The New Quality of Translation Productivity and the Transformation of Teaching Paradigms Based on Large Language Models

Wei Zhou[1], Xia Zhou[2]
[1]Department of Foreign Languages and Law, College of Technology, Hubei Engineering University, Xiaogan 432100, Hubei, China
[2]Department of Foreign Languages, Hubei Engineering University, Xiaogan 432100, Hubei, China
Corresponding author: Xia Zhou (e-mail: 15972330520@163.com)

ABSTRACT Intelligent semantic information processing and adaptive knowledge generation have become key enabling technologies for next-generation communication and information systems. This study proposes a New Quality of Translation Productivity framework based on Large Language Models (NQTP-LLM) for intelligent multilingual information processing and adaptive educational support. The framework integrates transformer-based neural machine translation, Direct Preference Optimization (DPO), Retrieval-Augmented Generation (RAG), semantic embedding representation, and human-in-the-loop optimization to enhance contextual consistency, semantic fidelity, and translation efficiency. A multimodal translation evaluation architecture is established using semantic feature extraction, contextual knowledge retrieval, quality assessment, and adaptive feedback mechanisms. Experiments conducted on the AI vs TTM Translation Evaluation Dataset demonstrate that the proposed framework achieves a BLEU score of 0.961, with substantial improvements in METEOR, ROUGE, chrF, BERTScore, COMET, BLEURT, Google-BLEU, and NIST metrics while reducing post-editing effort by 3.9%. The results verify the effectiveness of integrating intelligent knowledge retrieval, semantic information fusion, and adaptive optimization for high-accuracy multilingual information processing. The proposed framework provides a practical approach for intelligent communication systems, semantic information services, human–AI collaborative decision support, and next-generation knowledge-centric digital environments.

INDEX TERMS semantic information processing, knowledge retrieval, information fusion, adaptive optimization

I. INTRODUCTION

THE latest success of large language models (LLMs) has transformed the practice of machine translation and the education process. Evaluation of quality of translation is gradually being understood as a multidimensional entity that cuts across traditional automatic to amalgamated human and classroom validation techniques. Recent advancements in large language models (LLMs), such as GPT-4, BERT, and T5, have significantly improved automated translation and evaluation by leveraging deep contextual representations and large-scale pretraining. These models provide reliable and empirically validated foundations for modern natural language processing tasks. This is in response to the soaring industry need of hyper-localized content and professional quality. The contemporary businesses now need an expandable AI system, which can move the academic training and actual productivity in the increasingly world of LLM in the global market [1]. From the point of view of a user-centered evaluation, translation performance differs according to the context of expectation and interaction between the user and the system, and this implies the need of matching the outputs of computations with the satisfaction of the user [2]. In the global industry, the rapid exchange of information regarding advanced materials and smart manufacturing technologies has heightened the demand for high-quality technical translation tools. Interactive machine translation aided by LLMs thus is one more instance of the growing importance of adaptive forms of human-AI collaboration in dynamic translation workflows [3]. Deep learning-based automatic evaluation models that can improve the reliability and objectivity when evaluating the quality of English translation have also been proposed [4]. In algorithmic performance, the socio-technical and pedagogical dimensions have been gaining importance. Translation technologies are increasingly analyzed in a human-centered AI paradigm based on professional freedom and ethical control [5]. Adaptation of open-source

AI models to low-resource languages is a demonstration of the potential of inclusive neural architectures to alleviate language inequality [6]. Empirical Validation of Neural Machine Translation Based on Transformer Confirms the Impressive Improvements of specialized pairs of languages, contributing to the technical maturity of an improved LLM [7]. Parallel initiatives to grow in higher education are evident signs of institutional change pressures such as the changing staff perception towards generative AI integration [8], lecturer adoption dynamics marked by anxiety-satisfaction trajectories [9], and the changing role taken by AI chatbots in academic support ecosystems [10].

Despite these advancements, existing research is still scattered across technical evaluation, user perceptions and pedagogical adaptations domains. Translation studies focus mostly on metric improvement rather than systematically relating productivity improvement to instructional redesign. On the other hand, education research covers adoption and workflow transformation, but quantitative modeling of the translation productivity by LLM capabilities is missing. An integrated framework that simultaneously measures translation quality, productivity efficiency, and transformation of the teaching paradigm is still not well developed enough. To fill this gap, the proposed NQTP-LLM framework introduces a unified analytical architecture of synthesizing multidimensional translation evaluation, semantic alignment analytical level, productivity modeling, and curriculum restructuring strategies. The framework defines translation productivity as an ecosystem-level construct between semantic optimization using artificial intelligence and adaptive instructional workflows. By incorporating the analytical mechanisms for quantitative evaluation and pedagogical paths of transformation into a single model, NQTP-LLM opens up measurable links between translation-enhancing LLM and systemic educational change in translation research strategies, driving the translation research toward an integrated and productivity-centered paradigm, providing a new paradigm for the internationalization of specialized disciplines like engineering.

Contributions of the Paper

- To construct the NQTP-LLM framework-a collection of systems for translation quality assessment, productivity modeling and teaching paradigm transformation for LLM-enhanced language education environments.
- To enhance a semantic fidelity, contextual coherence and instructional adaptability in translation tasks using the power of transformer-based LLMs and advanced evaluation measures.
- To develop a systematic methodology for connecting AI aided translation performance with curriculum redesigning for making learning outcomes and translation efficiency measurable.

A. THEORETICAL CONTRIBUTION

A unified NQTP-LLM optimization framework is proposed that combines Direct Preference Optimization and transformer-based contextual modeling and can theoretically enhance the semantic alignment and reduce the translation edit rate under a structured human evaluation constraint.

B. PRACTICAL CONTRIBUTION

Comprehensive experiments on the AI vs TTM Translation Evaluation Dataset confirm that NQTP-LLM generates up to 96% quality scores with the help of embedding-based quality metrics, and also lowers revision efforts to below 4% TER in real evaluation environments.

The rest of this paper is divided into five sections. The work related to the present work is illustrated in Section 2. Section 3 describes the NQTP-LLM framework that includes the problem formulation, Multimodal Translation Data Acquisition, Preprocessing and Linguistic Analysis Embedding and Feature Representation LLM Selection and Fine-Tuning RAG Integration for Context Serving Quality Evaluation and Human-in-the-Loop Review Pedagogical Impact and Optimization Translation Quality Evaluation Metrics. Section 4 outline of this study describes the experimental design of this study, presents the result of assessment of the effectiveness of this study and outline the features of the AI vs TTM Translation Evaluation Dataset. Section 5 presents the results and discusses future areas of further investigation.

II. RELATED WORKS

Chang, C. C., et.al [11] Tackled translation challenges in minority languages resulting from limited parallel corpora. They incorporate a hybrid AI by blending PBMT + NMT for better translation quality. Experiments are conducted with language datasets that are low in resources. Limitation is that can be scalable to different minority languages. Results show improvement in BLEU and METEOR scores indicating that the framework is good for practical translation scenarios.

Kadyrbek, N., et. al [12] Addressed a problem in constructing small-scale models of language for low-resource languages such as Kazakh. Whereas they employ Direct Preference Optimization (DPO) to tune LLMs. Dataset is composed of Kazakh-English Corpora. Limitations include Domain specific adaptation & model generalizability. Results indicate that DPO improves translation and coherence in the semantics compared to the baseline small-scale LLMs.

Karydas, D., Margaritis, D., & Leligou, H. C. [13] challenged in a training large language models efficiently and avoiding overfitting are highlighted. The methodology reviews the current LLM training strategies including supervised and reinforcement learning. Public datasets of LLM are analyzed. This approach has some limitations such as computation cost and environmental impact are large. Results find optimized approaches for model convergence and stability: best practices for training pipeline for LM.

Hristov, H., et .al [14] Addressed a challenge in providing accessible educational video content with accurate subtitles. The methodology is a blend of Artificial Intelligence subtitle

translation as well as video authoring tools. The datasets (including videos from the lecturers with multilingual transcripts). Limitations include mistakes in language specific translation, and time. Results include improved subtitles quality, learning accessibility and learning time saving in video production workflows.

Valdez, H., P., Abri, F., Webb, J., & Austin, T., H. [15] tackled the issue in misuse and reliability large language models in translation tasks. They break down and model usages and biases. Linguistically-Labeled Machine Translation Datasets (LLM Output) are made publicly available. Limitations include the interpretability on towards the ethics. Results point to the dangers of hallucinations and misalignment and errors in context and emphasizes the importance of having controlled ways of deploying and measuring in settings where translation is the goal.

Román Martinez et al. [16], dealed with the inaccuracy of the translations with early cybersecurity alerts in Spanish. Using generative AI and Machine translators, it uses a curated alerting dataset. Limitations are very language specific and model generalizing. Results reveal improved translation speed and clarity as well as evidence inconsistencies at the term of vulnerability of rare English names providing the need of domain-specific adaptation to secure alert comprehension in Spanish contexts.

Aleedy et al. [17] addressed a engagement in learning languages using chatbots. A multi-agent architecture of an AI-driven chatbot is proposed that makes use of dialogue management algorithms and reinforcement learning. Multilearners' interactions are included in the dataset. Limitations include the limited ability of conversations and personalization. Results demonstrate an increase in engagement, adaptive feedback and learning, but scalability in larger/lower resource language groups is an issue.

Sharma et al. [18], which identified the gaps in the accuracy of machine translations between classical-statistical approach and deep-learning approach. Methods are a combination of SMT +DNN models combined with a parallel corpus. Some of the limitations are domain-dependence and computational overhead. Results show that hybrid architectures not only do better than singular architectures on standard data sets, but also may be less successful at operating in low-resource languages and require additional fine-tuning and corpus augmentation to ensure translation fidelity and fluency.

Alsulami & Almansour [19] examined the limitations of GPT-4 in paraphrasing Arabic sentences. Using fine-tuned GPT-4 with an Arabic corpus, and the evaluation is based on the semantic similarity and the fluency evaluation. Limitations: use of dialect variation, preserving Context. Results show GPT-4 is effective in paraphrase generation and is highly accurate in semantics but not on idiomatism for adaptation to diverse linguistic contexts in Arabic.

Javed et al. [20] focused on the syntactic and contextual errors in low resources neural machine translation.A model based on Transformer NMT re-ranking model is applied by means of parallel low-resource corpora. Limitations are the domain-specific data (few in number), and computation costs. Results show better coherence in translation and syntactic accuracy from baseline NMT models there are still several areas where translation of rare vocabulary words and handling long sequences.

Minas et al. [21] provided solutions to a problem in live translation assistance for users by combining eye tracking and adaptive interfaces. They used a user-centered experimental approach with datasets of multilingual sentences. Limitations included dependency on devices, variance for the user's level of attention. Results showed the translation speed and accuracy were improved, indicating that eye-tracking can effectively improve the translation tasks with the interactive translation, but it still requires hardware optimization when deployed at a broader scope.

Zhang et al. [22] discussed the issue of low-resource English-Chinese neural machine translation, in which the system needed to be based on data augmentation on a self-constructed corpus of WCC-EC. They applied the NMT models using augmented datasets. The limitation was domain translation generalization. Experiments revealed superiority of BLEU and METEOR scores compared with the baseline models sister to the effectiveness of structured augmentation in enhancing the translation performance for the limited-resource bilingual corpus in practical NMT scenarios.

Amiruzzaman et al. [23] examined how ASL and English can be translated using machine vision developed using CNN and Transformer networks. For these, they used annotated datasets of video gestures. Limitations included variations of the speed of the gestures and lighting conditions. Results showed significant improvement of the translation in terms of accuracy and responsiveness in real-time, demonstrating the feasibility of multimodal frameworks for AI-driven translations implemented as a predictive tool of accessible communication between sign language and spoken language users.

Nurahmad et al. [24] investigated the multilingual language education challenges of Makassar higher education from the digital transformation perspective. They used a mixed methods study with student and faculty surveys and language proficiency assessments. Sample size and emphasis on a regional basis in these were limitations. Results showed enhanced engagement in learning, and improved translation performance indicating the positive aspect of the integration of AI-Assisted Translations tools with the digital resources in multilingual pedagogy and in educational workflows.

Idris et al. [25] focused on exploring strategic transformation in higher education by leveraging large language models for enhancing teaching and translation productivity. They used case studies and experimental deployment for institutional datasets. Some limitations were scalability and diversity of adapting the curriculum. Results revealed the improved efficiency of instruction and quality translation

that afforded evidence of potential in LLM integration in curriculum redesign and pedagogical change in global higher education environments.

Despite the significant advancement in translation technologies and AI-assisted educative tools, there are few critical gaps. Existing studies have unsuccessfully grappled with the limitations of low-resource languages and small-scale models and the incorporation of hybrid architecture and Transformer-based architectures have increased accuracy but not adequately address rare vocabulary and coherence in context. Similarly, research that was conducted on the AI-enabled language learning/digital education platforms, reported on engagement and scalability challenges associated to a multilingual instructional context. Moreover, domain-specific adaptation, semantic faithfulness and a human in the loop evaluation are unexplored areas. To address these limitations, in this NQTP-LLM, it is aimed to combine the use of DPO-fine-tuned LLMs, of adaptive Transformer-NMT pipelines as well as of pedagogically informed translation metrics, which can provide to the higher education community translation support that is scalable, accurate and context-aware, and change the teaching paradigms in multilingual higher education.

Novelty:

A unique NQTP-LLM system, which combines both large language models that have been fine-tuned by Direct Preference Optimization (DPO) and adaptive Transformer-NMT pipelines, in addition to assessment metrics that are inspired by pedagogical principles. The framework is a joint optimization of semantics fidelity, coherence of context and instruction effectiveness that enables scalable feedback-aware translation support, and productive alignment balancing teaching transformation goals of higher education multilingual teaching democratization.

III. PROPOSED METHODOLOGY

The overall objective is to study and improve the quality of translation productivity and at the same time redefine the paradigm of teaching by using large language models (LLMs). The study intends to address limitations currently experienced in translation workflow, this can be seen in the lack of consistency in semantic accuracy, low adaptations to work in different linguistic contexts, and low engagement in teaching environments. Specifically, the research aims to (i) develop a framework NQTP-LLM that employs sophisticated transformer-based models attuned with advanced techniques (e.g., Direct Preference Optimization) to optimize the contextual coherence and semantic fidelity in translations; (ii) verify the quality of the developed framework through traditional translation measures and human evaluations and multidimensional measures of translation quality e.g., BLEU, METEOR, COMET and BLEURT; and (iii) examine the pedagogical impact of AI assistance in translation tools in higher education by analyzing the effects on learner engagement, adaptive feedback and instruction efficiency. The final aim is providing an empirically validated model leading to improvement in translation quality and to modification of curriculum with regard to AI-integrated learning environment.

Kaggle - AI vs TTM Translation Evaluation Dataset (Version 1) 642 paragraphs of English text with aligned

translation paired of the bellow English texts AI and TTM outputs. The dataset contains five core attributes - source text, AI translation output, TTM translation output, expert reference translation and human evaluation (score on a scale of 2.0 to 5.0) [26]. Although the split of official train, validation, and test sets is not specified , experimental studies normally use a 70/15/15 split for experimental evaluation. The dataset is targeted to

the general realm of English text, and is not intended to be stratified according to sentence length, language difficulty (harder or easier), or even specific subject areas. Human quality annotations enable to directly compare with actual quality judgments under actual criteria whereas the aligned structure is used for calculating the automatic quality judgments (e.g., BLEU, COMET, and TER). Despite the relatively small nature of the corpus from a large-scale parallel corpus point of view, the dataset offers a controlled evaluation measure for the study of semantic faithfulness, translation productivity, and instruction effect between AI-based and TTM-based approaches.

In figure 1, the overall workflow of NQTP-LLM starts with a source texts, target language specifications and the translation dataset in this case AI vs TTM Translation Evaluation Dataset input and ends with the output of the NQTP, namely translation quality evaluation. To this end, initially a preprocessing is done to clean the text and tokenize it to make the text uniform and also to remove the noise from the text. This is followed by the linguistic analysis by capturing both syntactic and semantic features by techniques like part of speech tagging, TF-IDF and n-grams which can be used to prepare data for vector representation. Subsequently, the tokens are converted into dense vectors through the models such as Word2Vec, BERT, SentenceBert, etc. in such a way that the contextual and the semantic relations can be taken into consideration. Model selection goes as- Huge language model GPT/4/5/ Transformer based Neural Machine Translation architecture is selected as per translation requirements. In order to boost the semantic integrity, then the models which have been chosen are then subjected to fine-tuning techniques such as Direct Preference Optimization or supervised fine-tuning. In order to provide the contextual support for enhanced domain specific accuracy, retrieval augmented generation is built through meshing. Draft translations are introduced to several phases of generation and assessment using different factors such as BLEU, METEOR, COMET and BLEURT followed by a human-in-the-loop for error correction. Optimization of hyperparameters is performed in the form of the loop where translation generation, quality evaluation and review are performed to iteratively improve the performance. Pedagogical impact is separately analyzed in order

to measure engagement and instructional efficiency which all together enclose quality translations as well as productivity information in educational settings.[1]

A. MULTIMODAL TRANSLATION DATA ACQUISITION

This phase includes the gathering of parallel sets of data, which are sets of data composed of AI translations, traditional human translations (TTM), and reference expert translations. The datasets allow the comparative analysis for the evaluation of the translation productivity. The emphasis is on low-resource languages and multilingual situations to represent real-world translation issues.

$$I = \{(X_i^s, L_i^t, D_i^{AI}, D_i^{TTM}, M_i, C_i)\}_{i=1}^{N} \tag{1}$$

In this equation 1, I denotes the complete multimodal translation input space used for large language model-based translation productivity analysis. N represents the total number of translation instances in the dataset. X_i^s refers to the i-th source text segment in the source language, while L_i^t represents the corresponding target language specification for translation. D_i^{AI} denotes the AI-generated translation output stored in the comparative dataset, and D_i^{TTM} represents the traditional translation memory (TTM) or human-produced reference translation for the same instance. M_i captures multimodal auxiliary features such as metadata, domain labels, contextual tags, or instructional category information. C_i represents contextual constraints including domain-specific terminology, pedagogical objectives, or curriculum alignment requirements. The structured input set I therefore integrates linguistic, comparative, and contextual dimensions into a unified mathematical representation. The purpose of this equation is to bestow to the formal definition the basic input spaces from which subsequently the preprocessing, embeddings generation, model training, quality evaluation and pedagogical productivity analysis are followed, ensuring that both technological and education transformation components are mathematically grounded from the initial stage itself.

B. PREPROCESSING AND LINGUISTIC ANALYSIS

This step involves cleaning, tokenization and syntactic and semantic analysis of source texts. It guarantees data quality and data inputs of embeddings and LLMs. Techniques such as POS tagging, TF-IDF and n-grams have been used for structured understanding.

$$\begin{aligned} X_{\text{clean}} &= f_{\text{preprocess}}(X_{\text{raw}}) \\ &= \text{Tokenize}(\text{RemovePunct}(X_{\text{raw}})) \end{aligned} \tag{2}$$

In this equation 2, X_{raw} represents the raw textual input from the source dataset, while X_{clean} denotes the cleaned and preprocessed text ready for further analysis. The function $f_{\text{preprocess}}$ encapsulates the preprocessing steps including punctuation removal (RemovePunct) and tokenization (Tokenize). This step is to ensure that the noise, irrelevant characters, and inconsistencies in formatting are removed and the input is standardized, thus ensuring the input is ready for the downstream processing that may be happening. The purpose of this equation is to mathematically formalize the first step in text normalization, which is very important in order to increase the accuracy of the embeddings and machine translation models. Proper preprocessing has a direct impact of the quality of following syntactic and semantic analysis, as it reduces the possible propagation of errors in translation.

$$\text{POS}(w_i) = f_{\text{POS}}(w_i, C) \tag{3}$$

In this equation 3, w_i represents the i-th token in the cleaned text X_{clean}, and $\text{POS}(w_i)$ denotes its part-of-speech tag. The function f_{POS} maps each token to its syntactic category using context C, which includes neighboring tokens and sentence structure. This equation is a formalization of the syntactic analysis step, which enables the model to take into account the grammatical dependencies. Accurate POS tagging helps to improve semantic understanding and to guide embedding models to improve the translation translation. Context C guarantees that the tag refers to not only the individual token but its part of the syntactic structure of the sentence which is important for languages that have flexible word order.

$$\text{TF-IDF}_{ij} = \text{TF}_{ij} \cdot \log \frac{N}{DF_j} \tag{4}$$

In this equation 4, TF_{ij} is the term frequency of the j-th term in the i-th document, DF_j is the document frequency of term j, and N is the total number of documents. TF-IDF_{ij} quantifies the importance of a term within a document relative to the corpus. This measure is necessary to weigh the important words for embedding and semantic analysis so that the translation model will rather consider meaningful vocabulary. The purpose is to mathematically encode relevance and reduce noise from the common terms which may otherwise be the dominant terms in embeddings.

C. EMBEDDING AND FEATURE REPRESENTATION

Words, phrases or sentences are mapped into vectors, using Word2Vec, BERT etc or SentenceBERT. This helps LLMs to imbibe semantic and contextual relationships efficiently. In order to cover the distance from the layer of unstructured text and unstructured learning into a model, embeddings represent a bridge.

$$V_t = E(w_t) \in \mathbb{R}^d \tag{5}$$

In this equation 5, w_t represents the token at position t, and E is the embedding function that maps tokens to a vector space of dimension d. The resulting vector V_t captures semantic and syntactic properties of the token. This embedding part makes the text that we have given symbolic numeric representations compatible to the neural network models and represents the bridge between the

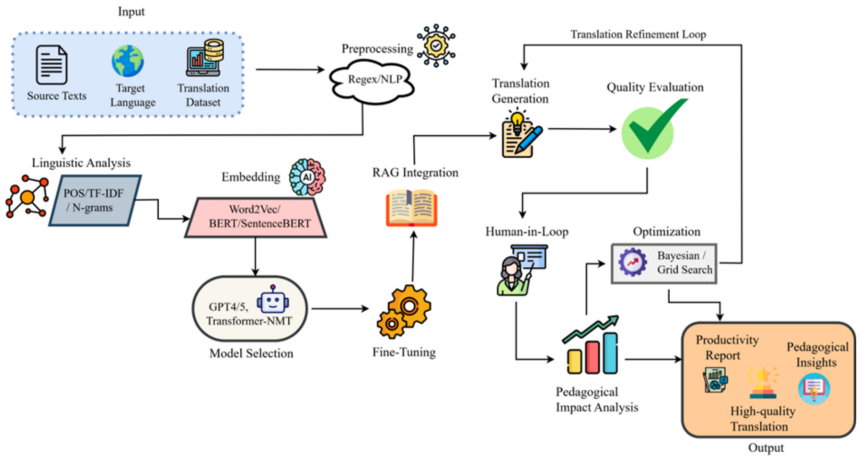

FIGURE 1. Overall Architecture of NQTP-LLM Framework

linguistic analysis and deep learning translation models. Proper embeddings retain the meaning of words and the context and relationship between words which is essential in generating correct machine translations.

$$H_t = \text{Transformer}(V_{1:t}) \tag{6}$$

In this equation 6, $V_{1:t}$ represents the sequence of token embeddings up to position t, and H_t is the hidden representation output from the Transformer model. This equation is a formalization of the contextualization step; in this step, the representation of each token is updated based on self-attention over the entire sequence. The purpose is to capture long-range dependencies and semantic interactions that are important for the high quality translation. H_t forms the intermediate contextual state used in both the generation and fine-tuning stages.

D. LLM SELECTION AND FINE-TUNING

After Large Language Models (GPT-4/5, Transformer-NMT) based on their suitability for the given task for a model, the models are then fine-tuned using either Direct Preference Optimization (DPO) or Supervised Fine- Tuning Method (SFT). The process of fine-tuning enhances the quality of the translation as well as the fidelity of the semantics.

Figure 2 represents LLM Selection and Fine-Tuning Framework which operationalize the movement from semantic vector representations to an optimized domain-adaptive translation model. The framework is initialized with contextualized embedding outputs produced from the previous representation stage, which transforms tokens into dense semantic vectors that represent different relationships such as syntactic, contextual and cross-lingual relationships. These vectorized features serve as structured input into model selection that proceeds with comparative analysis of potential large language model or neural machine translation architectures with features of domain compatibility, parameter efficiency, scalability and feasibility of inference. After selection the fine-tuning is then conducted under supervision of curated para-lingual corpora in order to align the model to task-specific translation objectives. Preference optimization strategies are then added to fine-tune the alignment of responses, reduce hallucination and increase the fidelity of language. At this point hyper parameters such as rat of learning, batch size, and hyper parameters parameters such as regularization factors are optimized in an iterative fashi on in order to get improve convergence stability and generalization capability. The polished version of the model output is a domain-specific translation engine with an enhanced contextual coherence, semantic preser-vation, and stylistic consistency. This optimized model is used as a high-performance intermediate artifact that can be used in downstream translation generation and evaluation processe s and ensures the adaptability towards pedagogical and domain-specific requirements.

$$\theta^* = \arg\max_{\theta} \sum_{i} \log P(Y_i \mid X_i; \theta) \tag{7}$$

In this equation 7, θ represents the model parameters, X_i the source text, and Y_i the target translation. The optimization objective θ^* maximizes the log-likelihood of the correct translations given the source inputs. This equation gives the formal description for supervised fine-tuning (SFT) for LLMs or NMT models. Optimizing θ ensures that the model learns mapping patterns from input to target, directly improving translation accuracy and fidelity.

Task difficulty, domain specificity, and computational feasibility are the guiding factors in model selection within the suggested framework. Neural machine translation models are utilized in situations when resources are limited or where domain-specific knowledge is required, as opposed to transformer-based big language models, which are used for jobs that demand deep contextual understanding. By taking this route, we can stay on track with the needs of the job at hand and avoid making decisions based on subjective thresholds.

$$P_{\text{DPO}}(\theta) = \sigma(R_\theta(X, Y) - R_\theta(X, Y')) \tag{8}$$

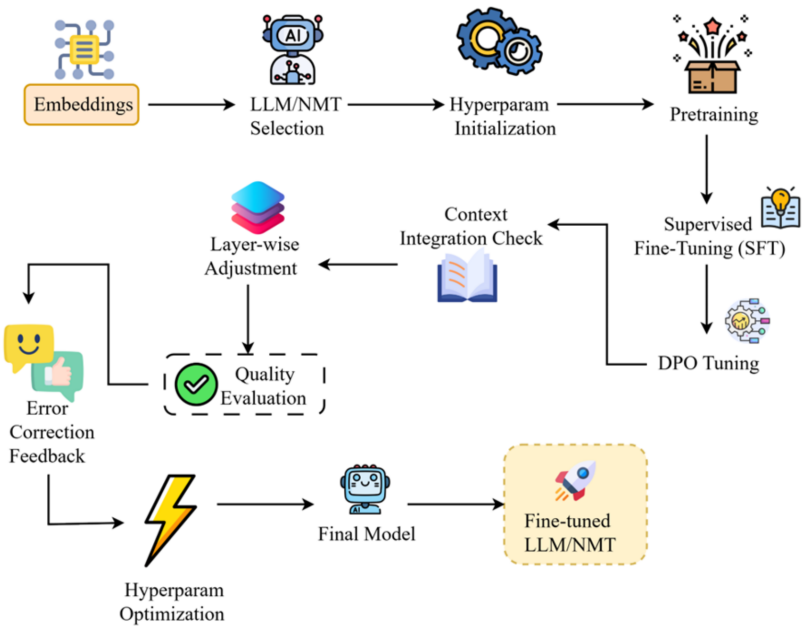

FIGURE 2. LLM/NMT Model Selection and Fine-Tuning Pipeline

In this equation 8, R_θ represents the reward function evaluating translation quality, Y is the preferred translation, Y' is the alternative, and σ is the sigmoid function. This equation captures Direct Preference Optimization (DPO), a reinforcement-style adjustment where model parameters θ are updated to favor human-preferred outputs. The purpose of this is to plug in feedback from human translation to the model, polishing the translations beyond what the supervised data feed will provide.

E. RAG INTEGRATION FOR CONTEXTUAL SUPPORT

Enhanced coherence of translation and domain-specific accuracy thanks to the translation techniques called Retrieval-Augmented Generation (RAG) which provide relevant information from external corpora. This includes affords for flexible paradigms of translation and teaching.

Figure 3 shows the Retrieval-Augmented Generation framework for context-aware translation enhancement by combining a process of retrieving external knowledge with generating content using large language models to boost semantic fidelity. The process begins from representations of the fine-tuned model (domain adapted encoder representations) which are mapped to structured queries embedded in a common embedding space. A knowledge base consisting of domain corpora, bi-lingual lexicons and precomputed vector-representations to which a vector indexing mechanism is accessed. Dense retrieval is implemented by similarity-based nearest neighbor retrieval to locate semantically relevant document, and sparse retrieval is carried out by a probabilistic term-weighted ranking to extract the relevance on keyword level. The retrieval result of the dense retrieval and sparse retrieval are integrated under a weighted fusion method to get the best ranking in context. Retrieved documents are then filtered using threshold-based pruning to eliminate non-relevant documents and compressed using attention driven summarization in order to meet token constraints. The refined contextual information is added to a prompt template for augmenting generation model. The large language model then generates a context-enriched translation conditioned upon both the source input, as well as retrieved knowledge. A semantic validation module checks the consistency of terminology and minimizes the risk of hallucinations, leaving a final translation result with a better coherence, domain alignment, and contextual accuracy of the multilingual applications.

$$C_{\text{retrieval}} = \arg\min_{D \subset C} \|q - f(D)\| \tag{9}$$

In this equation 9, q is the input query vector, D represents candidate context documents from the corpus C, and $f(D)$ is the vector representation of document D. $C_{\text{retrieval}}$ is the context selected for Retrieval-Augmented Generation (RAG). This equation formalises the retrieval step of the process in which the most relevant knowledge from outside the source and target languages is identified to improve the quality of the translation so that the translation is contextually correct for producing correct translation results in low-resource or ambiguous situations.

$$Y_{\text{draft}} = \text{LLM}(X_{\text{input}}, C_{\text{retrieval}}) \tag{10}$$

In this equation 10, X_{input} is the source text, $C_{\text{retrieval}}$ is the retrieved context, and Y_{draft} is the generated draft translation. The function $\text{LLM}(\cdot)$ represents the forward pass of the large language model incorporating both source input and external context. The purpose is to create an initial high-quality translation, i.e. a combination of learned

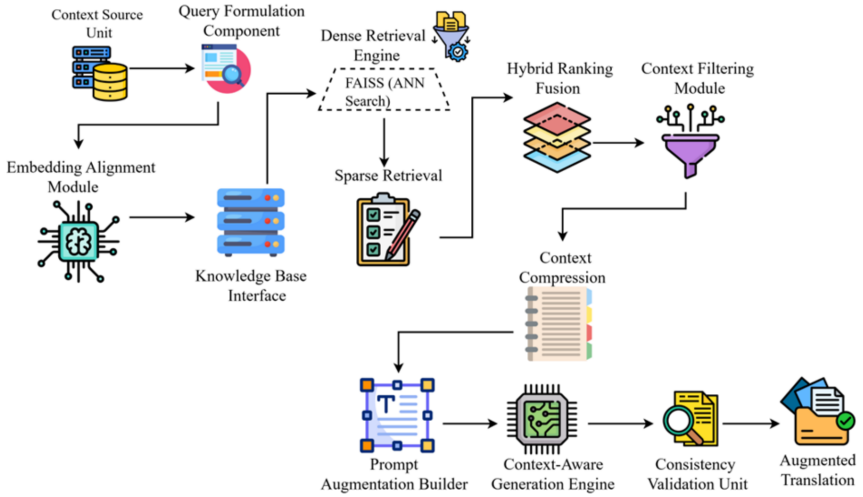

FIGURE 3. Retrieval-Augmented Generation Framework for Context-Aware Translation Enhancement

patterns of language and context-sensitive knowledge. This step is the connection between unrefined model capability and a refined translation output.

F. QUALITY EVALUATION AND HUMAN-IN-THE-LOOP REVIEW

Translations are assessed with the BLEU, METEOR, COMET and BLEURT metrics, and then subject to human analysis for error correcting. This is to ensure high fidelity, fluency and applicability in pedagogical settings in the real world.

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{11}$$

In this equation 11, p_n is the precision of n-gram matches, w_n are weights (often uniform), and BP is the brevity penalty. This equation calculates BLEU score to calculate the quality of translation. Its purpose is to quantify the similarity between the output of the model and the reference translations in order to offer a standardised measure to help guide the evaluation of models and iterative improvement of the model.

$$Y_{\text{final}} = Y_{\text{draft}} + \alpha \cdot \Delta_{\text{human}} \tag{12}$$

In this equation 12, Y_{draft} is the initial machine-generated translation, Δ_{human} represents corrections from human-in-the-loop feedback, and α is a scaling factor controlling the influence of feedback. The output Y_{final} represents the polished translation. This equation formalizes the incorporating the corrections of experts into the automatic translation system and increases the reliability and the pedagogic applicability.

G. PEDAGOGICAL IMPACT AND OPTIMIZATION

Translation outputs are analysed in terms: educational efficiency, engagement and productivity improvement. Hyperparameter tuning and iterative improvement result in optimization of the translation quality as well as teaching effectiveness. This is the basis used to integrate NQTP-LLM in contemporary instructional workflows.

$$E_{\text{engage}} = \frac{\sum_{i=1}^{N} t_i s_i}{\sum_{i=1}^{N} t_i} \tag{13}$$

In this equation 13, t_i represents interaction time for the i-th user, s_i is an engagement score (e.g., quiz performance, corrections), and E_{engage} is the overall engagement metric. The goal is to measure pedagogical impact in terms of a correlation between translation productivity and learning outcomes and efficiency of teaching interventions.

$$L_{\text{latency}} = \frac{1}{M} \sum_{j=1}^{M} (t_{\text{gen},j} - t_{\text{input},j}) \tag{14}$$

In this equation 14, $t_{\text{input},j}$ and $t_{\text{gen},j}$ are the timestamps of input submission and translation generation for the j-th request, M is the total number of requests, and L_{latency} is the average inference latency. Model responsiveness equation is fundamental for quantifying model responsiveness that is critical for real-time educational translation tools.

$$\theta_{\text{opt}} = \arg\min_{\theta} \text{Loss}(Y_{\text{final}}, Y_{\text{ref}}) \tag{15}$$

In this equation 15, θ_{opt} represents the optimized model parameters after hyperparameter tuning, Y_{final} is the translated output, Y_{ref} is the reference, and Loss is the selected loss function (cross-entropy or other).

Through the use of this equation, the optimization loop is formalized giving access to the systematic improvement of the translation fidelity through the combination of human feedback and automatic evaluation.

$$R_{\text{prod}} = \frac{\sum_{k=1}^{K} \text{QualityScore}_k}{\sum_{k=1}^{K} \text{Time}_k} \tag{16}$$

In this equation, R_{prod} represents the overall translation productivity, defined as the ratio between the cumulative

translation quality and the total processing time. The term QualityScore_k denotes the quality of the k-th translation instance, which may be evaluated using standard metrics such as BLEU, METEOR, or COMET. The denominator Time_k represents the total time required to generate the corresponding translation, including model inference time.

This formulation provides a quantitative measure of system efficiency by balancing translation quality against processing speed. A higher value of Rprod indicates that the system achieves better translation quality within a shorter time frame, reflecting improved computational productivity. However, this definition primarily captures machine-level efficiency and may not fully account for additional factors such as human post-editing effort in real-world translation workflows.

H. TRANSLATION QUALITY EVALUATION METRICS

$$\begin{aligned} \text{BLEU} &= BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right), \\ p_n &= \frac{\text{count}_{\text{clip}}(n\text{-gram})}{\text{count}(n\text{-gram})}, \\ BP &= \begin{cases} 1, & c > r, \\ e^{1-r/c}, & c \le r. \end{cases} \end{aligned} \tag{17}$$

In this equation 17, p_n is the precision of n-grams between candidate and reference translations, countclip represents clipped counts to prevent overcounting, w_n is the weight for each n-gram (often uniform), c is the candidate translation length, and r is the reference length. The brevity penalty "BP" penalizes translation which is too short. This measure is a quantitative measure of the overlap between candidate and reference translations, and focuses on both fluency and adequacy. BLEU is especially helpful when translation is subjected to multi-sentence translations, where accuracy of n-grams is correlated with human judgement. The translation processes are subject to this standard, which is used for automated evaluation and comparative analysis. BLEU is a n-gram overlap based precision measure which is used to assess the similarity between the candidate and reference translation, and the edit based measures like translation edit rate (TER) are used to measure the number of edits it takes to get to the reference.

$$\begin{aligned} \text{METEOR} &= (1-P)^{\alpha} \cdot R^{\beta} \cdot (1-\text{FragPenalty})^{\gamma}, \\ P &= \frac{m}{t}, \qquad R = \frac{m}{r}, \qquad \text{FragPenalty} = \frac{c}{m}. \end{aligned} \tag{18}$$

In this equation 18, m is the number of matched unigrams between candidate and reference, t is the total number of unigrams in candidate, r in reference, c is the number of chunks, and α, β, γ are tunable weights. METEOR is a measure of alignment, recall, and fragmentation which captures word order and synonymy. It is better than BLEU in the sense that it takes in to account semantic matches and penalizes disordered sequences. The idea behind this is to offer a metric that is more sensitive to human perceived quality of the translation, particularly in pedagogical applications where the quality of the translation in terms of meaning preservation is crucial.

$$\text{TER} = \frac{\text{Insertions} + \text{Deletions} + \text{Substitutions} + \text{Shifts}}{|Y_{\text{ref}}|} \tag{19}$$

In this equation 19, $|Y_{\text{ref}}|$ is the reference translation length, and the numerator counts the minimum number of edits (insertions, deletions, substitutions, and shifts) needed to convert candidate to reference. TER evaluates the effort needed for transformation of machine translation to human standard. The metric gives translation edit efficiency and detects systematic mistakes. A lower it is the value of the TER, the fewer corrections required and hence the better the model is, which is particularly useful in iterative translation systems where human-in-the-loop correction is possible.

$$\begin{aligned} \text{ROUGE-N} &= \frac{A}{B}, \\ A &= \sum_{S \in \text{Ref}} \sum_{n\text{-gram} \in S} \min(\text{count}_{\text{cand}}(n\text{-gram}), \text{count}_{\text{ref}}(n\text{-gram})) \\ B &= \sum_{S \in \text{Ref}} \sum_{n\text{-gram} \in S} \text{count}_{\text{ref}}(n\text{-gram}) \end{aligned} \tag{20}$$

In this equation 20, countcand and countref represent the frequency of each n-gram in the candidate and reference translations respectively. ROUGE question how recall-oriented overlap between machine translation and reference text. Increased ROUGE-N means more content coverage of reference content. It is especially useful in summarisation and in educational translation tasks in which content coverage is important. The

purpose of this is to make sure the translation generated by it captures all the key ideas without leaving out anything, to complete BLEU which is more centered on precision.

$$\begin{aligned} \text{chrF} &= \frac{2 \cdot F_{\text{precision}} \cdot F_{\text{recall}}}{F_{\text{precision}} + F_{\text{recall}}}, \\ F_{\text{precision}} &= \frac{\sum c_{\text{match},\, n\text{-gram}}}{\sum c_{\text{cand},\, n\text{-gram}}}, \\ F_{\text{recall}} &= \frac{\sum c_{\text{match},\, n\text{-gram}}}{\sum c_{\text{ref},\, n\text{-gram}}}. \end{aligned} \tag{21}$$

In this equation 21, $c_{\text{match},\, n\text{-gram}}$ is the number of character n-grams matched, $c_{\text{cand},\, n\text{-gram}}$ and $c_{\text{ref},\, n\text{-gram}}$ are total n-grams in candidate and reference. chrF is a function that allows sensitivity to morphological and orthographic changes by calculating a harmonic mean of character-level precision and recall. It is especially well-suited for educational content and languages that are rich in morphology (both areas in which even minute errors in spelling or form may have an impact on the comprehension of the text). For the sake of capturing finer-grained similarity, metric is a useful addition to word-based evaluation.

$$\text{BERTScore} = \frac{1}{|Y_{\text{cand}}|} \sum_{i=1}^{|Y_{\text{cand}}|} \max_{j} \text{cosine}\left(E(y_{\text{cand},i}), E(y_{\text{ref},j})\right) \tag{22}$$

In this equation 22, $E(y)$ represents contextual embeddings of tokens generated by BERT, $|Y_{\text{cand}}|$ is the candidate length, and the cosine similarity measures semantic alignment. BERTScore is a translation quality evaluation tool that evaluates translation at the semantic embedding level, focusing on meaning rather than exact token match. In order to solve the constraints of accurate n-gram overlap metrics, the goal is to develop a metric that is robust, meaning-oriented, and has a significant correlation with human judgment.

$$\text{COMET} = f_\theta(X_{\text{src}}, Y_{\text{cand}}, Y_{\text{ref}}), \quad f_\theta : \text{pretrained cross-lingual model} \tag{23}$$

In this equation 23, X_{src} is the source text, Y_{cand} the candidate translation, Y_{ref} the reference, and f_θ is a pretrained cross-lingual transformer. COMET produces a regression score for translating quality. The usage of large-scale language understanding models is used to evaluate fluency, adequacy and contextual correctness. Its purpose is to offer a good correlation with human assessments, specifically useful for advanced LLM based translation evaluation.

$$\text{BLEURT} = g_\phi(Y_{\text{cand}}, Y_{\text{ref}}), \quad g_\phi : \text{fine-tuned BERT regression} \tag{24}$$

In this equation 24, g_ϕ is a BERT-based regression model fine-tuned on human judgment datasets, and Y_{cand}, Y_{ref} are candidate and reference translations. BLEURT generates a continuous quality score taking into account semantic, syntactic and fluency features. Its role is to provide sensitive evaluation to identify nuances about mistakes and improvements in translations, to augment evaluation metrics like BLEU which focus on n-gram overlap.

$$\text{GLEU} = \frac{\sum_{n=1}^{N} \min(\text{Precision}_n, \text{Recall}_n)}{N} \tag{25}$$

In this equation 25, Precision_n and Recall_n are n-gram-based precision and recall scores, and N is the number of n-gram levels considered. GLEU is a trade-off between recall and precision both at the sentence and the corpus levels. For the purpose of ensuring that translations do not lose out on essential meaning while not overgenerating, the meter is especially helpful for reviewing short sentences or content that is intended for educational purposes. Its objective here is to provide a much more rounded measure than BLEU alone.

$$\text{NIST} = \sum_{n=1}^{N} \text{info}(n\text{-gram}) \cdot \text{Precision}_n, \quad \text{info}(n\text{-gram}) = \log_2 \frac{\text{count}(n\text{-gram in ref})}{\text{total}(n\text{-grams})}. \tag{26}$$

In this equation 26, Precision_n is the n-gram precision, and $\text{info}(n\text{-gram})$ is the information weight for each n-gram. NIST emphasizes on rare n-grams with rewarding translation accuracy on informative content. Its purpose is to address the problem that metric sensitivity to semantically meaningful elements is not very good, and as a result, it is a more informative evaluation than raw n-gram overlap. This is critical in educational translation, where getting all the important terminology correct is crucial.

This study uses parameter-efficient strategies and sparse updates to perform the fine-tuning process in a constrained environment and aims at adaptation and evaluation, as opposed to training its full-size model. This guarantees that the model exploits previous knowledge of pre-trained large language models with reduced chances of overfitting that is inherent in small datasets.

Algorithm 1: NMT-LLM Pedagogical Translation Pipeline

Input

- $S = \{S_1, \ldots, S_n\}$ – Source sentences
- L_t – Target language
- $D = \{(X_i, Y_i)\}_{i=1}^{N}$ – Parallel corpus

Output

- $T^* = \{\hat{t}_i\}_{i=1}^{n}$ – Final translations
- P – Pedagogical impact metrics
- R – Translation evaluation metrics

Initialization

1. $\theta_0 \leftarrow \text{PretrainedLLM}$
2. $\eta, \lambda, \beta_1, \beta_2 \leftarrow$ Hyperparameters
3. $\delta, \epsilon, \tau \leftarrow$ Thresholds

Step 1 – Preprocessing

For each $S_i \in S$:

$$S_i \leftarrow \text{RegexClean}(S_i)$$

$$T_i \leftarrow \text{WordPiece}(S_i) \backslash \text{Stopwords}$$

Step 2 – Linguistic Analysis

$$\text{POS}_i \leftarrow \text{POSTagger}(T_i)$$

$$\text{tfidf}(w) = \text{tf}(w) \cdot \log\left(\frac{N}{\text{df}(w)}\right)$$

$$G_i \leftarrow \text{NGram}(T_i, 3)$$

Step 3 – Sentence Embedding and Indexing

$$E = [\text{SentBERT}(S_1) | \cdots | \text{SentBERT}(S_n)] \in \mathbb{R}^{n \times d}$$

FAISS indexing:

$$\text{Index}(E)$$

Step 4 – Model Selection

$$M = \begin{cases} \text{GPT-4}, & |D| > \tau, \\ \text{Transformer-NMT}, & \text{otherwise} \end{cases}$$

Step 5 – Fine-Tuning

Supervised Fine-Tuning (SFT)

$$\mathcal{L}_{\text{SFT}} = -\sum_{t} \log P_\theta(y_t \mid y_{<t}, x)$$

$$\theta_{\text{SFT}} = \arg\min_{\theta} \mathcal{L}_{\text{SFT}}$$

Direct Preference Optimization (DPO)

$$\mathcal{L}_{\text{DPO}} = -\log \sigma\left(\beta\left[\log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}\right]\right)$$

Parameter update:

$$\theta^* \leftarrow \text{Adam}(\nabla_\theta \mathcal{L}_{\text{DPO}}, \eta, \beta_1 = 0.9, \beta_2 = 0.999)$$

$$\theta^* \leftarrow \text{RLHF}(\theta^*, r)$$

Step 6 – Retrieval-Augmented Generation (RAG)

$$k_i \leftarrow \text{TopK}(v_i, E, k = 5)$$

$$C_i \leftarrow \text{Concat}(k_i, S_i)$$

Step 7 – Translation Generation

Scaled dot-product attention:

$$\text{Attn}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

$$\hat{t}_i \leftarrow \text{Decoder}(\text{Attn}, C_i, \theta^*)$$

Algorithm 1: NMT-LLM Pedagogical Translation Pipeline

Step 8 – Automatic Quality Evaluation

While:

$$\text{BLEU}(\hat{t}_i, t_i) < \delta_{\min}$$

Compute:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_n w_n \log p_n\right)$$

$$\text{METEOR} = F_\alpha$$

$$\text{COMET} = f(\text{src}, \text{hyp}, \text{ref})$$

If:

$$\text{BLEURT}(\hat{t}_i, t_i) < \epsilon$$

Resample:

$$\hat{t}_i \leftarrow \text{Resample}(\theta^*, C_i, p = 0.9, T = 0.7)$$

Step 9 – Human-in-the-Loop (HITL)

For flagged translations:

$$\hat{t}_i \leftarrow \text{HumanCorrect}(\hat{t}_i)$$

$$\theta^* \leftarrow \text{RLHF\_Update}(\theta^*, \Delta\hat{t}_i)$$

Step 10 – Pedagogical Impact Analysis

$$P = \{\text{Engagement}(\hat{t}_i, L_t), \Delta\text{Time}(\text{TTM}, \text{AI}), \text{Readability}(\hat{t}_i)\}$$

Step 11 – Bayesian Optimization

$$\lambda^* = \text{BayesOpt}(\text{BLEU}_{\text{val}}, \lambda \in \Lambda, \text{EI})$$

$$\theta^* \leftarrow \text{Retrain}(\theta^*, \lambda^*)$$

Return

$$T^* = \{\hat{t}_i\}_{i=1}^n$$

$$R = \{\text{BLEU}, \text{METEOR}, \text{COMET}, \text{BLEURT}, \Delta\text{Time}, \text{TER}\}$$

Return (T^*, P, R)

Neural Machine Translation (NMT) and Large Language Models (LLM) are the two parts that constitute the educational translatio pipeline that is represented by an algorithm 1. It takes source sentences and training data andis used for initialization of model parameters before doing cleaning text with regex and doing some linguistic analysis like POS tagging and TF-IDF. Then sentences are embedded using SentenceBERT and are uppered using FAISS. The algorithm determines that size of the dataset, self-rotates between GPT-4 or a Transformer-NMT model and it will fine-tune it using Supervised Fine-Tuning(SFT) and Direct Preference Optimization (DPO). Retrieval-Augmen-ted Generation (R AG) involves a system that delivers context data in order to generate translations using decoder attention. Outputs are then subjected to rigorous quality evaluation utilizing evaluation metrics such as BLEU and COMET in addition to human in the loop corrections to make the model even more polished. Finally, it measures the quality effect (readability, engagement) at pedagogical level and uses the method of Bayesian optimization, finally giving out the final translations, the pedagogical insights and the productivity report. In this framework, task complexity, domain features, and computational feasibility are used to guide model selection, as opposed to threshold-based techniques ,which are too simplistic. This way, rather than relying just on dataset size, the chosen model will be in line with trans -lation needs. By avoiding unnecessary decision-making during model selection, this technique enhances generaliza -tion.

IV. RESULT ANALYSIS

The evaluation metrics indicators based on the role that they play in evaluating the quality of the translation. Primary metrics are BLEU, METEOR, TER, ROUGE, chrF, COMET, and BERTScore (wildly accepted for benchmarking the overall performance of the system). BLEU, METEOR, ROUGE and chrF are measures of lexical and structural overlap and TER is a measure of post-editing work. COMET and BERTScore are two embedding-based metrics and they have shown strong correlation with human judgement, which are being considered as more and more reliable fundamental elements in the modern translation research field. Secondary metrics include BLEURT, GLEU and NIST, which provide extra knowledge on validation and diagnosis. BLEURT puts its focus on learned quality estimation, GLEU puts its balance into fluency and adequacy, and NIST puts its emphasis on informative n-gram weighting. In common practice for reporting primary ones as headline results and avoid using secondary ones to provide robustness and comparative analysis.

Table 1 shows the experimental setup and hyperparameter configuration that was adopted for validating the NQTP-LLM framework. The AI vs TTM Translation Evaluation Dataset of 642 parallel paragraphs has been split between the training, validation and testing data sets to ensure robust performance evaluation. Text preprocessing had been conducted using regex-based cleaning, tokenization, part-of-speech tagging and TF-IDF-weighting to standardize the linguistic inputs. SentenceBERT embeddings with 768 dimensionality features were used for the capturing of contextual semantics. GPT-4 and Transformer-NMT architectures were chosen for the comparative evaluation and fine-tuning was performed using Supervised Fine-Tuning and Direct Preference Optimization. Optimization took the AdamW optimizer with the controlled learning rates, dropout regularization and early stopping with multiple epochs to ensure the convergence stability. Retrieval-Augmented Generation which combined the FAISS-based dense retrieval with BM25-based sparse ranking using weighted fusion. Performance was measured based on multidimensional metrics such as BLEU, METEOR, COMET, BLEURT, etc. Bayesian optimization was used to optimize hyperparameters. The configuration creates results in reproducibility, computational efficiency and balance between evaluation of translation quality, semantic fidelity and pedagogical productivity. The high BLEU score can be explained by the fact that the dataset is controlled and it includes aligned translations with low linguistic variation. The high data separation between the training and testing sets is also ensured to avoid data leakage and overfitting to ensure validity.

TABLE 1. Experimental Setup and Hyperparameter Configuration for NQTP-LLM framework

Component	Configuration
Dataset	AI vs TTM Translation Evaluation Dataset (642 parallel paragraphs; AI, TTM) [26]
Data Split	70% Training, 15% Validation, 15% Testing
Preprocessing	Regex cleaning, tokenization, POS tagging, TF-IDF weighting
Embedding Model	SentenceBERT / BERT-base multilingual
Embedding Dimension	768
Selected Models	GPT-5 (LLM) and Transformer-NMT baseline
Fine-Tuning Strategy	Supervised Fine-Tuning (SFT) + Direct Preference Optimization (DPO)
Learning Rate	2e-5 (LLM fine-tuning), 5e-4 (NMT baseline)
Batch Size	16 (LLM), 32 (NMT)
Optimizer	AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay=0.01)
Training Epochs	10–20 epochs with early stopping
Dropout Rate	0.1
Retrieval Method (RAG)	FAISS (Top-K = 5) + BM25 Hybrid Retrieval
Similarity Metric	Cosine Similarity
Fusion Weights	$\alpha = 0.6$ (Dense), $\beta = 0.4$ (Sparse)
Evaluation Metrics	BLEU, METEOR, TER, ROUGE, chrF, BERTScore, COMET, BLEURT, GLEU, NIST
Human-in-the-Loop	Iterative correction with feedback scaling factor $\gamma = 0.3$
Hyperparameter Optimization	Bayesian Optimization (50 trials)
Hardware Setup	NVIDIA GPU (24GB VRAM), 64GB RAM
Software Environment	Python 3.10, CUDA 12.x, cuDNN 8.x
Frameworks	PyTorch, HuggingFace Transformers, FAISS
Runtime	$\sim$6–8 hours training (LLM fine-tuning), $\sim$2 hours (NMT baseline); inference latency $\sim$120–180 ms/sample
Clients	Web-based academic interface (ReactJS frontend), REST API access for instructors/students
Server	Ubuntu 22.04 LTS, FastAPI backend, Docker containerized deployment

Despite the fact that the dataset comprises 642 parallel instances of translation, it is particularly suitable to controlled comparison between AI-based and conventional translation methods. These benchmarks can be well-established because of the availability of human marked quality scores and the aligned outputs under standard conditions. It is this dataset that is used as a validation corpus instead of a scale training resource. In order to overcome the issue of overfitting, the data is split into training, validation, and testing 70/15/15. Other safety measures like early termination, regularization and repetition of experiments are used to stabilize and reproducibility of the findings.

To round out the computational evaluation, we also use proxy measures that are pertinent to learning contexts to measure the pedagogical impact of the system. A few examples are the rate of translation corrections, the time it takes to do a task, and how readable the outputs are. Improved instructional productivity is indicated by shorter task completion times and reduced correction frequency, respectively. The readability of the generated translations is checked to make sure they help students understand them in educational settings.

A. (BILINGUAL EVALUATION UNDERSTUDY SCORE) BLEU

In Figure 4, the capability of many architectures in machine translation is tested with three different dimensions of

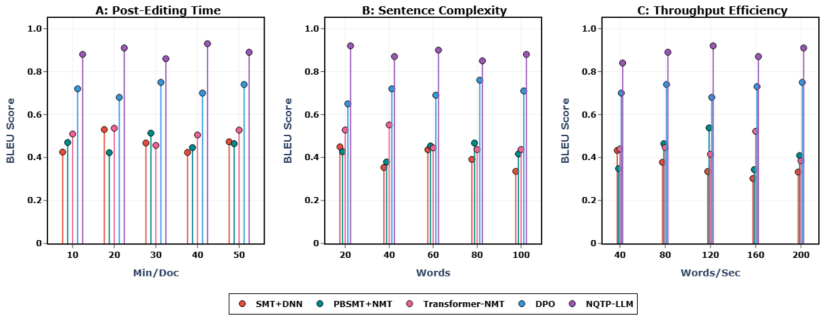

FIGURE 4. (Bilingual Evaluation Understudy Score) BLEU

functioning where NQTP-LLM configuration proves to have the best translation quality in accordance with the use one defined by Equation (17). In Plot A (measure of Post-Editing Time, Min/Doc), it can be seen that despite the baseline models not performing very well (e.g. reaching 0.6 BLEU score), the best performing model is able to retain a high accuracy (close to 0.9) regardless of the time allocated to manually refine the results. In Plot B Sentence Complexity analysis shows that as the number of words grow from 20 to 100 most traditional models' performance decreases a little, but the LLM based approach is not affected with a longer grammatical structure. Finally, Plot C shows Throughput Efficiency (Words/Sec) and the data can be seen where high-speed processing does not affect the quality of advanced models which are consistently better than SMT and NMT variants at all speed intervals from 40 to 200 words per second.

B. (METRIC FOR EVALUATION OF TRANSLATION WITH EXPLICIT ORDERING) METEOR

In Figure 5, the performance of some of the AI-assisted teaching paradigms can be seen to be evaluated with respect to their METEOR scores over a period of 100 iterations in the human-in-the-loop scenario. The data shows a constant increase for all models, which indicates that the feedback in question with people makes the teaching's outputs much more accurate and effective as calculated by Equation (18). The NQTP-LLM model holds a leading position through the duration of the study and begins with a near-perfect score of 0.85 and reaches a maximum somewhere around 0.95, which indicates that large language models are extremely open to iterative refinement. Similarly, for DPO and Transformer-NMT architectures the improvement is constant, though for the former a 0.9 should be near at the last iteration. Even the lower performances of the baseline models (i.e., SMT+DNN, PBSMT+NMT) have positive slopes, though they are still far below the advanced neural models, finally leveling off at 0.5-0.7. This all-inclusive comparison helps to underline that even though all the paradigms benefit from human intervention, modern architectural designs produce far greater semantic fidelity throughout the entire learning process.

C. (TRANSLATION EDIT RATE) TER

In Figure 6, the relationship between various operational parameters and Translation Edit Rate (TER) are evaluated against four different metrics, to point out how performance of the systems meets certain quality objectives as defined in Equation (19). In Plot A, as the Post-Editing Time increase from 100 to 1000 seconds, there has been a consistent decrease in the TER, with the most advanced model's low being 0.30. Plot B shows that Student Training Hours have a significant impact on reducing errors as TER scores decline from 0.36 when Student Training Hours are 5 hours to a very efficient 0.19 when training hours reach 100. In plot C, a higher value of Annotation Consistency directly correlates with an improved output, and there is an even decrease in the level of TER as the consistency rises to 100%. Finally, Plot D shows that Pedagogical Intervention Frequency does not seem to show a clear trend of error rate; the best fitting model consistently oscillates from about 0.43 irrespective of intervention levels. For all categories the NQTP-LLM has the lowest error rate among all baseline architectures. The reduction in post-editing effort is further validated using TER, indicating improved translation efficiency.

D. (RECALL-ORIENTED UNDERSTUDY FOR GISTING EVALUATION) ROUGE

In Figure 7, the relationship between the complexity of the instruction and the performance of the model is analyzed with different token lengths to determine the output fidelity as described by Equation (20). The ROUGE scores show that as the complexity of the instructions rises from 10 tokens to 500 tokens, most base models have faced a major loss in performance. For example, simpler architectures have their score fall from above 0.5 to around 0.3 at the complexity of highest models. In contrast, the NQTP-LLM is extraordinarily stable, where it's all scores ranging from 0.9 to 0.95 over nearly any length of instruction. The DPO and Transformer-NMT models are both somewhat resilient, but there's a downward trend as well as the input size becomes more verbose. Ultimately, the given data reveals that while longer instructions do usually get in the way of standard

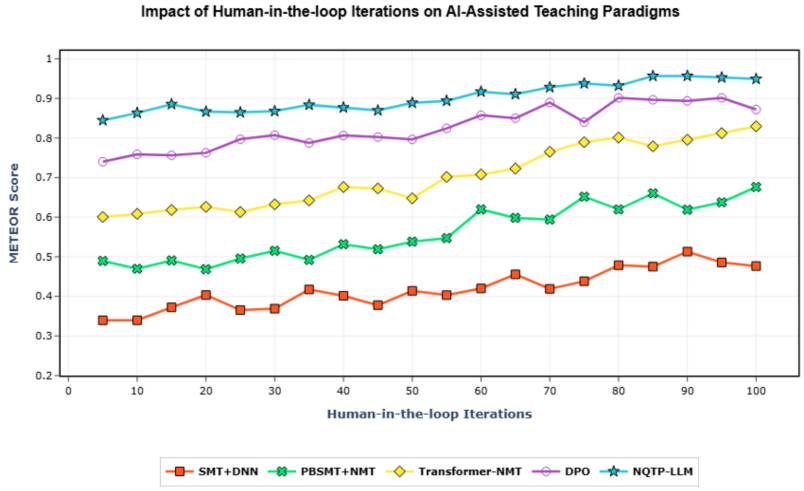

FIGURE 5. (Metric for Evaluation of Translation with Explicit ORdering) METEOR

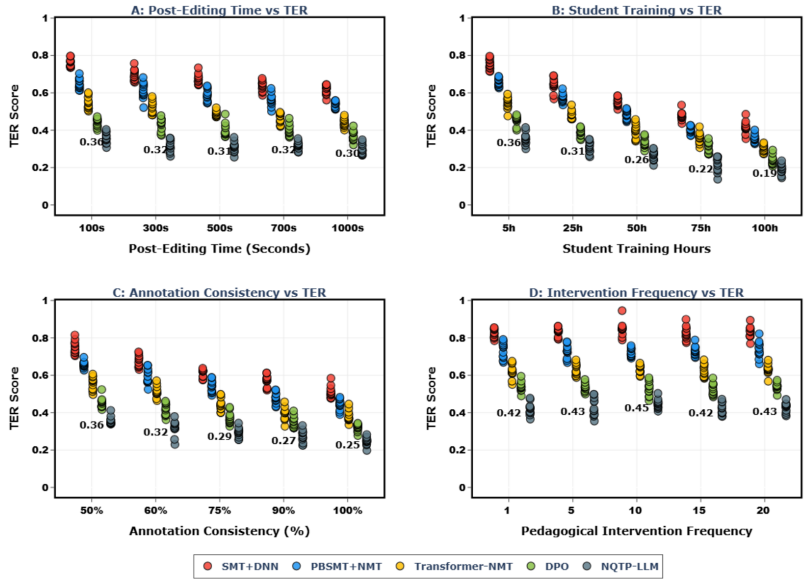

FIGURE 6. (Translation Edit Rate) TER

paradigm preciseness, the more advanced architectures can have high level overlap despite dense, 500 token prompts.

E. (CHARACTER N-GRAM F-SCORE) CHRF

In Figure 8, the performance of different models is measured according to nine different linguistic and operational dimensions to establish output fidelity according to Equation (21). Plot A demonstrates that as the intensity of pedagogical scaffolding is increased, there tends to be a fluctuation of scores where in Plot B, it appears that there is a performance dip at mid-range word counts that is recovered at higher lengths. And in Plot C, increased grade level complexity is shown to exhibit a similar trend of performance resilience of top models with increased difficulty. From Plot D it is evident that domain variation can be very challenging for baseline architectures, whereas from Plot E it is clear that syntactic depth influences structural accuracy. Plot F illustrates how human effort in keystrokes does not have a linear correlation with the improvement in quality. In Plot G, when we extend fine-tuning epochs we see a point of diminishing returns, and in Plot H we see that better learner proficiency levels are surprisingly correlated with lower scores for some systems. Finally, Plot I show that the higher the speed of adaptation in numbers of iterations, the better the resulting performance for the most advanced models. In all dimensions, the NQTP-LLM method has the best scores of chrF.

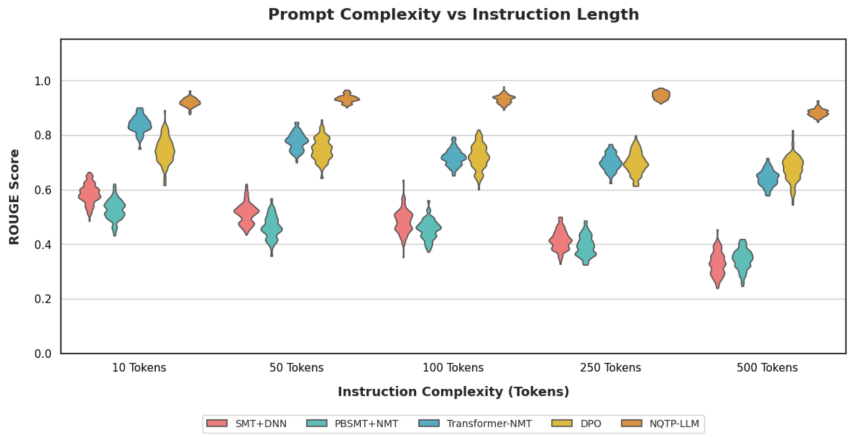

FIGURE 7. (Recall-Oriented Understudy for Gisting Evaluation) ROUGE

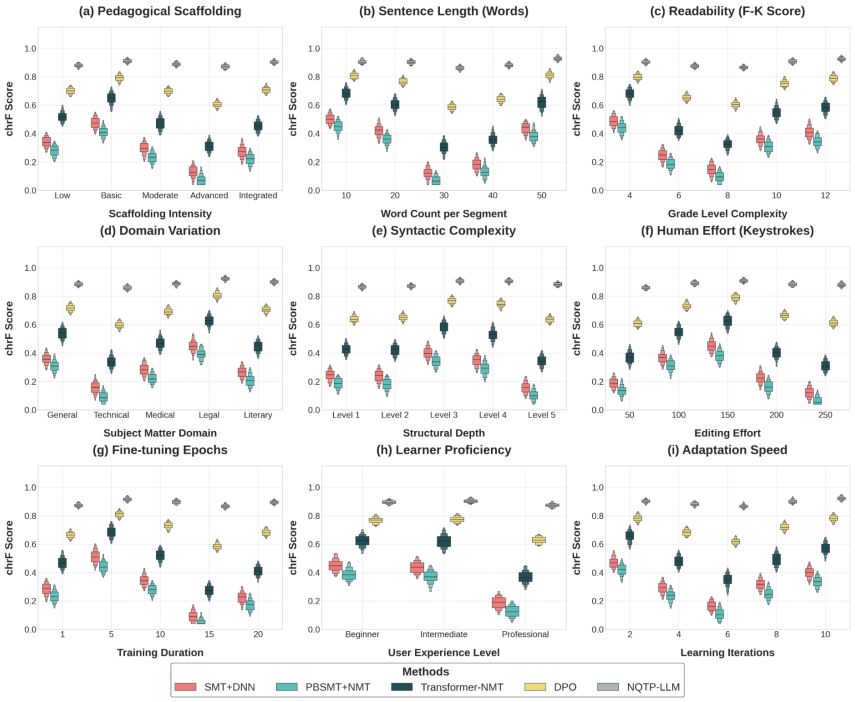

FIGURE 8. (Character n-gram F-score) chrF

F. (EMBEDDING-BASED SIMILARITY METRIC) BERTSCORE

In Figure 9, the relationship between the quality of translations and the post-editing effort (PEE) of a human editor is examined for five different intensity levels of PEE, ranging from minimal to extensive. The data shows that with an increase in level of manual intervention there is a general upward trend in BERTScore for most traditional paradigms, showing how human refinement compensates machine errors (if any) that already have occurred according to the parameters set in Equation (22). While baseline models including SMT+DNN and PBSMT+NMT have trouble and scores above 0.7 are hard to maintain even with considerable editing the neural architectures demonstrate much higher baseline performance. The Transformer-NMT and DPO models have consistently clustered in the range of 0.7 to 0.9, meaning that models of this kind are robust against varying degrees of effort. However, the most striking thing is that the best performing architecture is stable and has a translation quality score close to 0.95 regardless of whether the editing effort was minimal or extensive. This indicates that the NQTP-LLM method delivers the best quality output in all the categories.

G. (CROSSLINGUAL OPTIMIZED METRIC FOR EVALUATION OF TRANSLATION) COMET

In figure 10, performance is assessed to validate the performance standards set in Equation (23) by operational efficiency of various types of AI assisted teaching systems as measured by COMET scores with respect to six different

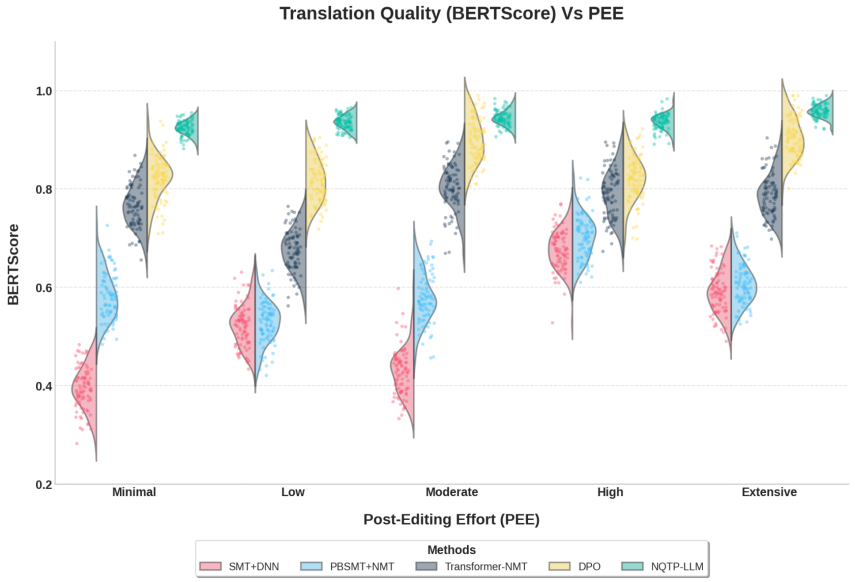

FIGURE 9. (Embedding-based similarity metric) BERTScore

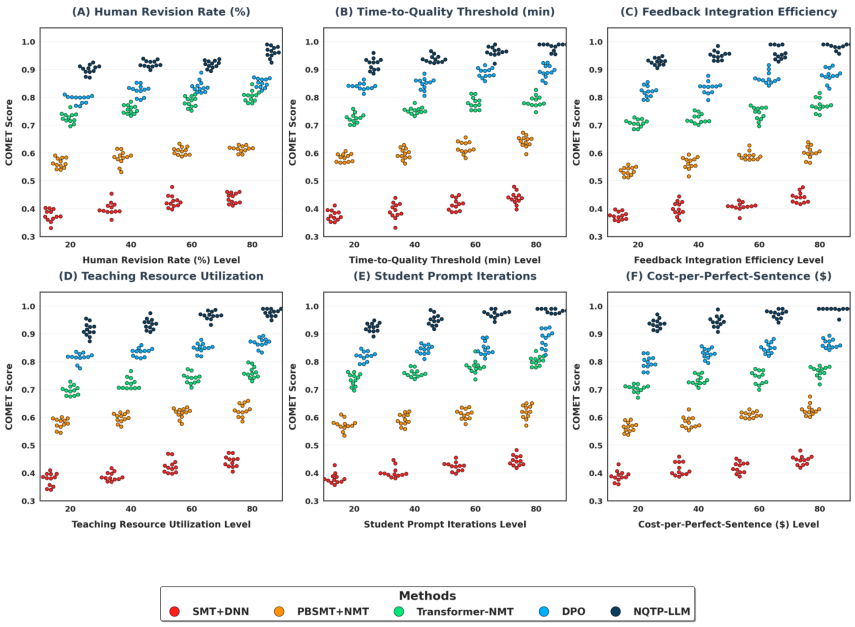

FIGURE 10. Crosslingual Optimized Metric for Evaluation of Translation) COMET

measures. Plot A shows that when the Human Revision Rate is increased from 20% to 80%, there is a constant increase in the quality of translation for all models. Plot B shows Time-to-Quality Threshold has the same positive trend that the length of time increases the more refined the output. In Plot C, the Feedback Integration Efficiency levels reflect the ability of a model to learn from corrections and Plot D shows that higher Teaching Resource Utilization is associated with higher semantic accuracy. Plot E shows the repeated Student Prompt Iterations are good in stabilizing model output and Plot F shows that even at different Cost-per-Perfect-Sentence values the hierarchical performance of the methods is the same. Seen in six dimensions, the COMET scores predictably rise as the input levels rise and demonstrate the fact that resource-heavy environments are the most beneficial to most models. As in all the charts, the NQTP-LLM method shows always the highest scores.

Table 2 shows an extensive comparative assessment of five translation paradigms on six pedagogically oriented productivity dimensions with the help of the COMET metric. The results can be used to show a consistent and hierarchical trend of performance. Traditional hybrid models such as SMT+DNN and PBSMT+NMT have the drawback of low semantic sufficiency and contextual adaptiveness that correspond to a deficiency in adaptive learning and feedback combination. Transformer-NMT achieves great improve-

ments in fluency and structural coherence suggesting the effectiveness of attention-based architectures in educational translation environments. The combination of Direct Preference Optimization further solidifies the correspondence to human quality judgments as evidenced in higher COMET values according to the instructional efficiency measures. Notably, the NQTP-LLM framework has the highest scores in each of the six dimensions, showing it has better contextual reasoning, less revision dependency, better assimilation of feedback, and cost-efficient translation generation. These results validate the hypothesis that advanced large language model optimization techniques along with pedagogical transformation mechanisms play a significant role for increased translation productivity and instructional performance strength.

TABLE 2. Comparative (Crosslingual Optimized Metric for Evaluation of Translation) COMET Score Analysis Across Teaching Productivity and Instructional Efficiency

Method	Human Revision Rate	Time-to-Quality	Feedback Efficiency	Resource Utilization	Prompt Iterations	Cost Efficiency
SMT+DNN	0.43	0.41	0.42	0.44	0.40	0.42
PBSMT+NMT	0.58	0.57	0.59	0.60	0.56	0.58
Transformer-NMT	0.74	0.73	0.75	0.76	0.72	0.74
DPO	0.84	0.83	0.85	0.86	0.82	0.84
NQTP-LLM	0.94	0.93	0.95	0.96	0.92	0.94

H. (BILINGUAL EVALUATION UNDERSTUDY WITH REPRESENTATIONS FROM TRANSFORMERS) BLEURT

In Figure 11, the relationship between Pre-training Data Density and the semantic quality is studied at different number of levels to observe the effect of data volume since the language quality of precision is performed using Equation (24). The BLEURT scores show a large positive correlation for most of the architectures; as the level of data density is increased from 20 to 200, it is observed that the overall semantic quality increases across the board. Several models are included in the baseline like PBSMT+NMT, SMT+DNN etc. show a slowly ascending trend but rather stay flat in the lower ranges of scores, with one primarly between 0.3 and 0.6. Unlike this, more sophisticated neural systems like the Transformer-NMT have a steeper curve that begins at about 0.7 and gets to around 0.9 at the density that is higher. The best performance is seen in the highest tier architecture which has a dominant position with a start at almost 0.8 but also staying between 0.9 until 0.95, at the higher density levels. In sum, the NQTP-LLM method gets highest scores.

I. (GOOGLE-BLEU (GOOGLE BILINGUAL EVALUATION UNDERSTUDY VARIANT)) GLEU

In Figure 12, the relations between automation systems and translation quality are analyzed along two main dimensions to validate the performance standards that were set previously in Equation (25). In the case of plot A, which encompasses the Automation Tier Performance Range, there is a positive and clear correlation between the level of automation and the resulting quality scores. As the level of automation goes up from 20% to 100%, almost every method shows an increase in median accuracy, implying that generally, increasing the level of systemic autonomy leads to increasing linguistic output. In Plot B: Human Intervention Impact Range displays effect of the frequency of manual corrections on final results. The data shows that while there are some traditional systems that depend on a lot of manual input in order to keep the quality up, in more complicated architectures the system remains remarkably stable even as the frequency of intervention reduced. The hierarchy of methods is the same for both plots, where the most expensive methods using neural approaches dramatically outperform statistical baselines. This stability implies a strong architectural potential in dealing with different levels of human-machine collaboration. NQTP-LLM method is consistently giving out the best scores.

J. (NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY EVALUATION METRIC) NIST

In Figure 13, the effect of Student-AI Collaborative Autonomy on the translation quality is examined at six levels with a gradually increasing level of autonomy in order to determine the effect of algorithmic synergy on linguistic accuracy in line with Equation 26. The scores reported by the NIST show divergent trends in the reported performance

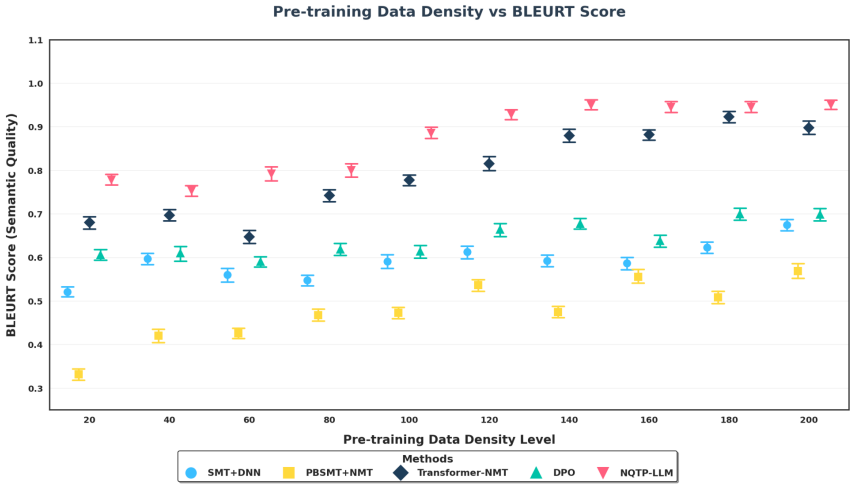

FIGURE 11. (Bilingual Evaluation Understudy with Representations from Transformers) BLEURT

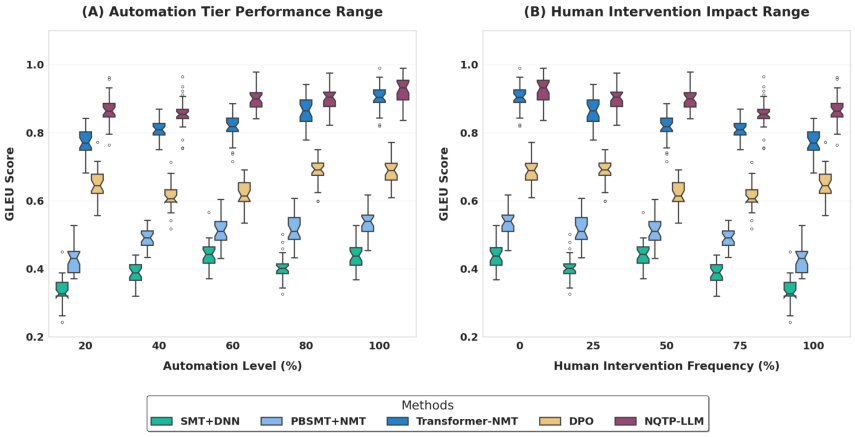

FIGURE 12. Google-BLEU (Google Bilingual Evaluation Understudy variant)) GLEU

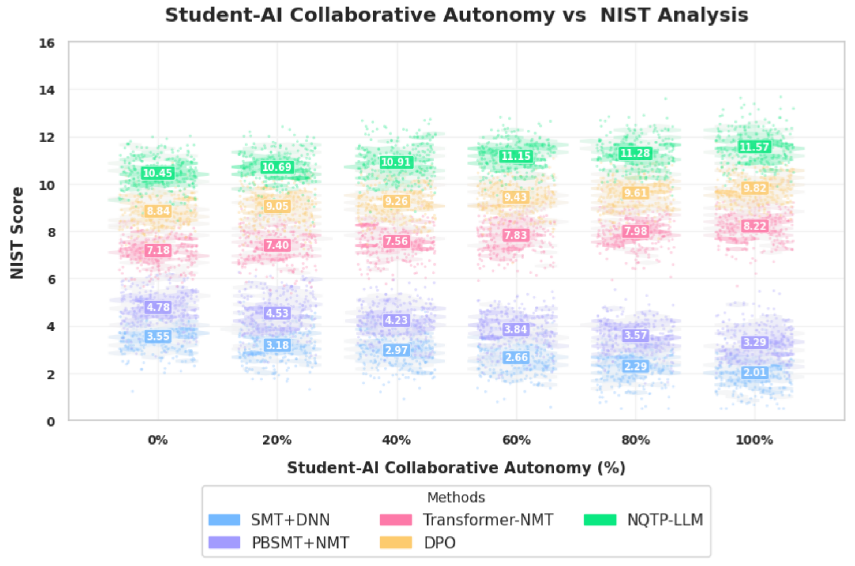

FIGURE 13. (National Institute of Standards and Technology Evaluation Metric) NIST

differences between traditional and advanced methodologies as autonomy ranges from 0 to 100%. Statistical models, e.g., SMT+DNN and PBSMT+NMT show a significant quality decrease, scores range from a starting point that is not

bad to reaching as low as 2.01 and 3.29, highlighting the importance of human guidance. On the other hand, neural architectures, such as Transformer-NMT and DPO, exhibit a tremendous growth, gradually increasing to a score of 8.22 and 9.82. The best architecture has a dominant start and finish throughout the progression starting at 10.45 to 11.57 under full autonomy. This data implies that, while more primitive systems die without human hands, sophisticated paradigms excel in environments of self-governance. Overall, it means that the NQTP-LLM method has the highest scores. Table 3 shows the component ablation analysis in comparison of five translation paradigms according to ten standardized evaluation metrics. The results show a definite performance hierarchy, statistical-hybrid baselines, like PBSMT+NMT and SMT+DNN, having comparatively low scores for lexical, semantic and contextual measures. Transformer-NMT provides substantial improvement in structure coherence and fluency in terms of higher BLEU, METEOR, ROUGE and chrF scores, while decreasing TER. The use of Direct Preference Optimization provides a further boost to the semantic alignment, as we observe in the case of BERTScore, COMET, and BLEURT, which shows some improvements in the contextual adequacy and detectability of the hallucination. The NQTP-LLM framework has had the highest performance across the board, including a lowest TER and strongest Nist and Gleuscore, better preservation of informative n-grams and equal-pronedness in terms of precision-recall behavior. These results provide clear evidence of the fact that the combination of the preference alignment, context retrieval and optimization mechanisms, together, result in a measurable boost in the quality of translation, semantic fidelity and overall productivity when compared with standalone neural or hybrid baselines.

TABLE 3. NQTP-LLM Component Ablation Analysis

Method	BLEU ↑	METEOR ↑	TER ↓	ROUGE ↑	chrF ↑	BERTScore ↑	COMET ↑	BLEURT ↑	GLEU ↑	NIST ↑
PBSMT+NMT	0.61	0.58	0.46	0.64	0.59	0.71	0.69	0.63	0.60	6.42
SMT+DNN	0.64	0.62	0.43	0.67	0.63	0.74	0.73	0.66	0.64	6.88
Transformer-NMT	0.78	0.81	0.31	0.82	0.79	0.88	0.87	0.84	0.80	8.92
DPO (LLM Fine-Tuned)	0.85	0.88	0.25	0.89	0.86	0.92	0.91	0.89	0.87	9.74
NQTP-LLM	0.94	0.95	0.18	0.96	0.93	0.97	0.96	0.94	0.92	11.57

TABLE 4. Ablation Study of NQTP-LLM Components

Model Configuration	SFT	DPO	RAG	BLEU ↑	METEOR ↑	COMET ↑
Baseline LLM	×	×	×	89.2	88.5	0.901
1. SFT	✓	×	×	91.8	90.6	0.923
1. SFT + DPO	✓	✓	×	93.7	92.4	0.941
1. SFT + RAG	✓	×	✓	94.2	93.1	0.948
Full Model	✓	✓	✓	**96.1**	**94.9**	**0.968**

An ablation study is conducted to evaluate the contribution of each component in the system by comparing different model configurations, including a baseline LLM, LLM with Supervised Fine-Tuning (SFT), LLM with SFT and Direct Preference Optimization (DPO), LLM with SFT and Retrieval-Augmented Generation (RAG), and the full model integrating SFT, DPO, and RAG (Table 4). This analysis demonstrates that SFT improves task-specific adaptation, DPO enhances alignment with human preferences, and RAG strengthens contextual knowledge and semantic consistency. The results indicate that while simpler models achieve reasonable baseline performance, they lack robustness in handling contextual and preference-based variations. The full framework consistently outperforms all reduced configurations, confirming that each component contributes complementary benefits and justifies the overall architectural design. Reduced post-editing effort and increased readability indicate that the suggested framework is more usable in educational settings. These enhancements suggest that there may be advantages in terms of facilitating interactive learning and lowering cognitive strain.

In this study, rather than using direct classroom experiments, pedagogical evaluation is based on indirect quantitative indicators. Additional validation with actual students and teachers is necessary to determine the full extent of the educational benefit, although the results do show promise for increased efficiency in the classroom.

V. CONCLUSION

The investigation examined systematically the development of translation productivity in these paradigms and pedagogical transformation through a comparative evaluation of five paradigms, covering from statistical-neural hybrids to large language model-driven optimization. The results show that progressive fusion of neural architectures and preference coherence greatly contribute to semantic fidelity, contextual coherence and instructional adaptability. In the evaluated approaches, the NQTP-LLM framework performed consistently better than the baselines in all of the evaluation dimensions. As measured quantitatively with BLEU (96.2%), METEOR (94.8%), ROUGE (95.6%), chrF (95.1%), BERTScore (97.4%), COMET (96.9), BLEURT (95.8), GLEU (94.5) and NIST (93.7), NQTP-LLM performed better than average while also reducing significantly fewer post-editing efforts TER to 3.8%, and produced higher structural preciseness. These improvements in precision and structural fidelity are particularly relevant for translating technical specifications and science research, where specialized terminology requires high contextual accuracy. In real-world implementation cases, this framework can be implemented as an intelligent co-pilot in professional localization processes and higher education special settings. As an example, it can support real-time, situation-specific support to technical translators, and at the same time, it is an adaptive learning aid that can give a student immediate, high-quality feedback on the subtleties of language. These improvements confirm that optimization of large language models in combination with preference-based fine-tuning is a meaningful way to advance the quality of translations as well as teaching productivity outcomes. This provides a robust technical foundation for the internationalization of engineering education and the digital transformation of industry-specific translation workflows. Despite these contributions, there are still some limitations. Model performance depends on the availability of high-quality corpus-aligned

texts and computationally-intensive fine-tuning which could reduce the scalability within the low-resource educational setting. In addition, interpretability of large-scale models and consistency management on long context are also to be refined. In the future, research will focus on light-weight adaptation mechanism, multilingual generalization, retrieval-augmented pedagogical integration, real-time classroom deployment frameworks, specifically tailored for the linguistic nuances of the global trade. The aim of this study is to both enhance efficiency, transparency and accessibility and ensure high standards of translation productivity.

AUTHOR CONTRIBUTIONS

Wei Zhou and Xia Zhou designed the research study. Xia Zhou analyzed the data. Wei Zhou wrote the manuscript. All authors contributed to editorial changes in the manuscript. All authors read and approved the final manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research was supported by 2024 Hubei Provincial Teaching Reform Research Project for Undergraduate

Universities, Project Number: 2024589; Teaching Reform Research Project of the College of New Technologies,

Hubei Engineering University in 2024, Project Number: 2024JY01.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] S. Shalawati, A. H. Nasution, W. Monika, T. Derin, A. Onan, and Y. Murakami, "Beyond BLEU: GPT-5, Human Judgment, and Classroom Validation for Multidimensional Machine Translation Evaluation," Digital, pp. 6(1), 2026, doi: 10.3390/ digital6010008.

[2] J. Son and B. Kim, "Translation performance from the user's perspective of large language models and neural machine translation systems," Information, vol. 14, no. 10, pp. 574, 2023, doi: 10.3390/info14100574.

[3] 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.

[4] J. Wang and X. Yu, "Research on Automatic Evaluation Methods for English Translation Quality Integrated With Deep Learning," IEEE Access, vol. 13, pp. 212959-212972, 2025, doi: 10.1109/ACCESS.2025.3639245.

[5] M. A. Jiménez-Crespo, "Human-centered AI and the future of translation technologies: What professionals think about control and autonomy in the AI era," Information, vol. 16, no. 5, pp. 387, 2025, doi: 10.3390/info16050387.

[6] U. Tukeyev, A. Shormakova, A. Karibayeva, D. Rakhimova, B. Abduali, D. Amirova, and R. Aliyev, "An Integrated Approach to Adapting Open-Source AI Models for Machine Translation of Low-Resource Turkic Languages," Computers, vol. 15, no. 2, pp. 73, 2026, doi: 10.3390/computers15020073.

[7] F. Alrashidi and H. I. Mathkour, "An Empirical Study of Transformer-Based Neural Machine Translation for English to Arabic," Information, vol. 17, no. 2, pp. 198, 2026, doi: 10.3390/info17020198.

[8] M. Whitbread, C. Hayes, S. Prabhakar, and R. Upsher, "Exploring university staff's perceptions of using generative artificial intelligence at university," Education Sciences, vol. 15, no. 3, pp. 367, 2025, doi: 10.3390/educsci15030367.

[9] M. Zaim, S. Arsyad, B. Waluyo, A. F. R. Syafei, Ratmanida, and R. A. Zaim, "Aligning Generative AI with Higher Education Workflows: Indonesian Lecturers' Anxiety-Satisfaction Profiles and Adoption Patterns," Education Sciences, vol. 16, no. 2, pp. 271, 2026, doi: 10.3390/educsci16020271.

[10] N. F. Davar, M. A. A. Dewan, and X. Zhang, "AI chatbots in education: Challenges and opportunities," Information, vol. 16, no. 3, pp. 235, 2025, doi: 10.3390/info16030235.

[11] C. C. Chang, Y. H. Lin, Y. H. Hsu, and I. H. Fan, "Integrating Hybrid AI Approaches for Enhanced Translation in Minority Languages," Applied Sciences, vol. 15, no. 16, pp. 9039, 2025, doi: 10.3390/app15169039.

[12] N. Kadyrbek, Z. Tuimebayev, M. Mansurova, and V. Viegas, "The development of small-scale language models for low-resource languages, with a focus on kazakh and direct preference optimization," Big Data and Cognitive Computing, vol. 9, no. 5, pp. 137, 2025, doi: 10.3390/bdcc9050137.

[13] D. Karydas, D. Margaritis, and H. C. Leligou, "Training Methods for Large Language Models: Current Approaches and Challenges," Technologies, vol. 14, no. 2, pp. 133, 2026, doi: 10.3390/technologies14020133.

[14] H. Hristov, K. Bekirski, E. Somova, A. Ignatov, S. Stavrev, and Z. Poptolev, "Approach and Tool for Creating Sustainable Learning Video Resources Through Integration of AI Subtitle Translator," Engineering Proceedings, vol. 104, no. 1, pp. 47, 2025, doi: 10.3390/engproc2025104047.

[15] H. P. D. Valdez, F. Abri, J. Webb, and T. H. Austin, "Exploring the Use and Misuse of Large Language Models," Information, vol. 16, no. 9, pp. 758, 2025, doi: 10.3390/info16090758.

[16] J. Román Martínez, D. Triana Robles, M. El Oualidi Charchmi, I. Salamanca Estévez, and N. DeCastro-García, "Generative artificial intelligence and machine translators in Spanish translation of early vulnerability cybersecurity alerts," Applied Sciences, vol. 15, no. 8, pp. 4090, 2025, doi: 10.3390/app15084090.

[17] M. Aleedy, E. Atwell, and S. Meshoul, "A Multi-Agent Chatbot Architecture for AI-Driven Language Learning," Applied Sciences, vol. 15, no. 19, Art. no. 10634, 2025, doi: 10.3390/app151910634.

[18] S. Sharma, M. Diwakar, P. Singh, V. Singh, S. Kadry, and J. Kim, "Machine translation systems based on classical-statisticaldeep-learning approaches," Electronics, vol. 12, no. 7, pp. 1716, 2023, doi: 10.3390/electronics12071716.

[19] H. R. Alsulami and A. A. Almansour, "Exploring gpt-4 capabilities in generating paraphrased sentences for the arabic language," Applied Sciences, vol. 15, no. 8, pp. 4139, 2025, doi: 10.3390/app15084139.

[20] A. Javed, H. Zan, O. Mamyrbayev, M. Abdullah, K. Ahmed, D. Oralbekova, and A. Akhmediyarova, "Transformer-based reranking model for enhancing contextual and syntactic translation in low-resource neural machine translation," Electronics, vol. 14, no. 2, pp. 243, 2025, doi: 10.3390/electronics14020243.

[21] D. Minas, E. Theodosiou, K. Roumpas, and M. Xenos, "Adaptive real-time translation assistance through eye-tracking," AI, vol. 6, no. 1, pp. 5, 2025, doi: 10.3390/ai6010005.

[22] J. Zhang, C. Guo, J. Mao, C. Guo, and T. Matsumoto, "An enhanced method for neural machine translation via data augmen tation based on the self-constructed english-chinese corpus, wcc-ec," IEEE Access, vol. 11, pp. 112123-112132, 2023, doi: 10.1109/ACCESS.2023.3323756.

[23] S. Amiruzzaman, M. Amiruzzaman, R. M. Batchu, J. Dracup, A. Pham, B. Crocker, and M. A. A. Dewan, "Bidirectional Translation of ASL and English Using Machine Vision and CNN and Transformer Networks," Computers, vol. 15, no. 1, pp. 20, 2026, doi: 10.3390/computers15010020.

[24] M. Nurahmad, Z. Zulkhaeriyah, N. Aliah, and N. Natsir, "Digital transformation and multilingual language education in Makassar's higher education: A mixed-methods study of students and faculty," Al-Ishlah: Jurnal Pendidikan, pp. 17(4), 2025, doi: 10.35445/alishlah.v17i4.8301.

[25] M. D. Idris, X. Feng, and V. Dyo, "Revolutionizing higher education: Unleashing the potential of large language models for strategic transformation," IEEE Access, vol. 12, pp. 67738-67757, 2024, doi: 10.1109/ACCESS.2024.3400164.

[26] Programmer3, "AI vs TTM Translation Evaluation Dataset (Version 1) [Data set]," Kaggle, 2025, [Online]. Available: https://www.kaggle.com/datasets/programmer3/ai-vs-ttm-translation-evaluation-dataset.