grounded Translation

Image-grounded (visual) multimodal MT uses an accompanying image to help disambiguate the source sentence. The Multi30K dataset provides English–German image descriptions and has been a standard resource for this line of work45. Image-grounded translation for Indian languages is still comparatively under-resourced and is listed as an open direction in Section XII.

IX. MACHINE TRANSLATION FOR INDIAN LANGUAGES

Linguistic Landscape and Challenges

India is characterised by extraordinary linguistic diversity, encompassing languages from the Indo-Aryan, Dravidian, Tibeto-Burman and Austro-Asiatic families. The Constitution lists 22 scheduled languages; this constitutional category should not be confused with the full set of languages spoken in India. MT research for Indian languages faces challenges including rich morphology in many languages, differing word-order patterns, script diversity, code-mixing and limited parallel corpora for numerous language pairs. Naskar and Bandyopadhyay provide an early overview of the use of MT in India46.

Early Systems

Indian MT began with rule-based and hybrid systems. AnglaBharti used an intermediate pseudo-target representation, rather than a fully language-independent interlingua, for translating English into Indian languages2. Anusaaraka is a language-accessor approach that aims to give readers of one Indian language access to text in another47. These systems were valuable for linguistic analysis and low-data settings, but shared the scalability limits of RBMT discussed in Section II.

Parallel and Monolingual Resources

Progress in neural Indian-language MT has been driven by data. Samanantar was introduced as the largest publicly available parallel corpus collection for 11 Indic languages48. The IndicTrans2 project then released the Bharat Parallel Corpus Collection (BPCC), which contains 230 million bitext pairs, of which 126 million were newly added, including 644 thousand manually translated sentence pairs9. For monolingual data and language-understanding evaluation, Doddapaneni et al. built monolingual corpora, a benchmark and models for Indic languages49. IndicBART is a pretrained sequence-to-sequence model for Indic natural language generation, including translation50. Table no 3 lists the main resources.

IndicTrans2

IndicTrans2 is the first open translation model reported to support all 22 scheduled Indian languages. Its authors addressed four gaps: training data (BPCC), evaluation (IN22, an n-way parallel benchmark covering all 22 languages), model coverage, and open release of models and data under permissive licences9. The model uses a multilingual Transformer architecture and is reported to achieve state-of-the-art results across many Indian-language pairs on the authors' benchmarks. Readers should note that these results were obtained on IN22 and related benchmarks at the time of publication, and that comparisons with newer LLM-based systems need to be re-run on common test sets.

Transliteration Resources

Transliteration remains important for Indian MT because named entities and technical terms frequently require script conversion. The Aksharantar dataset of Madhani et al. contains 26 million transliteration pairs for 21 Indic languages from three language families, written in 12 scripts; the authors report that it is 21 times larger than existing datasets, and they also release a 103k-word test set covering 19 languages and the IndicXlit transliteration model51. Aksharantar is primarily a training and evaluation resource for transliteration rather than a machine translation benchmark.

Benchmarks and Evaluation for Indian Languages

IN22 provides an n-way benchmark for translation into and out of all 22 languages9. IndicGenBench evaluates the generation capabilities of LLMs on Indic languages52. Because automatic metrics were developed mainly for English and other high-resource languages, IndicMT Eval provides a dataset for meta-evaluating MT metrics on Indian languages, that is, for checking how well metrics agree with human judgments53.

Resource	Type	Coverage	Key facts (as reported)	Source
FLORES-101	Evaluation benchmark	101 languages	Many-to-many evaluation set for low-resource and multilingual MT	32
FLORES-200	Evaluation benchmark	About 200 languages	Extension introduced with NLLB; used for the headline NLLB results	8
Multi30K	Multimodal dataset	English–German (image descriptions)	Standard resource for image-grounded translation	45
Samanantar	Parallel corpus	English and 11 Indic languages	Largest public parallel collection for these languages at publication	48
BPCC	Parallel corpus	22 scheduled Indian languages	230M bitext pairs; 126M newly added, including 644K manually translated	9
IN22	Evaluation benchmark	All 22 scheduled languages	n-way parallel translation benchmark	9
Resource	Type	Coverage	Key facts (as reported)	Source
IndicCorp v2 and IndicXTREME	Monolingual corpora; NLU benchmark	Indic languages	Corpora, benchmark and models for Indic language understanding	49
IndicGenBench	LLM generation benchmark	Indic languages	Evaluates LLM generation capabilities on Indic languages	52
Aksharantar	Transliteration dataset	21 Indic languages; 3 families; 12 scripts	26M pairs; 103k-word test set for 19 languages	51
IndicMT Eval	Metric meta-evaluation	Indian languages	Human judgments for validating automatic metrics	53
WMT24 and WMT25 general tasks	Shared tasks	Multiple language pairs	Annual community evaluations of MT systems	31, 54
WMT26 low-resource Indic shared task	Shared task / multidomain benchmark	10 English–Indicpairs across Northeast Indian language families	Bidirectional evaluation in healthcare, political, travel, sports and entertainment domains; based on INDICNE-CORP 2.0	37

Table no 3: Datasets and benchmarks relevant to machine translation, with emphasis on Indian languages.

Open Problems for Indian-Language MT

Based on the work reviewed above, several gaps stand out. First, large parallel corpora and models cover some Indian languages more effectively than others, so translation quality remains uneven; this is not a uniform property of all 22 sched uled languages. Second, code-mixed text and dialectal variation remain underrepresented in many benchmarks. Third, evaluation for Indian languages still depends heavily on metrics whose behaviour may vary across languages, which makes the validation work of IndicMT Eval53 important. Fourth, systematic comparisons between purpose-built models such as IndicTrans29 and general LLMs on shared Indian-language test sets remain limited. Recent work begins to address these gaps: SrcMix studies source-side mixing of related languages for extremely low-resource translation and releases an Angika MT dataset36, while the WMT 2026 low-resource Indic shared task evaluates ten English–Indic pairs across multiple Northeast Indian language families and five domains37. These task-specific findings are useful evidence of progress, but they should not be generalised to all Indian languages. These observations are the authors' synthesis of the literature rather than results reported in a single study.

X. EVALUATION OF MACHINE TRANSLATION

Machine translation evaluation remains a challenging problem because many different translations can be correct. Evaluation methods fall into surface-overlap metrics, learned metrics, human evaluation, and, more recently, LLM-based evaluators. Table no 4 summarises the main metrics.

Overlap-Based Metrics

BLEU is one of the earliest automatic metrics and remains widely used because it is simple and inexpensive to compute55. It primarily measures modified n-gram precision with a brevity penalty against reference translations and may fail to capture semantic equivalence. SacreBLEU provides a standardised implementation and reporting signature for BLEU; it is not a separate translation-quality metric. Because scores depend on tokenisation and configuration, Post recommends reporting the signature so that results can be reproduced and compared more reliably56. METEOR adds stemming and synonym matching to unigram alignment57, and TER estimates the edits needed to transform a hypothesis into a reference, including shift operations58. TER

can still penalise valid paraphrases and lexical alternatives. The character n-gram F-score chrF59 and its extension chrF++, which adds word n-grams60, are commonly used for morphologically rich languages, including many Indian languages.

Learned Metrics

Embedding-based and learned metrics were developed to correlate better with human judgments. BERTScore compares contextual embeddings of the output and reference61, and BLEURT fine-tunes a pretrained model on human ratings62. COMET leverages multilingual pretrained language models and evaluates a translation using the source, the reference and the hypothesis, and its authors report stronger correlation with human judgments than overlap metrics63. xCOMET extends this idea with fine-grained error-span detection in addition to a quality score64.

Human Evaluation

Human evaluation remains essential for assessing translation quality, particularly when automatic metrics may miss meaning, adequacy or pragmatic errors. The Multidimensional Quality Metrics (MQM) framework provides an error-annotation scheme in which trained annotators identify error spans, categories and severity levels. Freitag et al. compared expert-based evaluation with crowd-sourced ratings on large-scale WMT data and found that the two can lead to different conclusions about system quality65. Because human evaluation is costly, it is usually reserved for key claims and shared tasks. The WMT general tasks combine automatic metrics with human evaluation31, 54.

LLM-Based Evaluators

A recent development is prompting LLMs to judge translation quality. Kocmi and Federmann's GEMBA reports that GPT-based prompting reaches strong system-level evaluation accuracy on the evaluated WMT data66. This finding should not be interpreted as proof of reliable segment-level evaluation across all language pairs. A 2026 study further shows that translation errors in multilingual benchmarks can contribute to measurable drops in evaluated LLM accuracy, underlining that benchmark translation quality itself must be checked38. Open questions remain about evaluator reliability at the segment level, for low-resource languages and across prompt and model versions.

Benchmarks, Shared Tasks and Practical Recommendations

Shared tasks and benchmarks provide common ground for comparison: FLORES-101 and FLORES-200 for multilingual evaluation32, 8, IN22 for Indian languages9, the WMT general tasks31, 54, and the WMT 2026 low-resource Indic shared task37. The WMT25 findings call for moving away from evaluating on easy test sets54, while the 2026 Indic task illustrates the value of multidomain and bidirectional evaluation for underrepresented language pairs37. For authors of new MT papers, practical recommendations that follow from this literature are: report more than one metric, including at least one learned metric and one character-level metric; report metric versions and signatures; use human evaluation for the main claims; and state clearly which test sets and language directions were used.

Metric	Family	What it measures	Main limitations	Source
BLEU (SacreBLEU implementation)	N-gram overlap	Modified n-gram precision with brevity penalty against reference(s)	Weak on meaning and paraphrase; scores depend on tokenisation and configuration	55, 56
METEOR	Overlap with alignment	Unigram alignment with stemming and synonym matching	Needs language-specific resources	57
TER	Edit distance	Number of edits needed to match a reference	May penalise valid paraphrases and lexical alternatives; shift operations partly address reordering	58
chrF / chrF++	Character n-gram F-score	Character (and, for chrF++, word) n-gram overlap	Still surface-based	59, 60
BERTScore	Embedding similarity	Similarity of contextual embeddings	Depends on the language coverage of the underlying model	61
BLEURT	Learned	Pretrained model fine-tuned on human ratings	Depends on training data and languages	62
COMET	Learned	Quality score from source, hypothesis and reference using a multilingual encoder	Depends on training data; less transparent	63
xCOMET	Learned with error spans	Quality score plus error-span detection	Same dependence on training data	64
MQM	Human (expert)	Error spans with categories and severities	Costly and time-consuming	65
GEMBA	LLM-prompted	Quality judgments obtained by prompting an LLM	Sensitive to prompt and model; cost; reliability for low-resource languages not settled	66

Table no 4: Summary of MT evaluation metrics.

XI. CHALLENGES AND LIMITATIONS

Even though the research on MT has substantially progressed, several challenges persist.

Low-Resource Languages and Data Imbalance

Many languages lack sufficient parallel corpora for robust training. Large-scale efforts such as NLLB8 and, for India, Samanantar and BPCC9, 48 have expanded coverage, but quality remains uneven. LLMs inherit the imbalance of their pretraining data: the gap to commercial systems is largest for low-resource languages10.

Hallucination

Both NMT and LLM-based systems may generate fluent output that is unsupported by the source. Guerreiro et al. analysed hallucinations in large multilingual translation models and compared approaches to detecting and mitigating them67. Because hallucinated output reads fluently, it is harder for users to notice than an obvious error, and metrics such as BLEU may not penalise it appropriately.

Bias and Fairness

Training data may contain cultural, social and gender biases. Stanovsky et al. proposed an automatic evaluation of gender bias across several target languages and found that the tested systems were prone to stereotypical gender assignments 68. Differences in grammatical gender, pronoun systems and gender marking across languages can introduce translation errors, particularly when the source language does not explicitly encode information required by the target language. These phenomena vary by language and should not be generalised uniformly across Indian languages.

Ambiguity, Polysemy and Document-Level Translation

Words often have multiple meanings, and resolving them may require context beyond the sentence. Maintaining discourse consistency, such as terminology and pronoun reference, remains difficult for long texts. LLMs offer longer context windows and more flexible prompting, but whether they truly resolve these problems is still being studied44.

Evaluation Limitations

Automatic metrics are imperfect proxies for human judgment, and as systems improve, test sets can become too easy to separate them; the WMT25 findings argue for harder test sets54. Metrics developed for high-resource languages may not behave identically on Indian languages53, and LLM-based evaluators introduce their own sources of variation66. In addition, errors in translated multilingual benchmarks can affect downstream LLM evaluation, so benchmark translation quality and provenance should be documented38.

Privacy, Security and Cost

Cloud-based translation raises concerns about confidentiality and regulatory compliance, especially for legal, medical and governmental text. Large models are also expensive to train and serve, which limits access for smaller institutions and motivates open, moderate-size models such as those released for Indian languages9.

XII. FUTURE RESEARCH DIRECTIONS

The integration of foundation models, multilingual pretraining and multimodal learning is likely to shape the next generation of MT systems. Based on the gaps identified in this review, future research is expected to focus on:

1. Universal multilingual foundation models that treat translation, speech and understanding jointly, while evaluating quality separately by language and modality34, 40
2. Retrieval-augmented and terminology-aware machine translation for specialised domains
3. Stronger, common evaluation of LLM-based and purpose-built models for Indian languages, including code-mixed and extremely low-resource settings36, 37, 52, 53
4. Speech-to-speech and image-grounded translation for Indian languages, extending resources45
5. Low-resource and endangered-language preservation, building on large-scale data collection8
6. Explainable MT, source-grounded generation and reliable detection and mitigation of hallucination67
7. Human-in-the-loop translation and more reliable, affordable human and LLM-based evaluation65, 66
8. Bias mitigation techniques68
9. Privacy-preserving and on-device machine translation and
10. Trustworthy and responsible multilingual AI, including harder test sets, multilingual benchmark validation and transparent reporting38, 54.

XIII. CONCLUSION

Machine translation has moved from hand-written rules, through statistical models and recurrent neural networks, to Transformer-based multilingual systems and LLMs. Attention, the Transformer, multilingual transfer learning and, most recently, LLMs have improved translation quality and extended language coverage. Initiatives such as NLLB, IndicTrans2, SeamlessM4T and the 2026 Omnilingual MT family represent significant steps toward broader multilingual communication, while Indian-language MT has been strengthened by public datasets, benchmarks and recent low-resource shared-task research. Yet language coverage should not be equated with uniform translation quality: low-resource performance, hallucination, bias, evaluation validity and privacy continue to motivate further work. Recent studies on extremely low-resource transfer and multilingual benchmark

quality reinforce the need for language-specific evidence and careful evaluation. Future systems are likely to become increasingly multilingual, multimodal and context-aware, but progress should be assessed with transparent, domain-sensitive and human-informed evaluation.

REFERENCES

1. W. J. Hutchins, "The Georgetown-IBM experiment demonstrated in January 1954," in Machine Translation: From Real Users to Research (AMTA 2004), Lecture Notes in Computer Science, vol. 3265. Berlin, Germany: Springer, 2004, pp. 102–114.
2. R. M. K. Sinha, K. Sivaraman, A. Agrawal, R. Jain, R. C. Srivastava, A. Jain, et al., "ANGLABHARTI: A multilingual machine-aided translation project on translation from English to Indian languages," in Proc. IEEE Int. Conf. Systems, Man, and Cybernetics, vol. 2, 1995, pp. 1609–1614, doi: 10.1109/ICSMC.1995.538002.
3. P. F. Brown, J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin, "A statistical approach to machine translation," Computational Linguistics, vol. 16, no. 2, pp. 79–85, Jun. 1990.
4. P. Koehn, F. J. Och, and D. Marcu, "Statistical phrase-based translation," in Proc. Human Language Technology Conf. North American Chapter of the Association for Computational Linguistics (HLT-NAACL), Edmonton, Canada, 2003, pp. 127–133.
5. I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in Advances in Neural Information Processing Systems, vol. 27, 2014, pp. 3104–3112.
6. D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in Proc. Int. Conf. Learning Representations (ICLR), 2015, arXiv:1409.0473.
7. A. Vaswani et al., "Attention is all you need," in Advances in Neural Information Processing Systems, vol. 30, 2017, pp. 5998–6008, arXiv:1706.03762.
8. NLLB Team, "Scaling neural machine translation to 200 languages," Nature, vol. 630, pp. 841–846, Jun. 2024, doi: 10.1038/s41586-024-07335-x.
9. J. Gala et al., "IndicTrans2: Towards high-quality and accessible machine translation models for all 22 scheduled Indian languages," Trans. Machine Learning Research, 2023, arXiv:2305.16307.
10. W. Zhu, H. Liu, Q. Dong, J. Xu, S. Huang, L. Kong, J. Chen, and L. Li, "Multilingual machine translation with large language models: Empirical results and analysis," in Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, 2024, pp. 2765–2781, doi: 10.18653/v1/2024.findings-naacl.176.
11. Automatic Language Processing Advisory Committee, Language and Machines: Computers in Translation and Linguistics, Publication 1416. Washington, DC, USA: National Academy of Sciences–National Research Council, 1966.
12. A. Hurskainen and J. Tiedemann, "Rule-based machine translation from English to Finnish," in Proc. 2nd Conf. Machine Translation (WMT), Copenhagen, Denmark, 2017, pp. 323–329, doi: 10.18653/v1/W17-4731.
13. P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: Parameter estimation," Computational Linguistics, vol. 19, no. 2, pp. 263–311, Jun. 1993.
14. F. J. Och and H. Ney, "The alignment template approach to statistical machine translation," Computational Linguistics, vol. 30, no. 4, pp. 417–449, Dec.2004, doi: 10.1162/0891201042544884.
15. P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, et al., "Moses: Open source toolkit for statistical machine translation," in Proc. ACL Interactive Poster and Demonstration Sessions, Prague, Czech Republic, 2007, pp. 177–180.
16. P. Koehn and H. Hoang, "Factored translation models," in Proc. Joint Conf. Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), Prague, Czech Republic, 2007, pp. 868–876.
17. G. Foster and R. Kuhn, "Mixture-model adaptation for SMT," in Proc. 2nd Workshop on Statistical Machine Translation, Prague, Czech Republic, 2007,pp. 128–135.
18. N. Kalchbrenner and P. Blunsom, "Recurrent continuous translation models," in Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP), Seattle, WA, USA, 2013, pp. 1700–1709.
19. K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in Proc. EMNLP, Doha, Qatar, 2014, pp. 1724–1734, doi: 10.3115/v1/D14-1179.
20. T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP), Lisbon, Portugal, 2015, pp. 1412–1421, doi: 10.18653/v1/D15-1166.
21. Y. Wu et al., "Google's neural machine translation system: Bridging the gap between human and machine translation," arXiv preprint arXiv:1609.08144, 2016.
22. R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in Proc. 54th Annu. Meeting Association for Computational Linguistics (ACL), Berlin, Germany, 2016, pp. 1715–1725, doi: 10.18653/v1/P16-1162.
23. R. Sennrich, B. Haddow, and A. Birch, "Improving neural machine translation models with monolingual data," in Proc. 54th Annu. Meeting ACL, Berlin, Germany, 2016, pp. 86–96, doi: 10.18653/v1/P16-1009.
24. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Minneapolis, MN, USA, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.
25. A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in Proc. 58th Annu. Meeting ACL, online, 2020, pp. 8440–8451, doi: 10.18653/v1/2020.acl-main.747.
26. L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "mT5: A massively multilingual pre-trained text-to-text transformer," in Proc. NAACL-HLT, online, 2021, pp. 483–498, doi: 10.18653/v1/2021.naacl-main.41.
27. Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, "Multilingual denoising pre-training for neural machine translation," Trans. Association for Computational Linguistics, vol. 8, pp. 726–742, 2020, doi: 10.1162/tacl_a_00343.
28. M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, M. Hughes, and J. Dean, "Google's multilingual neural machine translation system: Enabling zero-shot translation," Trans. Association for Computational Linguistics, vol. 5, pp. 339–351, 2017, doi: 10.1162/tacl_a_00065.
29. A. Fan, S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, et al., "Beyond English-centric multilingual machine translation," J. Machine Learning Research, vol. 22, no. 107, pp. 1–48, 2021.
30. H. Xu, Y. J. Kim, A. Sharaf, and H. H. Awadalla, "A paradigm shift in machine translation: Boosting translation performance of large language models," in Proc. Int. Conf. Learning Representations (ICLR), 2024, arXiv:2309.11674.

31. T. Kocmi et al., "Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet," in Proc. 9th Conf. Machine Translation (WMT), Miami, FL, USA, 2024, pp. 1–46, doi: 10.18653/v1/2024.wmt-1.1.
32. N. Goyal, C. Gao, V. Chaudhary, P.-J. Chen, G. Wenzek, D. Ju, S. Krishnan, M. Ranzato, F. Guzmán, and A. Fan, "The Flores-101 evaluation benchmark for low-resource and multilingual machine translation," Trans. Association for Computational Linguistics, vol. 10, pp. 522–538, 2022, doi: 10.1162/tacl_a_00474.
33. NLLB Team et al., "No language left behind: Scaling human-centered machine translation," arXiv preprint arXiv:2207.04672, 2022.
34. B. Alastruey et al., "Omnilingual MT: Machine Translation for 1,600 Languages," in Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP), Budapest, Hungary, 2026, pp. 29177–29200, doi: 10.18653/2026.emnlp-main.1280.
35. S. Qian and Y. Scherrer, "Why do Large Language Models Fail in Low-resource Translation? Unraveling the Token Dynamics of Large Language Models for Machine Translation," in Proc. 26th Annu. Conf. European Association for Machine Translation (EAMT), vol. 1, Tilburg, The Netherlands, 2026, pp. 22–47.
36. S. Kumar, P. Jyothi, and P. Bhattacharyya, "SrcMix: Mixing of related source languages benefits extremely low-resource machine translation," in Findings of the Association for Computational Linguistics: EACL 2026, Rabat, Morocco, 2026, pp. 6306–6323, doi: 10.18653/v1/2026.findings-eacl.332.
37. P. Pakray et al., "Findings of WMT 2026 shared task on Low-resource Indic Languages Translation," in Proc. 11th Conf. Machine Translation (WMT), Budapest, Hungary, 2026, pp. 1091–1116, doi: 10.18653/2026.wmt-main.53.
38. K. Thellmann, B. Stadler, M. Färber, and J. Lehmann, "Quantifying the Impact of Translation Errors on Multilingual LLM Evaluation," in Proc. 64th Annu. Meeting Association for Computational Linguistics (ACL), vol. 1: Long Papers, San Diego, CA, USA, 2026, pp. 41311–41334, doi: 10.18653/v1/2026.acl-long.1916.
39. Seamless Communication et al., "SeamlessM4T: Massively multilingual and multimodal machine translation," arXiv preprint arXiv:2308.11596, 2023.
40. Seamless Communication et al., "Joint speech and text machine translation for up to 100 languages," Nature, vol. 637, pp. 587–593, Jan. 2025, doi:10.1038/s41586-024-08359-z.
41. OpenAI, "GPT-4 technical report," arXiv preprint arXiv:2303.08774, 2023.
42. A. Hendy, M. Abdelrehim, A. Sharaf, V. Raunak, M. Gabr, H. Matsushita, et al., "How good are GPT models at machine translation? A comprehensive evaluation," arXiv preprint arXiv:2302.09210, 2023.
43. C. Lyu, Z. Du, J. Xu, Y. Duan, M. Wu, T. Lynn, A. F. Aji, D. F. Wong, and L. Wang, "A paradigm shift: The future of machine translation lies with large language models," in Proc. Joint Int. Conf. Computational Linguistics, Language Resources and Evaluation (LREC-COLING), Torino, Italy, 2024, pp. 1339–1352.
44. J. Pang, F. Ye, D. F. Wong, D. Yu, S. Shi, Z. Tu, and L. Wang, "Salute the classic: Revisiting challenges of machine translation in the age of large language models," Trans. Association for Computational Linguistics, vol. 13, pp. 73–95, 2025, doi: 10.1162/tacl_a_00730.
45. D. Elliott, S. Frank, K. Sima'an, and L. Specia, "Multi30K: Multilingual English-German image descriptions," in Proc. 5th Workshop on Vision and Language, Berlin, Germany, 2016, pp. 70–74, doi: 10.18653/v1/W16-3210.
46. S. Naskar and S. Bandyopadhyay, "Use of machine translation in India: Current status," in Proc. Machine Translation Summit X: Posters, Phuket, Thailand, 2005, pp. 465–470.
47. A. Bharati, V. Chaitanya, A. P. Kulkarni, R. Sangal, and G. U. Rao, "Anusaaraka: Overcoming the language barrier in India," arXiv preprint arXiv:cs/0308018, 2003.
48. G. Ramesh, S. Doddapaneni, A. Bheemaraj, M. Jobanputra, R. AK, A. Sharma, et al., "Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages," Trans. Association for Computational Linguistics, vol. 10, pp. 145–162, 2022, doi: 10.1162/tacl_a_00452.
49. S. Doddapaneni, R. Aralikatte, G. Ramesh, S. Goyal, M. M. Khapra, A. Kunchukuttan, and P. Kumar, "Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages," in Proc. 61st Annu. Meeting ACL, Toronto, Canada, 2023, pp. 12402–12426, doi: 10.18653/v1/2023.acl-long.693.
50. R. Dabre, H. Shrotriya, A. Kunchukuttan, R. Puduppully, M. Khapra, and P. Kumar, "IndicBART: A pre-trained model for Indic natural language generation," in Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, 2022, pp. 1849–1863, doi: 10.18653/v1/2022.findings-acl.145.
51. Y. Madhani, S. Parthan, P. Bedekar, G. Nc, R. Khapra, A. Kunchukuttan, P. Kumar, and M. Khapra, "Aksharantar: Open Indic-language transliteration datasets and models for the next billion users," in Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, 2023, pp. 40–57, doi: 10.18653/v1/2023.findings-emnlp.4.
52. H. Singh, N. Gupta, S. Bharadwaj, D. Tewari, and P. Talukdar, "IndicGenBench: A multilingual benchmark to evaluate generation capabilities of LLMs on Indic languages," in Proc. 62nd Annu. Meeting ACL, Bangkok, Thailand, 2024, pp. 11047–11073, doi: 10.18653/v1/2024.acl-long.595.
53. A. Sai B, T. Dixit, V. Nagarajan, A. Kunchukuttan, P. Kumar, M. M. Khapra, and R. Dabre, "IndicMT Eval: A dataset to meta-evaluate machine translation metrics for Indian languages," in Proc. 61st Annu. Meeting ACL, Toronto, Canada, 2023, pp. 14210–14228, doi: 10.18653/v1/2023.acl-long.795.
54. T. Kocmi et al., "Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets," in Proc. 10th Conf. Machine Translation (WMT), Suzhou, China, 2025, pp. 355–413, doi: 10.18653/v1/2025.wmt-1.22.
55. K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in Proc. 40th Annu. Meeting ACL, Philadelphia, PA, USA, 2002, pp. 311–318, doi: 10.3115/1073083.1073135.
56. M. Post, "A call for clarity in reporting BLEU scores," in Proc. 3rd Conf. Machine Translation (WMT), Brussels, Belgium, 2018, pp. 186–191, doi: 10.18653/v1/W18-6319.
57. S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in Proc. ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Ann Arbor, MI, USA, 2005, pp. 65–72.
58. M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A study of translation edit rate with targeted human annotation," in Proc. 7th Conf.Association for Machine Translation in the Americas (AMTA), Cambridge, MA, USA, 2006, pp. 223–231.
59. M. Popović, "chrF: Character n-gram F-score for automatic MT evaluation," in Proc. 10th Workshop on Statistical Machine Translation (WMT), Lisbon, Portugal, 2015, pp. 392–395, doi: 10.18653/v1/W15-3049.
60. M. Popović, "chrF++: Words helping character n-grams," in Proc. 2nd Conf. Machine Translation (WMT), Copenhagen, Denmark, 2017, pp. 612–618, doi: 10.18653/v1/W17-4770.
61. T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," in Proc. Int. Conf.

Learning Representations (ICLR), 2020, arXiv:1904.09675.
62. T. Sellam, D. Das, and A. Parikh, "BLEURT: Learning robust metrics for text generation," in Proc. 58th Annu. Meeting ACL, online, 2020, pp. 7881–7892, doi: 10.18653/v1/2020.acl-main.704.
63. R. Rei, C. Stewart, A. C. Farinha, and A. Lavie, "COMET: A neural framework for MT evaluation," in Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP), online, 2020, pp. 2685–2702, doi: 10.18653/v1/2020.emnlp-main.213.
64. N. M. Guerreiro, R. Rei, D. van Stigt, L. Coheur, P. Colombo, and A. F. T. Martins, "xCOMET: Transparent machine translation evaluation through fine-grained error detection," Trans. Association for Computational Linguistics, vol. 12, pp. 979–995, 2024, doi: 10.1162/tacl_a_00683.
65. M. Freitag, G. Foster, D. Grangier, V. Ratnakar, Q. Tan, and W. Macherey, "Experts, errors, and context: A large-scale study of human evaluation for machine translation," Trans. Association for Computational Linguistics, vol. 9, pp. 1460–1474, 2021, doi: 10.1162/tacl_a_00437.
66. T. Kocmi and C. Federmann, "Large language models are state-of-the-art evaluators of translation quality," in Proc. 24th Annu. Conf. European Association for Machine Translation (EAMT), Tampere, Finland, 2023, pp. 193–203.
67. N. M. Guerreiro, D. M. Alves, J. Waldendorf, B. Haddow, A. Birch, P. Colombo, and A. F. T. Martins, "Hallucinations in large multilingual translation models," Trans. Association for Computational Linguistics, vol. 11, pp. 1500–1517, 2023, doi: 10.1162/tacl_a_00615.
68. G. Stanovsky, N. A. Smith, and L. Zettlemoyer, "Evaluating gender bias in machine translation," in Proc. 57th Annu. Meeting ACL, Florence, Italy, 2019,pp. 1679–1684, doi: 10.18653/v1/P19-1164.