CAAI Transactions on Intelligence Technology

WILEY

The Institution of Engineering and Technology

ORIGINAL RESEARCH OPEN ACCESS

Large Language Models With Contrastive Decoding Algorithm for Hallucination Mitigation in Low-Resource Languages

Zan Hongying[1] | Arifa Javed[1] | Muhammad Abdullah[1] | Javed Rashid[2] | Muhammad Faheem[3,4]

[1]School of Computing and Artificial Intelligence, Zhengzhou University, Zhengzhou, China | [2]Information Technology Services, University of Okara, Okara, Pakistan | [3]School of Technology and Innovations, University of Vaasa, Vaasa, Finland | [4]VTT Technical Research Center of Finland, Espoo, Finland

Correspondence: Muhammad Faheem (muhammad.faheem@uwasa.fi)

Received: 11 June 2024 | **Revised:** 11 October 2024 | **Accepted:** 24 October 2024

Funding: The authors are highly grateful to their affiliated universities and institutes for providing research facilities. The research work of M. Faheem is supported by VTT Technical Research Center of Finland.

Keywords: artificial intelligence | artificial neural network | computer vision | deep learning | deep neural networks | large language model

ABSTRACT

Neural machine translation (NMT) has advanced with deep learning and large-scale multilingual models, yet translating low-resource languages often lacks sufficient training data and leads to hallucinations. This often results in translated content that diverges significantly from the source text. This research proposes a refined Contrastive Decoding (CD) algorithm that dynamically adjusts weights of log probabilities from strong expert and weak amateur models to mitigate hallucinations in low-resource NMT and improve translation quality. Advanced large language NMT models, including ChatGLM and LLaMA, are fine-tuned and implemented for their superior contextual understanding and cross-lingual capabilities. The refined CD algorithm evaluates multiple candidate translations using BLEU score, semantic similarity, and Named Entity Recognition accuracy. Extensive experimental results show substantial improvements in translation quality and a significant reduction in hallucination rates. Fine-tuned models achieve higher evaluation metrics compared to baseline models and state-of-the-art models. An ablation study confirms the contributions of each methodological component and highlights the effectiveness of the refined CD algorithm and advanced models in mitigating hallucinations. Notably, the refined methodology increased the BLEU score by approximately 30% compared to baseline models.

1 | Introduction

Neural machine translation has significantly advanced due to deep learning and the development of pre-train models [1]. Recent advancements in large-scale multilingual machine translation have significantly advanced the goal of achieving a universal translation system. These sophisticated pre-trained models can manage a vast array of languages and translation directions as highlighted by Fan et al. [2]. At the same time, general-purpose large language models (LLMs) have demonstrated exceptional versatility and excelling in new tasks such as translation, where they are continually improving, as noted by Chowdhery et al. [3]. Unlike traditional bilingual models, these advanced systems offer substantial performance enhancements and streamline engineering processes by enabling a single model to handle all language pairs [4]. However, translating between low-resource languages is still difficult. These languages often lack enough training data and resources such as text corpora, annotated datasets, and linguistic tools [5]. The scarcity of training data results in poor

translation quality and makes models prone to generating content that appears plausible but is factually incorrect. Inconsistent or domain-specific data scarcity exacerbates these issues, causing models to produce inaccurate or misleading outputs.

Hallucinations in NMT happen when the translated sentence contains content not present in the original sentence, which leads to misleading or incorrect translations [6]. In other words, when there is a low contribution of the source sentence to the generation of the target sentence. It significantly undermines the trust in NMT systems [7]. Hallucinations in translation primarily stem from three main issues: insufficient context understanding, training data limitations, and overfitting. When models lack necessary contextual cues, they often produce translations that misrepresent the original source material. When models are trained on limited or noisy datasets, they are more prone to generating outputs that contain inaccuracies. Models that are overfitted to their training data may struggle to generalise effectively to new or varied inputs, leading to disconnected translations. Hallucination by LLM [8] represents an alarming barrier to the effective deployment and equitable impact of artificial intelligence technologies across globally diverse linguistic and cultural landscapes [9]. Recent studies have improved the understanding, detection, and mitigation of these pathological translations [10]. However, these studies have typically focused on high-resource languages or small bilingual models with less than 100 million parameters [11]. These models were often trained on a single English-centric, high-resource language pair [12, 13]. This problem is more common in low-resource language pairs like Chinese and Urdu because they face data scarcity. Previous work on mitigating hallucinations in low-resource NMT has largely focused on sampling translations and reranking them based on quality metrics [14]. Contrastive decoding is introduced to address issues like excessive repetition and low diversity in unconditional language models [15]. Other techniques such as data augmentation and noise-robust training have been introduced, but they have challenges such as overfitting or increased computational complexity.

This research aims to develop an NMT model for low-resource language pairs, Chinese and Urdu. A refined Contrastive Decoding algorithm is introduced along LLM. ChatGLM [16], a conversational language model, is adapted for translation tasks that enhance the translation's contextual understanding. LLaMA2 [17], a large multilingual model, is optimised for diverse language pairs to improve cross-lingual transfer and robustness. The CD Algorithm [15] is refined by dynamically adjusting the weight given to the amateur model based on the expert model's confidence. It generates multiple candidate solutions and selects the best one by using evaluation metrics. This ensures the final translation is accurate and relevant by reducing the hallucinations ratio. The major contributions are as follows:

- Utilisation of dynamic Contrastive Decoding algorithm that dynamically adjusts the weights of each translation segment generated by an amateur model based on the confidence scores of the expert model to reduce hallucinations and improve translation quality by controlling overfitting.
- Utilisation of large language models such as ChatGLM 2–6B [16] and LLaMA 65B and LLama 2 7B [17] as pre-train and fine-tuned models to improve contextual understanding and translation quality for low-resource language pairs as Chinese-Urdu.
- The Evaluation of multiple candidate solutions using BLEU score, semantic similarity, and NER accuracy to select the best translation. This ensured the translations reduced hallucinations while maintaining high contextual relevance and accuracy for named entities.

The proposed solution aims to maximise the difference between the log probabilities of the expert model and the amateur model to enhance the overall quality of the translation and reduce the likelihood of producing hallucinations. The remaining article is organised as follows: Section 2 reviews existing research on machine translation and hallucination for low-resource languages. Section 3 describes the material and method. Section 4 discusses the experimental setup. Section 5 discusses the results and discussion. Finally, look forward to concluding remarks and some potential research trends in the future.

2 | Literature Review

Multilingual NMT has emerged as a vital paradigm for building translation systems capable of handling numerous languages [6]. These approaches aim to translate directly with a single model for multiple language pairs without relying on any intermediate language. The dominant strategy for achieving these systems involves training large multilingual models on vast amounts of parallel data [18]. Data mining and data augmentation techniques are used for data acquiring with back-translation [19]. The multilingual capabilities of these systems result in significant improvements as compared to traditional bilingual models, particularly for low-resource and non-English-centric language pairs, which benefit the most from multilingual transfer [2]. An alternative and promising strategy adopts the emergent capabilities of large language models (LLMs). These models are pre-trained on massive corpora and then can be deployed to perform a variety of tasks [8, 20]. This approach has yielded impressive results across a wide range of NLP tasks [3, 21]. LLM can produce fluent and adequate translations, especially for high-resource English-centric language pairs, and these translations are competitive with those generated by dedicated supervised translation models [22, 23].

Hallucinations in machine translation represent a significant challenge, as accurate translation issues pose a critical threat to the safety and reliability of real-world applications. Notably, these hallucinations differ from natural language generation tasks, such as abstractive summarisation and generative question answering [7]. Hallucinations are substantially rarer and harder to observe in clean, unperturbed data. This rarity is possibly attributed to the more closed-ended nature of the task. Several previous studies have examined the properties of hallucinations by creating artificial scenarios where they are more likely to occur. For example, perturbations in the source text [18] or noise in the training data [24] have been introduced to study hallucinations. Hallucinations in machine translation are

categorised into hallucinations under perturbation and natural hallucinations.

According to Raunak et al. [24] and later extended by Guerreiro et al. [13] taxonomy, hallucinations are translations that contain content detached from the source text. Detecting hallucinations is necessary for improving machine translation reliability. Dale et al. [14], Guerreiro et al. [25], and Bawden and Yvon [23] evaluated various methodologies for identifying hallucinations and demonstrate the effectiveness of ALTI+ as a detection method. For hallucination mitigation, Guerreiro et al. [25] proposed a fallback model when a hallucination is detected, whereas other methods rely on sampling and re-ranking translations. For example, using COMET [26] helped to mitigate hallucinations by ranking sentences for the best performance. These approaches often depend on an additional external model and have been evaluated primarily on small de→en models [14, 25]. CD approach is used to mitigate hallucinations, evaluating them by counting segments with $chrF2 < 10$. They used the same model as the amateur but supplied it with randomly selected inputs [27]. We use different models as amateurs to provide more stable results. Additionally, our work compares different amateurs and techniques for combining expert and amateur distributions. This is complementarGy to the focus of Sennrich et al. [27] on off-target translations.

In reviewing the existing literature on NMT and hallucination detection and mitigation is summarised in Table 1. Despite significant advancements in NMT driven by deep learning and large-scale multilingual models, translating low-resource languages such as Chinese and Urdu remains a substantial challenge. The primary issue is the scarcity of training data, which leads to hallucinations—instances where the translated content diverges significantly from the source text. Existing methods to mitigate hallucinations are often plagued by overfitting and increased computational complexity. Moreover, no previous work specifically addresses hallucinations in Chinese-Urdu language pairs which highlights a significant gap in the current research. Therefore, there is an urgent need for a robust and scalable solution that can effectively reduce hallucinations and improve the overall quality of translations for low-resource languages.

3 | Materials and Methods

In this section, we describe the datasets, data preprocessing techniques, tokenisation, normalisation, and the proposed model. The detailed steps and algorithms used in our approach are outlined below and shown in Figure 1.

3.1 | Datasets

Because of data scarcity in low-resource languages, a diverse Chinese and Urdu parallel corpus was developed through human effort and web crawlers (ParaCrawl, Bitextor, Common Crawl and OpenNMT). ParaCrawl is a project that aims to build large-scale parallel corpora for machine translation by crawling multilingual websites. It uses sophisticated techniques to identify parallel texts on multilingual websites and provides open datasets for research and development in machine translation. Chinese and Urdu pairs are not available in open access. Bitextor is a popular tool designed specifically for crawling and collecting parallel corpora from the web. It automatically identifies and downloads bilingual websites, extracts and aligns parallel text segments. It can handle various text formats and encodings.

TABLE 1 | Summary of key literature on hallucination detection and mitigation in NMT.

Ref. Method	Datasets	Contribution	Limitation
Costa-jussà et al. [6] LLM with BT	Web-crawled corpora	Highlighted the potential of multilingual NMT	Limited focus on low-resource languages
Bapna et al. [18] pre-train GPT LLM	Flores-200, NLLB	Discussed data mining and augmentation strategies	Lack of robustness analysis
Fan et al. [2] M2M100	WMT, FLORES	Showed improvements in low-resource language pairs	Primarily English-centric
Brown et al. [20] GPT-3	WebText2, Books1	Demonstrated capabilities of LLMs in NLP tasks	High computational requirements
Chowdhery et al. [3], pre-train LLM	Multilingual corpora	LLMs' fluency in high-resource language pairs	Limited application to low-resource pairs
Ji et al. [7] NLG with LLM	Multilingual corpora	Reviewed hallucination issues in NLG tasks	General NLG focus, less on MT
Bapna et al. [18], MVE algorithm	IWSLT	Studied hallucinations under perturbation	Focus on artificial scenarios
Guerreiro, Voita, and Martins [13] NMT models	Manual datasets	Extended taxonomy of hallucinations	Challenges in natural hallucination detection
Guerreiro, Voita, and Martins [25], transformer	WMT'18	Evaluated hallucination detection methods	Primarily evaluated on small models
Sennrich et al. [27] CD and M2M 100-418	FLORES-101	Mitigated hallucinations using CD approach	Focus on off-target translations

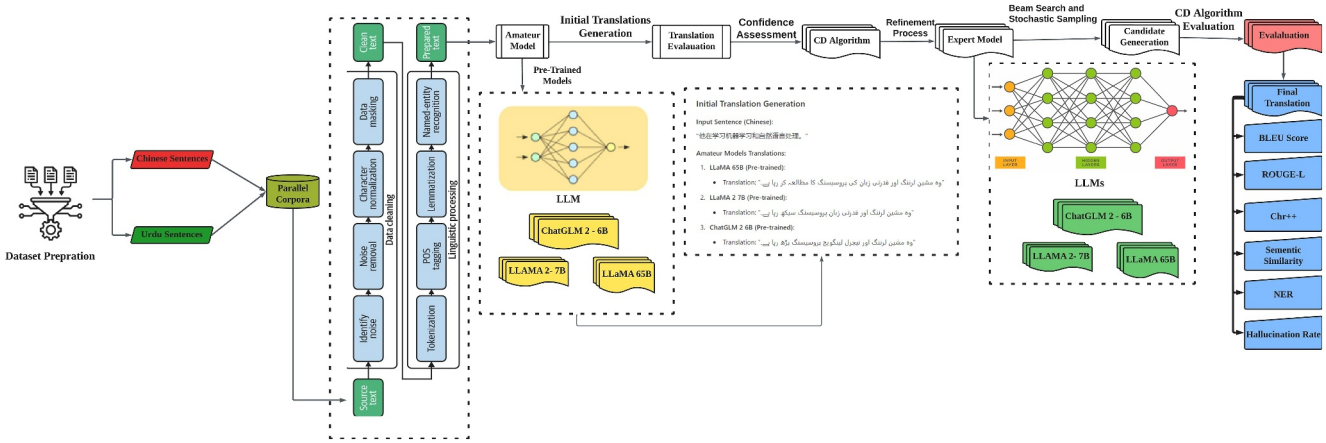

FIGURE 1 | The proposed model architecture.

Common Crawl provides a large, open repository of web crawl data. OpenNMT provides a suite of tools for machine translation, including utilities for data collection and preprocessing.

It is validated by native language and NLP experts and verified using multiple translators, including Google Translate and Baidu Translate, ensuring accuracy and reliability. Consistency checks included completeness, accuracy, uniform formats, and naming conventions. Redundancy checks removed duplicates, and label validation ensured correct and balanced datasets. Text consistency checks addressed spelling, grammar, and terminology. Sampling validation ensured accurate population representation for better model generalisation and bias detection. Text alignment and statistical analysis were performed to evaluate linguistic diversity and coverage across domains. In addition, we incorporated publicly available datasets.

- The OPUS dataset is a collection of parallel corpora for a wide range of languages, including Chinese and Urdu, which is widely used for machine translation and linguistic research [28]. The dataset is available at (http://opus.nlpl.eu/).
- The Workshop on Machine Translation (WMT) provides a benchmark in the field of machine translation. It includes a variety of parallel corpora for different language pairs and is used in annual machine translation competitions [29].
- WiLi 18 benchmark dataset of short text extracts from Wikipedia. It contains 1000 paragraphs in 235 languages, totalling 23,500 paragraphs. Each language in this dataset contains 1000 rows/paragraphs. After the same data selection and preprocessing, we selected the same Chinese Urdu 45 paragraph with the help of the middle language English [30].

These datasets are compiled and formatted in CSV format. Table 2 shows the details of the datasets. We consolidated all relevant datasets into a single comprehensive dataset in a unified format to facilitate experiments on large corpora. Additionally, we conducted dataset-wise experiments to ensure thorough analysis and validation. The size of the datasets for the training, testing, and validation is formed with a ratio of 70%, 15%, and 15%, respectively.

3.2 | Data Preprocessing

Data cleaning is performed for the removal of noise, such as special characters, HTML tags, punctuation, missing values, and unnecessary white space, as well as the standardisation of test cases and the correction of common misspellings. Consistency checks are conducted to ensure alignment between Chinese and Urdu sentences and prove the validity of this dataset using FastAlign. Statistical analysis evaluates the linguistic diversity and coverage across various domains. The correlation between the lengths of sentences in the source and target texts is analysed using the Pearson correlation coefficient [32]. A high correlation indicates good alignment and coherence in the dataset. SentencePiece tokeniser [33] is used to implement Byte-Pair Encoding (BPE) [34]. It breaks down rare words into subword units, which helps manage vocabulary size and improve translation accuracy for low-resource languages.

The process can be expressed as S be a source sentence, represented as a sequence of characters $S = \{c_1, c_2, ..., c_n\}$. The BPE [34] merges the most frequent pair of character sequences iteratively to form subword units. The tokenisation function T can be defined as: $T(S) = \{t_1, t_2, ..., t_m\}$ where t_m represents the subword units derived from S. The frequency of character pairs governs the merging process, and at each iteration, the most frequent pair (a, b) in the current vocabulary is merged into a new token ab. The updated rule for the vocabulary V at iteration k is represented in Equation (1) where $V^{(k)}$ is the vocabulary at iteration k.

$$V^{(k+1)} = \left(V^{(k)} \backslash \{a, b\}\right) \cup \{ab\} \tag{1}$$

To further refine normalisation, we incorporate an entropy-based normalisation function. The entropy H of the token distribution can be calculated as Equation (2), where $p(t_i)$ is the probability of token t_i in the dataset.

$$H(T) = -\sum_{i=1}^{n} p(t_i) \log p(t_i) \tag{2}$$

We adjust the token frequencies to ensure a more balanced token distribution. The entropy-normalised token frequency

TABLE 2 | Chinese-Urdu corpus data.

Corpus	Sentences	Zh tokens	Ur tokens	Training	Testing	Validation
OPUS Tiedemann et al. [28]	493,042	1,189,539	899,376	345,129	73,956	73,957
WMT Fonseca et al. [31]	608,405	1,348,494	1,106,496	425,883	91,260	91,262
Wi Li Thoma [30]	7938	33,357	56,894	5556	1190	1192
Custom	56,332	130,569	104,742	39,432	8449	8451
Total	1,165,717	2,701,959	2,167,508	815,000	174,855	174,862

$f'(t_i)$ can be expressed as Equation (3), where $f(t_i)$ is the original frequency of token t_i.

$$f'(t_i) = \frac{f(t_i)}{H(T)} \tag{3}$$

3.3 | Proposed Model

This research proposes a refined CD Algorithm [15] with large language models to enhance machine translation accuracy between Chinese and Urdu while minimising hallucinations. Back translation is employed to generate training data, where sentences from the target language are translated back into the source language to create additional parallel corpora. The amateur models consist of pre-trained LLaMA 65B, LLaMA 2 7B and ChatGLM 2 6B, whereas the expert models are their fine-tuned versions with enhancements. The process begins with amateur models generating initial translations, which serve as a baseline. These translations are evaluated using metrics such as BLEU score and lexical matching to identify areas for improvement. The CD algorithm plays a pivotal role in assessing the confidence levels of these initial translations. It dynamically adjusts the weights of each translation segment based on the confidence scores, giving higher weight to segments with higher confidence. This ensures that more reliable translations are prioritised. Subsequently, the expert models refine these weighted translations, correcting errors and enhancing quality through their fine-tuned capabilities. A beam search with varied parameters and stochastic sampling is then conducted by the expert models to generate multiple candidate translations. Each candidate represents a different possible translation, taking into account various confidence levels and model strengths. The CD algorithm evaluates these candidates using metrics like BLEU score, semantic similarity, and NER accuracy. The highest-weighted and most accurate segments from the expert models are combined to form the final translation.

3.3.1 | ChatGLM Model

The pre-trained ChatGLM 2 6B model [16] with 6 billion parameters offers enhanced capabilities in understanding and generating human-like text. We enhanced the attention mechanisms and integrated context-aware embeddings to boost translation accuracy and contextual understanding. Normally, models use static embeddings where each word has a fixed representation regardless of its context. This can lead to potential misunderstandings, especially with words that have multiple meanings depending on the context. Unlike static embeddings, context-aware embeddings are dynamically generated based on the entire sequence, allowing the model to capture nuanced meanings of words depending on their usage. For an input sequence $T = \{t_1, t_2, ..., t_n\}$, the context-aware embeddings for each token t_i are generated considering the full context of the sequence T. This can be represented as Equation (4):

$$E_{\text{ChatGLM}}(T) = [\text{ChatGLM}(t_1|CT), \text{ChatGLM}(t_2|CT), ..., \text{ChatGLM}(t_n|CT)] \tag{4}$$

where $CT = \text{context}(T)$ indicates the full contextual dependency of each token t_i on the entire sequence T. This modification enables the model to capture deep conversational contexts, aggregating all relevant information for the sequence, thereby improving its understanding of the text. The standard attention mechanism allows a model to focus on different parts of the input sequence when generating each part of the output. It calculates attention weights for each pair of tokens in the sequence. The attention score for a pair of tokens (t_i, t_j) is calculated using a scoring function, and these scores are then normalised to produce the attention weights. The standard scoring function is often a simple dot product between the query and key vectors projected through weight matrices. We enhanced the attention mechanism to ensure that the model focuses more effectively on the most relevant parts of the input sequence. The enhanced attention mechanism calculates attention for each token pair (t_i, t_j) as Equation (5):

$$\text{Attention}(t_i, t_j) = \frac{\exp(\text{score}(t_i, t_j))}{\sum_{k=1}^{n} \exp(\text{score}(t_i, t_k))} \tag{5}$$

The score function used in the attention calculation is defined as Equation (6):

$$\text{score}(t_i, t_j) = \frac{(W_q E_{\text{ChatGLM}}(t_i))(W_k E_{\text{ChatGLM}}(t_j))^T}{\sqrt{d_k}} \tag{6}$$

In these equations, W_q and W_k are weight matrices for the query and key projections, respectively, and d_k is the dimensionality of the key vectors. These enhancements collectively improve the model's ability to produce high-quality translations, especially for complex and low-resource language pairs such as Chinese and Urdu.

By implementing these custom enhancements, ChatGLM 2 6B's effectiveness in low-resource Chinese-Urdu bidirectional NMT tasks is significant. Its large parameter size and deep contextual

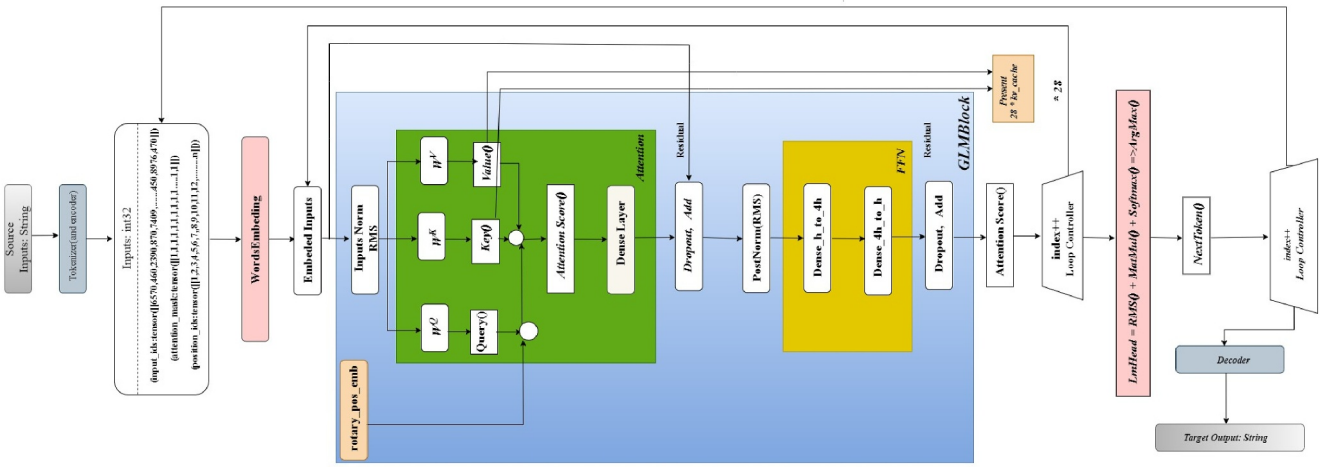

FIGURE 2 | ChatGLM model architecture.

embeddings enable better handling of complex language structures and nuances, reducing errors and improving overall translation quality. Figure 2 illustrates the model architecture for the ChatGLM-2 6B model.

3.3.2 | LLaMA

LLaMA 65B [17] is adopted for an experiment, which is a variant with 65 billion parameters. It provides enhanced capabilities for translation and contextually relevant tasks in text. It maintains a balance between model complexity and performance, which makes it suitable for our task. LLaMA's embeddings for input sequences T are used to handle the variability and scarcity of low-resource language data. The embeddings are represented as Equation (7).

$$E_{\text{LLaMA}}(T) = [\text{LLaMA}(t_1, CT), \text{LLaMA}(t_2, CT), ..., \text{LLaMA}(t_n, CT)] \quad (7)$$

The context CT includes different linguistic features. The encoder embeddings for an input sequence T are calculated as Equation (8):

$$H_{\text{enc}}^{65B} = \text{Encoder}_{\text{LLaMA 65B}}(T, \text{context}(T)) \quad (8)$$

LLaMA 2 7B is adopted for extensive analysis. It is pre-trained with 7 billion parameters. It offers greater model capacity and enables a more nuanced understanding and generation of text. It is particularly effective in handling complex language structures and improving translation quality in low-resource languages. The LLaMA architecture is shown in Figure 3. The encoder embeddings for an input sequence T are represented in Equation (9). The LLaMA model integrates morphological tags into the input features to enhance word embeddings. This modification enriches the linguistic data fed into the model by concatenating morphological tag embeddings with standard word embeddings. In addition, character-level embedding is used to capture detailed linguistic features, which is especially helpful for processing rare words that are underrepresented in the training dataset. To

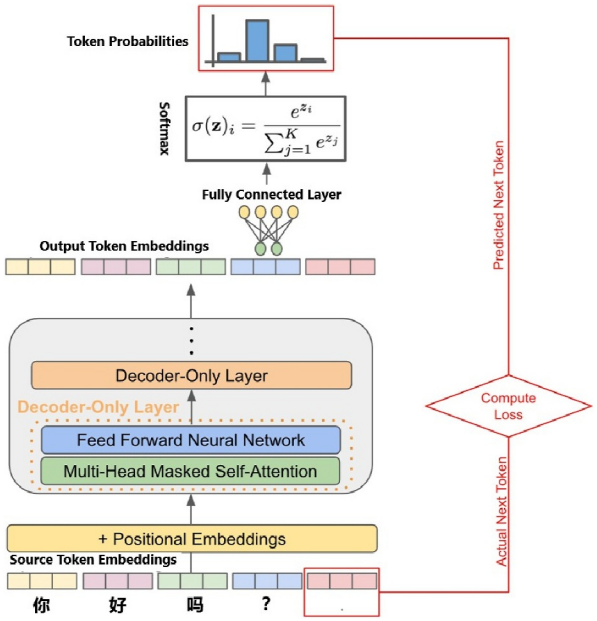

FIGURE 3 | LLAMA-2 model architecture.

handle the complexity of these enriched inputs, the encoder layers are adapted by incorporating specialised sub-networks. These sub-networks are designed to effectively process the morphological and character-level enhancements, ensuring they contribute optimally to the model's language understanding and generation capabilities. A multi-head attention mechanism is used to maintain precise alignment on different parts of the source and target texts. These mechanisms focus on monitoring the relevance and coherence of the output throughout the translation process. They regulate how much of the attention-modified inputs influence subsequent layers, adding an additional level of filtering and prioritisation. By strategically implementing this approach, the model not only focuses on the most pertinent parts of the data but also dynamically manages how these parts influence the learning and generation processes, ensuring the accuracy and relevance of the translated content.

$$H_{\text{enc}}^{7B} = \text{Encoder}_{\text{LLaMA 2 7B}}(T, \text{context}(T)) \quad (9)$$

3.3.3 | Refined Contrastive Decoding Algorithm

The CD algorithm [15] is developed to improve the quality of generated text by leveraging the outputs of both an expert model and an amateur model. It aims to refine the expert model's predictions by adjusting them based on the outputs of an amateur model. The CD algorithm helps in scenarios where enhancing diversity or preventing errors such as hallucinations is essential. We refined CD to address hallucinations in low-resource language pairs, Chinese and Urdu, along with LLM models. As introduced in Section 3.3, we set pre-trained ChatGLM, LLaMA 65B, and LLaMA 2 7B as amateur models and enhanced Fine-tuned ChatGLM 6B and LLaMA variants as an expert model by dynamically adjusting the weight given to the amateur model-based on the expert model's confidence. The CD score for a token i is given by Equation (10).

$$\mathrm{CD}(i) = \log P_{\text{expert}}(y_i|x) - \gamma \log P_{\text{amateur}}(y_i|x) \tag{10}$$

$P_{\text{expert}}(y_i|x)$ is the expert model's probability for token i after applying softmax, whereas $P_{\text{amateur}}(y_i|x)$ represents the probability assigned by the amateur model. A hyperparameter α within the range $0 < \alpha < 1$ is used to filter the probability distribution of the expert model. Tokens are included in the candidate set V_{thresh} if their probability is at least the maximum probability scaled by α. Formally, this can be expressed as Equation (11):

$$V_{\text{thresh}} = \Big\{ i \in V : \log\big(P_{\text{expert}}(y_i|x)\big) \geq \log(\alpha) + \max_j \log\Big(P_{\text{expert}}\Big(y_j|x\Big)\Big) \Big\} \tag{11}$$

The filtering process has two main purposes. Firstly, it prevents tokens with low probability from overwhelming the candidate set V_{thresh}. Secondly, it ensures that if the expert model exhibits high confidence, only the most probable token is included, allowing the candidate set to closely match the expert's preferred choice. The original CD algorithm used a constant weight γ at each time step by uniformly adjusting all probabilities. Instead of varying the number of candidates considered (as the threshold α does), we propose varying the degree of CD's influence on token generation by dynamically adjusting the weight γ. It is dynamically adjusted as: $\gamma = 1 - \max_i P_{\text{expert}}(x_i)^{\beta}$. This parameter handles the influence of the amateur model's log probability. The CD scores $\mathrm{CD}(i)$ replace the expert's scores during the beam search. Normalisation is applied to stabilise beam search and ensure consistent contribution across time steps. The normalisation constant N_{CD} is calculated as Equation (12):

$$N_{\text{CD}} = \frac{\sum_{i \in V_{\text{thresh}}} P_{\text{expert}}(i)}{\sum_{i \in V_{\text{thresh}}} \left(\frac{P_{\text{expert}}(i)}{P_{\text{amateur}}(i)} \right)^{\gamma}} \tag{12}$$

Dividing the CD scores by N_{CD} normalises them before scaling to the probability mass covered by V_{thresh}. This set of normalised CD scores is combined with the expert probabilities outside V_{thresh} to form a probability distribution. This method is referred to as NORMALISED CD. In our approach to CD, we prioritise preventing hallucinations in neural machine translation (NMT) over prioritising diversity. There are two main challenges to consider: hallucinations only occur in a small fraction of translated sentences, and NMT outputs need to be closely linked to the source sentence. As a result, CD should only modify outputs when hallucinations occur and have minimal effects on non-hallucinated outputs. These challenges are tackled using normalisation and dynamic weighting. By considering the proposed model, our approach aims to provide robust and accurate translations for low-resource language pairs, such as Chinese-Urdu, while effectively mitigating hallucinations. The Algorithm 1 summarises the proposed methodology.

ALGORITHM 1 | Hallucination mitigation algorithm.

1: **Input:** Source sentence x, Expert model P_{expert}, Amateur model P_{amateur}, Hyperparameter β, Models {ChatGLM, LLaMA,}
2: **Output:** Best translation y^*
3: Preprocess x: clean, tokenise, and normalise
4: Generate initial translation candidates $\{y_1, y_2, ..., y_n\}$ using beam search on P_{expert}
5: **for** each candidate translation y_i **do**
6: Calculate CD score:
$\mathrm{CD}(i) = \log P_{\text{expert}}(y_i|x) - \gamma \log P_{\text{amateur}}(y_i|x)$
7: Adjust weight γ dynamically: $\gamma = 1 - \max_i P_{\text{expert}}(x_i)^{\beta}$
8: Normalise scores using N_{CD}
9: **end for**
10: Evaluate candidates using BLEU score, semantic similarity, and NER accuracy
11: Select the best translation y^* based on the highest combined score
12: **Model Selection:** Select the model (ChatGLM 2 6B, LLaMA 65B and LLAMA 2 7B) that provides the best translation y^* based on evaluation metrics
13: **Return** y^*

The inclusion of NER [35] within the proposed model is effective in maintaining the integrity and accuracy of named entities in translation. When a token t_k in input sequence T is recognised as a named entity, the embedding is adjusted before translation as Equation (13):

$$E'_{\text{ChatGLM 2 6B}}(t_k) = \begin{cases} E_{\text{ChatGLM 2 6B}}(t_k) & \text{if } \neg \text{ is NE}(t_k) \\ E_{\text{NER}}(t_k, \text{source lang, target lang}) & \text{if is NE}(t_k) \end{cases} \tag{13}$$

This approach ensures that named entities pare accurately represented and translated across different languages.

4 | Experimental Setup

In our experimental setup for evaluating the performance of the proposed model, Python is chosen as the primary programming language due to its robust support for machine learning and natural language processing tasks. All experiments are conducted using the PyTorch framework. The computational tasks are executed on a cloud server featuring 1× A100 PCIe GPU, which provides the essential computational power for training large-scale translation models.

4.1 | Hyper-Parameters Fine Tuning and Model Training

Each model is fine-tuned on the hyperparameters such as learning rate, batch size, and number of epochs are optimised as shown in Table 3. The objective function for fine-tuning is given by Equation (14), where θ represents the model parameters, x is the source sentence, y is the target sentence, and D_{train} is the training dataset.

$$L(\theta) = -\sum_{(x,y)\in D_{\text{train}}} \log P(y|x;\theta) \tag{14}$$

4.2 | Evaluation Metrics

Semantic similarity is assessed using cosine similarity on embeddings generated by models. It ensures that the translations preserve the intended meaning of the source text. NER accuracy is measured to evaluate how well-named entities are maintained across translations. Other metrics are BLEU, METEOR, ROUGE-L, and chrF++. Flag translations as hallucinations if both BLEU and semantic similarity scores fall below-specified thresholds (e.g. $BLEU < 0.6, similarity < 0.75$).

5 | Results and Discussion

Results are reported in Table 4, which shows the performance of baseline and expert models across datasets. Firstly, the experiments are performed on a unified dataset. This comprehensive dataset is used to train and evaluate both the amateur and expert models. The initial translations are generated using amateur models and considered as baseline, which are then refined by expert models. Among these models, LLaMA 2–7B achieved the highest performance with a BLEU score of 47.0. The larger parameter size of LLaMA 2–7B enabled it to capture nuanced linguistic patterns more effectively. ChatGLM 2–6B also performed robustly and obtained a BLEU score of 46.0%, METEOR score of 38.0%, chrF++ score of 55.0%, and ROUGE-L score of 43.0%, demonstrating its capability in handling translation tasks efficiently. LLaMA 65B showed effectiveness with a BLEU score of 44.0(%). It showed slightly lower performance than the others, but its computational resource utilisation is good, indicating some limitations due to its smaller size and need for more training over epochs. The improved METEOR scores for fine-tuned models as 68.0% for LLaMA 2–7B indicate enhanced translation quality and contextual relevance. chrF++ (character n-gram F-score) measured the translation quality based on character n-gram, and it captured precision and recall. It is particularly effective for languages with complex morphology, such as Chinese and Urdu. The substantial increase in chrF++ scores for fine-tuned models as 56.0%–79.0% for LLaMA 2–7B highlighted the models' improvement in handling the morphological variations and fine-grained translation accuracy. The significant gains in ROUGE-L scores in fine-tuning (from 44.0% to 74.0% for LLaMA 2–7B) underscore its enhanced ability to produce coherent and fluent translations.

The LLaMA 2–7B as an expert model showed improvements in BLEU score of 79.1%, METEOR score of 68.0%, chrF++ score of 81.0%, and ROUGE-L score of 74.0% as compared to the baseline model. These results highlight the increased accuracy and reliability of translation. Similarly, ChatGLM 2–6B as an expert model got significant performance gains in a BLEU score of 76.0% and other metrics. The LLaMA 65B as an expert model achieved a BLEU score of 74.0, METEOR score of 63.0%, chrF++ score of 77.0%, and ROUGE-L score of 69.0%. The combination of diverse datasets in a standardised format ensured high-quality training data, and the multi-metric evaluation approach provided a comprehensive assessment of translation quality and reliability. These results verified the potential of advanced models and refined algorithms in addressing the above-mentioned challenges. In dataset-wise experiments, LLaMA 2–7B achieves the highest scores in all metrics for each dataset, with BLEU scores of 27.7 (%), 28.4 (%), and 19.6 (%) for OPUS, WMT, and Wili + Custom, respectively. ChatGLM 2–6B and LLaMA 65B also perform well but are consistently lower than LLaMA 2–7 B. The lower scores in the Wili + Custom dataset highlight the impact of limited and less diverse data on NMT performance, emphasising the challenge of translating low-resource languages. This variation underscores the need for comprehensive datasets to improve translation quality and reduce hallucinations.

TABLE 3 | Fine-tuning of hyperparameters for selected models.

Hyperparameter	LLaMA 65B	LLaMA 2 7B	ChatGLM 2 6B
Learning rate	5×10^{-6}	2×10^{-5}	2×10^{-5}
Batch size	64	32	24
Number of epochs	10	10	15
Warmup steps	1000	1000	750
Dropout rate	0.05	0.1	0.15
Weight decay	0.015	0.01	0.01
Gradient accumulation	2	2	2
Maximum sequence length	256	128	128
Optimiser	AdamW	AdamW	AdamW
Learning rate scheduler	Linear warmup & decay	Linear warmup & decay	Linear warmup & decay
Back Translation	Yes	Yes	Yes

Fine-tuning significantly enhances performance across all metrics. For example, LLaMA 2–7B's BLEU scores increase to 36.0, 39.0, and 23.6 for OPUS, WMT, and Wili + Custom, respectively. ChatGLM 2–6B and LLaMA 65B show similar improvements. The fine-tuned models exhibit better handling of diverse linguistic patterns and reduced hallucinations, particularly evident in the higher scores for OPUS and WMT datasets. However, the Wili + Custom dataset still shows relatively lower scores which highlights ongoing challenges with low-resource data. These results demonstrate the effectiveness of fine-tuning in enhancing model performance and the critical role of dataset quality in NMT systems.

5.1 | Performance Analysis Through NER Metrics

The performance of baseline and expert models is computed on NER and other translation quality metrics. Figure 4 provides valuable insights into how each model handles various aspects of translation quality and errors. Additionally, by including categories such as under-generation, fully detached, strongly detached, oscillatory, and other errors, we analysed where each model excels or struggles. In Figure 4, the first radar chart focuses on binary-score heuristics and evaluates models like Translation Neural Generation (TNG) and Reference Translation(RT). The second radar chart uses continuous scores for finer granularity. Comparing methods such as Attn-to-EOS, Attn-ign-SRC, TokHal-Model, Seq-Logprob, and MC-DSim are adopted. Attn-to-EOS measures the effectiveness of the attention mechanism towards the end of the sequence. Attn-ign-SRC indicates how often the model generates content by ignoring the source input. TokHal-Model evaluates token-level hallucinations. Seq-Logprob assesses the model's confidence in its translations. MC-DSim measures the dissimilarity between source and target sequences using Monte Carlo methods. It provides a more detailed analysis and is helpful to identify

TABLE 4 | Model Performance over Dataset and Combined datasdet.

		Baseline Model				Expert Model			
Corpus	**Model**	**BLEU (%)**	**METEOR (%)**	**chrF++ (%)**	**ROUGE-L (%)**	**BLEU (%)**	**METEOR (%)**	**chrF++ (%)**	**ROUGE-L (%)**
Combined	LLaMA 2–7B	47.0	38.0	56.0	44.0	79.1	68.0	81.0	74.0
Datasets	LLaMA 65B	44.0	36.0	53.0	41.0	74.0	63.0	77.0	69.0
	ChatGLM 2–6B	46.0	38.0	55.0	43.0	76.0	67.0	77.0	73.0
OPUS	ChatGLM 2–6B	22.8	24.1	46.7	30.2	32.5	34.0	53.3	39.7
	LLaMA 65B	24.6	26.0	40.1	31.7	31.0	32.0	49.0	37.0
	LLaMA 2–7B	27.7	29.0	50.4	34.0	36.0	37.0	58.0	43.0
WMT	ChatGLM 2–6B	27.1	28.5	52.2	34.8	30.0	32.4	54.6	38.5
	LLaMA 65B	25.6	27.0	40.8	31.5	26.3	27.1	50.4	33.5
	LLaMA 2–7B	28.4	29.3	53.10	35.3	39.0	41.2	62.0	46.3
Wili + custom	ChatGLM 2–6B	17.8	26.0	43.4	45.0	21.5	23.0	44.6	27.0
	LLaMA 65B	11.3	21.0	42.7	42.0	21.7	23.0	44.1	27.3
	LLaMA 2–7B	19.6	30.0	27.3	43.0	23.6	25.1	38.4	31.2

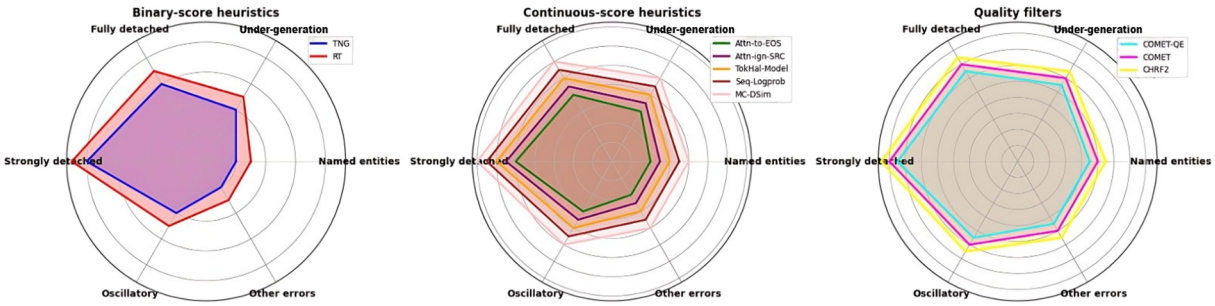

FIGURE 4 | NER calculations for NMT and hallucination mitigation through quality filters.

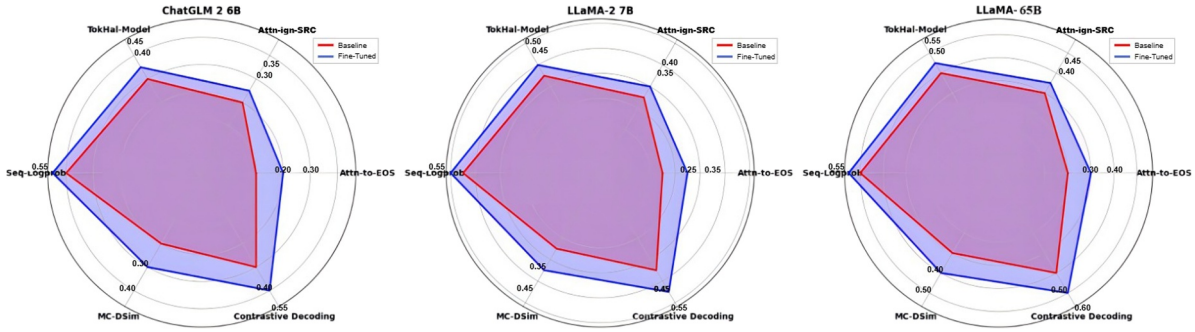

FIGURE 5 | Performance evaluation of Translation models with different quality metrics including contrastive decoding.

subtle differences in model performance. The third radar chart evaluates the same categories with quality filters such as COMET-QE, COMET, and CHRF2. It focuses on the quality of translations by indicating how well each model maintains translation quality across different aspects.

Figure 5 displays the performance improvements of proposed translation models across various metrics after fine-tuning. ChatGLM performs well on sequence log probability but needs improvement in attention-to-end-of-sequence and Monte Carlo dissimilarity. LLaMA 2 7B excels in attention-to-end-of-sequence and token hallucination but needs improvement in Monte Carlo dissimilarity. LLAMA 65B is strong in token hallucination and sequence log probability but weaker in attention-ignoring-source and Monte Carlo dissimilarity.

Figure 6 illustrates the reduction in hallucination rates for both baseline and expert models. The hallucination rate starts at 0.31 and gradually decreases to 0.15 in the baseline model. This shows a steady decline in the hallucination rate as the training progresses, indicating that the baseline model improves with more epochs, but the rate of improvement is relatively modest. The hallucination rate starts at 0.10 and decreases to 0.04 over 10 epochs in the expert model. This demonstrates a substantial decline compared to the baseline model that indicates that the proposed Algorithm 1 has a significant impact on reducing hallucinations quickly and effectively.

The cosine similarity between two vectors A and B is found as cosine_similarity$(A, B) = \frac{A \cdot B}{\|A\|\|B\|}$ where A and B are the embedding vectors of the source and target sentences respectively. The resulting value should be between -1 and 1. For semantic similarity, values closer to 1 indicate higher similarity. The results indicate that fine-tuned models significantly improve the semantic similarity scores as Figure 7. Specifically, LLAMA 2 7B showed the most substantial improvement as compared to LLaMA 65B and ChatGLM2 6B.

5.2 | Comparitive Analysis

The comparative analysis, as detailed in Table 5, highlights several studies on Chinese-Urdu neural machine translation (NMT) and focuses on the challenges of low-resource language pairs and issues related to hallucinations. Chen et al. developed a Chinese-Urdu NMT model integrating POS sequence prediction

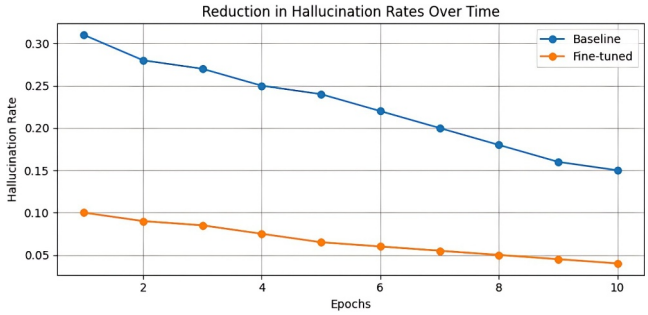

FIGURE 6 | Reduction in hallucination by proposed models.

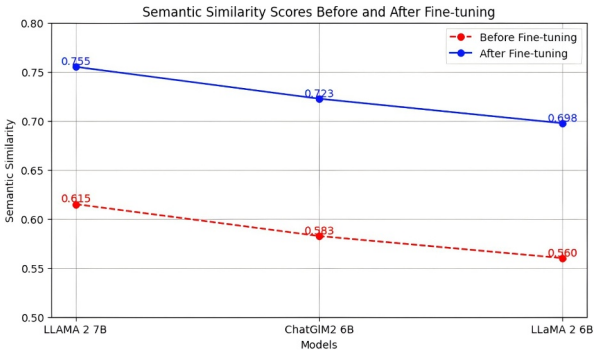

FIGURE 7 | Semantic similarity scores.

with the Transformer architecture, achieving a BLEU score of 0.36 [42]. Zeeshan et al. implemented the OpenNMT framework using LSTM and RNN-based models, attaining a lower BLEU score of 0.18 [41]. A further study by Zeshan et al. compared LSTM with Transformer models that demonstrates the superior performance of the Transformer, which significantly improved BLEU scores from 0.077 to 0.52, compared to 0.41 for LSTM [40]. The seq2seq NMT system is introduced for Chinese Urdu bidirectional translation by deploying a hybrid model as RNN with long short-term memory (LSTM) cells. The model gained a BLEU score of 0.42 [39]. The TranS2S model was applied to Chinese English translation by Zhou et al., and a Bleu score of 0.65 was obtained [36]. In low-resource bilingual translation, Li et al. showed the effectiveness of LLaMa and ChatGLM models [37]. The CD algorithm is used with a transformer-based M2m Model for low-resource languages. It improved the translation performance by 0.79% Bleu score and minimised the hallucination rate [38]. The proposed LLAMA 2 7B model stands out with the highest BLEU score of 79.17% among the studies focused on low-

TABLE 5 | Comparative analysis with state-of-the-art models.

Ref	Year	Model	Language pair	BLEU (%)
Zhou et al. [36]	2021	TranS2S	Chinese-English	0.65
J. Li et al. [37]	2024	LLAMA 2	Low resource bilingual	0.49
	2024	ChatGLM	Low resource bilingual	0.48
Waldendorf et al. [38]	2024	CD algorithm with M2M model	Low resource	0.79
J. Zeeshan et al. [39]	2021	Open NMT, LSTM and RNN	Chinese ↔ Urdu	0.18
Khan et al. [40]	2020	NMT, LSTM	Chinese ↔ Urdu	0.42
Z. A. Zeeshan and Jawad [41]	2020	LSTM	Chinese ↔ Urdu	0.41
Z. A. Zeeshan and Jawad [41]	2020	Transformer	Chinese ↔ Urdu	0.52
H. CHEN et al. [42]	2024	Transformer for POS	Chinese ↔ Urdu	0.36
Proposed		CD algorithm with ChatGLM	Chinese ↔ Urdu	0.76
Method		CD algorithm with LLAMA 65B	Chinese ↔ Urdu	0.74
		CD algorithm with LLAMA 2 7B	Chinese ↔ Urdu	0.79

resource settings, underscoring its effectiveness in mitigating hallucinations and enhancing translation quality for challenging language pairs like Chinese-Urdu.

In addition to focusing on low-resource language pairs such as Chinese-Urdu, we also tested our models on medium and high-resource languages to evaluate their generalisation and robustness across different language pairs. This approach ensures in Figure 8 that our refined model not only addresses the challenges specific to low-resource languages but also performs effectively across a broader spectrum of language pairs. The datasets used for these evaluations are mentioned above in Section 3.1. We computed hallucination rates for low, medium, and high-resource languages using thresholds. Translations are flagged as hallucinations if BLEU scores fall below 0.5. Translations are flagged as hallucinations if semantic similarity scores fall below 0.6. Figure 9 illustrates an overall improvements across the metrics for each model for quickly grasping how each model's performance and allowing for easy comparison of trends across different metrics. The improved NER metrics post-fine-tuning suggest that the models are better at detecting and handling hallucinations. It results in more accurate and reliable translations.

The BLEU scores exhibit a wide range indicating significant variability in translation quality. This variability can be attributed to the limited amount of training data available. The red dashed line represents the threshold below which translations are flagged as potential hallucinations. A substantial portion of the scores falls below this threshold. It suggests a higher likelihood of hallucinated translations. The distribution of BLEU scores for mid-resource languages shows higher median values than low-resource languages with a more concentrated spread. This indicates an improvement in translation quality as more training data becomes available. The threshold line shows fewer instances below it and signifies a reduction in hallucinated translations. The BLEU scores for high-resource languages are clustered towards the higher end of the scale, reflecting better translation performance. Most scores are above the threshold, indicating lower hallucination rates and demonstrating the effectiveness of abundant training data in improving translation quality.

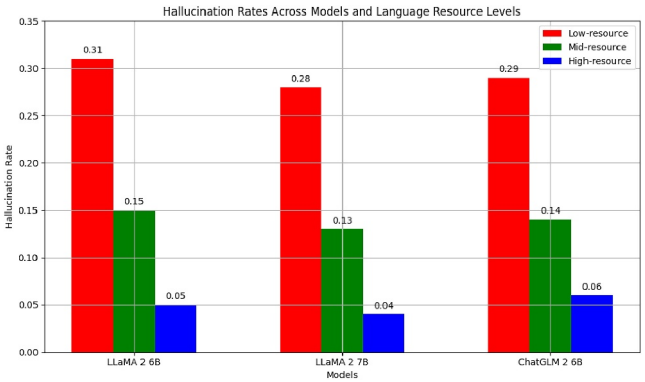

FIGURE 8 | Hallucination by resource levels.

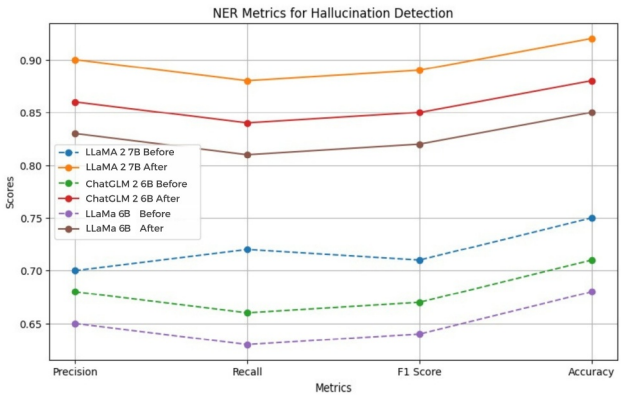

FIGURE 9 | Hallucination rate through NER metrics.

Figure 10 illustrates the training and validation loss curves for three models: LLAMA 65B, LLAMA 2 7B, and ChatGLM over 10 epochs. Each plot shows a consistent decrease in both training loss and validation loss as the number of epochs increases; it indicates that the models are learning effectively over time. The decreasing validation loss across all models suggests that the improvements generalise well to unseen data and confirm the effectiveness of the training process. There is no over-fitting in the model's evaluation.

5.3 | Discussion

The results demonstrate that the proposed approach significantly enhances translation quality for low-resource language pairs such as Chinese-Urdu. The integration of LLMs with Contrastive Decoding, dynamic weight adjustment, and multilingual embeddings contributed to these improvements. Dynamic weight adjustment helps the model to generalise better across different types of data. By focusing on rare and difficult instances, the model avoids overfitting to common patterns, which is a common cause of hallucinations. It ensures balanced learning by giving appropriate importance to different parts of the data. This prevents the model from being biased towards certain frequent phrases or structures. It improves the alignment between source and target texts. By focusing on aligning more complex or less frequent word pairs, the model reduces the likelihood of generating hallucinated sentences that do not correspond to the input text. Thresholds are defined to flag potential hallucinations. The results indicated that low-resource languages have the highest ratio of hallucination compared to mid- and high-resource languages. This trend underscores the impact of training data availability on translation quality. The higher variability and greater number of scores below the thresholds highlight the difficulties in achieving reliable translations with data scarcity. These results emphasise the critical role of training data in developing robust machine translation models and the effectiveness of defined thresholds in detecting hallucinations across different language resource levels. The combination of multiple evaluation metrics provides a comprehensive evaluation of translation quality. The scores for low-resource languages show a broad distribution and highlight the challenges in maintaining semantic integrity due to the scarcity of training data. The distribution of scores for high-resource languages is concentrated at the higher end, implying strong semantic fidelity in translations. The minimal number of scores below the threshold underscores the robustness of translations when ample training data is available, resulting in the lowest hallucination rates among the three categories. The experimental results validate the effectiveness of the proposed methodologies. In the future, the research may focus on accuracy and anomaly detection using proposed model in various appliations [43–46].

6 | Ablation Study

The research conducted an ablation study to understand the impact of various components on the proposed methodology. It is conducted on the same experimental setup, and a combined dataset is used. We measured the BLEU score, hallucination rate, and translation quality for each variant. The results are averaged across multiple runs to ensure consistency in Table 6 with different configurations.

The results of the ablation study highlighted the contributions of each component to the overall performance of the proposed model. This ensured that the selected translations were not only accurate but also contextually relevant. Combining all components yielded the highest performance across all metrics, with the BLEU score of 79.17%, and the hallucination rate dropped to 0.09%. This confirmed that the integration of all enhancements provides a synergistic effect that leads to robust and reliable translations.

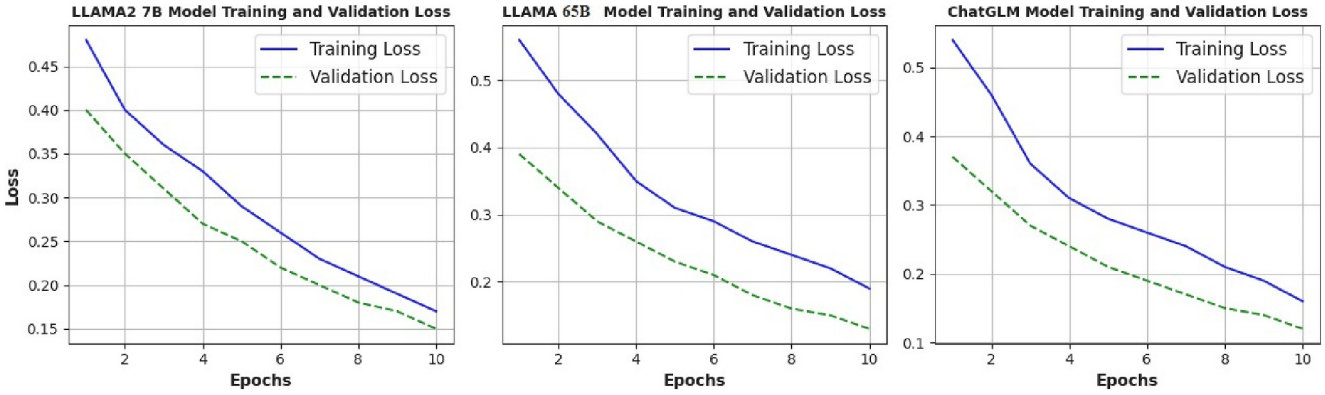

FIGURE 10 | Performance Evaluation through Models Training and Validation loss.

TABLE 6 | Ablation study results.

Configuration	BLEU (%)	Hallucination rate (%)	Translation quality (score)
Baseline	48.57	0.31	0.56
Baseline + refined CD algorithm	62.34	0.18	0.67
Fine-tuned + refined CD algorithm	68.45	0.15	0.72
Baseline + multi-metric evaluation	64.78	0.19	0.70
Fine-tuned + multi-metric evaluation	75.12	0.11	0.74
Full model	79.17	0.09	0.79

7 | Conclusion

This research presented a better approach to addressing hallucinations in low-resource NMT, specifically for Chinese-Urdu translation. We Utilised a refined CD algorithm with LLM models to enhance translation accuracy and reduce hallucinations. The methodology integrates comprehensive data preprocessing, large language model selection, and a robust evaluation framework that significantly improves the reliability and quality of translations. We deployed effective tokenisation using SentencePiece and normalisation techniques to ensure clean and consistent training data. We employed back translation to generate additional parallel corpora to enhance the robustness of the models. The proposed algorithm evaluates translations using BLEU score, semantic similarity, and Named Entity Recognition (NER) accuracy. The experimental results demonstrated substantial improvements and achieved scores of 79.17 (LLAMA 2 7B), 74.00 (LLaMA 65B), and 71.23 (ChatGLM 2 6B). The hallucination rate is reduced from 31% to 0.14% on the baseline and from 0.10% to 0.04% on fine-tuned models. Semantic Similarity and NER Accuracy metrics reflect better preservation of meaning and accurate handling of named entities. Future research would explore extending this methodology to other low-resource language pairs and refining the algorithm to enhance performance in more challenging translation scenarios. Emphasising data augmentation strategies, adopting community-shared resources, and advancing model architectures will be key to continually improving NMT systems.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data will be available upon request to the corresponding author.

References

1. X. Guan, Y. Liu, H. Lin, et al., "Mitigating Large Language Model Hallucinations via Autonomous Knowledge Graph-Based Retrofitting," in *Proceedings of the AAAI Conference on Artificial Intelligence* 38 (2024): 18126–18134, https://doi.org/10.1609/aaai.v38i16.29770.

2. A. Fan, S. Bhosale, H. Schwenk, et al., "Beyond English-Centric Multilingual Machine Translation," *Journal of Machine Learning Research* 22, no. 107 (2021): 1–48.

3. A. Chowdhery, S. Narang, J. Devlin, et al., "Palm: Scaling Language Modeling With Pathways," *Journal of Machine Learning Research* 24, no. 240 (2023): 1–113.

4. N. Arivazhagan, A. Bapna, O. Firat, et al., Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges (2019): arXiv preprint arXiv:1907.05019.

5. N. Goyal, C. Gao, V. Chaudhary, et al., "The Flores-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation," *Transactions of the Association for Computational Linguistics* 10 (2022): 522–538, https://doi.org/10.1162/tacl_a_00474.

6. M. R. Costa-jussà, J. Cross, O. Çelebi, et al., No Language Left behind: Scaling Human-Centered Machine Translation (2022): arXiv preprint arXiv:2207.04672.

7. Z. Ji, N. Lee, R. Frieske, et al., "Survey of Hallucination in Natural Language Generation," *ACM Computing Surveys* 55, no. 12 (2023): 1–38, https://doi.org/10.1145/3571730.

8. A. Radford, J. Wu, R. Child, et al., "Language Models Are Unsupervised Multitask Learners," *OpenAI blog* 1, no. 8 (2019): 9.

9. G. Wenzek, V. Chaudhary, A. Fan, et al., "Findings of the WMT 2021 Shared Task on Large-Scale Multilingual Machine Translation," in *Proceedings of the Sixth Conference on Machine Translation* (Association for Computational Linguistics (ACL), 2021), 89–99.

10. W. Xu, S. Agrawal, E. Briakou, M. J. Martindale, and M. Carpuat, "Understanding and Detecting Hallucinations in Neural Machine Translation via Model Introspection," *Transactions of the Association for Computational Linguistics* 11 (2023): 546–564, https://doi.org/10.1162/tacl_a_00563.

11. A. Hendy, M. Abdelrehim, A. Sharaf, et al., How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation (2023): arXiv preprint arXiv:2302.09210.

12. J. Ferrando, G. I. Gállego, B. Alastruey, C. Escolano, and M. R. Costa-jussà, Towards Opening the Black Box of Neural Machine Translation: Source and Target Interpretations of the Transformer (2022): arXiv preprint arXiv:2205.11631.

13. N. M. Guerreiro, E. Voita, and A. F. Martins, Looking for a Needle in a Haystack: A Comprehensive Study of Hallucinations in Neural Machine Translation (2022): arXiv preprint arXiv:2208.05309.

14. D. Dale, E. Voita, L. Barrault, and M. R. Costa-jussà, Detecting and Mitigating Hallucinations in Machine Translation: Model Internal Workings Alone Do Well, Sentence Similarity Even Better (2022): arXiv preprint arXiv:2212.08597.

15. X. L. Li, A. Holtzman, D. Fried, et al., Contrastive Decoding: Open-Ended Text Generation as Optimization (2022): arXiv preprint arXiv:2210.15097.

16. Z. Du, Y. Qian, X. Liu, et al., "Glm: General Language Model Pretraining With Autoregressive Blank Infilling," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, Vol. 1 (Association for Computational Linguistics (ACL), 2022), 320–335.

17. H. Touvron, L. Martin, K. Stone, et al., Llama 2: Open Foundation and Fine-Tuned Chat Models, 2023): arXiv preprint arXiv:2307.09288.

18. A. Bapna, I. Caswell, J. Kreutzer, et al., Building Machine Translation Systems for the Next Thousand Languages, 2022): arXiv preprint arXiv:2205.03983.

19. A. Siddhant, A. Bapna, O. Firat, et al., Towards the Next 1000 Languages in Multilingual Machine Translation: Exploring the Synergy between Supervised and Self-Supervised Learning, 2022): arXiv preprint arXiv:2201.03110.

20. T. Brown, B. Mann, N. Ryder, et al., "Language Models Are Few-Shot Learners," *Advances in Neural Information Processing Systems* 33 (2020): 1877–1901.

21. S. Zhang, S. Roller, N. Goyal, et al., Opt: Open Pre-trained Transformer Language Models (2022): arXiv preprint arXiv:2205.01068.

22. K. Peng, L. Ding, Q. Zhong, et al., Towards Making the Most of Chatgpt for Machine Translation, 2023): arXiv preprint arXiv:2303.13780.

23. R. Bawden and F. Yvon, *Investigating the Translation Performance of a Large Multilingual Language Model: The Case of Bloom* (2023): arXiv preprint arXiv:2303.01911.

24. V. Raunak, A. Menezes, and M. Junczys-Dowmunt, The Curious Case of Hallucinations in Neural Machine Translation (2021): arXiv preprint arXiv:2104.06683.

25. N. M. Guerreiro, E. Voita, and A. Martins, "Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation," in A. Vlachos and I. Augenstein, eds., *Proceedings of the*

17th Conference of the European Chapter of the Association for Computational Linguistics, May 2023 (Association for Computational Linguistics, 2023), 1059–1075, https://aclanthology.org/2023.eacl-main.75.

26. R. Rei, J. G. De Souza, D. Alves, et al., "Comet-22: Unbabel-Ist 2022 Submission for the Metrics Shared Task," in *Proceedings of the Seventh Conference on Machine Translation (WMT)* (Association for Computational Linguistics (ACL), 2022), 578–585.

27. R. Sennrich, J. Vamvas, and A. Mohammadshahi, Mitigating Hallucinations and Off-Target Machine Translation with Source-Contrastive and Language-Contrastive Decoding, 2023): arXiv preprint arXiv:2309.07098.

28. J. Tiedemann, M. Aulamo, D. Bakshandaeva, et al., "Democratizing Neural Machine Translation With Opus-Mt," *Language Resources and Evaluation* 58, no. 2 (2023): 1–43, https://doi.org/10.1007/s10579-023-09704-w.

29. T. Kocmi, E. Avramidis, R. Bawden, O. Bojar, A. Dvorkovich, C. Federmann, et al., "Findings of the 2023 conference on machine translation (WMT23): LLMs are here but Not quite there yet," in P. Koehn, B. Haddow, T. Kocmi, and C. Monz eds., *Proceedings of the Eighth Conference on Machine Translation, Dec. 2023* (Association for Computational Linguistics, 2023), 1–42, https://aclanthology.org/2023.wmt-1.1.

30. M. Thoma, The Wili Benchmark Dataset for Written Language Identification, 2018): arXiv preprint arXiv:1801.07779.

31. E. Fonseca, L. Yankovskaya, A. F. Martins, M. Fishel, and C. Federmann, "Findings of the WMT 2019 Shared Tasks on Quality Estimation," in *Proceedings of the Fourth Conference on Machine Translation*, Vol. 3, (Association for Computational Linguistics (ACL), 2019), 1–10: Shared Task Papers, Day 2, https://doi.org/10.18653/v1/w19-5401.

32. P. K. Buttar and M. K. Sachan, "A Review of the Approaches to Neural Machine Translation," *Natural Language Processing and Information Retrieval* (2023): 78–109, https://doi.org/10.1201/9781003244332-4.

33. S. Choo and W. Kim, "A Study on the Evaluation of Tokenizer Performance in Natural Language Processing," *Applied Artificial Intelligence* 37, no. 1 (2023): 2175112, https://doi.org/10.1080/08839514.2023.2175112.

34. S. Hellsten, Incremental Re-tokenization in BPE-Trained Sentencepiece Models, (Umeå University, 2024).

35. S. Chen, Y. Pei, Z. Ke, and W. Silamu, "Low-resource Named Entity Recognition via the Pre-training Model," *Symmetry* 13, no. 5 (2021): 786, https://doi.org/10.3390/sym13050786.

36. Zhou, C., Neubig, G., Gu, J. et al., "Detecting Hallucinated Content in Conditional Neural Sequence Generation." in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (Association for Computational Linguistics (ACL), 2021): 1393–1404, https://aclanthology.org/2021.findings-acl.122.

37. J. Li, J. Chen, R. Ren, et al., "The Dawn After the Dark: An Empirical Study on Factuality Hallucination in Large Language Models," *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2024): 10879–10899: arXiv preprint arXiv:2401.03205, https://doi.org/10.18653/v1/2024.acl-long.586.

38. J. Waldendorf, B. Haddow, and A. Birch, "Contrastive Decoding Reduces Hallucinations in Large Multilingual Machine Translation Models," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, Vol. 1 (Association for Computational Linguistics (ACL), 2024), 2526–2539.

39. J. Zeeshan, M. Zakira, and M. Niaz, "A Seq to Seq Machine Translation From Urdu to Chinese," *Journal of Autonomous Intelligence* 4, no. 1 (2021): 1–5, https://doi.org/10.32629/jai.v4i1.359.

40. Z. Khan, M. Zakira, W. Slamu, and N. Slam, "A Study of Neural Machine Translation From Chinese to Urdu," *Journal of Autonomous Intelligence* 2, no. 4 (2020): 29–36.

41. Z. A. Zeeshan and M. Z. Jawad, "Research on Chinese-Urdu Machine Translation Based on Deep Learning," *Journal of Autonomous Intelligence* 3, no. 2 (2020): 34–44, https://doi.org/10.32629/jai.v3i2.279.

42. H. H. Chen, J. Wang, and N. U. H. Muhammad, "Chinese-Urdu Neural Machine Translation Interacting Pos Sequence Prediction in Urdu Language," *Computer Engineering & Science* 46, no. 03 (2024): 518.

43. M. Faheem, M. A. Al-Khasawneh, A. A. Khan, and S. H. H. Madni, "Cyberattack Patterns in Blockchain-Based Communication Networks for Distributed Renewable Energy Systems: A Study on Big Datasets," *Data in Brief* 53, no. 5 (2024): 110212, https://doi.org/10.1016/j.dib.2024.11021250.66.

44. M. Faheem and A.-K. Mahmoud Ahmad, "Multilayer Cyber attacks Identification and Classification Using Machine Learning in Internet of Blockchain(IoBC)-Based Energy Networks," *Data in Brief* 54, no. 5 (2024): 110461, https://doi.org/10.1016/j.dib.2024.110461.68.

45. M. Faheem, B. Raza, M. S. Bhutta, and S. H. H. Madni, "A Blockchain-Based Resilient and Secure Framework for Events Monitoring and Control in Distributed Renewable Energy Systems," *IET Blockchain* (2024): 1–15, https://doi.org/10.1049/blc2.12081.69.

46. A. Akram, J. Rashid, M. A. Jaffar, M. Faheem, and R. Amin, "Segmentation and Classification of Skin Lesions Using Hybrid Deep Learning Method in the Internet of Medical Things." *Skin Research and Technology* 29, no. 11 (2023): e13524, https://doi.org/10.1111/srt.13524.