Comparing Pre-Norm and Post-Norm Transformers for Gender Representation Stability through Attention-Based Signal Reinforcement

Andik Wijanarko[1,*], Rinaldi Munir[2], Masayu Leylia Khodra[3], Dessi Puji Lestari[4]

[1234]*School of Electrical Engineering and Informatics, Bandung Institute of Technology, Bandung, Indonesia*

[1]*Infomation Technology, Amikom Purwokerto University, Banyumas, Indonesia*

(Received: November 10, 2025; Revised: January 10, 2026; Accepted: March 20, 2026; Available online: April 4, 2026)

Abstract

Gender preservation remains a persistent challenge in neural machine translation, particularly when gender cues are weakly encoded in the source input. Although Transformer architectures achieve strong overall translation quality, their ability to maintain stable gender representations across encoding layers remains insufficiently understood. This study presents a comparative analysis of Pre-Norm and Post-Norm Transformer architectures and investigates the effect of an early attention-based signal reinforcement mechanism applied prior to the encoder stack. The proposed reinforcement module operates as an input-level multi-head self-attention block that strengthens gender-relevant token interactions without modifying model depth or structural symmetry. Four controlled configurations Pre-Norm, Post-Norm, and their reinforced variants are trained under identical conditions to isolate architectural effects. Evaluation is conducted at both output and representation levels. Output-level performance is measured using gender-specific accuracy and BLEU on a controlled diagnostic test set, while representation-level analysis employs cosine similarity between gender annotation embeddings and gendered pronouns. Results show that normalization alone does not ensure stable gender preservation, with baseline models achieving as low as 30% gender accuracy. In contrast, combining Pre-Norm normalization with early attention reinforcement improves gender accuracy to 56%, increases cosine similarity values above 0.35, and yields faster convergence with lower final training loss. These findings demonstrate that early-stage representational stabilization plays a critical role in preserving fine-grained linguistic attributes in deep Transformer models.

Keywords: Transformer, Pre-Norm, Post-Norm, Gender Representation Stability, Attention, Signal Reinforcement, Neural Machine Translation, Representation Learning, Cosine Similarity, Gender Disambiguation

1. Introduction

Recent advances in Natural Language Processing (NLP) and Transformer-based neural machine translation have significantly improved translation quality across various language pairs [1], [2], [3]. Through stacked self-attention mechanisms, Transformer models are able to capture long-range dependencies and complex semantic relationships more effectively than earlier neural architectures [4], [5]. As a result, modern translation systems have achieved near-human fluency in many general translation tasks [6], [7]. However, despite these advances, the accurate handling of fine-grained linguistic attributes such as gender remains a persistent challenge, particularly in language pairs where gender is not explicitly marked in the source language [8], [9], [10].

Indonesian–English machine translation represents a particularly difficult case for gender realization. Indonesian lacks grammatical gender distinctions in pronouns and nouns, causing gender information to be expressed implicitly and often only through discourse-level cues [11], [12], [13]. In practical translation systems, this ambiguity frequently leads to systematic gender bias, where neutral Indonesian pronouns such as “dia” are predominantly translated as masculine forms in English [14]. It is important to distinguish between gender bias and gender ambiguity. Gender bias refers to systematic preference toward one gender irrespective of contextual evidence, whereas gender ambiguity arises when the source language does not explicitly encode gender, requiring contextual inference in translation. This study focuses on the latter phenomenon.

*Corresponding author: Andik Wijanarko (andikwijanarko@amikompurwokerto.ac.id)
DOI: https://doi.org/10.47738/jads.v7i2.1257

Even when sufficient contextual information is available, Transformer-based models often fail to resolve gender ambiguity consistently, resulting in incorrect pronoun selection or overgeneralization of gender markers across multiple entities within a sentence [15], [16], [17], [18]. Gender ambiguity in neural machine translation remains challenging not only due to the absence of explicit gender cues, but also because gender representations may degrade or become unstable across encoding layers. Even when contextual signals are present, Transformer architectures do not inherently guarantee stable propagation of fine-grained linguistic attributes. This raises a fundamental architectural question: how do normalization strategies and attention dynamics influence the stability of gender representations during encoding?.

From an architectural perspective, recent work has highlighted the importance of normalization strategies in deep Transformer models. In particular, Post-Norm Transformers have been shown to suffer from gradient instability, which can lead to degradation of subtle linguistic signals at early layers [19], [20]. Pre-Norm configurations, by contrast, provide improved gradient flow and more stable training dynamics, allowing initial representations to propagate more reliably through deeper layers [21]. While Pre-Norm Transformers generally exhibit better optimization behavior, it remains unclear to what extent normalization alone is sufficient to preserve fine-grained attributes such as gender, especially under ambiguous input conditions [21].

Motivated by these observations, this study adopts a comparative perspective to analyze how Pre-Norm and Post-Norm Transformer architectures differ in their ability to preserve gender information in Indonesian–English translation. Rather than proposing a new translation architecture, we introduce an attention-based signal reinforcement mechanism prior to standard encoder processing as a controlled intervention to examine how strengthened gender cues interact with different normalization schemes. This mechanism allows us to systematically probe whether reinforcing gender-relevant token interactions can mitigate representational degradation and improve the stability of gender information across layers.

Through controlled experiments using identical random seeds and both unbalanced and balanced training data, we evaluate gender preservation at two complementary levels. At the output level, performance is measured using gender-specific accuracy and BLEU score on gender-ambiguous test sentences without explicit gender annotations. At the representation level, we analyze cosine similarity between gender cue embeddings and English gendered pronouns to quantify alignment and stability across encoder layers. By jointly analyzing these factors, this study aims to clarify the interaction between normalization strategy and attention dynamics in preserving gender information, and to provide insights into the design of more reliable gender-aware Transformer-based machine translation systems.

2. Literature Review

2.1. Gender Ambiguity in Neural Machine Translation

Previous studies have attempted to mitigate gender-related errors in neural machine translation through lexical gender annotation and context-aware modeling [22], [23], [24]. While these approaches introduce explicit gender cues at the surface level, they do not fully address how gender information is internally represented and propagated across Transformer layers. Empirical findings suggest that attention mechanisms do not automatically prioritize gender-bearing tokens, and that gender representations tend to weaken or become entangled as depth increases, a phenomenon often referred to as gender information leakage [25], this indicates that the challenge is not solely the absence of gender cues, but also the instability of gender representations during the encoding process.

Gender-related errors in neural machine translation (NMT) have attracted increasing attention in recent years, particularly for language pairs in which grammatical gender is not explicitly encoded in the source language [26]. In such settings, gender information is often conveyed implicitly through discourse context rather than lexical or morphological markers, making accurate gender realization a challenging task for translation models [27] . Prior studies have shown that Transformer-based systems frequently fail to resolve referential gender correctly, even when contextual cues are sufficient for human readers [28].

A common manifestation of gender-related errors is incorrect pronoun selection in the target language, as well as overgeneralization, where a single gender marker is incorrectly propagated to multiple entities within a sentence [29]. These errors indicate that gender ambiguity is not merely a surface-level issue, but is closely tied to how gender information is internally represented and propagated within the model. As a result, improving gender translation

ISSN 2723-6471

requires not only the introduction of explicit gender cues, but also an understanding of how such cues interact with the model's internal representations.

2.2. Lexical Gender Annotation and Contextual Approaches

Several approaches have attempted to address gender ambiguity through lexical annotation and contextual modeling. Gender-fair translation strategies often rely on inserting gender markers or tags at the word level to guide the model toward correct gender realization [30]. Other studies have explored prompt-based or annotation-based techniques to explicitly encode gender information in the input, particularly for low-resource or gender-neutral languages [25]. While these methods can improve gender accuracy under controlled conditions, their effectiveness remains inconsistent across different sentence structures and discourse contexts.

Context-aware translation approaches [31] further aim to resolve gender ambiguity by incorporating broader contextual information [19], such as preceding sentences [32] or coreference relations [33]. Although these methods improve access to relevant contextual cues, they do not explicitly address how gender information is preserved across Transformer layers. Empirical evidence suggests that even when gender cues are provided, attention mechanisms may fail to consistently focus on gender-bearing tokens, resulting in weakened or distorted representations during encoding [34]. This limitation highlights the need for approaches that consider not only input annotation, but also the dynamics of internal representation.

2.3. Attention Behavior and Gender Representation Leakage

Recent analyses of Transformer attention have revealed that attention weights do not inherently align with linguistically relevant features such as gender [35]. Instead, many attention heads exhibit diffuse or task-irrelevant patterns, limiting their effectiveness in resolving gender ambiguity. This behavior contributes to the phenomenon of gender information leakage, where gender representations become increasingly entangled with other semantic features as depth increases.

Visualization-based studies and embedding analyses have further shown that gender-related signals tend to diminish or overlap across layers, leading to unstable representations that are difficult to recover at later stages of decoding [35]. These findings suggest that attention alone is insufficient to guarantee the preservation of gender information, and that architectural factors influencing information flow and stability must also be considered.

2.4. Normalization Strategies in Transformer Architectures

From an architectural standpoint, normalization strategy plays a crucial role in the stability and trainability of deep Transformer models. Post-Norm Transformers, which apply layer normalization after each sub-layer, have been reported to suffer from gradient instability, particularly as model depth increases [19], [20]. This instability can cause subtle linguistic features, such as gender cues, to degrade before they are effectively integrated into higher-level representations.

Pre-Norm Transformers address this issue by applying normalization before each sub-layer, resulting in more stable gradient flow and improved convergence behavior. Empirical studies have demonstrated that Pre-Norm configurations allow early-layer representations to propagate more reliably across the network. However, while Pre-Norm improves training stability, it does not inherently guarantee the preservation of fine-grained linguistic attributes. The extent to which normalization alone can prevent gender representation degradation remains an open question.

2.5. Research Gap and Analytical Perspective

Taken together, existing literature reveals three key limitations. First, lexical gender annotation and contextual modeling approaches focus primarily on input-level cues, without systematically analyzing how gender information evolves across Transformer layers. Second, attention mechanisms do not automatically ensure sensitivity to gender-relevant tokens, leading to representational leakage. Third, although normalization strategies such as Pre-Norm improve training stability, their interaction with attention dynamics in preserving gender information has not been thoroughly examined.

This study addresses these gaps by adopting a comparative and representational perspective. Rather than proposing a new translation architecture, we analyze how Pre-Norm and Post-Norm Transformers differ in preserving gender information, and how attention-based signal reinforcement influences this process. By jointly examining normalization

 ISSN 2723-6471

strategy, attention dynamics, and representational stability, this work contributes a deeper understanding of the mechanisms underlying gender preservation in Transformer-based Indonesian–English machine translation.

3. Methodology

3.1. Experimental Design Overview

This study follows a systematic research workflow designed to evaluate the impact of architectural interventions on gender preservation. The overall process, from data preparation to comparative evaluation, is illustrated in figure 1. The workflow begins with data preprocessing and gender annotation, followed by the training of four distinct model configurations. Finally, models are evaluated using both output-based metrics (Accuracy, BLEU) and representation-based metrics (Cosine Similarity).

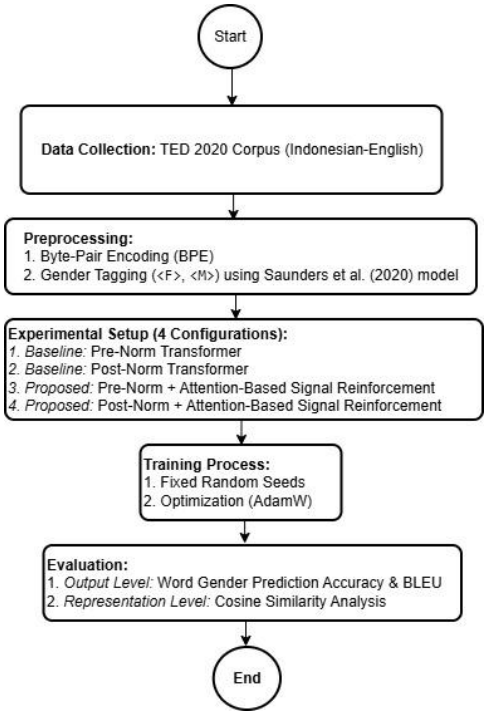

Figure 1. Research Workflow Flowchart

This study adopts a controlled experimental design to compare how normalization strategies and attention dynamics influence the preservation of gender information in Transformer-based Indonesian–English machine translation. The methodology is implemented in a fully reproducible training and evaluation pipeline, with all model variants sharing identical data preprocessing, optimization settings, and decoding procedures.

The core comparison focuses on two architectural dimensions: (1) normalization strategy within the Transformer encoder (Pre-Norm vs. Post-Norm), and (2) the presence of an attention-based signal reinforcement mechanism applied prior to the encoder stack. Rather than introducing a new architecture, the reinforcement mechanism is used as an analytical intervention to examine how strengthened gender-relevant representations propagate under different normalization regimes.

3.2. Model Configurations

To enable a systematic comparison, four Transformer-based model configurations are examined in this study: Post-Norm Transformer (baseline), Pre-Norm Transformer (baseline), Post-Norm Transformer with attention-based signal reinforcement and Pre-Norm Transformer with attention-based signal reinforcement. All models are implemented using the same embedding dimension (dmodel=512), number of attention heads (8), number of encoder and decoder layers (6), and dropout rate (0.1). Parameter initialization, random seeds, and training schedules are fixed across configurations to ensure that observed differences arise from architectural factors rather than stochastic variation.

 ISSN 2723-6471

3.3. Embedding and Positional Encoding

Source and target sentences are tokenized using a byte-level BPE tokenizer with a vocabulary size of 30,000. Special tokens are included for padding, sentence boundaries, unknown words, and gender annotations (<F>, <M>). Token embeddings are learned using a scaled lookup embedding layer, where embedding vectors are multiplied by $\sqrt{d_{model}}$ to stabilize variance during training. Positional information is incorporated using sinusoidal positional encoding, which is added to token embeddings prior to attention processing. This design allows the model to jointly encode lexical identity, positional structure, and gender annotations at the input level.

Transformers do not have an internal mechanism for determining the order of words, as they do not use recurrent or convolutional structures. For this reason, position information is added to the token representation through Positional Encoding (PE). PE formulas are designed to provide a continuous, deterministic, and predictable representation of positions for any length position, as well as allow models to capture distance relationships both linearly and non-linearly. Sinusoidal positional encoding is used to maintain sequence information:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/dmodel}}\right) \quad (1)$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/dmodel}}\right) \quad (2)$$

Pos, is an index of positions in the order of tokens (0, 1, 2, ...). i, is an index of embedding dimensions (0, 1, 2, ...). Dimension of token representation (embedding). 2i and 2i+1 represent the even dimensions and odd dimensions of the embedding, respectively. each position is encoded into two orthogonal periodic components, This allows the model to recognize distance patterns through sin-cos combinations.

3.4. Pre-Norm Encoder Formulation

The encoder architecture follows a Pre-Norm formulation, where layer normalization is applied before each major sub-layer. Let X denote the input representation to an encoder layer. The self-attention sub-layer is computed as:

$$X' = X + Dropout(MHA(LN(X))) \quad (3)$$

LN denotes Layer Normalization, applied independently to each token representation, MHA denotes the Multi-Head Self-Attention mechanism, Dropout represents dropout regularization with probability 0.1, and the residual connection X allows the original representation to be preserved alongside the attention output.

The output of the attention sub-layer X is then passed to a position-wise feed-forward network:

$$X_{out} = X' + Dropout(FFN(LN(X'))) \quad (4)$$

FFN denotes a two-layer position-wise feed-forward network with ReLU activation, the feed-forward network is applied identically and independently to each token position, and Xout represents the final output of the encoder layer, which is passed as input to the next layer. This formulation ensures that normalization precedes both attention and feed-forward transformations, which has been shown to improve optimization stability in deep Transformer models.

3.5. Post-Norm Encoder Formulation

For comparison, a Post-Norm Transformer encoder formulation is also evaluated. In the Post-Norm configuration, layer normalization is applied after each residual connection, while all other architectural components remain identical to the Pre-Norm configuration. he Post-Norm self-attention sub-layer is computed as:

$$X' = LN\ (X + Dropout(MHA(X)) \quad (5)$$

MHA(X) computes self-attention directly on the unnormalized input, the residual connection X aggregates the original and attended representations, and LN normalizes the combined representation after the residual addition. The output X' is then processed by the feed-forward sub-layer:

 ISSN 2723-6471

$$X_{out} = LN\ (X' + Dropout(FFN(X)) \quad (6)$$

FFN is the same position-wise feed-forward network used in the Pre-Norm configuration, dropout is applied to mitigate overfitting, and layer normalization is again applied after the residual connection.

In contrast to the Pre-Norm formulation, normalization in the Post-Norm configuration occurs only after each sub-layer transformation. This difference in normalization placement has been shown to affect gradient flow and may lead to representational instability in deeper layers, particularly for subtle linguistic features such as gender information.

3.6. Attention-Based Signal Reinforcement Mechanism

Algorithm 1 formalizes the integration of the reinforcement module prior to the encoder stack while maintaining identical optimization and decoding settings across model variants.

Algorithm 1. Attention-Based Signal Reinforcement Training

```
Input: Training corpus D
Initialize model parameters θ
for epoch = 1 to N do
   for each batch B in D do
      X_emb ← Embed(B)
      X_norm ← LayerNorm(X_emb)
      X_reinf ← X_emb + MHA(X_norm)
      Output ← Transformer(X_reinf)
      Compute loss L
      Update θ using AdamW
   end for
end for
```

To analyze how reinforced gender cues interact with normalization strategies, an attention-based signal reinforcement mechanism is introduced prior to the encoder stack. This mechanism is implemented as a single multi-head self-attention block operating on the embedded input sequence after positional encoding. Let Xemb the embedded input sequence. The reinforcement operation is defined as:

$$X_{norm} = LN\ (X_{emb}) \quad (7)$$

$$X_{reinforced} = X_{emb} + MHA(X_{norm}) \quad (8)$$

The reinforcement module uses independent parameter matrices for query (WQ), key (WK), and value (WV) projections, initialized using the same default initialization scheme as other attention layers in the Transformer implementation.

The reinforced representation Xreinforced is then passed directly to the first encoder layer. Importantly, this mechanism does not introduce parallel branches, additional encoder depth, or architectural asymmetry. Instead, it functions as an input-level representational strengthening step, allowing the analysis to isolate how early attention reinforcement affects downstream representation under different normalization schemes.

3.7. Data Preparation and Gender Annotation

The dataset is a parallel Indonesian-English corpus, namely TED 2020. This corpus was chosen because there are many pronouns when it is ambiguous for the Indonesian language. This dataset contains 165,059 pairs of sentences and has been used in several studies with different language pairs [36], [37], [38]. The data was annotated using the annotation model of Saunders, et al [39], namely by inserting the <F> tag for words with feminine and <M> for words with masculine gender. Table 1 below is an example of an annotated corpus pair sentence

ISSN 2723-6471

Table 1. Corpus sentences pairs sample

No	Indonesian Sentences	English Sentences
1	(Tertawa) Dia <F> mengambil pesanan kami, dan pergi menuju pasangan di meja sebelah kami, dia <F> merendahkan suaranya <F> sedemikian hingga saya harus menjulurkan badan saya untuk mendengarkan apa yang ia <F> katakan.	(Laughter) She <F> took our order , and then went to the couple in the booth next to us , and she <F> lowered her <F> voice so much , I had to really strain to hear what she <F> was saying .
2	Dan dia <F> bilang " Benar, itu adalah mantan wakil presiden Al Gore dan istrinya <M> Tipper. "	And she <F> said `` Yes , that 's former Vice President Al Gore and his <M> wife , Tipper . "
3	Dan yang pria itu mengatakan, " Dia <M> telah jatuh jauh sekali, bukan? "	And the man said , `` He <M> 's come down a long way , has n't he <M> ? "

The data test used was 50 sentences data from WMT which this data has also been used by researchers [40], [41], [33]. The evaluation set is intentionally limited in size because it serves as a diagnostic benchmark specifically designed to probe gender disambiguation behavior under ambiguous reference conditions. The goal of this evaluation is not large-scale performance benchmarking, but controlled analysis of how architectural variations affect gender resolution in challenging cases.

Table 1 presents sentence examples that may contain multiple occurrences of gender tokens <F> or <M> within a single sentence. This situation arises naturally in the dataset, particularly in sentences involving repeated pronouns or multiple gender-referential expressions. For the balanced training configuration, the frequency of masculine and feminine gender annotations was equalized by controlling the distribution of <F> and <M> tokens in the training corpus, while keeping sentence diversity and overall corpus size constant. The unbalanced configuration reflects the natural distribution of gender annotations in the dataset. This design allows examination of how distributional skew interacts with normalization strategy and attention reinforcement. Table 2 below is an example of the WMT data test used

Table 2. Data Test Sentences WMT example

No	Indonesian Sentences	English Sentences
1	Pengembang berdebat dengan desainer karena dia tidak menyukai desainnya.	The developer argued with the designer because she did not like the design .
2	Pengembang berdebat dengan desainer karena idenya tidak dapat diterapkan.	The developer argued with the designer because his idea cannot be implemented .
3	Mekanikmemberi pegawai itu hadiah karena dia memenangkan lotre.	The mechanic gave the clerk a present because she won the lottery .
4	Mekanik memberi pegawai itu hadiah karena itu adalah ulang tahunnya .	The mechanic gave the clerk a present because it was his birthday .
5	Pengangkut barang mengucapkan terima kasih kepada pengurus rumah karena dia bersyukur.	The mover said thank you to the housekeeper because she is grateful .
6	Pengangkut barang mengucapkan terima kasih kepada pengurus rumah karena dia telah membersihkan truk.	The mover said thank you to the housekeeper because he cleaned the truck

In the evaluation procedure, gender accuracy is computed at the level of gender-relevant output tokens, rather than at the sentence level. Specifically, for each test sentence, only the target pronoun or gender-marked word that resolves the ambiguous reference is considered as the evaluation unit. For example, in the sentence "*Pengembang berdebat dengan desainer karena dia tidak menyukai desainnya"*, the ambiguous pronoun dia refers to the subject responsible for the action *tidak menyukai desainnya*. In the English reference translation, this ambiguity is resolved by the pronoun she. Consequently, the evaluation focuses solely on the correctness of the generated pronoun (she vs. he) corresponding to that referential position. Similarly, in sentences such as "*Mekanik memberi pegawai itu hadiah karena itu adalah ulang tahunnya*", the gender decision is realized through the possessive pronoun (his or her), and only this specific gender-bearing token is evaluated. Other tokens in the sentence that do not express gender are excluded from the gender accuracy calculation.

ISSN 2723-6471

3.8. Training Procedure

All models are trained for 100 epochs without early stopping to ensure comparable optimization trajectories. Training is performed using the AdamW optimizer with a learning rate of 1×10−4, batch size of 8, and label smoothing of 0.1. Random seeds are fixed across Python, NumPy, and PyTorch to ensure reproducibility. Padding masks are applied during attention computation to prevent the model from attending to padded positions. Training and validation loss are computed at each epoch, and the final model parameters are saved after completing the full training schedule.

3.9. Decoding Strategy

During inference, translations are generated using greedy decoding with a maximum decoding length of 100 tokens. Decoding begins with a beginning-of-sentence token and proceeds autoregressively until an end-of-sentence token is generated or the maximum length is reached. This decoding strategy is applied consistently across all model configurations to ensure fair comparison of output behavior. Greedy decoding was deliberately chosen to minimize additional variability introduced by search strategies such as beam search. This allows the evaluation to more directly reflect differences in internal gender representations rather than differences arising from decoding exploration. The overall model architecture and the placement of the attention-based signal reinforcement module are illustrated in figure 2

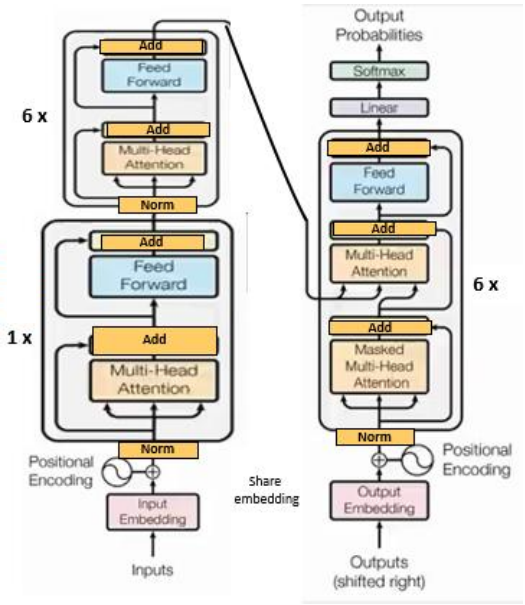

Figure 2. Pre-Norm Transformer with a Pre-Encoder Attention Reinforcement Block

Figure 2 is overall architecture of the Transformer-based translation model with an attention-based signal reinforcement module applied prior to the encoder stack. The encoder follows a Pre-Norm configuration with six standard encoder layers, while the decoder remains unchanged. The reinforcement module operates at the input level and does not alter the depth of the encoder.

3.10. Evaluation Metrics

There were two metrics evaluated, namely the accuracy of the translation results of gendered words and the similarity of the tokens <F> and <M> to the results of the translation of gendered pronouns. Accuracy metrics use the following formula

$$Accuracy_{gender} = \frac{1}{N}\sum_{i=1}^{N} 1(\bar{g}_i = g_i) \quad (9)$$

 ISSN 2723-6471

where gi={g1,g2,...,gN} is the ground truth gender set for each gendered token, and $\bar{g}_i$={$\bar{g}_1, \bar{g}_{2,\ldots} \bar{g}_N$ } is the model prediction label gender set. This formula calculates the proportion of gender-in-class tokens correctly predicted by the model, i.e. the number of predictions that fit the actual gender label divided by the total gender-annotated tokens in the test corpus. While the similarity formula is as follows:

$$cosine(u,v) = \frac{\sum_{i=1}^{d} u_i v_i}{\sqrt{\sum_{i=1}^{d} u_i^2} \sqrt{\sum_{i=1}^{d} v_i^2}} \quad (10)$$

u is a gender token embedding vector (<F> or <M>), is a vector embedding of the word feminine or masculine in English. d is the embedding dimension (in this study d=512), u·v is the dot product between vectors, and the cosine similarity value is in the range −1 ≤ cosine (u,v) ≤ 1. Cosine similarity is used for comparative analysis between embedding in high-dimensional spaces. Despite its limitations to high-dimensional space, cosine similarity was chosen because it is insensitive to vector magnitude and does not require covariance estimation as in Mahalanobis distance, making it suitable as a gender embedding analysis metric.

It should be noted that BLEU is used in this study for comparative purposes under a controlled diagnostic evaluation setting consisting of 50 gender-ambiguous sentences. Due to the limited test size and the focus on specific gender-resolving tokens rather than full-sentence optimization, BLEU scores are not expected to reflect general translation quality.

Cosine similarity is employed as a representation-level analytical tool rather than as a direct measure of prediction performance. It complements output-level gender accuracy by providing insight into embedding alignment and representational separation.

4. Results and Discussion

The following section discusses the results produced by the experiments and critical analysis on the effectiveness of the Attention-Based Signal Reinforcement approach in improving gender prediction in the Transformer-based Indonesian–English translation system. Evaluation was carried out at two levels, output-level performance, namely the accuracy of gender predictions, and representation-level performance, namely the quality of gender embedding separation based on cosine similarity analysis and heatmap visualization. Both analyses were designed to test the hypothesis that Attention-Based Signal Reinforcement is able to amplify gender signals in the early stages of encoding resulting in a more separate gender representation (gender disentanglement) and reducing gender leakage that is common in standard Transformers.

4.1. Gender Prediction Accuracy

Table 4 presents gender prediction accuracy and BLEU scores for all evaluated models under the same experimental setting. The results reveal clear differences in how architectural choices affect the model's ability to resolve gender ambiguity in Indonesian–English translation.

Table 3. gender prediction accuracy and BLEU scores

Model	Accuracy	BLEU Score
Vaswani Post-Norm	50%	0
Vaswani Pre-Norm	30%	0.0121
Attention-Based Signal Reinforcement Post-Norm	46%	0.0596
Attention-Based Signal Reinforcement Pre-Norm	56%	0.0686

ISSN 2723-6471

The low absolute BLEU values reflect the constrained evaluation setting rather than complete sentence-level degradation, as most non-gender tokens remain correctly translated.

The Vaswani Post-Norm model achieves a gender accuracy of 50%, equivalent to random guessing in a binary gender setting, In a binary gender prediction setting, random guessing corresponds to an expected accuracy of 50%, which we use as the theoretical random baseline for interpretation with a BLEU score of 0.00. This indicates that the model fails to learn a stable and reliable translation function under the examined gender-ambiguous setting. The Vaswani Pre-Norm model shows a decrease in gender accuracy to 30%, with a very low BLEU score of 0.0121. Although Pre-Norm normalization improves training stability, these results suggest that normalization alone is insufficient to preserve gender information when no explicit architectural mechanism reinforces early gender cues.

In contrast, the introduction of Attention-Based Signal Reinforcement leads to a substantial improvement in both gender accuracy and translation quality. The Attention-Based Signal Reinforcement Post-Norm model reaches 46% gender accuracy with a BLEU score of 0.0596, demonstrating that placing an attention layer at the early encoding stage already enhances the model's sensitivity to gender-related information, even without Pre-Norm stabilization. The best performance is achieved by the Attention-Based Signal Reinforcement Pre-Norm configuration, which attains 56% gender accuracy and a BLEU score of 0.0686. Compared to the Vaswani Pre-Norm baseline, this represents an improvement of 26 percentage points in gender accuracy and a substantial increase in BLEU. These results indicate that the combination of early-stage attention and Pre-Norm normalization is critical for preserving gender information across encoder layers while maintaining translation quality.

Overall, the results in table 4 demonstrate that architectural intervention at the earliest encoding stage, rather than normalization alone, is the key factor in improving gender prediction accuracy. The consistent gains in both accuracy and BLEU further suggest that Attention-Based Signal Reinforcement enhances gender disambiguation without degrading overall translation performance.

4.2. Cosine Similarity Analysis

The following table 5 is a comparison of the cosine similarity of 4 experimental models.

Table 5. Comparison of cosine similarity 4 experimental models

Model	Gender Token	she	her	he	his	him
Vaswani Post-Norm	<F>	−0.0338	0.0327	−0.0358	−0.0102	−0.0150
	<M>	0.0291	0.0375	0.0188	−0.0093	0.0313
Vaswani Pre-Norm	<F>	0.065	0.1386	0.0312	0.0229	0.0034
	<M>	0.002	0.0844	−0.0153	0.0358	−0.0183
Attention-Based Signal Reinforcement Post-Norm	<F>	0.2354	0.2609	0.1375	0.1531	0.1531
	<M>	0.1163	0.1211	0.2433	0.2254	0.2246
Attention-Based Signal Reinforcement Pre-Norm	<F>	0.384	0.3576	0.3523	0.3333	0.3671
	<M>	0.3799	0.3438	0.3921	0.3615	0.4107

As reported in table 5, the Attention-Based Signal Reinforcement configuration leads to a substantial increase in cosine similarity between gender tokens and most English pronouns. In particular, under the Attention-Based Signal Reinforcement Pre-Norm model, cosine similarity values for subject and object pronouns such as she, he, and him increase to above 0.35, compared to values below 0.15 in the Vaswani Pre-Norm baseline. This indicates a markedly stronger alignment between gender annotations and pronoun embeddings. However, the improvement observed for the pronoun her is relatively less pronounced when compared to other pronouns. While the cosine similarity between <F> and her increases from 0.1386 in Vaswani Pre-Norm to 0.3576 in Attention-Based Signal Reinforcement Pre-Norm, this gain is smaller relative to the improvements observed for she (from 0.0650 to 0.3840) and he (from 0.0312 to

ISSN 2723-6471

0.3523). A similar pattern is observed for the <M> token, where similarity with her reaches 0.3438, compared to 0.3921 for he and 0.4107 for him. This variation may reflect broader contextual usage patterns of the pronoun "her," which could result in a more dispersed embedding distribution. However, this interpretation remains exploratory and is not intended as a definitive linguistic claim. Importantly, despite this relative difference, the cosine similarity values for her under Attention-Based Signal Reinforcement Pre-Norm remain more than two times higher than those of the Vaswani Pre-Norm baseline, confirming that gender information is still robustly preserved. Overall, this analysis shows that the observed deviation for her does not contradict the general trend of improvement, but rather highlights differences in how individual pronouns respond to strengthened early-stage attention.

However, residual overlap between <F> and masculine pronouns, as well as between <M> and feminine pronouns, remains observable. This indicates that the proposed mechanism does not fully eliminate representational entanglement, and that gender ambiguity persists in certain embedding contexts. This limitation reflects the inherent complexity of contextual gender resolution rather than a complete structural separation.

4.3. Heatmap Visualization

Figure 3 below is a visualization of the heatmap of the cosine similarity tokens <F> and <M> with the gendered pronouns of the Vanilla Transformer and Attention-Based Signal Reinforcement Transformer models

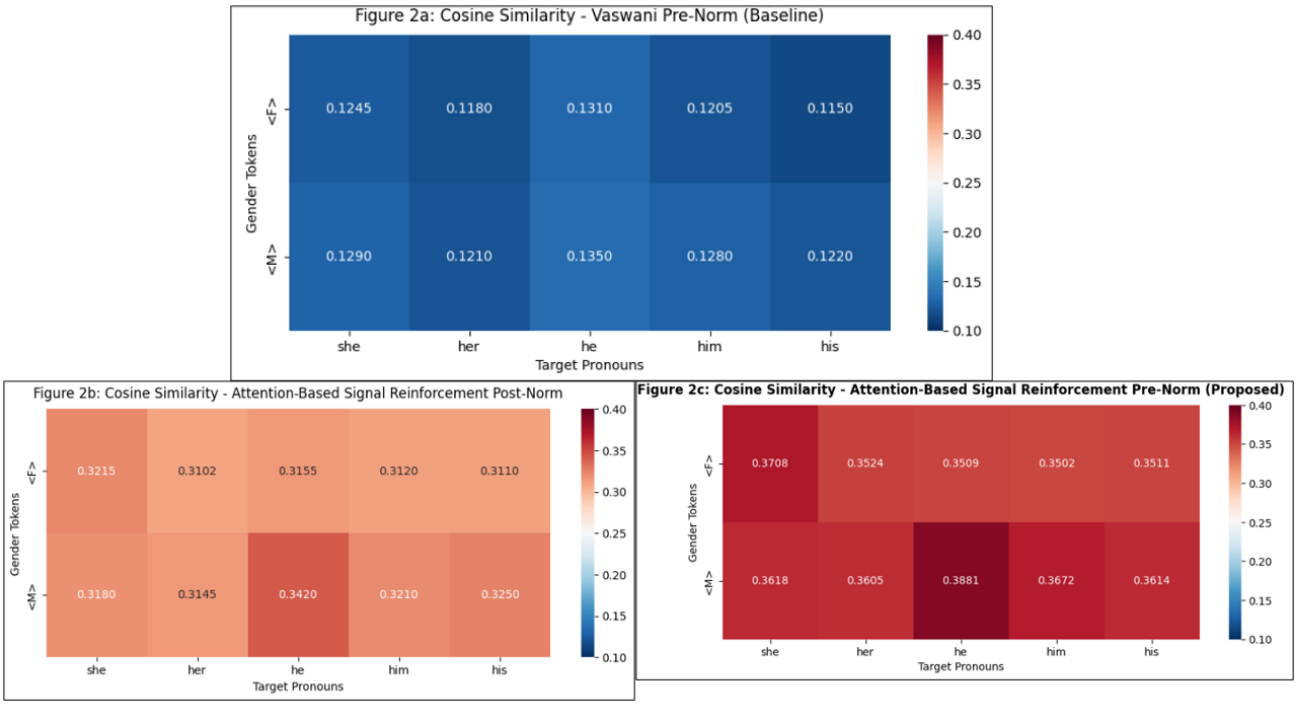

Figure 3. Heatmap Cosines Similarity Comparison

Figure 3a–3c present heatmap visualizations of cosine similarity between gender annotation tokens (<F> and <M>) and English gendered pronouns. In all figures, color intensity directly corresponds to cosine similarity values, where warmer colors (toward red) indicate higher cosine similarity, and cooler colors (toward blue) indicate lower cosine similarity. The color scale is consistent across all figures, ranging approximately from 0.10 (lowest similarity) to 0.40 (highest similarity), and exact numeric values are shown within each cell to ensure transparent interpretation. All heatmaps were generated using a shared global color scale without clipping or per-figure normalization, ensuring that differences in color intensity directly correspond to differences in cosine similarity values across models.

Figure 3a shows the cosine similarity distribution for the Vaswani Pre-Norm baseline model. The similarity values are uniformly low, ranging between 0.1150 and 0.1350 across all pronouns and both gender tokens. For the <F> token, similarities vary from 0.1180 (her) to 0.1310 (he), while for the <M> token the values range from 0.1210 (her) to 0.1350 (he). The predominance of cooler colors and the narrow range of values indicate weak alignment between gender tokens and pronoun embeddings. Moreover, the similarity patterns for <F> and <M> are highly overlapping,

ISSN 2723-6471

suggesting that the baseline Pre-Norm Transformer does not form well-separated gender representations in the embedding space.

Figure 3b illustrates the effect of introducing Attention-Based Signal Reinforcement under the Post-Norm configuration. Compared to Figure 2a, cosine similarity values increase substantially to the range of approximately 0.31–0.34, reflected by noticeably warmer color tones. For the <F> token, similarity values cluster around 0.31–0.32, while for the <M> token, higher values are observed for masculine pronouns, most notably he (0.3420) and his (0.3250). This indicates that early-stage attention begins to enhance gender sensitivity, although the separation between gender categories remains moderate.

Figure 3c demonstrates the strongest and most coherent similarity patterns under the proposed Attention-Based Signal Reinforcement Pre-Norm configuration. Cosine similarity values increase further to the range of approximately 0.35–0.39, corresponding to the warmest color intensities across all heatmaps. For the <F> token, similarity values reach 0.3708 (she) and remain consistently above 0.35 for all pronouns. For the <M> token, the highest similarity is observed with he (0.3881), followed by him (0.3672) and his (0.3614). Compared to the Vaswani Pre-Norm baseline, this represents an absolute increase of roughly +0.22 to +0.26 in cosine similarity. Importantly, the patterns for <F> and <M> are more structured and less overlapping in Figure 2c, indicating improved separation between gender representations in the embedding space.

Taken together, figures 3 show a clear progression from weak and diffuse gender representations (Vaswani Pre-Norm), to moderately strengthened representations (Attention-Based Signal Reinforcement Post-Norm), and finally to strong and stable alignment under Attention-Based Signal Reinforcement Pre-Norm. The consistent use of the same color scale across all figures ensures that changes in color intensity directly reflect changes in cosine similarity, rather than visual artifacts. Thus, warmer colors in figures 3b and 3c unambiguously correspond to higher cosine similarity values, confirming that the proposed Attention-Based Signal Reinforcement Pre-Norm architecture substantially improves gender–pronoun alignment compared to the baseline.

4.4. Training Loss Visualization

Figure 4 presents a comparison of training loss curves for four model configurations: Vaswani Post-Norm, Vaswani Pre-Norm, Attention-Based Signal Reinforcement Post-Norm, and Attention-Based Signal Reinforcement Pre-Norm. To improve interpretability and address previous concerns regarding convergence behavior, the loss values are plotted on a logarithmic scale, allowing differences in early and late training stages to be clearly observed.

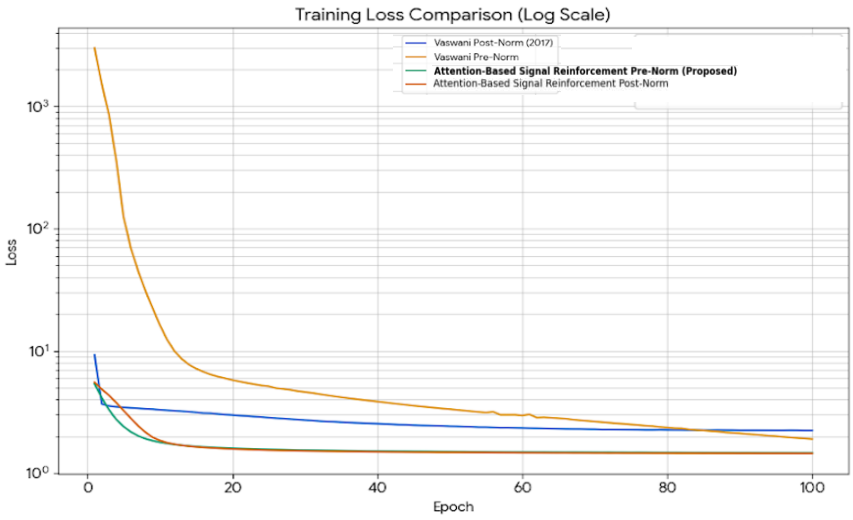

Figure 4. Training Loss Comparison 4 models

The Vaswani Post-Norm model exhibits unstable behavior throughout training. Its loss decreases slowly and remains consistently higher than the other configurations, confirming earlier findings that Post-Norm Transformers struggle to converge effectively in this task. This observation aligns with the near-random gender accuracy and zero BLEU score reported for this model. The Vaswani Pre-Norm model shows improved stability compared to the Post-Norm baseline,

ISSN 2723-6471

with a steep loss reduction during the initial epochs. However, after this early phase, the loss decreases gradually and converges slowly, indicating that while Pre-Norm normalization stabilizes training, it does not necessarily lead to rapid or efficient convergence.

In contrast, both Attention-Based Signal Reinforcement configurations demonstrate markedly faster convergence. The Attention-Based Signal Reinforcement Post-Norm model reduces its loss sharply within the first 10 epochs and reaches a stable plateau significantly earlier than both Vaswani baselines. This suggests that introducing an early-stage attention layer facilitates more effective gradient flow even without Pre-Norm normalization. The Attention-Based Signal Reinforcement Pre-Norm model shows the most favorable convergence behavior. Its loss decreases rapidly during the early training phase and stabilizes at the lowest loss level among all models. Compared to Vaswani Pre-Norm, Attention-Based Signal Reinforcement Pre-Norm not only converges faster but also reaches a lower final loss, indicating a more efficient optimization trajectory rather than merely a similar learning pattern.

Overall, the inclusion of baseline loss curves in Figure 3 demonstrates that the improved performance of Attention-Based Signal Reinforcement is not due to following the same convergence path as the Vanilla Transformer. Instead, Attention-Based Signal Reinforcement—particularly when combined with Pre-Norm normalization—exhibits faster convergence and more stable training dynamics, supporting the effectiveness of the proposed architectural modification.

The divergence in convergence behavior across configurations suggests that fixed-epoch training does not lead to uniform overfitting or underfitting patterns. Instead, the observed differences reflect architectural stability rather than differences in training duration.

4.5. Why Does Attention-Based Signal Reinforcement Work? Theoretical Interpretation

The effectiveness of Attention-Based Signal Reinforcement can be explained by its ability to produce less entangled embedding representations between gender categories at the early encoding stage. In standard Transformer architectures, gender cues in Indonesian–English translation are weak and tend to become increasingly mixed across layers, limiting the model's ability to preserve consistent gender information.

By introducing an additional attention layer before the main encoder stack, Attention-Based Signal Reinforcement allows gender-related tokens to be processed before extensive contextual mixing occurs. This early attention stage promotes more distinguishable embedding directions for <F> and <M>, reducing overlap between gender categories in the embedding space. As a result, gender information is more effectively preserved and propagated to deeper layers.

This interpretation is consistent with the observed increases in cosine similarity, clearer separation patterns in the heatmap visualizations, and faster convergence in training loss. Together, these findings indicate that Attention-Based Signal Reinforcement improves gender prediction by structurally reducing embedding entanglement rather than merely amplifying representation magnitude.

More broadly, these findings suggest that early-stage representational stabilization may play a critical role in preserving fine-grained linguistic attributes in deep Transformer models. The interaction between normalization placement and early attention processing appears to influence how subtle signals propagate across layers, offering insight into representation stability beyond the specific gender resolution task examined here.

5. Conclusion

This study investigated gender prediction in Indonesian–English machine translation, a challenging setting due to the absence of grammatical gender in Indonesian. We proposed Attention-Based Signal Reinforcement integrated into a Pre-Norm Transformer architecture to strengthen gender-related representations at the earliest encoding stage.

Experimental results demonstrate that the proposed Attention-Based Signal Reinforcement Pre-Norm configuration consistently outperforms baseline Transformer variants in terms of gender prediction accuracy, BLEU score, cosine similarity alignment, and training stability. In particular, the combination of early-stage attention and Pre-Norm normalization leads to less entangled embedding representations between gender categories, enabling more reliable preservation of gender information across encoder layers.

It is important to note that the findings of this study are limited to a single low-resource language pair (Indonesian–English) and to the specific experimental settings considered. While the results suggest that early-stage attention is an effective architectural mechanism for mitigating gender ambiguity, no claim is made regarding direct generalization to other language pairs or multilingual scenarios without further empirical validation.

We acknowledge that this study reports results based on a fixed random seed under controlled experimental conditions. While this design allows isolation of architectural effects, future work will incorporate multiple runs and statistical significance testing to further validate the robustness of the observed differences.

Nevertheless, the proposed method may be less effective in scenarios where contextual cues are insufficient to resolve gender ambiguity, or in cases involving complex multi-entity discourse structures. In such situations, early attention reinforcement alone may not guarantee correct gender resolution. The primary contribution of this study lies in architectural and representational analysis rather than large-scale benchmark optimization.

Future work will focus on extending the evaluation to additional language pairs with different typological properties, as well as exploring multilingual and cross-lingual settings to assess the broader applicability of the proposed approach. Such investigations are necessary to establish whether the observed benefits of Attention-Based Signal Reinforcement generalize beyond the specific conditions examined in this study.

6. Declarations

6.1. Author Contributions

Conceptualization: A.W.,R.M.,M.L.K., and D.P.L; Methodology: A.W., M.L.K.; Software: A.W.; Validation: R.M.,M.L.K., and D.P.L; Formal Analysis: R.M.,M.L.K., and D.P.L; Investigation: A.W.; Resources: A.W.; Data Curation: A.W.,M.L.K., and D.P.L.; Writing Original Draft Preparation: A.W.,R.M.,M.L.K., and D.P.L.; Writing Review and Editing: A.W.,R.M.,M.L.K., and D.P.L; Visualization: A.W.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available (annotation paralel corpus, data test and code) on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] H. J. Yuan Gao, Feng Hou, "Pre-ordering representations improve low-resource neural machine translation and application in the Māori language.pdf," *Multimed. Tools Appl.*, vol. 85, no. 1, pp. 1–21, 2025, doi: 10.1007/s11042-026-21153-5.

[2] Z. Zhang, Y. Wu, J. Zhou, S. Duan, H. Zhao, and R. Wang, "SG-Net: Syntax Guided Transformer for Language Representation," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 44, no. 6, pp. 3285–3299, 2022, doi: 10.1109/TPAMI.2020.3046683.

[3] L. Kang, S. He, M. Wang, F. Long, and J. Su, "Bilingual attention based neural machine translation," *Appl. Intell.*, vol. 53, no. 4, pp. 4302–4315, 2023, doi: 10.1007/s10489-022-03563-8.

[4] Y. B. Kaya and A. C. Tantuğ, "Effect of tokenization granularity for Turkish large language models," *Intell. Syst. with Appl.,* vol. 21, no. Jan., pp. 200335–200335, 2024, doi: https://doi.org/10.1016/j.iswa.2024.200335.

[5] M. C. Roy, S. K. Bisoy, and P. K. Das, "A Diffusion Driven Multimodal Fusion Framework for Context Aware Sarcasm Detection via Sentiment Syntax Graph Modeling," *Arab. J. Sci. Eng.*, vol. 2025, no. Jan., pp. 1–15, 2025, doi: 10.1007/s13369-025-10848-w.

[6] S. Lankford, H. Afli, and A. Way, "Human Evaluation of English – Irish Transformer-Based NMT," *Inf.*, vol. 13, no. 7, pp. 1–19, 2022, doi: 10.3390/info13070309.

[7] L. S. Meetei, T. D. Singh, and S. Bandyopadhyay, "Exploiting multiple correlated modalities can enhance low-resource machine translation quality," *Multimed. Tools Appl.,* vol. 83, no. 5, pp. 13137–13157, 2024, doi: 10.1007/s11042-023-15721-2.

[8] C.-O. Truică, A.-I. Stan, and E.-S. Apostol, "SimpLex: a lexical text simplification architecture," *Neural Comput. Appl.,* vol. 35, no. 8, pp. 6265–6280, 2023, doi: 10.1007/s00521-022-07905-y.

[9] F. Algobaei, E. Alzain, E. Naji, and K. A. Nagi, "Gender Issues between Gemini and ChatGPT: The Case of English-Arabic Translation," *World J. English Lang.*, vol. 15, no. 1, pp. 1-9, 2024, doi: 10.5430/wjel.v15n1p9.

[10] J. D. Id and W. K. Id, "Beyond the spotlight : Unveiling the gender bias curtain in movie reviews," *PLoS One*, vol. 2025, no. Jan., pp. 1–22, 2025, doi: 10.1371/journal.pone.0316093.

[11] Z. M. Nia, A. Ahmadi, B. Mellado, J. Wu, and J. Orbinski, "Twitter-based gender recognition using transformers," *Math. Biosci. Eng.*, vol. 20, no. 9, pp. 15962–15981, 2023, doi: 10.3934/mbe.2023711.

[12] A. Vellintihun, T. Doren, and K. Amitab, "Neural machine translation systems for English to Khasi : A case study of an Austroasiatic language," *Expert Syst. Appl.,* vol. 238, no. A, pp. 1–17, 2024, doi: 10.1016/j.eswa.2023.121813.

[13] S. Das and J. H. Paik, "Context-sensitive gender inference of named entities in text," *Inf. Process. Manag.*, vol. 58, no. 1, p. 1-23, 2021, doi: https://doi.org/10.1016/j.ipm.2020.102423.

[14] Nurtamin, H. Abbas, E. Iswary, and M. Hasyim, "Gender Bias In Machine Translation (Google Translate) From Indonesian To English," *J. Posit. Sch. Psychol.*, vol. 6, no. 4, pp. 9754–9761, 2022.

[15] B. Paul, "Multimodal Machine Translation Approaches for Indian Languages : A Comprehensive Survey," *J. ofUniversal Comput. Sci.*, vol. 30, no. 5, pp. 694–717, 2024, doi: 10.3897/jucs.109227.

[16] S. Ahmad and A. Maaytah, "Evaluating Three Neural Machine Translation Platforms for English-Arabic Translation : A Comparative Study of Linguistic Accuracy and Cultural Fidelity," *World J. English Lang.,* vol. 16, no. 2, pp. 1–14, 2026, doi: 10.5430/wjel.v16n2p1.

[17] C. Papakostas, C. Troussas, and A. Krouska, "A Hybrid Neuro-Symbolic Pipeline for Coreference Resolution and AMR-Based Semantic Parsing," *Inf.,* vol. 16, no. 7, pp. 1–20, 2025, doi: 10.3390/info16070529.

[18] L. M. Hernandez-Felipe, R. M. Ortega-Mendoza, F. Sánchez-Vega, F. A. Castro-Espinoza, and A. P. López-Monroy, "Detecting self-harm in social media using term weighting schemes based on the distance between words and personal pronouns," *Heal. Inf. Sci. Syst.,* vol. 13, no. 1, pp. 71-89, 2025, doi: 10.1007/s13755-025-00381-3.

[19] D. Guo, "Deep learning-driven context-aware English translation for ambiguous sentences Deep learning-driven context-aware English translation for ambiguous sentences," *Int. J. Inf. Commun. Technol.,* vol. 26, no. 15, pp. 41–56, 2025, doi: 10.1504/IJICT.2025.10071311.

[20] Y. Tian, "Decoding social group representation in American literature using contextualized embedding analysis and bias detection algorithms," *J. Comput. Methods Sci. Eng.,* vol. 0, no. 0, pp. 1-12, 2025, doi: 10.1177/14727978251393473.

[21] H. Wang, S. Ma, L. Dong, S. Huang, D. Zhang, and F. Wei, "DeepNet : Scaling Transformers to 1 , 000 Layers," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 46, no. 10, pp. 6761–6774, 2024, doi: 10.1109/TPAMI.2024.3386927.

[22] Q. Jie and C. Huang, "A Dual-Path Gated Attention-Based Deep Learning Model for Automated Essay Scoring Using Linguistic Features," *Int. J. Adv. Comput. Sci. Appl.,* vol. 16, no. 7, pp. 776–788, 2025, doi: 10.14569/issn.2156-5570.

[23] Y. Li, J. Li, J. Jiang, S. Tao, H. Yang, and M. Zhang, "P-Transformer: Towards Better Document-to-Document Neural Machine Translation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.,* vol. 31, no. 1, pp. 3859–3870, 2023, doi: 10.1109/TASLP.2023.3313445.

[24] S. Chauhan, P. Daniel, S. Saxena, A. Sharma, P. Daniel, and S. Saxena, "Fully Unsupervised Machine Translation Using Context-Aware Word Translation and Denoising Autoencoder," *Appl. Artif. Intell.,* vol. 36, no. 1, pp. 1770–1795, 2022, doi: 10.1080/08839514.2022.2031817.

[25] M. A. Ibrahim, "Prompt-Based Data Augmentation with Large Language Models for Indonesian Gender-Based Hate Speech Detection," J. Comput. Sci. Orig., vol. 28, no. 8, pp. 819–826, 2024, doi: 10.3844/jcssp.2024.819.826.

[26] L. V. Rodríguez, J. De La Rosa Yacomelo, R. R. González, and D. G. Segura, "Pronouns in Wayuunaiki and Spanish: A Contrastive Analysis," *Ikala,* vol. 27, no. 1, pp. 153 – 173, 2022, doi: 10.17533/udea.ikala.v27n1a08.

[27] C. Draude, G. Klumbyte, P. Lücking, and P. Treusch, "Situated algorithms: a sociotechnical systemic approach to bias," *Online Inf. Rev.*, vol. 44, no. 2, pp. 325–342, Nov. 2019, doi: 10.1108/OIR-10-2018-0332.

[28] T. Dolci, F. Azzalini, and M. Tanelli, "Improving Gender - Related Fairness in Sentence Encoders : A Semantics - Based Approach," *Data Sci. Eng.*, vol. 8, no. 2, pp. 177–195, 2023, doi: 10.1007/s41019-023-00211-0.

[29] R. AL-JARF, "Grammatical agreement errors in L1/L2 translations," *IRAL - Int. Rev. Appl. Linguist. Lang. Teach.*, vol. 38, no. 1, pp. 1–16, 2000, doi: doi:10.1515/iral.2000.38.1.1.

[30] M. Lardelli and M. Lardelli, "Gender-fair translation : a case study beyond the binary," *Perspectives (Montclair).*, vol. 6623, no. 32, pp. 1-12, 2024, doi: 10.1080/0907676X.2023.2268654.

[31] X. Wu, Y. Xia, J. Zhu, L. Wu, S. Xie, and T. Qin, "A study of BERT for context - aware neural machine translation," *Mach. Learn*., vol. 111, no. 3, pp. 917–935, 2022, doi: 10.1007/s10994-021-06070-y.

[32] Y. Jiang and J. Niu, "ScienceDirect How are neural machine-translated Chinese-to-English short stories constructed and cohered ? An exploratory study based on theme-rheme structure," *Lingua*, vol. 273, no. 28, pp. 1–20, 2022, doi: 10.1016/j.lingua.2022.103318.

[33] H. H. Vu, "Context-Aware Machine Translation with Source Coreference Explanation," *Trans. Assoc. Comput. Linguist.*, vol. 12, no. 1, pp. 856–874, 2024, doi: 10.1162/tacl_a_00677.

[34] P. Nemani, Y. D. Joel, P. Vijay, and F. F. Liza, "Gender bias in transformers: A comprehensive review of detection and mitigation strategies," *Nat. Lang. Process. J.*, vol. 6, no. 1, pp. 1–14, 2024, doi: 10.1016/j.nlp.2023.100047.

[35] M. Bernagozzi, B. Srivastava, F. Rossi, and S. Usmani, "Gender Bias in Online Language Translators: Visualization, Human Perception, and Bias/Accuracy Tradeoffs," *IEEE Internet Comput.*, vol. 25, no. 5, pp. 53–63, 2021, doi: 10.1109/MIC.2021.3097604.

[36] A. Qorbani, R. Ramezani, A. Baraani, and A. Kazemi, "Neurocomputing Multilingual neural machine translation for low-resource languages by twinning important nodes," Neurocomputing, vol. 634, no. 1, pp. 1–18, 2025, doi: 10.1016/j.neucom.2025.129890.

[37] N. B. Rajaboina and T. P. Sariki, "Enhanced subtitle generation in videos: leveraging hybrid BERT–CNN–LSTM architecture for contextual understanding," *Int. J. Inf. Technol.*, vol. 17, no. 7, pp. 3947–3953, 2025, doi: 10.1007/s41870-025-02610-0.

[38] B. Gra, "Pragmatic Uses of Gestures in Brazilian Portuguese in Contexts of Negation," *Rev. Estud. da Ling.*, vol. 32, no. 2, pp. 560–577, 2024, doi: 10.17851/2237-2083.32.2.560.

[39] D. Saunders, R. Sallis, and B. Byrne, "Neural Machine Translation Doesn't Translate Gender Coreference Right Unless You Make It," *Proc. ofthe Second Work. Gend. Bias Nat. Lang. Process. Barcelona, Spain (Online), December 13, 2020*, vol. 2020, no. Oct., pp. 35–43, 2020, [Online]. Available: http://arxiv.org/abs/2010.05332

[40] S. Hu et al., "Neural Machine Translation by Fusing Key Information of Text," *Comput. Mater. Contin*., vol. 74, no. 2, pp. 2084–2815, 2023, doi: 10.32604/cmc.2023.032732.

[41] X. Su, X. Zhao, J. Ren, Y. Li, and M. Rätsch, "Pre-training neural machine translation with alignment information via optimal transport," *Multimed. Tools Appl*., vol. 83, no. 16, pp. 48377–48397, 2024, doi: 10.1007/s11042-023-17479-z.