A comprehensive overview of LLM-based approaches for machine translation

Bhuvaneswari Kumar[1], Varalakshmi Murugesan[2]
[1]School of Computer Science Engineering and Information Systems, Vellore Institute of Technology, Tamil Nadu, India
[2]School of Computer Science and Engineering, Vellore Institute of Technology, Tamil Nadu, India

Article Info

Article history:

Received Jul 18, 2024
Revised Oct 21, 2024
Accepted Oct 30, 2024

Keywords:

Large language models
Low-resource languages
Machine translation
Prompting
Quality estimation
Tuning

ABSTRACT

Statistical machine translation (SMT) used parallel corpora and statistical models, to identify translation patterns and probabilities. Although this method had advantages, it had trouble with idiomatic expressions, context-specific subtleties, and intricate linguistic structures. The subsequent introduction of deep neural networks such as recurrent neural networks (RNNs), long short-term memory (LSTMs), transformers with attention mechanisms, and the emergence of large language model (LLM) frameworks has marked a paradigm shift in machine translation in recent years and has entirely replaced the traditional statistical approaches. The LLMs are able to capture complex language patterns, semantics, and context because they have been trained on enormous volumes of text data. Our study summarizes the most significant contributions in the literature related to LLM prompting, fine-tuning, retrieval augmented generation, improved transformer variants for faster translation, multilingual LLMs, and quality estimation with LLMs. This new research direction guides the development of more efficient and innovative solutions to address the current challenges of LLMs, including hallucinations, translation bias, information leakage, and inaccuracy due to language inconsistencies.

Corresponding Author:

Varalakshmi Murugesan
School of Computer Science and Engineering, Vellore Institute of Technology
Tamil Nadu, India
Email: mvaralakshmi@vit.ac.in

1. INTRODUCTION

A technique that uses algorithms to translate text from one language to another automatically is referred to as machine translation (MT). Earlier MT models relied on statistical approaches implemented with large parallel corpora of text [1]. Statistical models and intricate feature engineering were the foundation of the statistical machine translation (SMT) system, which employed syntax-based or phrase-based models in translating source sentences into target sentences. SMT ensures that every word in the source sentence is translated into a semantically relevant target phrase [2]. The development of translation rules, dictionaries, and parallel corpora for SMT systems required a significant amount of human intervention. In SMT, all translations are distinctly memorized, which includes rare words, and every word is treated as a discrete symbol [3]. Despite these merits, SMT systems struggled to produce fluent translations and handle long-range dependencies. In order to overcome the limitations of SMT, neural machine translation (NMT) systems employ sequence-to-sequence (seq2seq) models that are based on the encoder-decoder architecture. The encoder neural network converts the source sentence into a vector representation, and this encoded vector is fed into the decoder that applies teacher-forcing to use the ground truth instead of previously decoded words and generates one word at a time during the target translation. With the advent of transformers with attention mechanisms [4], the models trained with large parallel corpora learned to maximize the probability of producing the correct target translations for the given source sentence. NMT models learn from data directly to generate more accurate and coherent translations with an improved ability to manage long-distance dependencies [5], capture contextual information, and adapt to new language pairs or domains when fine-tuned. Though attention mechanisms, transformer architectures, and multi-source NMT models have surpassed traditional SMT methods in enhancing the quality of machine-translated data, they still struggle with rare words, proper nouns, out-of-vocabulary terms, training resource-limited language pairs, and out-of-domain data; furthermore, it is computationally expensive to train with large datasets and deploy larger models.

"Large" language models, as their name suggests, are pre-trained transformer models trained on huge volumes of data for various tasks such as text summarization, translation, question answering, and sentiment analysis. These models are further fine-tuned or prompt-tuned with smaller, task-specific datasets to meet the tailored requirements of stakeholders. Language models have evolved substantially from rule-based systems to generative artificial intelligence (AI) models like GPT-3, and bidirectional encoder representations from transformers (BERT). The availability of voluminous data, innovative architectural designs, and technological enhancements enriched further breakthroughs in generative models. Through unsupervised training on vast amounts of data, pre-trained large language models (LLMs) can capture a wide range of knowledge, identify the statistical patterns and relationships in the training data, and improve language comprehension to produce human-like text. LLMs reduce computational costs and resource requirements to generate more accurate and coherent translations. The rapid advancement of LLMs has led to a significant paradigm shift in MT due to their generation capabilities, contextual understanding, fluency, and accuracy. LLMs can reshape the MT research by integrating with different NLP applications, such as summarization and question-answering, exploring rare, unseen languages, and making wider room for the research community. In the efforts to advance MT with LLMs [6]-[11], researchers have explored efficient fine-tuning strategies, prompting approaches, multi-lingual models, low-resource settings, and domain adaptation techniques to get more fluent and quality translations.

Gaps identified - the existing literature reviews focus only on SMT and NMT that investigate several approaches for enhancing translation quality [12]-[16] but they fail to analyze the numerous works published on employing LLMs for MT. Sensing this significant gap in the literature due to the absence of a comprehensive survey on LLM-based MT, this paper presents an in-depth review of the latest approaches in LLM translation. It aims to serve as a pivotal resource for researchers seeking to deepen their understanding and knowledge in this domain. The exhaustive survey spearheads diverse LLM research domains geared towards addressing the inherent challenges of LLMs including hallucinations, translation bias, inconsistencies in the languages that impact translation accuracy, and information leakage leading to data security breaches. In the application side, this review is of great help to the research groups that explore the possibilities of using LLMs for the translation of real time conversation that supports seamless interaction and real time translation of social media posts in multiple languages, multilingual generation of subtitles, captions and dubbed audio for videos without trading off the original emotions, translation of literary works and legal documents with high accuracy, improved translation of sign language and spoken languages and multilingual chatbot to interact with customers in their preferred languages.

This review paper presents a total of 80 existing scholarly articles on LLMs for MT, published during the period 2017 to 2024. The articles are grouped into five major sections based on their key research focus and contributions – pioneering research on LLM-based MT, LLM prompting, LLM fine-tuning, LLMs for low-resource languages, and quality estimation with LLM. Figure 1 depicts how LLMs for MT can be categorized based on their research areas.

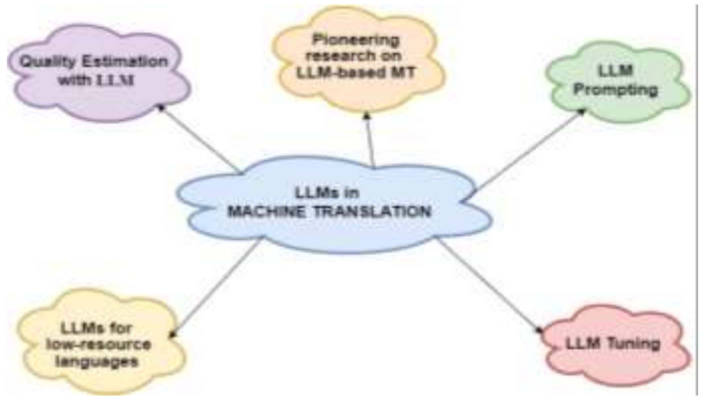

Quality Estimation with LLM
Pioneering research on LLM-based MT
LLM Prompting
LLMs in MACHINE TRANSLATION
LLMs for low-resource languages
LLM Tuning

Figure 1. Major categories in this review paper

2. PIONEERING RESEARCH ON LLM-BASED MT

2.1. Sequence-to-sequence models

In MT, LLMs are typically trained on a parallel corpus of translation data using seq2seq models. Some of the earliest and most influential works in this area include a comprehensive toolkit designed for sequence modeling tasks such as MT, abstractive document summarization, story generation, error correction, and multilingual embeddings that can be applied to research settings. This toolkit has enabled rapid inference for non-recurrent models by increasing prediction rates through incremental decoding and model state caching [17]. A translation model for low-resource languages uses a multilingual NMT model based on transfer learning to find shared patterns, structures, and features in high-resource languages. This resolves the problem of data scarcity issues in low-resource languages [18].

2.2. Non–autoregressive models

While seq2seq models have been successful in NMT, they suffer from slow inference speed as they generate translations token-by-token. In order to address this issue, an edit-based transformer model, "EDITOR," was developed and trained through reinforcement learning [19]. This model aims to improve translation quality by enabling source token repositioning during translation and incorporating soft lexical constraints tailored to the user's preferences, which helps deal with word choice and complex reordering problems. Unlike constrained beam search methods, this edit-based model accelerates decoding considerably by enabling parallel editing during the decoding process, resulting in faster translation speeds without sacrificing translation quality. The Levenshtein transformer [20] enhances sequence generation by incorporating insertion and deletion operations. This approach offers versatility, efficiency, and adaptability in various tasks, including MT, text summarization, and refinement processes. The model demonstrates its adaptability through its ability to apply a MT-trained Levenshtein transformer directly to automatic post-editing tasks without any changes. These models introduce techniques like iterative refinement and sequence-level distillation to enable non-autoregressive, rapid translation while maintaining high quality.

2.3. Retrieval-augmented models

Researchers explore fuzzy matching techniques to identify similar translations and incorporate them into training data. Xu *et al.* [21] compares the results of these techniques with a baseline model that does not use augmentation. Furthermore, he focuses on augmenting models by explicitly retrieving information from a translation memory or a database of translation examples. The retrieve-edit-rerank framework aims to improve the quality of final outputs in text generation tasks. This approach involves retrieving potentially relevant outputs for each input, modifying each candidate individually, and then reranking the edited candidates to determine the best output based on post-generation ranking [22]. The framework trains a transformer-based seq2seq editing model by concatenating the input with the retrieved output. This method demonstrates enhanced performance on text generation when tested on MT datasets and particularly improves translation quality for resource-limited language pairs or domain-specific terminology.

2.4. Multilingual LLMs

Researchers explore multilingual MT by leveraging the massively multilingual nature of some LLMs like mT5 and multilingual bidirectional and auto-regressive transformer (mBART). This approach allows a single model to translate between multiple language pairs. The mBART model, a seq2seq architecture, is pre-trained on extensive monolingual corpora in various languages. It uses word-span masking and sentence permutation techniques to enhance denoising and translation accuracy [23]. To fine-tune the model for translation tasks, researchers feed source sentences into an encoder and decode each target sentence. The model's performance is evaluated on sentence-level and document-level MT tasks. Recent studies on MT strive to eliminate language barriers globally. However, despite such efforts, many resource-limited languages remain underserved. In order to address this challenge, researchers are developing a sparsely gated mixture of expert conditional models for low-resource language datasets. These datasets are created using novel mining tools [24]. The approach reduces the performance gap between high-resource and low-resource languages. By implementing various architectural and training enhancements, these models outperform previous state-of-the-art systems. This progression paves the way for the development of a universal translation system.

3. LLM PROMPTING

LLM prompting involves crafting source text to elicit desired output from language models. Effective prompting can effectively enhance LLM performance without further training. The ChatCite framework [25] incorporates human workflow guidance and reflective incremental mechanisms. It extracts salient elements from the related literature and generates comprehensive summaries. By utilizing carefully crafted prompts, researchers and practitioners can guide LLMs to produce more accurate, relevant, and targeted responses. The knowledge-prompted estimator method [26] enhances segment-level estimation in MT by integrating various prompting techniques. This approach combines three one-step prompting techniques: perplexity, token-level similarity, and sentence-level similarity. Additionally, it incorporates two chain-of-thought (CoT) prompting evaluations: perplexity-token prompting and perplexity-token-sent prompting. To evaluate MT quality at the segment level, researchers employ different scoring methods. These include scalar scoring, 5-star scoring, and 5-category scoring. These scoring methods allow for a comparison of the performances of different prompting techniques, potentially leading to more effective translation systems and evaluation methods. Researchers investigate the efficacy of in-context learning within LLMs for MT tasks. It includes experimenting with various types of task instructions, examining perturbations within in-context demonstrations, analyzing directionality effects, and studying misalignment susceptibility [27]. To enhance the cultural awareness and accuracy of MT systems, recent studies explore the importance of leveraging innovative metrics, culturally specific datasets, and prompting strategies [28].

The conversational SIMULMT framework [29] enhances the efficiency of LLM-based translation. This framework demonstrates strong performance by optimizing the inference process, reducing latency, and maintaining translation quality in real-time simultaneous translation tasks. Researchers suggest the DecoMT approach [30], a decomposed prompting strategy to enhance the MT between related languages. It leverages monotonic alignment and incorporates context-aware translation, resulting in more precise and robust translations than traditional MT methods. The study [31] investigates gender bias in translations using LLMs and compares them to traditional NMT models. By leveraging specific prompt templates and relevant in-context examples (ICEs), LLMs produce tailored outputs that are more precise and robust. The LLMs demonstrate improved performance without requiring additional training or finetuning. This is achieved using deep syntax-level knowledge during the in-context example selection process. The top-k syntactically similar examples are chosen based on a polynomial distance metric and an ensemble strategy that integrates word-level closeness and syntax-level similarity [32]. ChatGPT's translation capabilities are enhanced by incorporating translation task information, context domain information, and part-of-speech (POS) tag components [33]. Subsequently, ChatGPT outperforms Google Translate and DeepL Translate in MT tasks. A pre-edit scheme and a two-step prompt strategy are introduced to incorporate linguistic knowledge and customized prompts. It guides the ChatGPT model in effectively handling the complexities of translating attributive clauses in low-resource scenarios [34]. While LLMs and transfer learning play a vital role in advancing low-resource MT, challenges such as data scarcity, domain mismatches, and difficulties with distant language pairs remain. Ongoing efforts focus on developing more robust and efficient models, utilizing data augmentation techniques, and applying linguistic knowledge to further enhance performance. With the advent of more advanced prompting strategies and tools, LLM prompting can effectively leverage the power of LLMs in the future. Table 1 in Appendix lists some papers on LLM prompting with their corresponding datasets, models, methods, and language pairs used in each paper.

4. LLM TUNING

LLMs trained on vast datasets require tuning to perform specific tasks efficiently. Tuning enables models to adapt to specialized applications, resulting in improved performance. Effective tuning techniques utilize limited computational resources and data. LLMs trained on extensive datasets require tuning to efficiently perform specific tasks. This tuning allows models to adapt to specialized applications, leading to enhanced performance. Effective tuning techniques utilize limited computational resources and data. The BigTranslate model [35], a multilingual translation model, begins by training on a large volume of monolingual Chinese data, followed by a vast parallel dataset. This process incorporates an incremental data sampling strategy with 1,000 parallel sentence pairs for each language pair. By addressing the issue of unbalanced language proficiency, the model achieves mastery across 102 languages during its multilingual learning journey. ML50 benchmarks are developed to create multilingual translation models by combining multilingual pretraining with monolingual data, particularly for languages with limited bitext resources. These models are later fine-tuned with parallel data [36]. A two-stage fine-tuning algorithm [37] enhances the ability of LLMs to follow instructions. At first, the LLM is fine-tuned on a translation dataset using the maximum likelihood estimation loss. The second stage introduces an extra unlikelihood loss to learn from instruction-conflicting examples, where correct translations are randomly replaced with incorrect ones. The LLMs-based E-commerce machine translation (LEMT) approach [38] focuses on utilizing LLMs, gathering e-commerce resources (including a parallel corpus for e-commerce domains and specialized term pairs), optimizing the tokenizer, and implementing a rigorous two-stage fine-tuning and self-contrastive enhancement process. This approach enables the model to learn e-commerce translation features effectively. A multi-step approach leverages LLMs for generating synthetic bilingual terminology data. This process integrates technical terms into the translation model. Later, a generic encoder-decoder MT model undergoes fine-tuning by combining the synthetic terminology with the original training data. This combination allows the model to generate high-quality translations that are specifically well-suited for specialized domains [39]. The simultaneous translation (SimulMT) demonstrates impressive performance during SimulMT inference by employing more intricate decoding techniques and various prompting approaches [40]. Pre-trained LLMs fine-tuned on a resource-constrained dataset can perform both simultaneous translation and input segmentation [41].

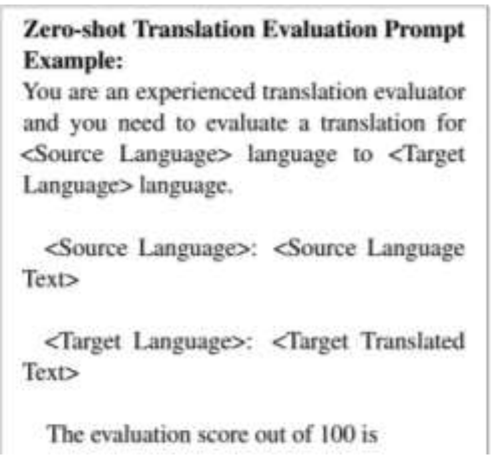
This approach ensures that source words are causal relative to their corresponding target words, providing a highly effective and efficient method for direct supervision. SiLLM, an integrated LLM, utilizes the correlation between translation and policy-decision agents to achieve SiMT. It helps to overcome the vocabulary mismatch problem [42]. This approach leverages the strengths of LLMs in understanding context and generating coherent translations while addressing the specific challenges of simultaneous translation, such as latency and accuracy trade-offs. Fine-tuning Mistral 7B can enhance its in-context learning capability through a combination of zero-shot and one-shot prompts for adaptive MT [43]. This approach shows notable improvements in translation quality when tested on specific domains with limited translation pairs. For domain-specific MT tasks, LlamaIT uses lonf range (LoRA) prompt-tuning on the Llama2-7B model. By integrating domain-specific bilingual vocabulary into the input source sentence, it reduces the need for post-processing or in-context examples [44]. Two translation approaches use different instruction formats. The first uses bilingual pairs and the Alpaca dataset for fine-tuning. The second, Llama2-7B, undergoes continuous pretraining on concatenated translation pairs and is fine-tuned using the Alpaca dataset. These methods leverage existing datasets and fine-tuning them to enhance translation capabilities [45]. The contrastive preference optimization (CPO) approach [46] develops high-quality preference data for MT models. This enables the models to generate high-ranking translations and reject flawed ones, helping to avoid inadequate translations and overcome the limitations of supervised fine-tuning. Fine-tuning with adapters proves to be an effective method for guiding language models (LLMs) in enhancing translation tasks. By adding a few-shot examples during the fine-tuning process, this approach not only matches the performance of traditional fine-tuning but also reduces computational costs [47]. A fine-tuned LLM creates a dataset from cybercrime chats by employing eight LLM models to translate messages. This method achieves quick, more precise translations by encapsulating the subtleties of the language, yielding high-quality translations at considerably lower costs than a human translator [48].

A multiplicative joint scaling law proposes a systematic study of various scaling factors by selecting the best fine-tuning strategies that impact the performance of fine-tuning LLMs in resource-limited scenarios [49]. The optimal fine-tuning method is highly task- and data-dependent, whereas parameter-efficient tuning fosters better zero-shot transfer than full model tuning. For document-level machine translation (DOCMT) tasks across multiple languages, LLMs show better generalization to out-of-domain text and context awareness through well-designed prompt-efficient fine-tuning, context structure, and natural instructions [50]. A new generative paradigm called "GenTranslate" leverages the strong reasoning abilities of LLMs to integrate diversified translation variants from the N-best list to produce high-quality outputs [51]. To generate different responses to instructions, an instruction-tuned LLM is constructed that effectively distinguishes quality translations and learns from contrasting examples by fine-tuning LLMs [52]. Using a resource-constrained parallel corpus to generate high-quality translation data, the LLMs are fed with examples of correct and incorrect translations for the same input and employ preference comparison for better regularization. LLM tuning harnesses the power of advanced models for specific tasks. As the field advances, more efficient and effective tuning methods strike a balance between task-specific performance and general potentialities. Listed in Table 2 in Appendix are a few papers about LLM tuning, with corresponding datasets, models, methods, language pairs, and metrics.

5. LLMS FOR LOW-RESOURCE LANGUAGES

Preserving global multilingualism and ensuring technological inclusion is imperative in developing LLMs for low-resource languages with constrained text data. Innovative methods and approaches are making strides in this area to overcome challenges such as developing efficient pre-training methods, cross-lingual knowledge transfer, multimodal integration, and incorporating linguistic knowledge into model architectures. A cross-search approach comprises antagony-cross search and similarity-cross search techniques. The antagony-cross search uses token-level control to produce monolingual data closely aligned with the target domain. Similarity-cross search generates target language content that is more semantically related to the source language. It employs a similarity score in back translation to maintain alignment between source and target sentences [53]. LLMs enhance Ge'ez translation quality and consistency through domain-specific vocabulary, user feedback integration, and similarity-based sentence retrieval from a parallel corpus. These sentences are used as context samples with LLMs to create translations for new source sentences from Ge'ez [54]. A probability-driven meta-graph prompter (POMP) enhances LLMs' ability to translate low-resource languages by sampling the language-specific directed acyclic meta-graph to generate multiple translation paths [55]. It prompts LLMs to generate target sentences and updates the likelihood of auxiliary languages in different directions based on backpropagated reward scores. Contrastive alignment instructions (AlignInstruct) on LLMs effectively translate unseen languages using MTInstruct (model fine-tuning via MT instructions). Its focal point is cross-lingual supervision, which employs a cross-lingual discriminator constructed from statistical word alignments [56].

Multi-lingual large language models (MLLMs) demonstrate improved performance by fine-tuning adaptMLLM, which is trained on two low-resource, in-domain language pairs [57]. It simplifies the process of optimizing multilingual language models by designing a tailored, user-friendly web interface for harnessing models as a translation service within the application. The DIPMT approach simplifies translations for languages with limited resources by effectively incorporating dictionary knowledge into the prompt and adding a few-shot illustration to acquaint the model with a specific framework [58]. A new programmer-interpreter technique improves LLM performances by harnessing the interpreter's domain generalization expertise and encoding task-specific knowledge through the programmer's competence [59]. LLMs translate Ukrainian folktales into English while preserving their meaning and literary style by including an additional layer of culturally relevant data and testing different prompt techniques on the LLMs [60]. In order to overcome off-target translations and hallucinations, source-contrastive and language-contrastive decoding methods are introduced by providing the correct input and language indicator [61]. Figures 2 and 3 illustrate the prompt for zero-shot reference-less translation evaluation and the prompt for example-based in-context learning with LLM [62]. Table 3 in Appendix shows some papers in LLMs for low-resource languages that have been tabulated with their datasets, models, methods, and language pairs, as well as the results from their research.



Zero-shot Translation Evaluation Prompt Example:

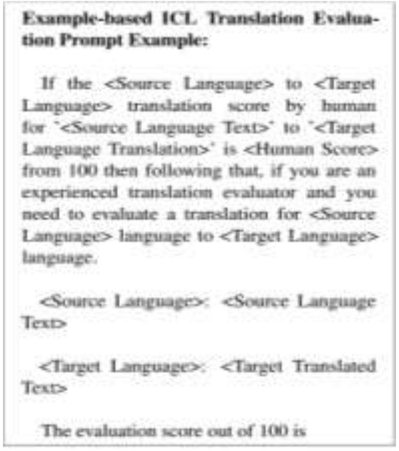
You are an experienced translation evaluator and you need to evaluate a translation for <Source Language> language to <Target Language> language.

<Source Language>: <Source Language Text>

<Target Language>: <Target Translated Text>

The evaluation score out of 100 is

Figure 2. Prompt for zero-shot reference-less translation evaluation [62]



Example-based ICL Translation Evaluation Prompt Example:

If the <Source Language> to <Target Language> translation score by human for '<Source Language Text>' to '<Target Language Translation>' is <Human Score> from 100 then following that, if you are an experienced translation evaluator and you need to evaluate a translation for <Source Language> language to <Target Language> language.

<Source Language>: <Source Language Text>

<Target Language>: <Target Translated Text>

The evaluation score out of 100 is

Figure 3. Prompt for example-based in-context learning with LLM [62]

6. QUALITY ESTIMATION WITH LLM

Human language is complex and nuanced, making it challenging to estimate the quality of MT. MT quality estimation focuses on the model rather than its output. LLM-based MT systems are often benchmarked using automated metrics like BLEU and human evaluation, which typically measure adequacy and fluency. QE techniques have changed over time and are crucial for evaluating the quality of machine-translated content at various granularities, ranging from words to entire documents. A comprehensive analysis of MT quality estimation (MTQE) research throws light on different methodologies with handcrafted features for deep learning and LLMs in QE [63]. Researchers construct challenge sets containing word swap, hallucination, coreference, and unit conversion errors to evaluate the ability of MT metrics to distinguish between accurate and inaccurate translations [64]. Researchers examine the impact of multilingual embeddings, metric sensitivity, and the need to integrate language-specific information into the evaluation process. These assessments are conducted at both the phenomenon and language levels to gauge the capability of MT metrics. To generate accurate translations between many language pairs, the KG-BERTScore (a reference-free metric) and the HWTSC-EE-Metric (a reference-based metric) offer segment-level and system-level scoring for quick evaluation and comparison of MT systems on large corpora [65]. A reference-free approach, EvLP (evaluation via LLMs polishing), where LLMs are prompted and used as annotators to "polish" the translated text by post-editing. The potential bias of LLM is investigated to enhance the assessment of MT quality following human intervention and refinement [66].

A perturbation-based QE technique is making MT system outputs more flexible, adaptive, and domain-independent across different language pairs and directions. This is achieved by perturbing source sentences and assessing how different source words influence the generation of translated words [67]. A tuned encoder-based model produces better results than a tuned decoder-based model. It captures context and surface word sequences in MT and semantic textual similarity tasks [68]. GPTSCORE, an evaluation framework, scores the fluency and accuracy of translated texts by leveraging generative pre-trained models like GPT-3. Using instruction prompt templates with annotated examples, these models calculate the conditional probability of producing high-quality translated text [69]. Researchers create specific prompts for the LLM performance predictor (LLM-PP) to evaluate how well deep neural network architectures perform in MT tasks [70]. They use this information to build a multilayer perceptron (MLP) regression model that remains effective while also reducing costs. Coarse-grained and fine-grained prompts are used to evaluate the performance of different LLMs in four distinct input modes. This approach examines how LLMs utilize source and reference information to assess translations [71]. A Div-ref method is suggested to evaluate generated texts and improve the correlation between automatic metrics and human evaluation results. It incorporates diversified reference sentences into different expressions while maintaining semantic consistency to eliminate bias and insufficiency associated with single references [72]. The optimized LLMs predict the need for post-editing in MT tasks and detect the best model configuration and size. This aims to provide accurate and productive outcomes in evaluating MT quality [73]. AUTOmatic multidimensional quality metrics (AUTOMQM) is developed to leverage LLMs' reasoning and in-context learning skills. This technique assesses the quality of MT systems by producing more accurate and contextually relevant feedback than human experts, without the need for further training or fine-tuning [74]. LLMs now perform more in-depth translation analysis, including locating specific error spans and categorizing faults according to the MQM framework. An unsupervised QE framework is being developed that relies on LLM's zero-shot ability for MT quality estimation. This framework eliminates the need for extensive training data, supervision, or references and closely aligns with human assessments [75]. A deep interaction-based evaluation paradigm enables the assessment of LLMs in dynamic real-world scenarios. LLMs adopt the writer and editor roles and partake in a writing-polishing process where the results are compared and assessed for the simultaneous evaluation of writing and polishing skills. Based on semantic consistency and polishing accuracy, a judge model assesses the LLMs' translation and proofreading effectiveness in the MT task [76]. Llama 2 LLM employs rich semantic embeddings to compute the cosine similarity between semantic embeddings using the Embed_Llama metric [77]. It suggests that adding more layers to the Llama 2 model may help to comprehend words better and evaluate translations more effectively. A "QE-fusion" approach uses computationally efficient quality estimation metrics to fuse translation candidates into a synthesized output. This approach demonstrates substantial improvements in generating divergent outputs from LLMs compared to NMT systems [78].

An INSTRUCTSCORE framework provides a numerical score and a detailed qualitative diagnostic report on the generated text [79]. Through fine-tuning feedback mechanisms, this metric aligns with human judgments and is being tested across multiple domains and tasks, resulting in a more accurate assessment of text generation quality. Researchers are experimenting with diverse prompt templates on various GPT models and evaluating them using the GPT estimation metric-based assessment (GEMBA) tool [80]. GEMBA is compared with other quality estimation metrics, where the GPT-4 model using GEMBA outperforms other metrics at the segment level and effectively analyses translation quality. Recent research examines the benefits and drawbacks of LLMs for MT using test sets designed to investigate specific language phenomena, domain resilience, and other skills. Questions remain about optimizing pre-trained LLMs for MT and precisely calculating the amount of in-domain training data required to achieve high-quality outcomes in specialized fields like technical, medical, or legal translation. LLMs have difficulties with consistently translating pronouns, understanding colloquial idioms, and preserving context during lengthy text passages. These challenges highlight areas for further improvement in LLM-based MT systems. Table 4 in Appendix outlines the datasets, models, methods, language pairs used, and results of a few papers in quality estimation with LLMs.

7. RESULTS AND DISCUSSION

This study reveals that LLMs represent a significant advancement in the field of MT, offering more versatile, comprehensive, and high-quality translation capabilities. However, the existing literature is limited to only SMT and NMT approaches for enhancing translation quality and falls short of investigating the extensive works on employing LLMs for MT. Our work attempts to close this gap by presenting a comprehensive analysis of LLM-based approaches for MT.

This paper suggests that LLMs can be trained on data from multiple languages simultaneously without having to build separate models. Owing to their context awareness, they demonstrate an enhanced capability for open-vocabulary translation in handling neologisms and rare and unseen words. LLMs' strong language modelling abilities contribute to enhanced fluency, producing more accurate, natural-sounding, and fluent outputs compared to phrase-based SMT approaches.

The findings of this review are distinct from the existing literature that predominantly addresses SMT and NMT approaches. The current literature pays no attention to LLM-based MT works. This study encapsulates the most noteworthy recent contributions in the literature related to LLM prompting, fine-tuning, retrieval augmented generation, improved transformer variants for faster translation, multilingual LLMs, and quality estimation with LLMs.

LLMs face several challenges and limitations in MT tasks regardless of their befitting outcomes. Notably, the lack of explicit word or phrase alignments between languages may impact accuracy, in contrast to SMT approaches. Furthermore, LLMs may generate factually inconsistent translations and are prone to "hallucinating" facts discordant with the input. It is a significant concern for high-stakes domains requiring precise translations. Additionally, the computational requirements of huge LLMs may restrict scalability for production MT deployments. Finally, LLMs require extensive multilingual training datasets, which may not be readily available for all language pairs, leaving them "data hungry" and limiting their potential. This data scarcity can limit the models' effectiveness across diverse languages with limited resources and domains.

Our collection of choicest LLM articles is a handy resource for quick reference for future LLM researchers. It serves as a beacon, guiding the development of more efficient and innovative solutions for future works to address the current challenges of LLMs that include hallucinations, translation bias, information leakage, and inaccuracy due to language inconsistencies. Future studies should advance towards the application-based research works with LLMs including, but not limited to using LLMs for the translation of real-time conversation that supports seamless interaction, real-time translation of social media posts in multiple languages, multilingual generation of subtitles, captions, and dubbed audio for videos without trading off the original emotions, translation of literary works and domain-specific documents with high accuracy, improved translation of sign language and spoken languages and multilingual chatbots.

8. CONCLUSION

This review is a maiden attempt towards providing a detailed and comprehensive analysis of the existing literature on employing LLMs for MT. To the best of our knowledge, LLM based translation works have not been reviewed in the literature, to date. Our study highlights the most significant contributions made in LLM prompting and fine-tuning that are regarded as the two powerful techniques for LLM performance enhancement. It also covers the other major works related to LLMs including retrieval augmented generation, improved transformer variants for faster translation, multilingual LLMs, and automatic prediction of the quality of machine translated output.

APPENDIX

Table 1. Datasets, models, methods, language pairs, and metrics of a few papers in LLM prompting

Paper	Dataset	Models	Language pairs	Method/prompts	Score
[28]	OPUS	NLLB	EN-ZH, ZH-EN	Zero-prompts	CultureSpecificItems-
	Samanantarv0.2	NLLB-A	EN-FR, FR-EN	Two-shot prompts	CSI-Match [NLLB-R]
		NLLB-R	EN-ES, ES-EN		78.7, 79.8
		LLAMA2	EN-HI, HI-EN		92.6,92.1
		LLAMA2-A	EN-TA, TA-EN		94.0,95.2
		LLAMA2-R	EN-TE, TE-EN		83.6, 98.3
		CHATGPT			81.6, 97.9
		GOOGLE			89.8,94.7
[30]	FLORES	BLOOM	hin↔mal,	Few-shot prompting	SPBleu[mT5]
		7.1B	hin↔mar,		3.0
		mT5 3.7B	hin↔guj,		3.6
		XGLM 7.5B	hin↔tel,		3.2
			ind↔zsm,		3.6
			rus↔ukr		4.9
					4.5

Table 1. Datasets, models, methods, language pairs, and metrics of a few papers in LLM prompting *(Continued)*

Paper	Dataset	Models	Language pairs	Method/prompts	Score
[31]	LINGUALHOLISTICBIAS FLoRes's test set WinoMT & BUG	NLLB LLaMa	Arabic, Cyrillic Latin, Tamil, Greek, Thai Devanagari	In-context examples (ICE-5,16,32)	BLEU [avg] 0.31 [5-ICE] 0.63 [16-ICE] 1.02 [32-ICE]
[32]	FLORES+ WikiMatrix v1	XGLM7.5B	DE↔EN ES↔EN FR↔EN JA↔EN RU↔EN ZH↔EN	In-context examples (ICE-2,4,16) Polynomial, BM25 + Polynomial, Polynomial + BM25	BLEU [avg into EN] 30.09 30.98 30.79 [avg out of EN] 23.31 24.35 24.39
[33]	Flores-101	Google DeepL	En↔ Es En↔Fr Es↔Fr	Few-shot prompts [0,1,5 shots]	BLEU[Google] 23.49, 25.32 54.75, 49.66 26.89, 22.48

Table 2. Datasets, models, methods, language pairs, and metrics of a few papers in LLM tuning

Paper	Dataset	Models	Language pairs	Method/prompts	Score
[37]	FLORES-200 WMT'21 WMT'22 WMT'23	ALMA-13B-LoRA GPT-4	cs↔en, de↔en, is↔en, zh↔en, ru↔en	Fine-tuning contrastive reference optimization+	Avg [En-XX] 83.34 [KIWI-22] 85.74 [KIWI-XXL] 94.05 [XCOMET]
[42]	OPUS	NLLB 3.3B gpt-3.5-turbo Mistral 7B Mistral7B+Fine-tuned	Spanish→English	zero-shot one-shot Fine-tune	BLEU [1-shot] 47.42 48.34 47.35 49.69
[43]	Flores-101 OPUS-100 IT XFIT24	Llama2-7B LlamaIT	Chinese→English and English→Chinese	fine-tuning with LoRA zero-shot prompting dictionary-based prompt	BLEU [LlamaIT] 22.04,32.60 35.91,37.79 36.24,40.41 55.16,63.76
[46]	OPUS Flores-200 WMT22	LLaMA 7B and LLaMA 13B	nl↔ en fr↔ en de↔en pt↔en ru↔en	Fine-tuning Zero-shot Few-shot	BLEU [zh↔en] 32.44 [Format1] 32.62 [Format2] 32.39 [Format3]
[52]	FLORES-200	BLOOMZ-7b-mt LLaMA-2-7b Alpaca MT TIM	Zh⇒En En⇒Zh De⇒En En⇒De	Tuning with low-rank matrices Tuning with embedding fixed. Tuning full parameters	BLEU [MT-FixEmb] 26.41 33.80 42.14 32.23

Table 3. Datasets, models, methods, language pairs, and metrics of a few papers in LLMs for low-resource languages

Paper	Dataset	Model	Language pair	Method/prompts	Score
[54]	Opus corpus and the AAU Ethiopian Languages corpus	Bilingual, Multilingual, NLLB-200, GPT-3.5 text-davinci-003	en-gez, gez-en	- - Fine-tuning Few-shot	BLEU 4.1, 9.91 13.07, 16.67 0.2, 3.8 9.2
[55]	OPUS, WMT-News-v2019, CCAligned, wmt19test and Flores-200 Testset	Cross-lingual transfer NMT model, Language-specific Meta-Graph	Gu-En, Kk-En, Si-En	POMP + In-Context Learning	BLEURT 75.20 71.84 70.17
[56]	OPUS-100, Flores-200	BLOOMZ-7b1	OPUS en-xx OPUS xx-en FLORES en-xx FLORES xx-en	Fine-tuning MT+Align+Hint+Revise	BLEU 12.00 19.68 3.40 11.67
[57]	LoResMT2021	adaptMLLM	mr-en-tuned en-mr-tuned en-ga-tuned ga-en-tuned	Fine-tuning	BLEU 52.6 26.4 41.2 75.1

Table 3. Datasets, models, methods, language pairs, and metrics of a few papers in LLMs for low-resource languages *(Continued)*

Paper	Dataset	Model	Language pair	Method/prompts	Score
[61]	M2M-100 SMaLL-100	Llama 2 model family	X-branch [C $_{src+lang}$] M2M-100 SMaLL-100	source-contrastive decoding, language- contrastive decoding	spBLEU 9.3 11.2
[62]	Llama-2-13b-Adpt	Bharat parallel corpus collection (BPCC)	English to 4 Indian Languages (Hindi, Gujarati, Marathi, Tamil, and Telugu)	Fine-tuned LLM and COMET-QE with reference less translation evaluation task	0.4574 [Spearman's Rank] 0.53744 [Pearson Rank] 0.3437 [Kendall's Rank]

Table 4. Datasets, models, methods, language pairs, and metrics of a few papers in quality estimation with LLMs

Paper	Dataset	Models	Language pairs	Method/prompts	Score
[64]	WMT20, WMT21, STS-B, SICK	RoBERTa-large, Cerebras-GPT	English to Japanese	RoBERTa fine-tuning, LLM LoRA-tuning, In-context learning	[Kendall's correlation] 0.699, 0.663 0.391, 0.383 0.737, 0.625 0.658, 0.483
[65]	MQM-2020	GPT3, GPT2, FT5-small, OPT	Chinese to English	Few-shot with demonstration	[Spearman correlation-Avg] 31.0 [Vanilla] 32.1 [Instruction] 33.3 [Inst+demo]
[69]	WMT22	gpt-3.5-turbo-instruct	Zh→En, En→De, En→Ru	single reference diversified reference	GEMBA [single/div] 36.3/37.0 29.5/29.7 32.1/33.9
[72]	WMT'22 WMT'19	PaLM and PaLM-2	en→de, zh→en, en→ru, en↔gu, en↔kk	AUTOMQM prompting In-context learning fine-tuning	segment-level prompt 0.275 [unicorn] 0.252 [unicorn] 0.209 [unicorn] 0.523, 0.334 [unicorn] 0.536, 0.433 [unicorn]
[79]	WMT22 WebNLG20 Flicker3K-CF Commongen BAGEL	GPT-4 LLaMA-7B	German to English	Fine-tune Fine-tune+refinement	InstructScore [Kendall and Pearson] 40.3/51.9 39.5/59.0 30.1/34.6 58.2 25.6/34.2

REFERENCES

[1] P. Koehn, F. J. Och, and D. Marcu, "Statistical phrase-based translation," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL '03*, 2003, vol. 1, pp. 48–54, doi: 10.3115/1073445.1073462.

[2] X. Wang, Z. Tu, and M. Zhang, "Incorporating statistical machine translation word knowledge into neural machine translation," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 26, no. 12, pp. 2255–2266, Dec. 2018, doi: 10.1109/TASLP.2018.2860287.

[3] X. Wang, Z. Lu, Z. Tu, H. Li, D. Xiong, and M. Zhang, "Neural machine translation advised by statistical machine translation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Feb. 2017, vol. 31, no. 1, pp. 3330–3336, doi: 10.1609/aaai.v31i1.10975.

[4] C. Park, Y. Yang, K. Park, and H. Lim, "Decoding strategies for improving low-resource machine translation," *Electronics*, vol. 9, no. 10, p. 1562, Sep. 2020, doi: 10.3390/electronics9101562.

[5] D. Bahdanau, K. H. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014, doi: 10.48550/arXiv.1409.0473.

[6] Y. Wang, J. Zhang, T. Shi, D. Deng, Y. Tian, and T. Matsumoto, "Recent advances in interactive machine translation with large language models," *IEEE Access*, vol. 12, pp. 179353–179382, 2024, doi: 10.1109/ACCESS.2024.3487352.

[7] K. Chen *et al.*, "General2Specialized LLMs translation for e-commerce," in *Companion Proceedings of the ACM Web Conference 2024*, May 2024, pp. 670–673, doi: 10.1145/3589335.3651510.

[8] B. Zhang, B. Haddow, and A. Birch, "Prompting large language model for machine translation: a case study," *Proceedings of Machine Learning Research*, vol. 202, pp. 41092–41110, 2023.

[9] Q. Luo, W. Zeng, M. Chen, G. Peng, X. Yuan, and Q. Yin, "Self-attention and transformers: driving the evolution of large language models," in *2023 IEEE 6th International Conference on Electronic Information and Communication Technology, ICEICT 2023*, Jul. 2023, pp. 401–405, doi: 10.1109/ICEICT57916.2023.10245906.

[10] P. J. Barclay and A. Sami, "Investigating markers and drivers of gender bias in machine translations," in *2024 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*, Mar. 2024, pp. 455–464, doi: 10.1109/SANER60148.2024.00054.

[11] Z. He *et al.*, “Exploring human-like translation strategy with large language models,” *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 229–246, Mar. 2024, doi: 10.1162/tacl_a_00642.

[12] C. S. Devi and B. S. Purkayastha, “An empirical analysis on statistical and neural machine translation system for English to Mizo language,” *International Journal of Information Technology (Singapore)*, vol. 15, no. 8, pp. 4021–4028, Sep. 2023, doi: 10.1007/s41870-023-01488-0.

[13] S. Maruf, F. Saleh, and G. Haffari, “A survey on document-level neural machine translation: methods and evaluation,” *ACM Computing Surveys*, vol. 54, no. 2, pp. 1–36, Mar. 2021, doi: 10.1145/3441691.

[14] S. Shi, X. Wu, R. Su, and H. Huang, “Low-resource neural machine translation: methods and trends,” *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 21, no. 5, pp. 1–22, Sep. 2022, doi: 10.1145/3524300.

[15] B. Haddow, R. Bawden, A. V. M. Barone, J. Helcl, and A. Birch, “Survey of low-resource machine translation,” *Computational Linguistics*, vol. 48, no. 3, pp. 673–732, Sep. 2022, doi: 10.1162/coli_a_00446.

[16] M. R. Costa-Jussà and M. Farrus, “Statistical machine translation enhancements through linguistic levels: a survey,” *ACM Computing Surveys*, vol. 46, no. 3, pp. 1–28, Jan. 2014, doi: 10.1145/2518130.

[17] M. Ott *et al.*, “Fairseq: a fast, extensible toolkit for sequence modeling,” in *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT (Demonstrations Session)*, 2019, pp. 48–53, doi: 10.18653/v1/n19-4009.

[18] N. Arivazhagan *et al.*, “Massively multilingual neural machine translation in the wild: findings and challenges,” *arXiv preprint 1907.05019*, Jul. 2019, doi: 10.48550/arXiv.1907.05019.

[19] W. Xu and M. Carpuat, “Editor: an edit-based transformer with repositioning for neural machine translation with soft lexical constraints,” *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 311–328, Mar. 2021, doi: 10.1162/tacl_a_00368.

[20] J. Gu, C. Wang, and J. Zhao, “Levenshtein transformer,” *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[21] J. XU, J. Crego, and J. Senellart, “Boosting neural machine translation with similar translations,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, vol. 2, pp. 1580–1590, doi: 10.18653/v1/2020.acl-main.144.

[22] N. Hossain, M. Ghazvininejad, and L. Zettlemoyer, “Simple and effective retrieve-edit-rerank text generation,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 2532–2538, doi: 10.18653/v1/2020.acl-main.228.

[23] Y. Liu *et al.*, “Multilingual denoising pre-training for neural machine translation,” *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 726–742, Dec. 2020, doi: 10.1162/tacl_a_00343.

[24] M. R. Costa-jussà *et al.*, “No language left behind: scaling human-centered machine translation,” *arXiv preprint 2207.04672*, Jul. 2022, [Online]. Available: http://arxiv.org/abs/2207.04672.

[25] Y. Li, L. Chen, A. Liu, K. Yu, and L. Wen, “ChatCite: LLM agent with human workflow guidance for comparative literature summary,” *arXiv preprint 2403.02574*, 2024, doi: 10.48550/arXiv.2403.02574.

[26] H. Yang, M. Zhang, S. Tao, M. Wang, D. Wei, and Y. Jiang, “Knowledge-prompted estimator: a novel approach to explainable machine translation assessment,” in *2024 26th International Conference on Advanced Communications Technology (ICACT)*, Feb. 2024, pp. 305–310, doi: 10.23919/ICACT60172.2024.10471974.

[27] P. A. Chitale, J. Gala, and R. Dabre, “An empirical study of in-context learning in LLMs for machine translation,” *arXiv preprint 2401.12097*, Jan. 2024.

[28] B. Yao, M. Jiang, T. Bobinac, D. Yang, and J. Hu, “Benchmarking machine translation with cultural awareness,” *arXiv preprint 2305.14328*, May 2023, doi: 10.48550/arXiv.2305.14328.

[29] M. Wang, T.-T. Vu, Y. Wang, E. Shareghi, and G. Haffari, “Conversational SimulMT: efficient simultaneous translation with large language models,” *arXiv preprint 2402.10552*, Feb. 2024, doi: 10.48550/arXiv.2402.10552.

[30] R. Puduppully, A. Kunchukuttan, R. Dabre, A. T. Aw, and N. Chen, “DecoMT: decomposed prompting for machine translation between related languages using large language models,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 4586–4602, doi: 10.18653/v1/2023.emnlp-main.279.

[31] E. Sánchez, P. Andrews, P. Stenetorp, M. Artetxe, and M. R. Costa-jussà, “Gender-specific machine translation with large language models,” *arXiv preprint 2309.03175*, Sep. 2023, doi: 10.48550/arXiv.2309.03175.

[32] C. Tang, Z. Wang, and Y. Wu, “Going beyond word matching: syntax improves in-context example selection for machine translation,” *arXiv preprint 2403.19285*, Mar. 2024, doi: 10.48550/arXiv.2403.19285.

[33] Y. Gao, R. Wang, and F. Hou, “How to design translation prompts for ChatGPT: an empirical study,” *arXiv preprint 2304.02182*, Apr. 2023, doi: 10.48550/arXiv.2304.02182.

[34] W. Gu, “Linguistically informed ChatGPT prompts to enhance Japanese-Chinese machine translation: a case study on attributive clauses,” *arXiv preprint 2303.15587*, Mar. 2023, doi: 10.48550/arXiv.2303.15587.

[35] W. Yang, C. Li, J. Zhang, and C. Zong, “BigTranslate: augmenting large language models with multilingual translation capability over 100 languages,” *arXiv preprint 2305.18098*, May 2023, doi: 10.48550/arXiv.2305.18098.

[36] Y. Tang *et al.*, “Multilingual translation with extensible multilingual pretraining and finetuning,” *arXiv preprint 2008.00401*, Aug. 2020, doi: 10.48550/arXiv.2008.00401.

[37] C. Zan, L. Ding, L. Shen, Y. Zhen, W. Liu, and D. Tao, “Building accurate translation-tailored LLMs with language aware instruction tuning,” *arXiv preprint 2403.14399*, Mar. 2024, doi: 10.48550/arXiv.2403.14399.

[38] D. Gao *et al.*, “LLMs-based machine translation for E-commerce,” *Expert Systems with Applications*, vol. 258, p. 125087, Dec. 2024, doi: 10.1016/j.eswa.2024.125087.

[39] Y. Moslem, G. Romani, M. Molaei, J. D. Kelleher, R. Haque, and A. Way, “Domain terminology integration into machine translation: leveraging large language models,” in *Proceedings of the Eighth Conference on Machine Translation*, 2023, pp. 902–911, doi: 10.18653/v1/2023.wmt-1.82.

[40] V. Agostinelli, M. Wild, M. Raffel, K. A. A. Fuad, and L. Chen, “Simul-LLM: a framework for exploring high-quality simultaneous translation with large language models,” in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 10530–10541, doi: 10.18653/v1/2024.acl-long.567.

[41] R. Koshkin, K. Sudoh, and S. Nakamura, “TransLLaMa: LLM-based simultaneous translation system,” in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024, pp. 461–476, doi: 10.18653/v1/2024.findings-emnlp.27.

[42] S. Guo, S. Zhang, Z. Ma, M. Zhang, and Y. Feng, “SiLLM: large language models for simultaneous machine translation,” *arXiv preprint 2402.13036*, Feb. 2024, doi: 10.48550/arXiv.2402.13036.

[43] Y. Moslem, R. Haque, and A. Way, “Fine-tuning large language models for adaptive machine translation,” *arXiv preprint 2312.12740*, Dec. 2023, doi: 10.48550/arXiv.2312.12740.

[44] J. Zheng *et al.*, "Fine-tuning large language models for domain-specific machine translation," *arXiv preprint 2402.15061*, Feb. 2024, doi: 10.48550/arXiv.2402.15061.

[45] J. Pang *et al.*, "Salute the classic: revisiting challenges of machine translation in the age of large language models," *arXiv preprint 2401.08350*, Jan. 2024, doi: 10.48550/arXiv.2401.08350.

[46] H. Xu *et al.*, "Contrastive preference optimization: pushing the boundaries of LLM performance in machine translation," *arXiv preprint 2401.08417*, Jan. 2024, doi: 10.48550/arXiv.2401.08417.

[47] D. M. Alves *et al.*, "Steering large language models for machine translation with finetuning and in-context learning," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 11127–11148, doi: 10.18653/v1/2023.findings-emnlp.744.

[48] V. Valeros, A. Širokova, C. Catania, and S. Garcia, "Towards better understanding of cybercrime: the role of fine-tuned LLMs in translation," in *Proceedings - 9th IEEE European Symposium on Security and Privacy Workshops, Euro S and PW 2024*, Jul. 2024, pp. 91–99, doi: 10.1109/EuroSPW61312.2024.00017.

[49] B. Zhang, Z. Liu, C. Cherry, and O. Firat, "When scaling meets LLM finetuning: the effect of data, model and finetuning method," *arXiv preprint 2402.17193*, Feb. 2024, doi: 10.48550/arXiv.2402.17193.

[50] M. Wu, T.-T. Vu, L. Qu, G. Foster, and G. Haffari, "Adapting large language models for document-level machine translation," *arXiv preprint 2401.06468*, 2024, doi: 10.48550/arXiv.2401.06468.

[51] Y. Hu *et al.*, "GenTranslate: large language models are generative multilingual speech and machine translators," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 74–90, doi: 10.18653/v1/2024.acl-long.5.

[52] J. Zeng, F. Meng, Y. Yin, and J. Zhou, "Teaching large language models to translate with comparison," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 17, pp. 19488–19496, Mar. 2024, doi: 10.1609/aaai.v38i17.29920.

[53] M. Zhang, M. Tu, F. Zhang, and S. Liu, "A cross search method for data augmentation in neural machine translation," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2024, pp. 11071–11075, doi: 10.1109/ICASSP48485.2024.10447171.

[54] A. K. Wassie, "Machine translation for Ge'ez language," *arXiv preprint 2311.14530*, Nov. 2023, doi: 10.48550/arXiv.2311.14530.

[55] S. Pan *et al.*, "POMP: probability-driven meta-graph prompter for LLMs in low-resource unsupervised neural machine translation," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 9976–9992, doi: 10.18653/v1/2024.acl-long.537.

[56] Z. Mao and Y. Yu, "Tuning LLMs with contrastive alignment instructions for machine translation in unseen, low-resource languages," in *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, 2024, pp. 1–25, doi: 10.18653/v1/2024.loresmt-1.1.

[57] S. Lankford, H. Afli, and A. Way, "adaptMLLM: fine-tuning multilingual language models on low-resource languages with integrated LLM playgrounds," *Information (Switzerland)*, vol. 14, no. 12, p. 638, Nov. 2023, doi: 10.3390/info14120638.

[58] M. Ghazvininejad, H. Gonen, and L. Zettlemoyer, "Dictionary-based phrase-level prompting of large language models for machine translation," *arXiv preprint arXiv:2302.07856*, Feb. 2023, doi: 10.48550/arXiv.2302.07856.

[59] Z. Li, L. Haroutunian, R. Tumuluri, P. Cohen, and G. Haffari, "Improving cross-domain low-resource text generation through LLM post-editing: a programmer-interpreter approach," *arXiv preprint 2402.04609*, Feb. 2024, doi: 10.48550/arXiv.2402.04609.

[60] O. Burda-Lassen, "Machine translation of folktales: small-data-driven and LLM-based approaches," in *Proceedings of the 2023 CLASP Conference on Learning with Small Data (LSD)*, 2023, pp. 68–71.

[61] R. Sennrich, J. Vamvas, and A. Mohammadshahi, "Mitigating hallucinations and off-target machine translation with source-contrastive and language-contrastive decoding," *arXiv preprint 2309.07098*, Sep. 2023, doi: 10.48550/arXiv.2309.07098.

[62] V. Mujadia, P. Mishra, A. Ahsan, and D. M. Sharma, "Towards large language model driven reference-less translation evaluation for English and Indian languages," *arXiv preprint 2404.02512*, Apr. 2024, doi: 10.48550/arXiv.2404.02512.

[63] H. Zhao *et al.*, "From handcrafted features to LLMs: a brief survey for machine translation quality estimation," in *2024 International Joint Conference on Neural Networks (IJCNN)*, Jun. 2024, pp. 1–10, doi: 10.1109/ijcnn60899.2024.10650457.

[64] N. Moghe *et al.*, "Machine translation meta evaluation through translation accuracy challenge sets," *arXiv preprint 2401.16313*, Jan. 2024, doi: 10.48550/arXiv.2401.16313.

[65] Z. Wu *et al.*, "Empowering a metric with LLM-assisted named entity annotation: HW-TSC's submission to the WMT23 metrics shared task," in *Proceedings of the Eighth Conference on Machine Translation*, 2023, pp. 822–828, doi: 10.18653/v1/2023.wmt-1.70.

[66] Y. Wang, "Large language models evaluate machine translation via polishing," in *2023 6th International Conference on Algorithms, Computing and Artificial Intelligence*, Dec. 2023, pp. 158–163, doi: 10.1145/3639631.3639658.

[67] T. A. Dinh and J. Niehues, "Perturbation-based QE: an explainable, unsupervised word-level quality estimation method for blackbox machine translation," in *MT Summit 2023 - Proceedings of 19th Machine Translation Summit*, 2023, vol. 1, pp. 59–71.

[68] T. Kasahara and D. Kawahara, "Exploring automatic evaluation methods based on a decoder-based LLM for text generation," *arXiv preprint 2310.11026*, Oct. 2023, doi: 10.48550/arXiv.2310.11026.

[69] J. Fu, S. K. Ng, Z. Jiang, and P. Liu, "GPTScore: evaluate as you desire," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2024*, 2024, vol. 1, pp. 6556–6576, doi: 10.18653/v1/2024.naacl-long.365.

[70] G. Jawahar, M. Abdul-Mageed, L. V. S. Lakshmanan, and D. Ding, "LLM performance predictors are good initializers for architecture search," *arXiv preprint 2310.16712*, Oct. 2023, doi: 10.48550/arXiv.2310.16712.

[71] X. Huang, Z. Zhang, X. Geng, Y. Du, J. Chen, and S. Huang, "Lost in the source language: how large language models evaluate the quality of machine translation," in *Findings of the Association for Computational Linguistics ACL 2024*, 2024, pp. 3546–3562, doi: 10.18653/v1/2024.findings-acl.211.

[72] T. Tang *et al.*, "Not all metrics are guilty: improving NLG evaluation by diversifying references," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2024*, 2024, vol. 1, pp. 6596–6610, doi: 10.18653/v1/2024.naacl-long.367.

[73] S. Gladkoff, G. Erofeev, I. Sorokina, L. Han, and G. Nenadic, "Predictive data analytics with AI: assessing the need for post-editing of MT output by fine-tuning OpenAI LLMs," *AMTA2023: Generative AI and the Future of Machine Translation*, 2023.

[74] P. Fernandes *et al.*, "The devil is in the errors: leveraging large language models for fine-grained machine translation evaluation," in *Conference on Machine Translation - Proceedings*, 2023, pp. 1064–1081, doi: 10.18653/v1/2023.wmt-1.100.

[75] H. Huang *et al.*, "Towards making the most of LLM for translation quality estimation," in *CCF International Conference on Natural Language Processing and Chinese Computing*, 2023, pp. 375–386, doi: 10.1007/978-3-031-44693-1_30.

[76] J. Li, R. Li, and Q. Liu, "Beyond static datasets: a deep interaction approach to LLM evaluation," *arXiv preprint 2309.04369*, Sep. 2023, doi: 10.48550/arXiv.2309.04369.

[77] S. Dreano, D. Molloy, and N. Murphy, “Embed_Llama: using LLM embeddings for the metrics shared task,” in *Proceedings of the Eighth Conference on Machine Translation*, 2023, pp. 738–745, doi: 10.18653/v1/2023.wmt-1.60.

[78] G. Vernikos and A. Popescu-Belis, “Don’t rank, combine! Combining machine translation hypotheses using quality estimation,” *arXiv preprint 2401.06688*, Jan. 2024, doi: 10.48550/arXiv.2401.06688.

[79] W. Xu *et al.*, “INSTRUCTSCORE: explainable text generation evaluation with finegrained feedback,” *arXiv preprint 2305.14282*, May 2023, doi: 10.48550/arXiv.2305.14282.

[80] T. Kocmi and C. Federmann, “Large language models are state-of-the-art evaluators of translation quality,” *arXiv preprint 2302.14520*, Feb. 2023, doi: 10.48550/arXiv.2302.14520.

BIOGRAPHIES OF AUTHORS

Bhuvaneswari Kumar is a research scholar in the School of Computer Science Engineering and Information Systems, VIT University. She received her B.E.(CSE) degree from Bharathiyar University and M.Tech.(CSE) degree from Anna University. She has about 12 years of academic experience. Her research interests include natural language processing, deep learning, and machine learning. She can be contacted at email: bhuvaneswari.k@vit.ac.in.

Varalakshmi Murugesan is an associate professor in the School of Computer Science and Engineering, VIT University. She received her B.E.(CSE) degree from Madras University and M.Tech. (IT) degree (Gold medalist) from VIT. She has about 18 years of academic experience. Her research interests include natural language processing, deep learning, and high-performance computing. She is particularly interested in large language models. She currently works on a project funded by ISRO and an Indo-Russian joint research project, funded by DST. She also works on consultancy projects for startups. She can be contacted at email: mvaralakshmi@vit.ac.in.