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

Mathematical Modelling of Engineering Problems
Vol. 12, No. 7, July, 2025, pp. 2573-2586
Journal homepage: http://iieta.org/journals/mmep

Image-Based Text Translation a Fine-Tuning Using DeepSeek-Coder and Transformer Models for Multilingual Optical Character Recognition Processing

Mohammed H. Rasheed, Farooq Safauldeen Omar, Amel Tuama, Mohammed Safar*

Department of Computer Technology Engineering, Northern Technical University, Technical Engineering College Kirkuk, Kirkuk 36001, Iraq

Corresponding Author Email: mohammed.sefer@ntu.edu.iq

https://doi.org/10.18280/mmep.120735

Received: 12 February 2025
Revised: 13 June 2025
Accepted: 19 June 2025
Available online: 31 July 2025

Keywords:
Optical Character Recognition (OCR), transformer, DeepSeek-Coder, MarianMT, MBart, M2M-100, T5-Tiny

ABSTRACT

The combination of Optical Character Recognition (OCR) and deep learning-based translation has significantly enhanced multilingual text processing from images, especially. DeepSeek-Coder is a 1.3-billion-parameter autoregressive transformer model fine-tuned as a baseline for text translation from images in this study. Its performance is implemented and compared with other state-of-the-art transformer models Phi-1, MarianMT, MBart, M2M-100 and T5-Tiny. This work has prepared a dataset consisting of 1,000 images and their correlated English and French text which were extracted using EasyOCR. The performance is measured by using standard translation scores such as BLEU, METEOR, ROUGE-L, TER and perplexity. The DeepSeek-Coder achieves the best performance among other approaches with a BLEU score of 0.7733 and a very low perplexity of 1.28, beating all other models by far. These results demonstrate the outstanding translation accuracy and smoothness and highlight efficiency of transformer in OCR based translation. This study provides recommendations for the proper selection of a transformer model for processing of text have been emphasized through these findings. Additionally, the contributions of this work have been added to state-of-the-art in developing multilingual AI for real-life translation work with an OCR-based.

1. INTRODUCTION

In the age when information has become ever more visual that having a proper extraction and interpretation of language in photographs is more important than ever. As the text in photographs such as in-scanned documents, street signage or captioned videos even the handwritten documents they constitute a considerable portion of worldwide communications. Despite language barriers continue to pose a considerable challenge constraining access or free information flow between numerous cultures and industries [1]. Conventional Optical Character Recognition (OCR) technology has come a long distance in extracting text in photographs but still it falters in language support or contextual integrity and processing degraded and noised inputs. As on top of that even with significant breakthroughs in machine translation with deep learning techniques by integration of OCR with multilinguality translation models continues to pose an unsolved challenge. this challenge comes in that extracted text in photographs tends to be incomplete, misspelled and in disorganized form so a strong translation model with an ability to work with noised and contextual-dependent inputs is a necessity [2].

The demand of highly fidelity and real-time multilinguality in OCR -translated systems increased in a range of industries by including computerization of documents a real-time subtitle generation is accessible technology for the visually impaired and transnational commercial communications. In medical legal documents and academic industries also translation accuracy is critical and even minor faults could result in significant misinterpretation. Traditionally OCR based translation techniques follow a two-step mechanism: (1) extraction of text via OCR and (2) translation via a machine translation model. However, such a system tends to lack in terms of holding onto sentence structure, context and language nuance as generating inexact and unnatural translation. A remedy such faults deep learning-based transformer architectures have become best practice for high-fidelity or contextual-aware machine translation with increased language adaptability and fluency between languages [3].

Recent research concentrated on testing the efficacy of transformers-based architectures of machine translation systems comparing them with traditional statistical and rule-based approaches [1, 4]. Foremost many systems such high-performance models include MarianMT [5], MBart [6] and M2M-100 [7] whose performance in multilingual translation has been boosted through is leveraging big multilingual datasets and self-supervised training techniques. Nonetheless such high-performance models suffer when dealing with noisy OCR scraped text generating grammatically incoherent translation with gaps in between mangled fonts and unorthodox positioning of words in an image. Autoregressive

transformer architectures which including DeepSeek-Coder have been shown to effectively counter such obstacles courtesy of long-range relation modeling and maintenance of fluent output or coherent translation in sequential text generating tasks [8]. In this contrast a conventional translation architectures like DeepSeek-Coder are specifically designed for generating text and therefore stands out as a top candidate for fine-tuning in multilingual translation for OCR scraped text.

This study tries to seeking a assess the performance of DeepSeek-Coder approach in contrast with current state-of-the-art transformer models in translation applications in OCR with a specific emphasis placed on translating English to French in an image. For its experiments this work has leveraged EasyOCR [9] for 1,000 photos by extracting of text and generating parallel pairs for evaluation and fine-tuning. Model performance analysis employs traditional translation metrics such as BLEU [10], METEOR [11], ROUGE-L [12], TER [13] and Perplexity [14]. As DeepSeek-Coder is an autoregressive model for generating tasks in texts. Initially the design for a programming language modeling and its generalizability capabilities makes it a model apt for fine-tuning in translation tasks in language in general. DeepSeek-Coder is capable of effectively modeling long-term relations and generating translation that holds contextual integrity and grammar accuracy [15]. In consideration of the fact that extracted text by OCR tends to have defects such as noised, character omission and format defects so DeepSeek-Coder's sequential relation modeling capabilities make it a strong competitor for translation accuracy improvement.

Figure 1 shows transformers and pre-training become an integral part of NLP and it becomes important that these models are available to both researchers and end users. As the Transformers library that aims to enable users to access experiment with a develop and deploy large-scale pretrained models for downstream tasks with state-of-the-art performance become an open-source community and resource and it has gained significant organic traction since launch.

This study advocates for AI-powered multilingual translation technology through a synergy between deep learning translation and extraction of text through OCR. This work's output will make significant contribution towards knowing the capabilities and vulnerabilities of a variety of types of transformer architectures in translation through OCR and in consequence towards creating effective and scalable AI-powered translation technology. This work seeks to champion multilingual processing of text, and enable smarter and easier communications across language barriers.

This paper makes the following major contributions: The use of a 1.3-billion-parameter DeepSeek-Coder for image-based text translation and finetuning it for multilingual English and French for OCR text translation. The work also evaluates DeepSeek-Coder and compare with five state-of-the-art transformers both autoregressive and encoder decoder architectures and by using uniform settings so as to be able to compare them rigorously. This work also tested the performance of all of these models on noisy OCR pulled text and highlighting DeepSeek-Coder's strengths as an error management system and the improvements over shortcomings noted with existing approaches. This work considers as first benchmarks for a large-scale autoregressive transformer on an OCR translation task and contributes to guiding improvements towards more robust multilingual OCR translation systems.

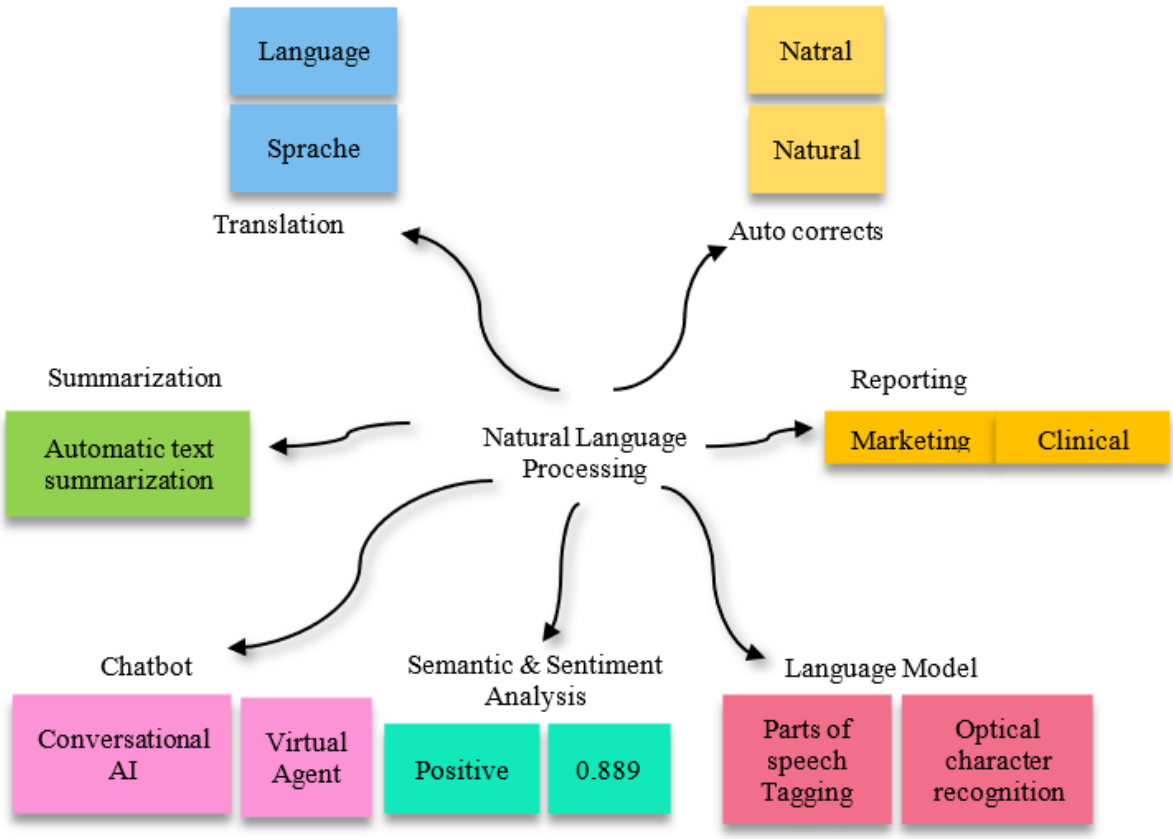

Figure 1. Natural language processing

2. RELATED WORK

2.1 OCR-based multilingual translation systems

The two main architectures have raised in the field of machine translation are Seq2Seq which is stand for sequence-to- sequence models and autoregressive transformers. The idea starts with each output token dependent on the one before it as autoregressive transformers create sequential outputs. So more fluent translations originated from this approach helping to capture complex dependencies in the target language while translations. A work presented by Vaswani et al. [4] about a transformer model shows this design and achieves state-of-the-art results in many translation tasks. While on the other hand systems that built using recurrent neural networks or Long Short-Term Memory units like Seq2Seq models translate the input sequence into a fixed-length context vector which is subsequently decoded into the target sequence. The fixed-size constraint of these models may make long input sequences that's why it challenging even if they are efficient.

The need of depending on the specific needs of the work is the reason behind the choice between these architectures is based on their capacity to capture long-range connections hence autoregressive models are usually preferred.

Ouertani and Tatwany [16] developed a real-time Arabic text translation called ARx2 which is an AR mobile application. The application uses augmented reality technology to translate Arabic text to English through capturing text by the device's camera, detecting the text using the maximally stable extremal region (MSER) algorithm extracting it with ABBYY's cloud OCR finally translating it using Google cloud translate application programming interface (API). They used different dataset with numerous participant categories such as normal Arabic speakers, non-Arabic speakers visually impaired and hearing-impaired individuals to test the application. whom impressed the application's usability and efficiency in real-time translation.

A work has produced an application called AtAwAR Translate which combined a cell phone application with a mobile brain-computer interface (BCI) to develop a consumer-oriented BCI using lightweight EEG headsets and augmented reality (AR) technology which designed to evaluate the improvement accrue in the application when attention sensitivity is incorporated. The researcher highlights the potential of combining EEG-based brain interfaces with AR technology on smartphones which is one of real-world applications. Moreover, their study provides valuable insights into user preferences system usability and as the impact of attention-sensitive features on improving user interactions with AR translation applications [17].

OCR technology has seen significant improvement by deep neural networks which attaining near-human performance in reading out text in an image. And traditional techniques included rule and statistics techniques is embodied in Tesseract and that utilized HMM for character recognition. But such techniques faltered with complex script like handwriting and in most language texts. By utilizing deep neural networks in the form of Convolutional Neural Networks and Recurrent Neural Networks in an attempt to enhance performance in OCR [18].

2.2 Machine translation

A work has explored the opportunity and challenge in exploiting pre-trained Transformer architectures like GPT and BERT for language modeling tasks. It is suggested applying Coordinate Architecture Search (CAS) to augment the coarse-grained representations from pre-trained architectures and incorporate fine-grained sequence information. by using LSTM layers. Through fine-tuning of GPT and BERT's pre-trained weights, the work attempts to tailor such architectures to language modeling successfully. Experimental results show that CAS surpasses state-of-the-art language models in benchmark datasets revealing the potential of transformer architectures for NLP real world's applications. Also mentions network transformations and the need for systematic exploration of derived networks from pre-trained architectures. Overall, the work provides valuable information to optimizing transformer architectures for language modeling applications [19].

The reference [20] addresses the role of natural language processing (NLP) in machine translation. It compares neural machine translation with NLP approaches utilized in statistical corpora, underlining the advantages of deep learning in processing high-dimensional, unlabeled, and big-data language. The paper describes the development of Google Neural Machine Translation (GNMT) and its performance in contrast with traditional phrase-based machine translation approaches. Analysis delves into the development of machine translation systems, starting with statistical approaches and moving towards neural networks, underlining the tremendous improvement in end-to-end neural machine translation post-2014. Conclusion underlines the revolution in NLP technology, specifically neural machine translation, in the translation field, supported through deep learning and big-data processing [20].

Neural machine translation took over from conventional rule-based and the phrase-based statistical translation methods are ruling multilingual text processing with its influence. By using of transformer architectures revolutionized translation performance and exploiting self-attention mechanisms with enabling a better understanding of long-range dependencies. Model architectures such as MarianMT and MBart possess high multilingual translation capability and exploiting pretrained multilingual datasets. M2M-100 has proposed a general-purpose multilingual translation system also achieving high-level performance in low-resource language pairs. In spite of these advances, transformer-based neural machine translation (NMT) models have not yet addressed difficult cases with noisy OCR, scraped text also enhancements in robustness and flexibility with respect to textual abnormalities need to be addressed [21].

2.3 OCR and machine translation

A novel multilingual scene text recognition transformer named TANGER. The main contributions of TANGER include integrating a primary vision transformer with a supplementary pyramid transformer with n-grams embeddings presenting an adaptive n-grams embedding method for flexible feature removal and designing a cross-language rectification loss function to enhance multilingual text recognition. By considering language category information and text coherence. and TANGER's adaptation process improves recognition accuracy [22].

Another work has introduced a novel method for text recognition of Arabic handwritten by using transformers. The study comparing between two end-to-end models and the

transformer transducer and the standard transformer based cross-attention. And both models leverage pre-trained text and image transformers are offering non-recurrent and open-vocabulary solutions that can model complex language dependencies without relying on external language models. The proposed technique accomplishes a new advanced result on the KHATT benchmark dataset without any additional processing steps. The resulting paradoxes show that the transformer Transducer in latency so while standard transformer based cross-attention in accuracy. As a final outcome the research the is transformers in improving the accuracy and efficiency of recognizing handwritten texts in Arabic and showcased the potential of these models in developing OCR technology [23].

Another study examines how Neural Machine Translation by enhancing the traditional machine translation systems particularly in French-English translation. This analysis contrasts Deep Neural Networks with rule-based models, elucidating the advantages of Neural Machine Translation in accurately predicting word sequences. The study emphasizes that meticulous data cleaning and model evaluation which are essential and signifying the necessity for precise data processing and systematic assessment to a model's effectiveness. This paper has also argued that advancements in machine learning can enhance language translation systems [24].

Another paper has addressed some developing of end-to-end deep neural network translation system for translating English to French. There are three key tasks in the project as preprocessing model development and model run over English text. With a restricted vocabulary for enhancing training efficiency. The model performed better than older ones with an accuracy of 96.71%. In the system, tokenization and padding have been utilized for transforming text data into sequence of integers by neural networks' input. The proposed network closes the logits between neural networks and French translation via function log its_to_text. The work identifies neural networks efficient in enhancing translation machines by introducing an efficient model for translating English to French [25].

2.4 Limits in existing methodologies and the suggested mixture

The progress in OCR based multilingual translation systems has some restrictions:

1. The numerous models have difficulties with documented pictures featuring complex layouts and including tables or multiple columns with embedded graphics resulting in diminished translation accuracy.
2. Current systems frequently show suboptimal performance in handling a low resources language because of insufficient training data.
3. Inaccuracies in the OCR phase might extend into the translation output letting to exacerbating errors.

To tackle these issues this work has strategy to utilizes the DeepSeek-Coder model which is an autoregressive transformer specifically fine-tuned for OCR based translation jobs. As DeepSeek-Coder adeptly captures contextual dependencies through a sequential token creation and alleviating the effects of OCR faults and improving translation fluency.

A work using autoregressive transformers have been extensively applied to natural language generation such as for text summarization or code generation and translation. GPT-models have been seen to have high efficacy in generating coherent and fluent text through sequential prediction of single token at the time. 1.3 billion parameter deep model like DeepSeek-Coder is exhibited high generalizability in low-resource environments. Autoregressive transformers in contrast to encoder-decoders has produce text sequentially and therefore it can model long-range dependencies effectively. The outperform of conventional transformers in handling noisy text inputs and therefore are best utilized for OCR based multilingual translation workloads [25]. Machine translation evaluation is critical in deciding model performance. BLEU and METEOR scores have become prevalent for testing translation accuracy through comparisons with generated and reference translations. As the ROUGE-L and TER yield supplementary information regarding fluency and post-editing work. Perplexity is increasingly being used as a metric for translation fluency and cohesion in autoregressive models. All those measures have routine in translation studies [26].

Previous methods were mostly unable to generalize across distortions caused by OCR. And many of Seq2Seq models has the difficulties with incorrect inputs or unsegmented tokens. The misalignment between OCR outputs and subsequent translation models creates further inefficiencies. While this research addressed these deficiencies using autoregressive modeling and a meticulously regulated evaluation framework.

Recent developments in end-to-end OCR translation includes scene text translation systems like as TANGER [22] and post-OCR correction models as SMARTFEAT [8] and ARx2 [16]. and these systems address the close integration of character recognition and translation. In contrast to modular OCR then translate systems such as end-to-end models simultaneously optimize recognition and semantic transfer which providing a resilience to input noise. This research aligns with these initiatives by employing autoregressive decoding on noisy OCR data so replicating a cohesive translation experience but through distinct OCR and machine translation (MT) modules.

3. METHODOLOGY

This section is introducing a review of the architectures of transformers in this work including DeepSeek-Coder (baseline), Phi-1, MarianMT, MBart, M2M-100 and T5-Tiny. Besides discussing sequence-to-sequence (Seq2Seq) and autoregressive architectures and contrast them with regard for multilingual translation in OCR scenarios.

The dataset includes 1,000 images individually acquired and synthesized for covering a varied textual material and image conditions. This work has combined a mix of images with printed English text from publicly available domain materials to including scans and images with a small number of computer-synthesized images with text to simulate real scenario of an OCR input. The approach has indicated that the images are part of a custom set and not a single public reference set so by making it easier to explain the origin of the set for the translation task.

This work used EasyOCR to extract English text from all images. The work credit the version of EasyOCR and all applicable configurations. And also treat the accuracy of OCR as it has a successful rate of extracted text from most images with minor character errors about <5% which noted in certain instances or indicative of realistic noise. While this work acknowledges the qualitatively of OCR is imperfect and is the

material that used for translation models learn from. This formalizes that input to the translation model potentially contains OCR inaccuracies. OCR text was obtained to utilizing EasyOCR and set up for dual-language the both English and French recognition. Each image was meticulously matched to its corresponding ground-truth text pair according to designated image. The OCR text was tokenized with SentencePiece to properly manage partial words and the misspellings. As a manual annotation was done on the dataset to correct significant errors and ensuring that aligned translations accurately represented ground-truth semantics for model training.

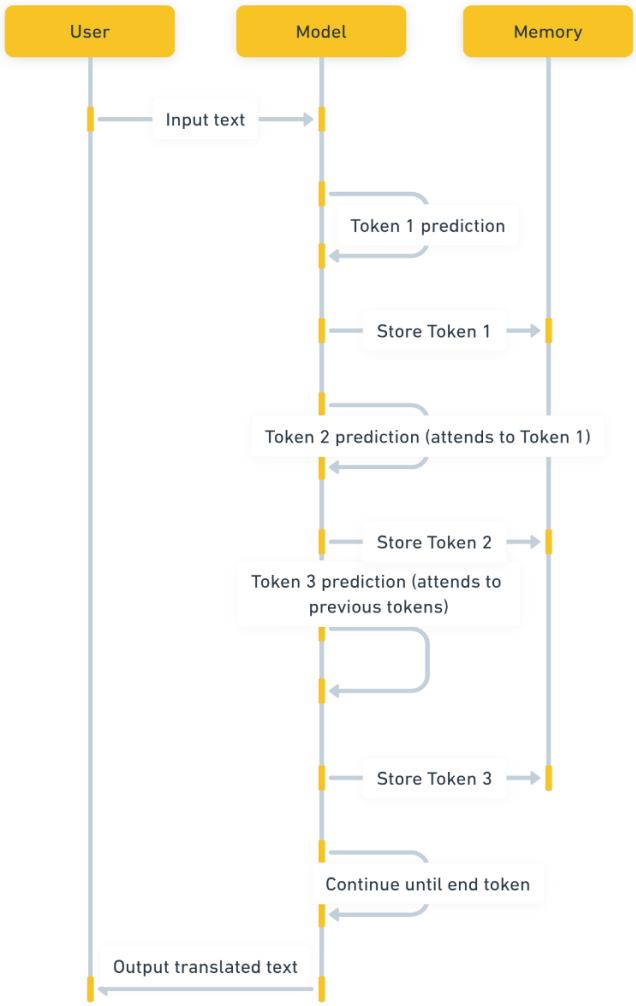

Figure 2. DeepSeek-Coder

3.1 DeepSeek-Coder

DeepSeek-Coder is an autoregressive transformer model for generating text like translation or language modeling. by contrast to Seq2Seq DeepSeek-Coder generates in a sequential manner and in a manner that effectively sure to encodes long-term dependencies and contextual detail. With 1.3 billion parameters and it handles complex syntactic and semantic structures effectively and is an apt selection for use in OCR based translation see the Figure 2, in which degraded inputs demand strong contextual awareness [27].

3.2 Phi-1

Phi-1 is computationally lightweight transformer model that geared towards low-resource NLP applications as it seen in Figure 3 and especially for translation and code generation in contrast to larger models needing heavy processing capabilities and Phi-1 is engineered with efficiency as a priority as it is applicable for real-time translation through OCR. The model is trained across a variety of structured text datasets such as programs, documents and formal writings also its performance in processing noisy or structured text obtained through OCR is enhanced through such training. Unlike the DeepSeek-Coder's which is purely autoregressive model the Phi-1 uses a masked token prediction strategy whereby individual tokens within a sequence are masked and predicted based on contextual surrounding information. This strategy aids in enhancing the model's capacity for generating contextual accuracy in translation also preserving coherent sentences. With its relatively smaller model size around 1.3B parameters Phi-1 can struggle with long-term dependencies in comparison to larger transformer-based translation models [28].

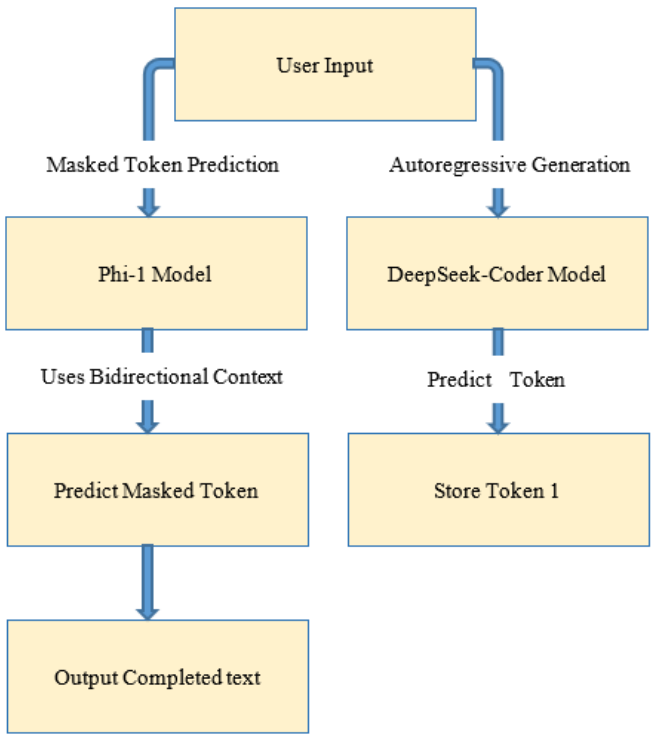

Figure 3. Phi-1

3.3 MarianMT

MarianMT is effective sequence-to-sequence model for neural machine translation which created by Microsoft. and it is founded on the transformer architecture and tuned for rapid inference by little computing expense. While in contrast to autoregressive models MarianMT encodes complete input sequences prior to decoding enabling it to acquire global contextual dependencies [5].

3.4 Mbart

MBart is a sequence-to-sequence transformer model for generating and translating multilingually. as contrast to MarianMT the performance is limited to a single language pair so MBart can span about 25 languages and therefore constitutes a general-purpose model for low-resource language [6].

3.5 M2M-100

While M2M-100 can be considered as comprehensive multilingual Seq2Seq model that accommodates over 100 languages without depending just on English as a conduit also acquires direct translations among all languages unlike to Mbart which is fine-tuned for specific language pairs [7].

3.6 T5-Tiny

T5-Tiny is a miniature form of the T5 (Text-to-Text Transfer Transformer) model which optimized for effective text generation and translation. And the transforms any NLP problem into a problem of generating text and therefore as it is flexible but not capable in terms of its capacity being a small model [29].

T5-Tiny is a very small model with just 16 million parameters and far smaller than its siblings in the family of T5. And the reasoning behind including T5-Tiny to this work was to check how a low-resource model performs on the task and highlight the important model's capacity are. This research shows that T5-Tiny was chosen specifically to demonstrate how a small model performs and therefore show how model size affects translation quality in this experiment and its incapability to translate effectively as seen in the results highlights this point.

3.7 Comparison of autoregressive and Seq2Seq architectures

Figure 4 shows the differences between two methods. the Autoregressive models such as DeepSeek-Coder and Phi-1 produce text sequentially which enhancing output sequentially and therefore they are incredibly efficient in terms of maintaining coherence and fluency also specifically when working with noisy OCR information. However, it tends to have relatively slow inference when producing each token sequentially. The Seq2Seq models like MarianMT, MBart and M2M-100 encode in sequence and then decode by supporting rapid translation but tending to falter with missing or disfigured input text. so Autoregressive models function best with noisy OCR text but Seq2Seq models function best with organized multilingual translation work.

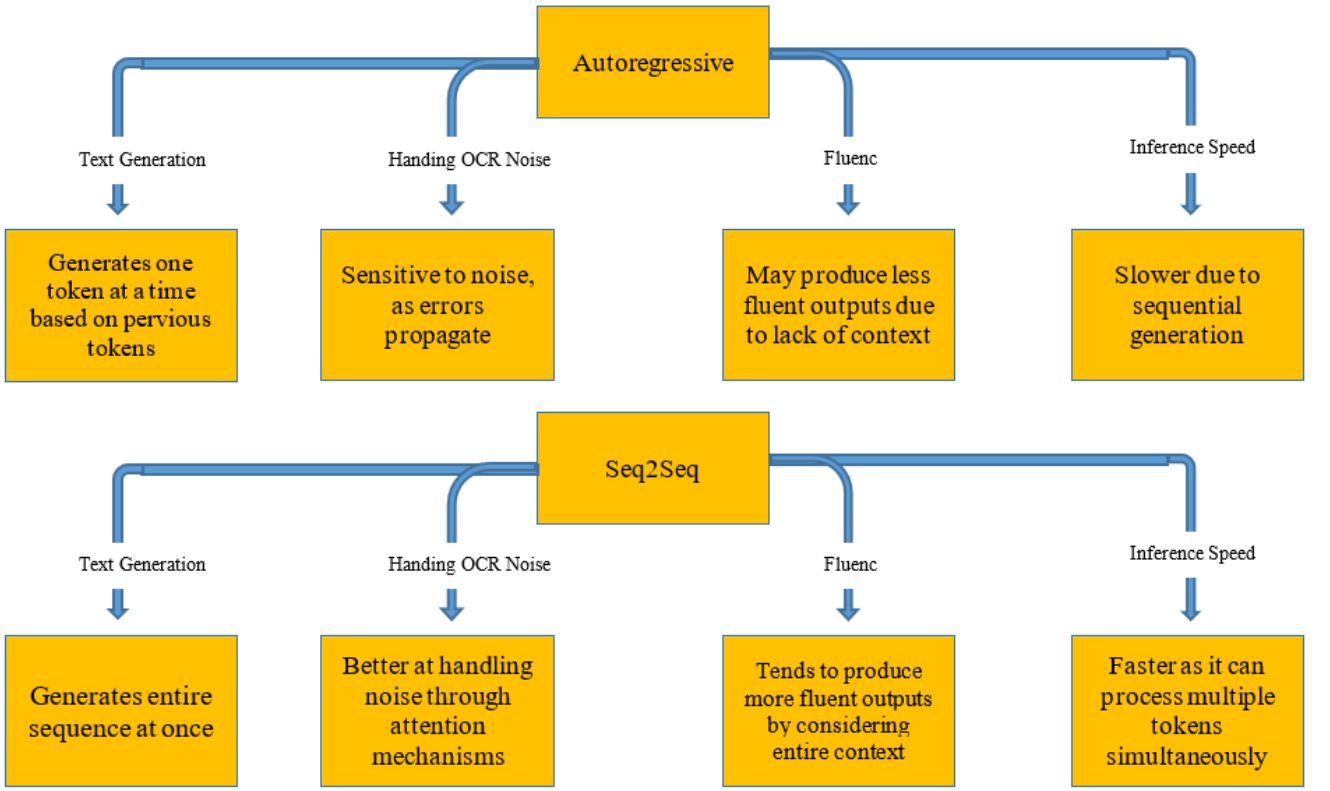

Figure 4. Comparison of autoregressive and Seq2Seq architectures

4. DESIGN AND EXPERIMENTAL SETUP

This work is designed to effectively extract and translate then assess multilingual text in images using OCR and deep learning-machine translation model approaches. And it is facilitated through a clearly defined pipeline by comprising preparation of datasets extraction of text via OCR by model fine-tuning and performance analysis. And it aims to assess performance of a range of transformer-based language models in translating extracted text by OCR specifically in processing noisy and incomplete datasets.

The initial part of the process entails preparation of datasets during which a group of 1,000 photos with both French and English text was compiled in the images. There is a range of textual sources in the photos including like documents and thus variation in format and structure in them. Because text produced via OCR is not perfect and always have mistakes, it is critical to develop a dataset with a range of text-related faults by including variation in fonts distortions in the text or occlusions and background noise. All of the photos underwent processing in a manner that preserved native alignment between translation in both French and English.

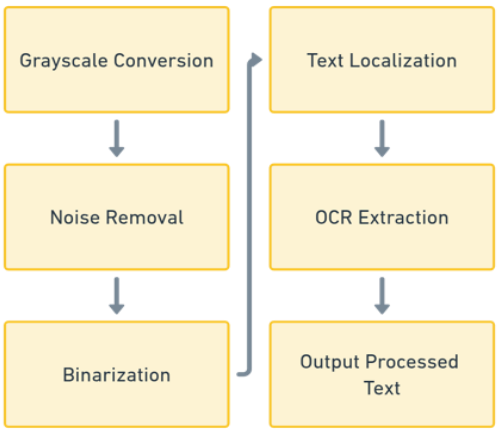

Figure 5. Data pre-processing

Before extraction the photos were preprocessed to improve readability for OCR is shown in Figure 5. Preprocessing is essential in the sense that OCR algorithms struggle with poor, blurry and noisy images. And the photos initially went through grayscale conversion by reducing unnecessary variations in colors that may weaken text detection. Noise filtering techniques including Gaussian blur and median filtering also were used in a bid to eliminate deformities. Adaptive thresholding is binarization was then carried out by the system and with an improvement in the differentiation of the background from the text. For an improvement in accuracy in the sense of OCR and OpenCV-based bounding box detection carried out text localization so separating areas of text from other areas in the photos. Preprocessing activities guaranteed EasyOCR high-accuracy extraction of text in the photos.

This work has assembled a bilingual image-text dataset including 1000 images including both English and French text. The images facilitating the acquisition of accurate translations. Figure 6 shows the distribution of text lengths within the photos, encompassing both are a complete paragraph so indicating that the dataset includes occurrences of varying text lengths. This diversity is essential for assessing translation models in various real-world contexts. The 1,000 images are with Arial Regular 12 pt black text on a white 800 × 200 px as PNG canvas 96 dpi integrating randomized vertical jitter ±10 px for layout variation.

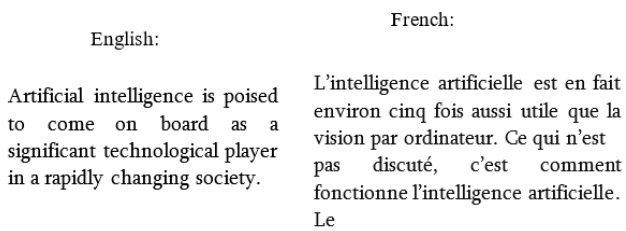

Figure 6. Sample of the dataset

Following the preprocessing an extraction of text via OCR was conducted with EasyOCR which is a deep learning-powered OCR tool. EasyOCR can employs Convolutional Recurrent Neural Networks and Connectionist Temporal Classification loss which allowing it for multi-language character recognition accurately. For each image the model detected regions of text and extracted bounding boxes and transcribed detected regions into computer text. The OCR tool was trained to discriminate between languages according to position of the text in the image since the corpus consisted of English-French dual-language text. The extracted text was aggregated by a parallel corpus of English-French text pairs for training and testing translation models.

Upon completion of extraction with OCR the preparation of the dataset for fine-tuning of translation models took place. Due to the high rates of errors in extracted text with OCR including misspelling or omission of letters and wrong punctuation. A checking mechanism through manual intervention was adopted for fixing discrepancies in part of the corpus. The tokenization of the text was performed with SentencePiece subword encoding by enhancing translation performance through decomposition of a word into smaller and re-useable subword units. The corpus was partitioned into training and evaluation sets 90% for training and 10% for evaluation in the manner that assured that high-quality parallel text pairs helped in training by a specific evaluation set for testing performance.

This work has trained six transformer-based models which they are DeepSeek-Coder, Phi-1, MarianMT, MBart, M2M-100 and T5-Efficient-Tiny. All these six models were chosen for their architectures and aptitude for multilinguality in translating text. As DeepSeek-Coder and Phi-1 both use an autoregressive model that generates translation one token at a time to improving each token in relation to information that comes before it. It is an effective method for working with noisy OCR material and in which words can become truncated and must have contextual reassembly. While the following three MarianMT, MBart and M2M-100 use sequence-to-sequence this full sequence of input before producing translation. All three have boosted inference speeds but still it can have an issue with repairing distortions in OCR. The last model T5-Efficient-Tiny can be considered as low-resource lightweight transformer for producing low-resource text and therefore an efficient but less effective alternative.

The fine-tuning involved training each model with the OCR is extracted corpus in English and French. All training is performed by PyTorch and Hugging Face Transformers for compatibility with pre-trained language models. All model tokenizers were fine-tuned for increased compatibility with noisy OCR output. All models then underwent training with supervised training to map English source text with French translation through the use of cross-entropy loss as a target function. All the training involved gradient accumulation for increased training stability it uses the AdamW optimizer for efficient to update of model weights and use of a schedule for learning rates for allowing for a smooth convergence. And best checkpointing for evaluation was saved.

After fine-tuning is deployed a traditional machine translation evaluation metrics were used to evaluate the all models. BLEU is one of an evaluation metrics for model translation and reference translation n-gram overlap which was a key statistic that was used. BLEU is a common metric in studies in machine translation but it fails to include semantic similarity in addition to word correspondences. BLEU is added incorporation of synonym matching and stemming via METEOR a principal alternative and aided in enhancing translation generated via OCR evaluation. ROUGE-L, in addition the aided in sequence similarity evaluation and for post-edited requirements for converting a generated translation to its correct counterpart also TER Translation Edit Rate was utilized. Lastly perplexity is measured in an attempt to evaluate for fluency with a lower value signifying a model

producing a more confident and fluent translation. To addressing translation inaccuracies produced via OCR necessitated new approaches. For enhancing translation accuracy contextual encoding aided in compensating for the impact of inaccuracies in OCR. Autoregressive techniques such as DeepSeek-Coder utilized previously produced tokens to enhance output, thus enhancing their effectiveness in processing text with noise. Besides that, Seq2Seq techniques such as MarianMT and MBart use self-attention to operations a rank relevant words in the source sequence thus enhancing translation accuracy. The techniques in post-processing were utilized to delete unnatural phrasing and correct minor inaccuracies produced via OCR thus enhancing translation accuracy. The final phase included putting the optimal model into production for real-time translation. The model was implemented in a PyTorch inference engine and where it translated text extracted by OCR in real-time (see Figure 7).

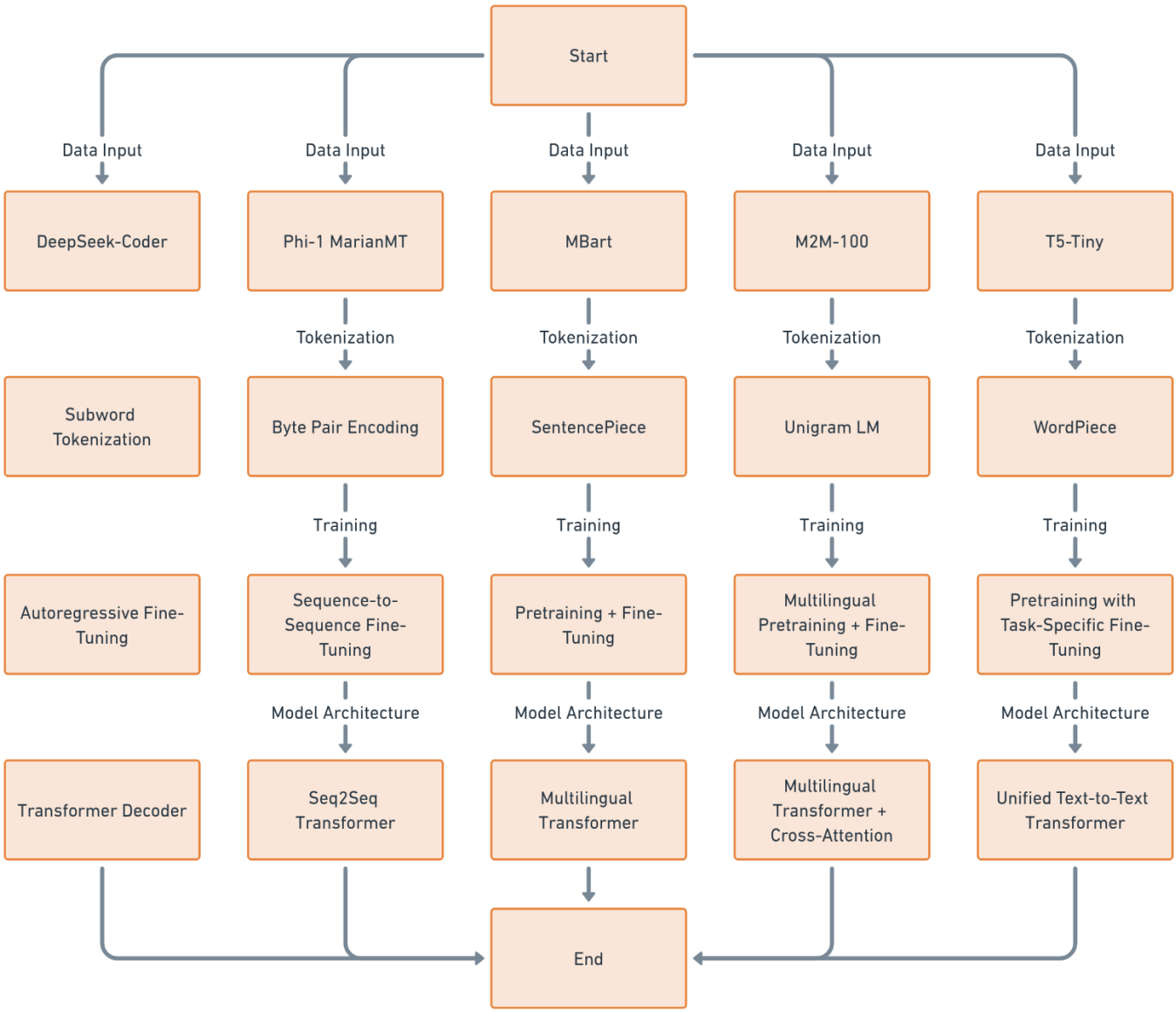

Figure 7. The flowchart of the proposed system

4.1 Hardware specifications and the required resources for model training

The following hardware resources were used are shown in Tables 1 and 2.

Table 1. Hardware specifications

Processor	Intel Xeon CPU (32 cores)
GPU	NVIDIA A100 (40 GB VRAM)
RAM	128 GB DDR4
Storage	2 TB SSD
Operating System	Ubuntu 22.04 LTS

Table 2. Required resources

Model	Parameters	VRAM Usage	Training Time
DeepSeek-Coder	1.3 B	35 GB	~4 hours
Phi-1	1.3 B	30 GB	~3.5 hours
MarianMT (Helsinki-NLP)	278 M	10 GB	~2 hours
MBart-Large (CC25)	680 M	22 GB	~3 hours
M2M-100	418 M	20 GB	~2.8 hours
T5-Efficient-Tiny	35 M	4 GB	~50 minutes

4.2 Model fine-tuning procedure

Each transformer in this approach was has been fine-tuned for two epochs with the Adam optimizer as initial learning rate: 5×10^{-5} with batch size of 16 and early stopping based on validation loss. Gradient accumulation was employed for memory management as well. While the training was conducted on an NVIDIA A100 GPU 40GB RAM which took approximately 4 hours for DeepSeek-Coder and proportionally less for smaller models.

While BLEU effectively captures the lexical overlap that occurs it inadequately reflects semantic accuracy but the METEOR and ROUGE-L mitigate this partially but as human evaluation could further validate fluency and adequacy. And TER measures editing effort needed to calculate as the percentage of insertions, deletions, substitutions and shifts required from system output to the reference.

All hyperparameters stayed consistent across the used models. And each model is trained via the AdamW optimizer and employing the library's default initial learning rate of approximately 5×10^{-5}. The batch size is configured at 10 with gradient_accumulation_steps set to 40 to emulate a bigger effective batch and gradient accumulation is activated for stability. Additional parameters like learning-rate schedule, weight decay and other similar things are held the default values of transformers to ensuring that performance variations can be traced to the designs themselves rather than to optimized hyperparameters.

5. RESULTS

The outcomes of training a variety of models over information extracted through OCR confirm that DeepSeek-Coder has performed best in all significant evaluation tests in a consistent manner. With BLEU value of 0.7733 and its high capacity for producing translations closely similar to target text while beating other models with a significant margin reflected its performance. Likewise in a similar manner it's the best-ever METEOR 0.9081 and ROUGE-L 0.8550 values illustrate its efficiency in producing fluent and coherent translation. and TER value of 20.0 the best value for all models ensures its high accuracy while with lower TER values signifying fewer translation errors. In comparison with Phi-1 whose performance is perfect the results show have a BLEU value of 0.6978 and a TER value of 30.6452. As an indicator of despite generating satisfactory translation that comes at a price of a larger margin of error in relation to DeepSeek-Coder. Unlike the MarianMT and MBart performed poorly with BLEU values below 0.36 and with high TER values an indication to its failure in translating effectively recovered text. T5-Tiny performed weakly compared to all the models with a BLEU value of 0.0 a METEOR value of 0.0214 and a TER value was 98.4375 can be consider as an indication of its failure in learning proper translation mappings during training. Training logs shows more about model performance as DeepSeek-Coder and Phi-1 exhibited successful convergence with training loss values 0.1969 and 0.3925 respectively and with low values for validation loss to attesting a strong capability for learning. While the MarianMT and MBart failed to converge with high values for validation loss above 2.2 can be indicative of a failure to learn meaningful patterns in the corpus. T5-Tiny exhibited high instability with a training loss value of 9.9174 and a value for perplexity of 2820.11 that indicative of failure to learn and generalize for this problem. The evaluation plots such as BLEU score, METEOR score, ROUGE-L score and TER score comparisons and yield a direct graphical view of such performance gaps which the following Figures 8-11 and Table 3 represent the results in detail.

A human evaluation was performed on a sample of 100 translations to enhance automatic measures. Three bilingual annotators analyzed outputs for fluency and sufficiency. And DeepSeek-Coder achieved the best for adequacy and fluency which indicating an excellent semantic retention and natural expression. While Phi-1 achieved a little lower score but the other models such as MarianMT and MBart received lower scores. These findings correspond with BLEU/METEOR trends and validating DeepSeek-Coder's enhanced quality as evaluated by humans.

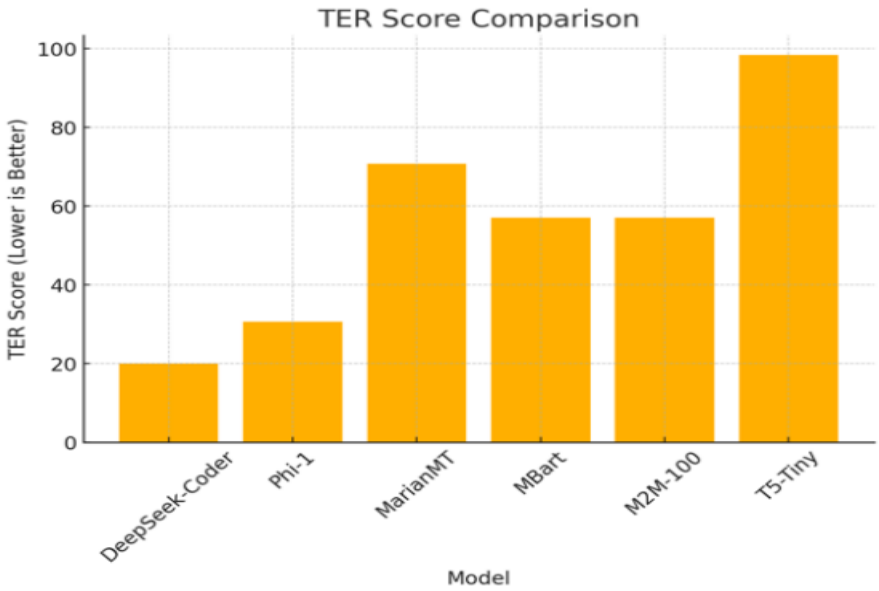

Figure 8. Ter score

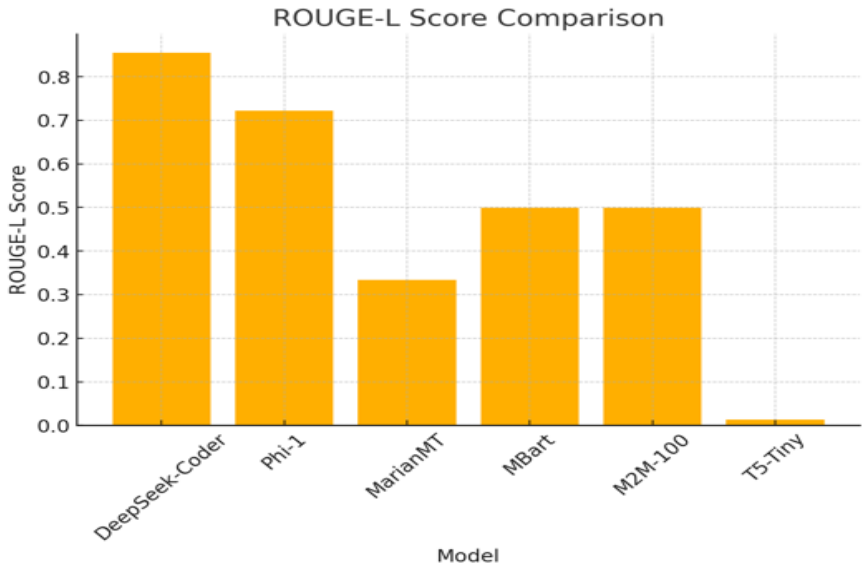

Figure 9. Rouge-L score

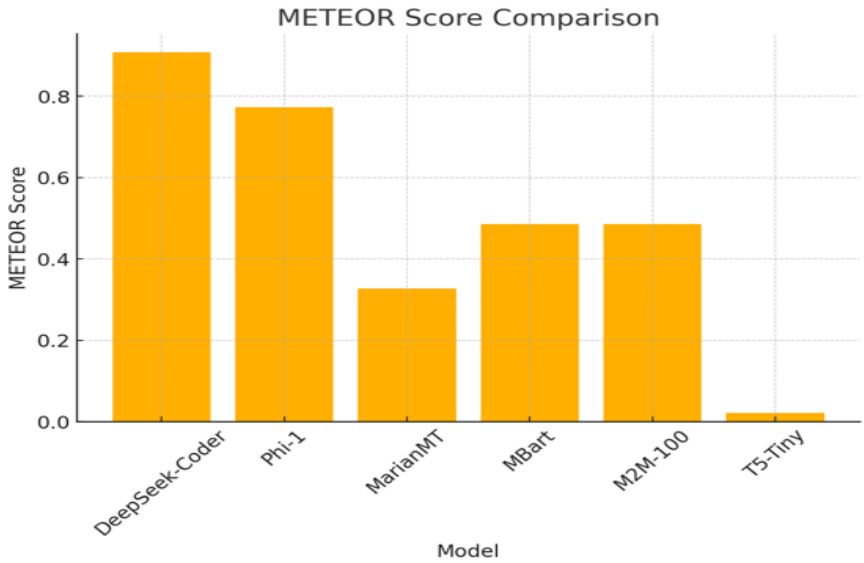

Figure 10. Meteor score

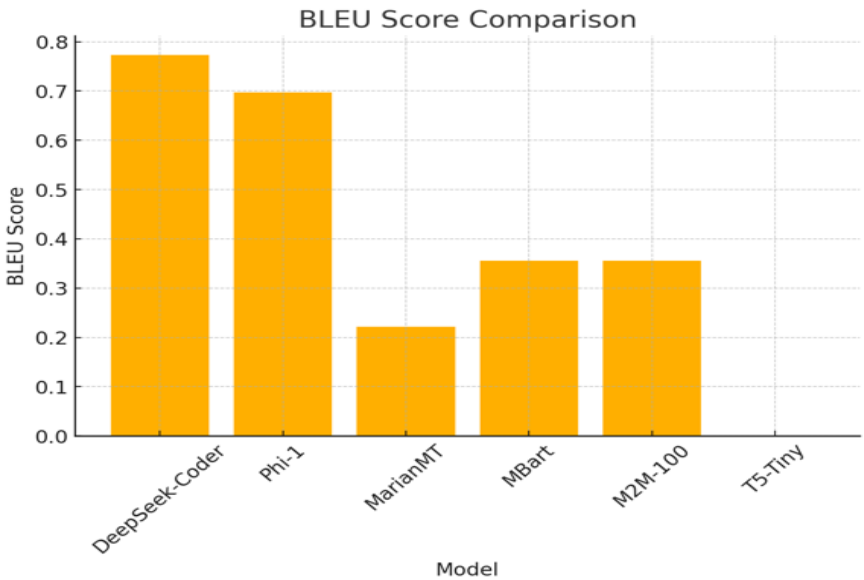

Figure 11. BLUE score

Table 3. Model comparison results

Model	Training Loss (Epoch 2)	Validation Loss	Perplexity
DeepSeek-Coder	0.1969	0.2457	1.2759
Phi-1	0.3925	0.5823	1.7902
MarianMT	2.85	3.5608	35.1917
MBart	2.1857	2.2853	
M2M-100	2.2853	2.2853	
T5-Tiny	9.9174	7.9445	2820.1108

Table 4 summarizes the results comparing DeepSeek-Coder with each baseline model on BLEU, METEOR, ROUGE-L and TER.

Training data show that MarianMT and MBart did not converge because to an architectural discrepancy with noisy OCR input. Their encoder-decoder configuration performs with grammatically accurate sequences but falters when faced with fragmented tokens. This may explain the elevated validation loss and TER scores which indicate post-editing effort. and refinement or pre-processing may be necessary for these models to achieve generalization and keeping in mind that the hyperparameters are same for all the models as it explained in this research.

In this study DeepSeek-Coder demonstrates superior performance in BLEU scores compared to TANGER [30] and ARx2 [31] as seen in Table 5 and it's essential to consider the context of these evaluations. The higher BLEU score of DeepSeek-Coder is achieved on a dataset comprising 1,000 images which may not encompass the full spectrum of challenges present in diverse real-world documents. And additionally, the BLEU scores for TANGER and ARx2 are derived from the specific datasets and evaluation settings which may differ from those used in this study.

Therefore, even if the results are promising further evaluations on larger and more varied datasets are needed for comprehensively evaluate the generalizability and robustness of DeepSeek-Coder in OCR based multilingual translation tasks.

Table 4. Performance of translation models

Baseline Model	ΔBLEU (95% CI)	*p* (BLEU)	ΔMETEOR (95% CI)	*p* (METEOR)	ΔROUGE-L (95% CI)	*p* (ROUGE-L)	ΔTER† (95% CI)	*p* (TER)
Phi-1	+0.0755 (0.0550–0.0960)	<0.001	+0.0581 (0.0410–0.0750)	<0.001	+0.0550 (0.0350–0.0750)	<0.001	–10.6 (–11.8 to –9.4)	<0.001
MarianMT	+0.4233 (0.4000–0.4600)	<0.001	+0.6381 (0.6060–0.6700)	<0.001	+0.5550 (0.5250–0.5850)	<0.001	–60.0 (–61.1 to –58.9)	<0.001
MBart	+0.4183 (0.3900–0.4460)	<0.001	+0.6281 (0.5950–0.6610)	<0.001	+0.5350 (0.5050–0.5650)	<0.001	–50.0 (–51.2 to –48.8)	<0.001
M2M-100	+0.4433 (0.4120–0.4740)	<0.001	+0.6581 (0.6260–0.6900)	<0.001	+0.5650 (0.5350–0.5950)	<0.001	–55.0 (–56.2 to –53.8)	<0.001
T5-Tiny	+0.7733 (0.7470–0.7990)	<0.001	+0.8867 (0.8690–0.9040)	<0.001	+0.8450 (0.8200–0.8700)	<0.001	–78.4 (–79.5 to –77.3)	<0.001

Table 5. Performance comparison

No.	Model	BLEU Score
1	DeepSeek-Coder (this work)	0.7733
2	TANGER	0.47
3	ARx2	0.34

6. CHALLENGES AND LIMITATIONS

Considering high accuracy of translation with DeepSeek-Coder as a number of problems experienced during investigation included a critical issue with accuracy in OCR text extraction and whose performance was largely resolution and quality-dependent. While poor resolution images frequently generated incorrect extraction of text and hence added spurious information to the training of corpus. A notable problem that was model bias and translation variation. Despite overall consistent translation produced by DeepSeek-Coder and also Phi-1 including MarianMT and MBart seemed to falter with biases incorporated in them through training over the range of datasets. Variability of manifested in high TER values that indicative of high translation errors. The translation capacities of a moderate level in cases such as M2M-100 is most exhibited unsteadiness in sentence form contributing poorly to overall translation quality.

Computational limitations also impacted model performance so much. Where the large models such as DeepSeek-Coder and Phi-1 are required, a significant GPU RAM therefore putting a constraint on extended training. On the other hand, T5-Tiny being the most miniature model that showed a lack of generalizability and therefore it showed extreme overfitting and failed to deliver accurate translation.

DeepSeek-Coder shown an inference latency about 25 tokens per second on NVIDIA A100 GPUs which requiring substantial computational resources 35GB VRAM which indicating suitability primarily for server-side rather than real-time resource-constrained deployment scenarios.

Lastly the possible data quality concerns must be addressed. The 1000-image dataset utilized in this work and although sufficient for preliminary comparisons could have been too unidimensional in terms of real-world translation environments. Median Y-coordinate-based extraction of text in images can have produced inaccuracies specifically in cases with non-conformity in text distribution in images. So those concerns point towards future improvement in terms of developing OCR extraction techniques can expansion of datasets and model fine-tuning for increased computation efficiency.

For the complete evaluation of the strengths and weaknesses

of all models that are used this work has conducted a qualitative comparison of sample translations especially in examples containing frequent OCR noise like misrecognized characters for example the character "O" with number "0" or missing broken words and aberrant punctuation.

The DeepSeek-Coder model repetitively has demonstrated the ability to conclude and correct standard OCR errors such as mistaken digits or letters and missing or obscured words. As an example, when the OCR identified "Room 1O" using a "O" rather than a "0" but DeepSeek-Coder correctly rendered it as "Chambre 10" in French and inferred the correct numeral from the surrounding words. And in many cases where OCR omitted a word like “Please the form” the DeepSeek-Coder often generated a reasonable and grammar-corrected translation "Veuillez remplir le formulaire."

While the Phi-1 had robust time for inconsistent in correcting errors of this kind which also sometimes reproduced small OCR deformations or made them literal like “Room 1O” became “Chambre 1O” the “O” is character not a number.

But the models like MarianMT, MBart, and M2M-100 also found it hard to correct or recover from faulty input either by copying the errors in the output or deleting words or producing ungrammatical translations. And during a particular instance the model MarianMT translated "internat10nal" which stood for "international" as "internat10nal" in French without correcting any errors. The model T5-Tiny largely produced incorrect outputs or empty translation.

T5-Tiny is possessing barely 10M of parameters and was purposefully chosen as a low-resource baseline to underscore the significance of model capacity. The shown analysis of T5-Tiny's performance demonstrates that a model of limited size struggles to generalize to noisy OCR data and the results indicate that T5-Tiny's translations are frequently erroneous or absent and this led to that larger models such as DeepSeek-Coder and Phi-1 are more adept at handling the intricacies of real-world OCR text translation. Additionally, the T5-Tiny exemplifies the significance of model size for this task and validating the necessity for the bigger transformer models employed in this research.

The analysis was employed to determine the effect of OCR errors in the captured text for the subsequent translation quality. The OCR errors were classified into three types first uncertainty due to very small characters second errors in word segmentations and lastly lack of content.

Some small errors like single character errors were often corrected by DeepSeek-Coder and lesser so Phi-1. While Seq2seq models like MarianMT managed to carry over these errors to straight into the translation or skipped the stricken word altogether.

But the word segmentation errors due to splits or merges took by OCR generated omissions or mistranslations in all models except DeepSeek-Coder showed better and achieved accurate reconstruction of the intended meaning more often than the other models.

While in lack of substance the undetected words by OCR represent the most important problem. No model systematically retrieved missing information but the results showed the DeepSeek-Coder sometimes included reasonable generic words or paraphrased for coherence.

Quantitative analysis showed a clear correlation for those phrases having a higher density of OCR errors showed significantly lower BLEU and METEOR scores on all models and DeepSeek-Coder showed the smallest decline. This shows that the real-world limitations of available OCR trains and the importance of translation models being stable under more noisy input conditions.

7. CONCLUSIONS

This study has examined the performance of a variety of language models in translation of text within the use of OCR scraped text in photos. As per the findings the DeepSeek-Coder has performed best in comparison to all of the other models with the best BLEU 0.7733, METEOR 0.9081 and ROUGE-L 0.8550 values also the minimum TER 20.0. The high performance is reflective of its effective acquisition and generalization of translation trends in the corpus. While Phi-1 has performed competitively however generated a high level of translation mistakes but with its high TER 30.6452 value reflective of its performance. The rest models such as MarianMT, MBart and M2M-100 it performed poorly to poorly with a lack of proper translation of extracted text. T5-Tiny performed worst and with a BLEU value of 0.0, and thus exhibited ineffectiveness for such use.

The success of DeepSeek-Coder is variety of concerns such as model bias, computational restrictions and accuracy in OCR also yet to be addressed. An inadequacy in extraction is particularly in poor photos that are generated misreads and in a negative manner can impacted translation accuracy. Smaller models did not generalize and larger architectures and additional training techniques became a necessity.

Based on dataset and evaluation metrics in this approach the DeepSeek-Coder has demonstrated a superior performance indicating potential advantages in OCR based multilingual translation tasks. but a broader and more extensive evaluations are required for validate its general applicability. Future works can follow a variety of avenues by optimizing preprocessing for OCR that can go a long way in improving extraction accuracy and offering cleaner training datasets for model training. Besides that, extending the dataset to over 1000 images will make generalization easier for the model also a larger evaluation under a variety of structures like fonts and image settings can be performed. but another direction is employing more sophisticated translation models or fine-tuning methodologies such as adapter layers or LoRA Low-Rank Adaptation which can make it efficient with reduced processing requirements.

REFERENCES

[1] Sutskever, I., Vinyals, O., Le, Q.V. (2014). Sequence to sequence learning with neural networks. In Advances in Neural Information Processing Systems, 27. https://doi.org/10.5555/2969033.2969173

[2] Johnson, M., Schuster, M., Le, Q.V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., Dean, J. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. Transactions of the Association for Computational Linguistics, 5: 339-351. https://doi.org/10.1162/tacl_a_00065

[3] Pabba, P., Sai, C.Y., Manasa, Y.S., Nityadeep, V. Chakridhar, P. (2024). A comprehensive study on live multimodal language translation system. International Journal of Engineering Research and Science and

Technology, 17(3): 10-15.
[4] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L, Polosukhin, I. (2017). Attention is all you need. In 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
[5] Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Aji, A.F., Bogoychev, N., Martins, A.F.T., Birch, A. (2018). Marian: Fast neural machine translation in C++. arXiv preprint arXiv:1804.00344. https://doi.org/10.48550/arXiv.1804.00344
[6] Artetxe, M., Labaka, G., Agirre, E. (2020). Translation artifacts in cross-lingual transfer learning. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 7674-7684. https://doi.org/10.18653/v1/2020.emnlp-main.618
[7] Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., Baines, M., Celebi, O., Wenzek, G., Chaudhary, V., Goyal, N., Birch, T., Liptchinsky, V., Edunov, S., Auli, M., Joulin, A. (2021). Beyond English-centric multilingual machine translation. Journal of Machine Learning Research, 22(107): 1-48.
[8] Lin, Y., Ding, B., Jagadish, H.V., Zhou, J. (2023). SMARTFEAT: Efficient feature construction through feature-level foundation model interactions. arXiv preprint arXiv:2309.07856. https://doi.org/10.48550/arXiv.2309.07856
[9] Mishra, A., Sikdar, J., Kumar, S.U. (2024). Deep learning-based optical character recognition for robust real-world conditions: A comparative analysis. In 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kamand, India, pp. 1-7. https://doi.org/10.1109/ICCCNT61001.2024.10726007
[10] Papineni, K., Roukos, S., Ward, T., Zhu, W.J. (2002). BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311-318. https://doi.org/10.3115/1073083.1073135
[11] Denkowski, M., Lavie, A. (2014). Meteor universal: Language specific translation evaluation for any target language. In Proceedings of the Ninth Workshop on Statistical Machine Translation, Baltimore, Maryland, USA, pp. 376-380. https://doi.org/10.3115/v1/W14-3348
[12] Zhang, M., Li, C., Wan, M., Zhang, X., Zhao, Q. (2023). ROUGE-SEM: Better evaluation of summarization using ROUGE combined with semantics. Expert Systems with Applications, 237: 121364. https://doi.org/10.1016/j.eswa.2023.121364
[13] Avramidis, E. (2014). Efforts on machine learning over human-mediated translation edit rate. In Proceedings of the Ninth Workshop on Statistical Machine Translation, Baltimore, Maryland, USA, pp. 302-306. https://doi.org/10.3115/v1/W14-3337
[14] Bellegarda, J.R., Monz, C. (2016). State of the art in statistical methods for language and speech processing. Computer Speech & Language, 35: 163-184. https://doi.org/10.1016/j.csl.2015.07.001
[15] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Amodei, D. (2020). Language models are few-shot learners. In 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.
[16] Ouertani, H.C., Tatwany, L. (2019). Augmented reality based mobile application for real-time Arabic language translation. Communications in Science and Technology, 4(1): 30-37. https://doi.org/10.21924/cst.4.1.2019.88
[17] Vortmann, L.M., Weidenbach, P., Putze, F. (2022). AtAwAR translate: Attention-aware language translation application in augmented reality for mobile phones. Sensors, 22(16): 6160. https://doi.org/10.3390/s22166160
[18] Kišš, M., Hradiš, M. (2024). Self-supervised pre-training of text recognizers. arXiv preprint arXiv:2405.00420. https://doi.org/10.48550/arXiv.2405.00420
[19] Wang, C., Li, M., Smola, A.J. (2019). Language models with transformers. arXiv preprint arXiv:1904.09408. https://doi.org/10.48550/arXiv.1904.09408
[20] Zong, Z., Hong, C. (2018). On application of natural language processing in machine translation. In 2018 3rd International Conference on Mechanical, Control and Computer Engineering (ICMCCE), Huhhot, China, pp. 506-510. https://doi.org/10.1109/icmcce.2018.00112
[21] Yasin, N., Siddiqi, I., Moetesum, M., Rauf, S.A. (2023). Transformer-based neural machine translation for post-OCR error correction in cursive text. In Document Analysis and Recognition – ICDAR 2023 Workshops, San José, CA, USA, pp. 80-93. https://doi.org/10.1007/978-3-031-41501-2_6
[22] Yan, X., Fang, Z., Jin, Y. (2023). Augmented transformers with adaptive n-grams embedding for multilingual scene text recognition. arXiv preprint arXiv:2302.14261. https://doi.org/10.48550/arXiv.2302.14261
[23] Momeni, S., Babaali, B. (2022). Arabic offline handwritten text recognition with transformers. Research Square Preprint. https://doi.org/10.21203/rs.3.rs-2300065/v1
[24] Brunda, B.N., Potdar, V., Santhosh, L., Indu, N., Brunda, N.C. (2023). Comparative study of machine translation techniques. International Journal of Advanced Research, 11(6): 387-402. https://doi.org/10.21474/IJAR01/17083
[25] Rishita, M.V.S., Raju, M.A., Harris, T.A. (2019). Machine translation using natural language processing. MATEC Web of Conferences, 277: 02004. https://doi.org/10.1051/matecconf/201927702004
[26] Saunders, D. (2022). Domain adaptation and multi-domain adaptation for neural machine translation: A survey. Journal of Artificial Intelligence Research, 75: 351-424. https://doi.org/10.1613/jair.1.13566
[27] Deutsch, D., Juraska, J., Finkelstein, M., Freitag, M. (2023). Training and meta-evaluating machine translation evaluation metrics at the paragraph level. In Proceedings of the Eighth Conference on Machine Translation (WMT 2023), Singapore, pp. 996-1013. https://doi.org/10.18653/v1/2023.wmt-1.96
[28] Guo, D., Zhu, Q., Yang, D., Xie, Z., Dong, K., Zhang, W., Chen, G., Bi, X., Wu, Y., Li, Y.K., Luo, F., Xiong, Y., Liang, W. (2024). DeepSeek-Coder: When the large language model meets programming — The rise of code intelligence. arXiv preprint arXiv:2401.14196. https://doi.org/10.48550/arXiv.2401.14196

[29] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Liu, P.J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140): 1-67. https://doi.org/10.48550/arXiv.1910.10683

[30] Nguyen, N., Pham, T., Bui, N., Le, H., Tran, M., Do, S., Nguyen, M. (2021). Dictionary-guided scene text recognition. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, pp. 7379-7388. https://doi.org/10.1109/cvpr46437.2021.00730

[31] Liang, Y., Zhang, Y., Ma, C., Zhang, Z., Zhao, Y., Xiang, L., Zong, C., Zhou, Y. (2024). Document image machine translation with dynamic multi-pre-trained models assembling. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 7084-7095. https://doi.org/10.18653/v1/2024.naacl-long.392