PAMUKKALE ÜNİVERSİTESİ
Mühendislik Bilimleri Dergisi

Araştırma makalesi | Research paper

PAMUKKALE UNIVERSITY
Journal of Engineering Sciences

Cilt | Vol: XX • Sayı | No: X • ss | pp: XXXX-XXXX • Yıl | Year: 20XX

Assessing the role of key and value dimensionality in multi-head attention mechanisms

Ramazan Katırcı, Hilal Çelik, Nusret Gön

ARTICLE INFO

Dates:
Received: 31.03.2026
Accepted: 26.06.2026

Doi:
10.65206/pajes.1920115

Corresponding author:
Hilal Çelik
(hilalcelik@sivas.edu.tr)

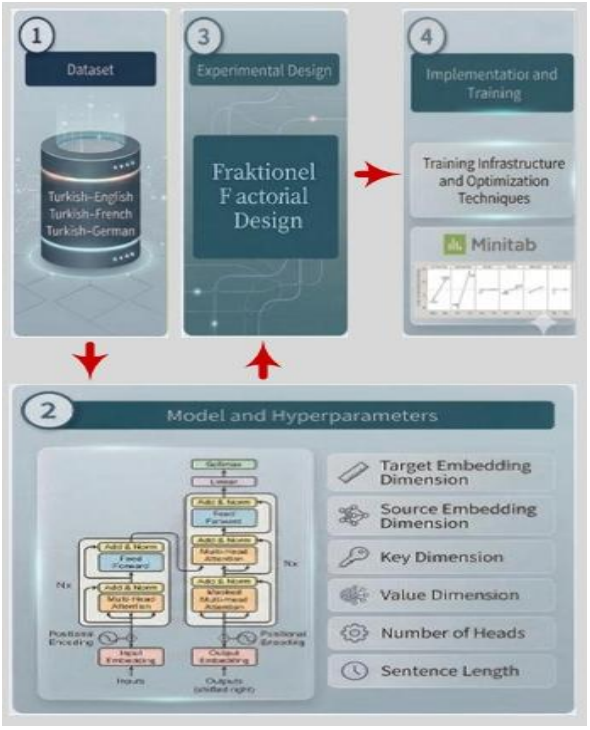
Author addresses:
Department of Computer Engineering, Faculty of Engineering and Natural Sciences, Sivas University of Science and Technology, Sivas, 58000, Türkiye
(ramazankatirci@sivas.edu.tr; hilalcelik@sivas.edu.tr; ngon@sivas.edu.tr)

ABSTRACT

Context—The transformer architecture has occupied a prominent position in recent research due to its attention-driven design; however, the role of asymmetric key and value dimensions within its attention mechanism remains largely underexplored in the literature despite their potential impact on model behavior and performance.

Objective—This study aims to systematically analyze the effects of key architectural parameters in the Transformer architecture on neural machine translation performance. These parameters include asymmetric key and value dimensions within the multi-head attention mechanism, embedding dimension, number of attention heads and sentence length.

Method—This study analyses the architectural components of the Transformer model across three neural machine translation datasets: English–Turkish, French–Turkish, and German–Turkish. Six fundamental hyperparameters—source and target embedding dimensions, key and value dimensions, the number of attention heads, and sentence length—were assessed at two levels each. To address the computational complexity of the Transformer's modular structure, a fractional factorial experimental design was employed, reducing the 64 possible configurations to 16 experimental runs. This approach significantly reduced the use of computational resources while ensuring a comprehensive analysis of both individual parameters and their pairwise interactions. All models were trained for 150 epochs using PyTorch with Distributed Data Parallel (DDP) and Automated Mixed Precision (AMP) in an HPC environment with NVIDIA A100 GPUs. Finally, statistical evaluation and performance visualization were conducted using Minitab.



Results—The results show that while establishing equivalent key and value dimensions serves as a stable default, it does not guarantee optimal performance across linguistic contexts. The model's response to asymmetrical configurations was highly language-specific, with notable variability observed in the interaction between the key dimension and the number of attention heads. Furthermore, increasing the target embedding dimension consistently improved generalization, whereas a larger source embedding dimension led to overfitting across all language pairs. These effects resulted from specific combinations of parameters that improved training accuracy but reduced validation performance, reflecting limitations in the generalization of language-dependent processes.

Conclusion—These results emphasize that optimal Transformer performance cannot be achieved through a universal hyperparameter configuration and must instead be tailored to the linguistic characteristics of each dataset.

Key Words—Asymmetrical dimensions, Deep Learning, Hyperparameter optimization, Key dimension, Multi-head attention, Neural machine translation, Transformer architecture, Value dimension.

I. INTRODUCTION

The transformer architecture [1] has emerged as one of the most influential breakthroughs in natural language processing (NLP). This architecture enhances the model's ability to represent semantic relationships and to handle complex linguistic structures across diverse NLP tasks [2]. In recent years, its strong ability to model long-range dependencies and its efficiency in training have made it the dominant framework in neural machine translation (NMT) [3], [4]. In this context, several studies have specifically evaluated the performance of transformer-based translation systems across various language pairs [5], highlighting their versatility and effectiveness in multilingual settings.

NMT has evolved significantly over the past decade. Early studies focused on recurrent neural network (RNN)-based encoder/ decoder architectures, which achieved promising results in sequence-to-sequence tasks [6], [7]. The introduction of the attention mechanism further enhanced these models by allowing the decoder to selectively focus on relevant parts of the input sequence [8]. However, a major paradigm shift occurred with the introduction of the transformer architecture proposed by Vaswani et al. [1], which replaced recurrence and convolution with self-attention. This design enabled more efficient parallelization and improved the ability to model long-range dependencies. Since its introduction, the transformer has become the backbone of state-of-the-art NMT systems and has strongly affected the development of large-scale language models [3], [9].

A considerable number of studies have focused on understanding the role of attention heads within the multi-head attention mechanism. The main objective of these works is to analyze how the heads contribute to the overall performance of the transformer. Some studies reported that many heads are not critical and can be removed without significant degradation in performance [10], whereas others found that certain heads are essential while the rest can be pruned, emphasizing that not all heads contribute equally [11]. In recent years, research has increasingly focused on disentangling and rethinking the design of queries, keys, and values within attention modules. Ainslie et al. (2023) propose multi-query attention (MQA) for faster decoding, while introducing grouped-query attention (GQA) to maintain quality with intermediate single key-value (KV) heads [5]. Another study by Meng et al. (2024) proposes a novel approach in which the key component in the self-attention mechanism of vision transformers is disentangled from the query and value components and modeled independently. Instead of a standard linear transformation, the key is represented as a manifold composed of multiple local charts, each processed by multilayer perceptrons (MLPs) and fused into a meta-representation. This geometry-aware formulation enables richer key representations and improves model performance [12]. Lu et al. (2024) leverage multi-head attention in LONGHEADS by distributing important context chunks across heads based on query-key correlations, enabling LLMs to efficiently process much longer inputs without extra training [13]. Jin et al. (2025) introduced the Mixture-of-Head (MoH) mechanism, in which each Query dynamically selects the most relevant heads, only the corresponding KV pairs are utilized, and the resulting Value outputs are integrated through a weighted summation [14]. More recently, Ji et al. (2025) offered the Multi-Head Latent Attention (MLA) architecture to reduce memory and computational costs during inference. Compared to standard Multi-Head Attention (MHA) and its variants (e.g. GQA), MLA achieves more economical performance by compressing the KV cache into a latent vector representation, which serves as a lower-dimensional and summarized form of the original KV cache [15].

Table 1 provides a comparative overview of recent variants of the multi-head attention mechanism. While the original transformer (Vaswani et al., 2017) employs fixed-sized heads operating in parallel, subsequent studies have introduced modifications to improve efficiency, scalability, or contextual representation. For example, Ainslie et al. (2023) offered the GQA approach to reduce memory usage, whereas Lu et al. (2024) and Wang et al. (2024) enhanced head utilization for long contexts and richer key representations, respectively. More recent designs, such as MLA and MoH, focus on compressing KV pairs and enabling dynamic head selection to further optimize performance.

Understanding the internal dynamics of deep learning architectures, particularly the impact of individual components, has become an increasingly vital research focus in recent years. This focus stems from the fact that model performance is affected not only by individual components but also by their complex interactions with one another and with hyperparameters. For example, in GRU-based architectures, İşlek et al. (2024) analyzed how vocabulary size, corpus size, the choice of optimizer (Adam or SGD), number of epochs, batch size, number of units, learning rate and embedding dimension jointly influence model accuracy [16]. Another line of research extends beyond factorial effect analysis by incorporating Response Surface Methodology (RSM) for hyperparameter optimization. For instance, Lujan-Moreno et al. (2022) employed a designed experimental framework (such as central composite designs) to construct a second-order regression model that captures both interaction and nonlinear effects of hyperparameters on model performance. This approach enables not only the assessment of parameter interactions but also the identification of optimal hyperparameter configurations via response surface modelling [17]. Similarly, Pannakkong et al. (2022) applied RSM to ANN, SVM, and DBN models, where a second-order response surface was constructed to model the relationship between hyperparameters and prediction error, facilitating efficient hyperparameter optimization [18]. This highlights the worth of considering both the effects of individual parameters and the combined effects of multiple parameters in deep learning studies. Building upon previous findings, this study focuses on transformer-specific components, including target embedding dimension, key dimension, value dimension, number of heads, and sentence length, and investigates both their individual impacts and pairwise interactions on translation quality, addressing a gap in the analysis of transformer architectures. This approach provides a more comprehensive understanding of the transformer's internal mechanisms.

Table 1. Comparison of multi-head attention variants.

Study	Approach	Query dimension	Key dimension	Value dimension	Innovation
Vaswani et al. – MHA	Transformer	d_{model} / h	d_{model} / h	d_{model} / h	Fixed size, all heads used in parallel
Ainslie et al. – GQA	Multi-Query	d_{model} / h	Shared (K)	Shared (V)	Converts multi-head to GQA, improves memory efficiency
Lu et al. – LONGHEADS	Long context	d_{model} / h	modified (correlation-based)	standard	Efficient head utilization for long sequences
Meng et al. – Manifold Key	Vision Transformer	d_{model} / h	manifold-expanded	$d_{m}od_{el}$ / h	Represents keys on a manifold for richer feature learning
Ji et al. – MLA	Latent KV	standard	standard	compressed latent representation	Reduces KV cache size for efficient inference
Jin et al. – MoH	Mixture-of-Head	d_{model} / h	d_{model} / h	d_{model} / h	Dynamic query–head selection, value-weighted routing

The attention mechanism is represented by three vectors for each token: a query (Q), a key (K), and a value (V). The transformer model employs a scaled dot-product operation to compute the similarity between queries and keys, which is then normalized using a SoftMax function to determine the attention weights assigned to the corresponding values. This process produces a weighted sum that forms the output representation. The multi-head attention mechanism extends this computation by utilizing multiple attention heads, each attending to a different contextual aspect. The integration of these parallel representations enables the model to jointly capture diverse linguistic dependencies, thereby enhancing its semantic understanding [1], [19].

In view of the advantages of holistic optimization in deep learning architectures, recent research has increasingly emphasized the need to analyze the joint behavior of multiple hyperparameters rather than isolating them [20]. However, within the transformer architecture, such joint analyses remain limited due to the model's inherent complexity and the intricate dependencies among its components. Consequently, the majority of extant studies have focused on individual hyperparameters, such as the learning rate [21]-[23] or the activation function [24], [25], without analyzing how architectural parameters collectively influence model performance. While analyzing the interactions among components is relatively straightforward in conventional deep learning models [20], the same process becomes considerably more challenging in transformer architectures due to their large number of interdependent components. To address this complexity, the present study employed a fractional factorial design, which reduced 64 possible parameter combinations to 16 experimental runs, enabling an efficient yet comprehensive analysis.

The novelty of this study lies in its systematic exploration of asymmetric key and value dimensions within the Transformer's multi-head attention mechanism. Whilst preceding studies generally presuppose uniform dimensions, this investigation explores the impact of disparate KV configurations on translation efficacy. The analysis is further expanded to encompass additional architectural factors, including embedding dimension, number of heads and sentence length. This design-based approach facilitates a more holistic understanding of how parameter interactions shape the performance of Transformer-based NMT models across multiple language pairs.

The remainder of this paper is organized as follows: Section II reviews related studies on MHA and hyperparameter optimization. Section III explains the methodology and experimental design. Section IV presents the results and analysis, and Section V discusses key findings. Section VI concludes the paper and outlines future research directions.

II. METHODS

This section presents a detailed description of the datasets, performance metrics, experimental design, and analytical methodology employed to evaluate the impact of transformer components on model performance and stability.

A. Dataset and setup

In this study, three NMT datasets were used for training and evaluation: the English–Turkish (EN–TR), French–Turkish (FR–TR), and German–Turkish (DE–TR) translation datasets. For each dataset, a random subset of 200,000 samples was selected to ensure balanced and representative data. Reproducibility was maintained by setting random_state = 42, a constant commonly adopted in the literature. The data were then divided into two subsets, with 80% used for training and 20% for validation, to facilitate model training and performance assessment. Each model was trained for 150 epochs to ensure stable convergence and consistent evaluation across all experiments.

B. Data preprocessing

The textual data was first cleaned using a predefined filter mask to remove punctuation and noise, and the dataset was constrained to a slice of 100,000 lines. Following this text cleaning step, a word-level tokenization approach was applied separately to the source and target languages, ensuring vocabulary consistency across translation pairs. The vocabulary size was deterministically established by the unique tokens remaining after the filtering process. To handle sequence boundaries, padding, and out-of-vocabulary words, special tokens—specifically <START>, <END>, <PAD>, and <UNK>—were incorporated into the vocabulary. Finally, to ensure uniform sequence lengths within each batch, padding or truncation was applied to fix the sequence length at a maximum of 40 tokens. Padded tokens were explicitly excluded from the loss computation during training.

C. Training configuration

The models were trained using the Adam optimizer with an initial learning rate of 10^{-4}. The learning rate was adaptively adjusted using a ReduceLROnPlateau scheduler, which reduces the learning rate by a factor of 0.5 if the validation loss does not improve for 10 consecutive epochs, with a minimum learning rate threshold of 10^{-5}. The batch size was set to 128, and training was conducted for 150 epochs to ensure stable convergence across all experimental configurations. Cross-entropy loss with label smoothing of 0.1 was used as the objective function, while padded tokens were excluded from loss computation.

The proposed transformer architecture consists of 6 layers for both the encoder and decoder stacks, utilizing a multi-head attention mechanism with 4 parallel attention heads. The dimensionality of the key and value representations, as well as the source and target embedding sizes, were uniformly configured to 256. The feed-forward network features a forward expansion of 2 resulting in a hidden layer dimension of 512 and employs the GELU (Gaussian Error Linear Unit) activation function. To ensure robust regularization, a dropout rate of 0.1 was applied to the attention layers and residual connections.

D. Transformer hyperparameters and experimental levels

Table 2 presents a systematic evaluation of the transformer hyperparameters and components, including target embedding dimension, key dimension, value dimension, number of heads, sentence length, and training-related settings. Both individual effects and pairwise interactions of these components were analyzed to assess their influence on translation performance, providing insights into how specific design choices impact model accuracy, training stability, and overall efficiency.

E. Experimental design

Given the transformer's highly modular architecture, which comprises multiple interdependent components such as embedding layers, attention heads, and feed-forward networks, conducting a full combinatorial exploration of all possible hyperparameter settings becomes computationally infeasible. Therefore, integrated analysis of its components remains a significant research challenge. To address this limitation, this study adopts a Design of Experiments (DOE) framework based on a fractional factorial design to systematically investigate the

Table 2. Hyperparameters and their levels.

Hyperparameters	Level 1	Level 2
Target Embedding Dimension	256	512
Source Embedding Dimension	256	512
Key dimension	256	512
Value dimension	256	512
Number of Heads	4	8
Sentence length	40	80

effects of transformer hyperparameters. A $2^{(6-2)}$ fractional factorial design was employed, reducing 64 full factorial combinations to 16 experimental runs. The analysis focused on main effects and two-factor interaction effects of the hyperparameters, which were estimated using ANOVA-based factorial decomposition.

F. Implementation and training

The models were implemented using PyTorch, chosen for its robust parallelization capabilities. Distributed Data Parallel (DDP) and Automatic Mixed Precision (AMP) were utilized to facilitate efficient training and optimize GPU memory usage.

In this study, statistical analysis and data visualization were performed using Minitab to evaluate the impact of architectural and optimization parameters on model performance. The experiments were conducted on a high-performance computing (HPC) system running Rocky Linux 8.5, powered by AMD EPYC 7543 processors and equipped with NVIDIA A100 GPUs connected via HDR InfiniBand for high-speed data transfer.

G. Evaluation metrics

Model performance was evaluated training accuracy, validation accuracy, training loss and validation loss. These metrics were selected because the primary objective of this study is not to maximize translation quality, but rather to systematically analyze the effects of transformer architectural components, particularly key and value dimensionality, on learning behavior. Accordingly, accuracy and loss metrics were considered appropriate for isolating and quantifying the effects of architectural variations across experimental configurations.

III. RESULTS and DISCUSSION

This section presents the results of hyperparameter optimization experiments conducted on transformer-based neural machine translation models for DE–TR, EN–TR, and FR–TR language pairs. The analysis focuses on six key hyperparameters and highlights the impact of asymmetrical Key (Dim_Key) and Value (Dim_Val) dimensions on model performance. A particular focus is placed on analyzing the impact of asymmetrical dimensions for the Dim_Key and Dim_Val vectors within the MHA mechanism, deviating from their conventional identicality. The analysis is based on four key response variables: training accuracy, validation accuracy, training loss, and validation loss.

A. Analysis of main effects for the DE-TR pair

As illustrated in Fig. 1, the plots of the main effects presented demonstrate a clear hierarchy of influence for the tested hyperparameters on model performance. The analysis of training and validation metrics indicates that the dimensions of the embedding layers are the most critical factors, while other parameters exhibit more subtle effects.

The Target Embedding Dimension (Targ_Emb_Dim) emerged as the most influential factor across all four response variables. As demonstrated in the plot (see Fig. 1), increasing Targ_Emb_Dim from 256 to 512 resulted in the most significant performance gain: it dramatically increased both training accuracy (from approx. 55% to 69.5%, see Fig. 1a) and validation accuracy (from approx. 29.0% to 30.0%, see Fig. 1b). Concurrently, it was observed that the sharpest decrease in both training loss and validation loss. This finding underscores that a richer representation space for the target language (Turkish) is crucial for the model's learning and generalization capabilities in the DE-TR context. The effect of Source Embedding Dimension (Sour_Emb_Dim) presents a clear case of overfitting. Increasing the Sour_Emb_Dim from 256 to 512 led to a strong improvement in training metrics (higher accuracy, lower loss). However, this trend was reversed for validation metrics, where the same change resulted in a significant decrease in validation accuracy and an increase in validation loss. This suggests that while a larger source embedding dimension helps the model memorize the training data, it fails to generalize well to unseen data for this language pair, potentially capturing noise rather than meaningful linguistic features.

In the context of the DE-TR task, both Key and Value Dimensions (Dim_Key and Dim_Val) demonstrated minimal impact on performance, especially when compared to the embedding dimensions. Their main effects plots for both training and validation metrics are nearly flat, hovering close to the mean. For instance, Dim_Key shows a slight negative effect on validation accuracy, while Dim_Val has almost no effect on validation loss. This suggests that, for DE-TR translation, the model's performance is not highly sensitive to the internal dimensions of

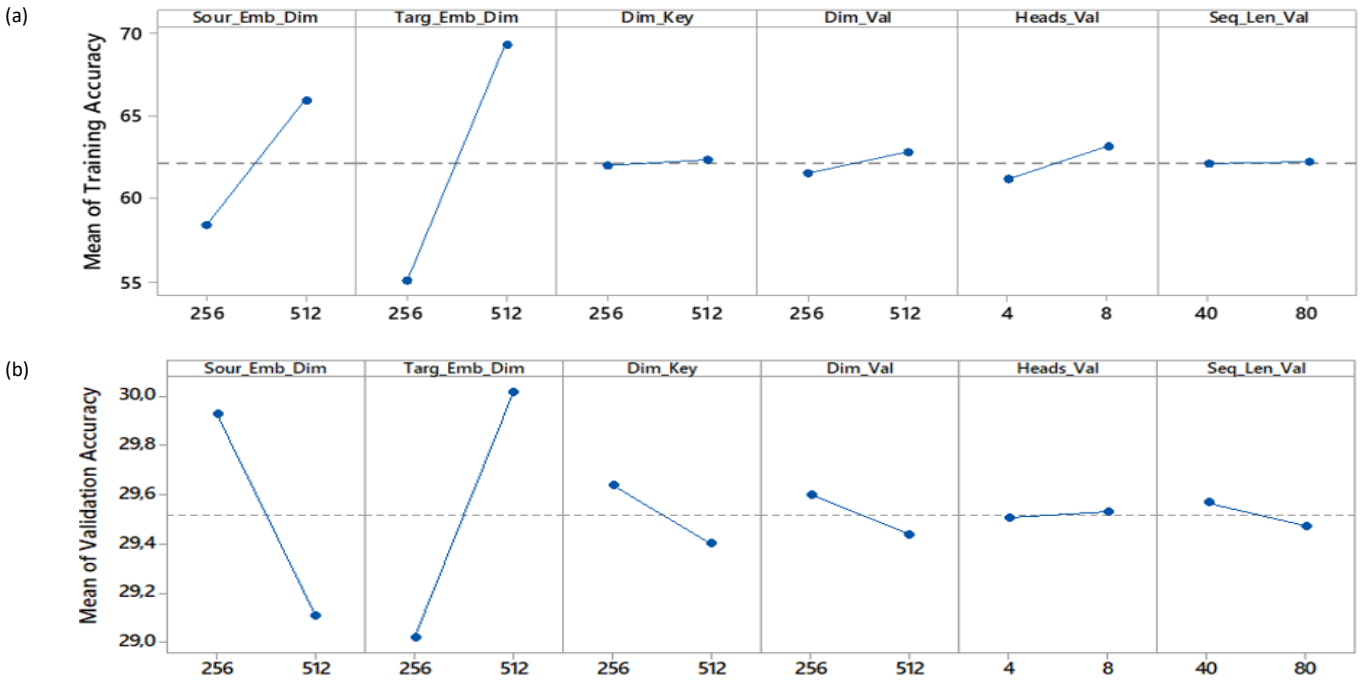


Fig. 1. Main effects plots of training and validation accuracy for the DE–TR model: (a) Training accuracy, (b) Validation accuracy.

the attention mechanism's Dim_Key and Dim_Val vectors within the tested range.

Increasing the number of attention heads from 4 to 8 yielded a moderate improvement in training accuracy and a corresponding decrease in training loss. However, its effect on validation metrics was negligible. This indicates that more heads help the model learn more complex patterns in the training set but do not necessarily translate to better generalization performance.

The sequence length (Seq_Len_Val) showed almost no discernible effect on any of the performance metrics, with its effect line remaining flat across all plots. This implies that within the tested range of 40 to 80 tokens, the model's performance is robust and largely independent of sentence length for the DE-TR dataset.

B. Analysis of main effects for the EN-TR pair

The hyperparameter sensitivity analysis for the EN-TR translation model, detailed illustrated in Fig. 2, reveals trends that are broadly consistent with the DE-TR pair, yet with subtle distinctions. The embedding dimensions continue to be the primary drivers of performance, while the internal dimensions of the attention mechanism show a more nuanced impact.

The analysis of embedding dimensions reveals distinct and contrasting roles for the target and source languages. The Target Embedding Dimension (Targ_Emb_Dim) is unequivocally the most powerful hyperparameter. The value increased from 256 to 512 results in the largest performance enhancement across the board, substantially elevating both training accuracy (from approx. 64.5% to 77.8%, see Fig. 2a) and validation accuracy (from approx. 33.4% to 34.5%, see Fig. 2b). This change also corresponds to the most significant reduction in training and validation loss. This reinforces the conclusion that a high-dimensional representation of the target language is a critical prerequisite for achieving optimal performance.

In contrast, the Source Embedding Dimension (Sour_Emb_Dim) exhibits an overfitting behavior in the EN-TR model. Increasing this dimension from 256 to 512 improves training accuracy and reduces training loss, but it degrades validation performance, as evidenced by a drop in validation accuracy and a rise in validation loss. This consistent pattern across both DE-TR and EN-TR datasets suggests that for translating into Turkish, a more compact source language representation may promote better generalization by preventing the model from learning spurious correlations in the training data.

The influence of the Key and Value dimensions is more discernible in the EN-TR model compared to the DE-TR model, though still secondary to embedding sizes. Dim_Key has a negligible effect on training metrics but shows a slight negative impact on validation accuracy and a moderate positive impact (worsening) on validation loss. In contrast, Dim_Val has a moderate positive effect on training accuracy and a negative effect on training loss, yet it shows a clear negative impact on validation accuracy while having no effect on validation loss. This asymmetrical and frequently detrimental impact on validation suggests that larger internal attention dimensions might introduce unnecessary complexity, hindering generalization for this specific task.

The number of attention heads has a moderately positive effect on all metrics, including a slight improvement in validation accuracy and a slight decrease in validation loss. This indicates that, for the EN-TR pair, increasing representational subspaces via more heads is beneficial not only for fitting the training data but also for generalization. Similar to the DE-TR model, Seq_Len_Val exerts almost no influence on model performance, with its effect plots remaining largely flat across all metrics.

C. Analysis of main effects for the FR-TR pair

The analysis of the FR-TR model (Fig. 3) introduces distinct patterns, particularly concerning the internal architecture of the attention mechanism, while reaffirming the foundational role of embedding dimensions.

Consistent with the previous language pairs, Targ_Emb_Dim remains a critical factor for generalization, being the only parameter to cause a substantial decrease in validation loss and a corresponding increase in validation accuracy (see Fig. 3b). The overfitting pattern associated with Sour_Emb_Dim also persists; increasing its size improves training metrics (see Fig. 3a) at the expense of validation performance. For the FR-TR pair, both source and target embedding dimensions exhibit a very strong positive effect on training accuracy, highlighting their importance in the initial learning phase.

A unique trend emerges with Dim_Key in the FR-TR model. Unlike in the other pairs where its effect was negligible on training, increasing Dim_Key from 256 to 512 here leads to a noticeable *decrease* in training accuracy and an *increase* in training loss. This suggests that a larger key dimension actively hinders the model's ability to learn from the FR-TR training data. On the validation

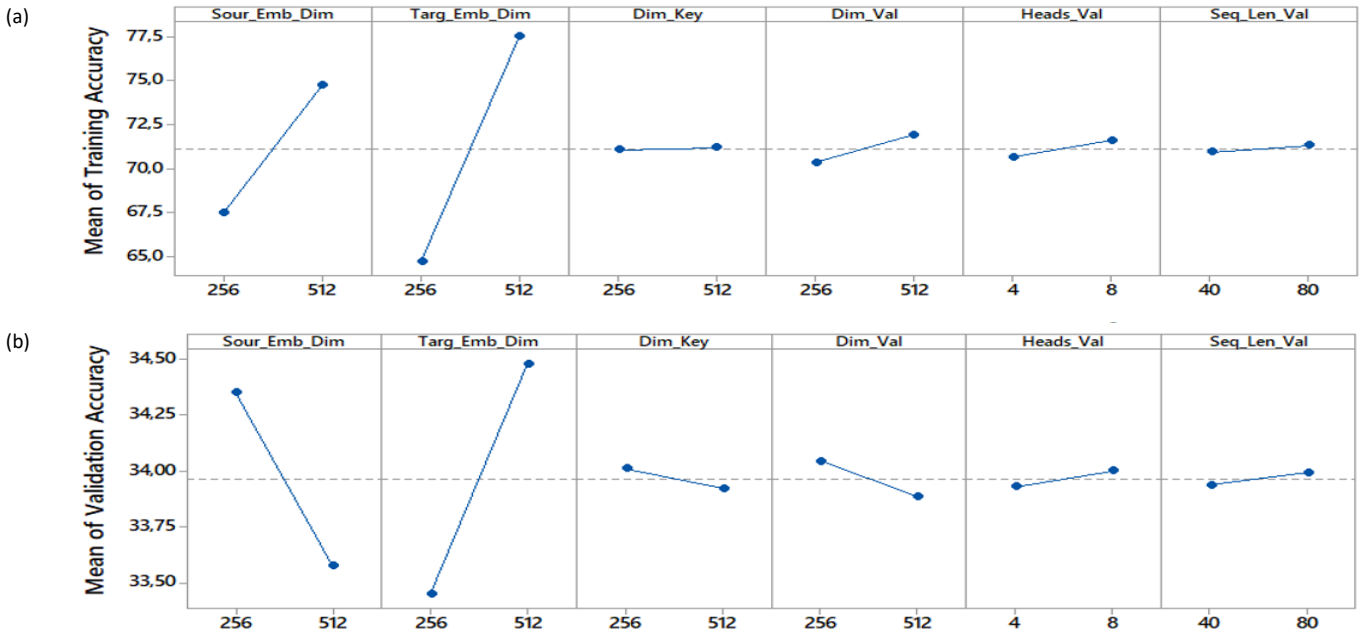


Fig. 2. Main effects plots of training and validation accuracy for the EN–TR model: (a) Training accuracy, (b) Validation accuracy.

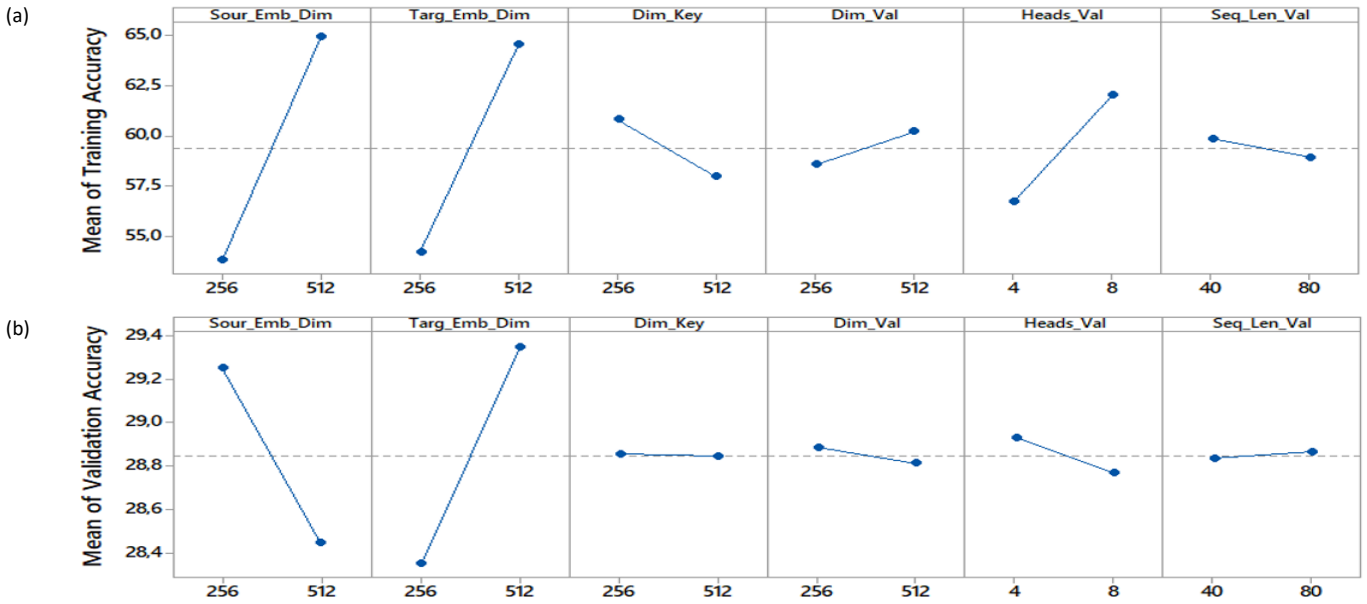


Fig. 3. Main effects plots of training and validation accuracy for the FR–TR model: (a) Training accuracy, (b) Validation accuracy.

set, its effect is neutral on accuracy and slightly detrimental to the loss. The effect of Dim_Val is comparatively minor. It provides a slight improvement in training metrics but has a small negative impact on validation accuracy, with a negligible effect on validation loss.

This parameter shows the most pronounced overfitting behavior besides Sour_Emb_Dim. Increasing the number of heads from 4 to 8 provides a strong boost to training performance, second only to the embedding dimensions. However, this gain is completely reversed on the validation set, where more heads lead to a clear decrease in accuracy and an increase in loss. This indicates that for the FR-TR dataset, a more complex attention configuration with 8 heads learns dataset-specific noise, and a simpler 4-head architecture generalizes more effectively. The sequence length shows a very slight negative trend on training accuracy but has no significant impact on any other metric, confirming its limited influence on overall performance within the tested range.

D. Comparative analysis of hyperparameter impact across language pairs

A comparative analysis of the main effects across the DE-TR, EN-TR, and FR-TR language pairs reveals both universal principles and language-specific sensitivities in the transformer model's architecture. While the foundational importance of embedding dimensions is a consistent theme, the optimal configuration of the internal attention mechanism appears to be highly contingent on the source language.

Across all three datasets, two hyperparameters consistently demonstrated dominant and predictable effects:

- *Primacy of Target Embedding Dimension (Targ_Emb_Dim)*: The most significant and universally positive factor for model performance was the target embedding dimension. For all three language pairs, increasing Targ_Emb_Dim from 256 to 512 resulted in the most substantial improvements in both training and validation accuracy, along with the sharpest reductions in loss. This universally strong correlation underscores that a rich, high-dimensional representation space for the target language (Turkish) is a non-negotiable prerequisite for effective translation, regardless of the source language's linguistic family.
- *Overfitting Tendency of Source Embedding Dimension (Sour_Emb_Dim)*: A clear and consistent pattern of overfitting was associated with the source embedding dimension. In every language pair, increasing Sour_Emb_Dim from 256 to 512 improved the model's performance on the training set but invariably degraded its performance on the validation set. This robust finding suggests a general principle for translating into Turkish from these Eurepean languages: A more compact source representation (256-dim) fosters better generalization, likely by forcing the model to learn more robust features instead of memorizing training set artifacts.
- *Neutrality of Sequence Length (Seq_Len_Val)*: The sequence length was consistently found to be a neutral factor, showing no significant impact on any performance metric for any of the language pairs within the tested 40-80 token range.

The most striking differences among the models were observed in their sensitivity to the hyperparameters governing the MHA mechanism (Dim_Key, Dim_Val, Heads_Val).

- *The Inconsistent Role of Dim_Key*: The effect of the key dimension varied significantly. For the DE-TR pair, its impact was negligible. For the EN-TR pair, it had a minor negative effect on validation metrics. Most notably, for the FR-TR pair, a larger Dim_Key actively hindered the model's learning process, uniquely causing a drop in *training* accuracy. This suggests that the complexity required for the query-key matching process may vary across the evaluated language pairs within the scope of this study.
- *The Overfitting Behavior of Heads_Val*: The number of attention heads also showed divergent effects on generalization. While more heads (8 vs. 4) moderately improved training performance across all pairs, their effect on validation varied:
  - DE-TR: Neutral effect on validation.
  - EN-TR: Slightly positive effect on validation, the only case where more heads aided generalization.
  - FR-TR: Strong negative effect on validation, indicating that a more complex, multi-headed attention structure led to significant overfitting for this pair.

This variability implies that the optimal number of representational subspaces (heads) is not universal but is instead tied to the specific linguistic characteristics of the source-target language pair.

E. Discussion of potential hyperparameter interactions and overfitting

The interaction plots provide critical insights into how pairs of hyperparameters jointly influence model performance, revealing complex relationships that are not apparent from main effects

alone. The visualizations facilitate the identification of specific parameter combinations that contribute to overfitting by illustrating how the impact of one hyperparameter varies depending on the level of another. The complete set of interaction plots for both training and validation accuracy across all tested hyperparameters is provided for all language pairs analyzed in this study, in the file designated as the 'Supplementary Materials'. A consistent pattern observed during training across all language pairs is a synergistic interaction between the target embedding dimension (Targ_Emb_Dim) and other parameters. The beneficial effects of increasing other parameters are often amplified when Targ_Emb_Dim is set to its higher level (512). For example, in the DE-TR training accuracy plot, the positive impact of increasing Dim_Key is significantly more pronounced when Targ_Emb_Dim is 512 compared to when it is 256. Similarly, for the FR-TR model, the performance boost from increasing Heads_Val to 8 is substantially greater with a Targ_Emb_Dim of 512. This suggests that a large target representation space is a foundational element that enables the model to effectively leverage other architectural complexities during the learning phase.

The most crucial findings emerge from comparing the training interactions with their validation counterparts, as this directly exposes the parameter combinations that hinder generalization.

The FR-TR dataset provides the clearest example of a detrimental interaction leading to overfitting. A strong, crossing interaction is visible between Heads_Val and Dim_Key (see Fig. 4).

- During training, the combination of a low Dim_Key (256) and a high Heads_Val (8) results in a significant accuracy increase.
- However, this relationship reverses dramatically for validation accuracy. For the validation set, the combination of a high Dim_Key (512) and a high Heads_Val (8) is the most harmful, causing the steepest decline in performance. This demonstrates that overfitting in the FR-TR model is not merely an additive effect of complex parameters but is specifically triggered by the interaction of a high-dimensional key dimension (Dim_Key) with a large number of attention heads.

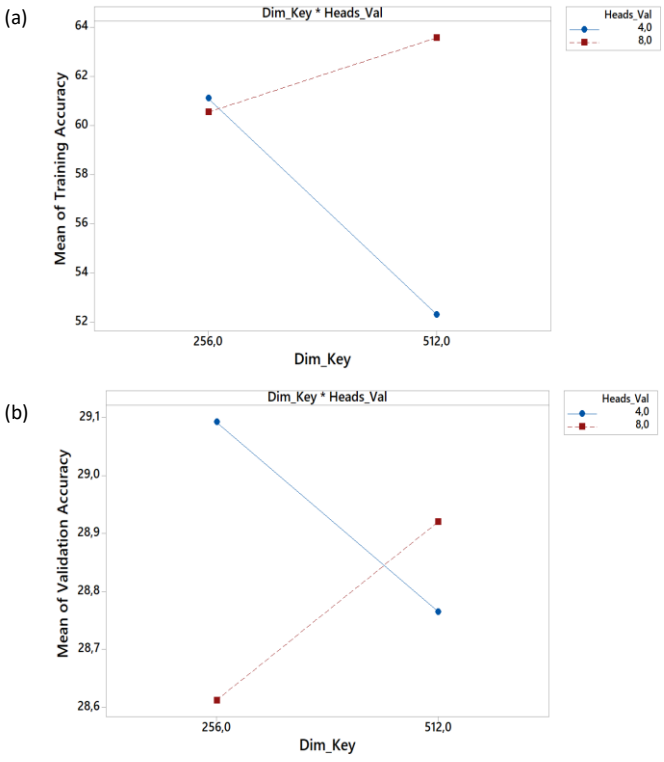


Fig. 4. Interaction plots of the FR–TR model, illustrating the joint effect of key dimension (Dim_Key) and number of attention heads (Heads_Val): (a) Training accuracy, (b) Validation accuracy

The interaction analysis between the key dimension (Dim_Key) and the number of attention heads (Heads_Val) reveals a significant joint influence on model performance. Similar interaction patterns in the FR–TR model are also evident in the EN–TR and DE–TR models. This consistency is reflected in both training and validation accuracies, suggesting that the combined effect of Dim_Key and Heads_Val exhibits a stable pattern across the three evaluated language pairs, with a relatively consistent influence on overall performance within the scope of this study.

F. Discussion on the role of asymmetrical key-value dimensions

A central motivation for this study was to challenge the conventional architectural assumption in the transformer model where the dimensions of the Key (d_k) and Value (d_v) vectors are typically kept identical. By decoupling these two hyperparameters, our experiments reveal that the impact of asymmetrical KV dimensions is not uniform but is highly dependent on the linguistic characteristics of the source-target language pair, offering new avenues for model optimization.

The results indicate that there is no universal benefit or detriment to using asymmetrical dimensions; rather, the optimal configuration is context specific. For the DE-TR model, the performance was largely insensitive to the dimensions of either Key or Value vectors. The main effects plots for both parameters were nearly flat, suggesting that for this language pair, the model is robust to such architectural variations, and asymmetrical configurations do not compromise performance.

In contrast, for the EN-TR pair, the effects became more pronounced and were generally detrimental to generalization. Increasing either Dim_Key or Dim_Val tended to degrade validation accuracy, suggesting that for EN-TR, larger and potentially asymmetrical internal attention dimensions introduce a complexity that hinders, rather than helps, the model's ability to generalize.

The most compelling evidence for the importance of this decoupling emerged from the FR-TR model. This was the only language pair where a specific architectural choice—a larger Dim_Key—was found to be actively harmful even during the training phase, uniquely causing a drop in training accuracy. Furthermore, the interaction analysis revealed that a specific combination of a high Dim_Val (512) and a high Heads_Val (8) was the primary trigger for severe overfitting in the FR-TR model. This strongly suggests that for certain language pairs, a more constrained attention mechanism, potentially with smaller or carefully chosen asymmetrical dimensions, is crucial for preventing the model from learning spurious correlations.

In summary, this study provides empirical evidence that the practice of setting d_k = d_v should be viewed as a safe default rather than a strict necessity. The performance sensitivity to asymmetrical Key and Value dimensions shows variation across the three evaluated language pairs, indicating that the observed effects are dependent on the specific architectural configurations considered in this study. Our findings indicate that for some translation tasks, particularly those prone to overfitting like FR-TR, deliberately constraining the model by using smaller or asymmetrical KV dimensions could be a viable strategy to improve generalization. This highlights the potential of exploring the d_k and d_v space independently as a valuable, albeit complex, dimension of hyperparameter tuning.

IV. CONCLUSION

The aim of this study was not to achieve state-of-the-art accuracy, but to conduct a foundational, exploratory investigation into the behavior of key hyperparameters within the Transformer architecture for neural machine translation. The primary objective was to systematically analyze the individual and joint

effects of these parameters on model performance, particularly in the context of limited data availability for DE-TR, EN-TR, and FR-TR language pairs.

Our findings reveal a clear distinction between universally impactful parameters and those with language-specific sensitivities. We consistently observed two dominant, opposing forces affecting generalization across all language pairs: a large target embedding dimension (Targ_Emb_Dim) was universally beneficial, acting as a prerequisite for effective learning, while a large source embedding dimension (Sour_Emb_Dim) was universally detrimental, consistently inducing overfitting.

The most novel contribution of this work lies in the deliberate decoupling and analysis of asymmetrical Key (d_k) and Value (d_v) dimensions—an area largely unexplored in existing literature. The results demonstrate that the conventional practice of setting d_k = d_v is a safe but not necessarily optimal default. The model's sensitivity to asymmetrical configurations was found to be highly language-pair dependent. For instance, the FR-TR model showed a unique vulnerability to specific combinations of attention-related hyperparameters, where a larger Dim_Key hindered learning and its interaction with other parameters triggered severe overfitting. This suggests that for certain linguistic contexts, a more constrained or deliberately asymmetrical attention mechanism could be a key strategy for improving generalization.

A limitation of this study is that the evaluation does not include explicit translation-quality metrics such as BLEU or chrF. While the selected metrics (training and validation accuracy and loss) are appropriate for analyzing architectural effects on learning dynamics and stability, they may not fully capture the linguistic quality of generated translations. Consequently, the reported findings primarily reflect model behavior from an optimization and convergence perspective rather than direct translation quality. Future work may integrate BLEU, chrF, or related metrics to enable a more comprehensive evaluation of translation performance and to complement the architectural analysis presented in this study.

Ultimately, this research emphasizes that hyperparameter tuning is not a one-size-fits-all process. Instead of presenting an optimized set of parameters, the study sheds light on the complex interplay and trade-offs inherent in the transformer architecture. The study serves as an empirical guide, emphasizing the importance of understanding these individual and interactive effects to design powerful models that are well-suited to the specific characteristics of the translation task at hand.

AUTHOR STATEMENT

Plagiarism Check—The article has been scanned with iThenticate and found to be compliant with the journal's plagiarism policy.
Conflict of Interest—There is no conflict of interest with any person/organization.
Ethics Committee Approval—Ethics committee approval is not required for this article.
Use of Artificial Intelligence Tools— Artificial intelligence tools were used only for minor language editing, grammar checking, and improving the clarity of the text. No AI tools were used for generating original scientific content, data analysis, results interpretation, or drawing conclusions.
Funding— This research received no external funding.
Data availability— The English–Turkish (EN–TR), German–Turkish (DE–TR), and French–Turkish (FR–TR) translation datasets used in this study were obtained from the Hugging Face Datasets repository.
CRediT Author Contribution— Conceptualization, Methodology, Supervision, Investigation, Software (**Ramazan Katırcı**); Formal analysis, Data Curation, Visualization, Writing – Original Draft, Writing – Review & Editing (**Hilal Çelik**); Validation, Data Curation (**Nusret Gön**).
Acknowledgment— The research utilized computational resources provided by the TÜBİTAK ULAKBİM High Performance and Grid Computing Center (TRUBA) and the Lütfi Albay Artificial Intelligence and Robotics Laboratory at Sivas University of Science and Technology.

REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, "Attention is all you need", *Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS)*, Long Beach, CA, USA, 04-09 December 2017.

[2] Y. Zhang, M. X. Tuo, Q. Y. Yin, L. Qi, X. X. Wang, T. Liu, "Keywords extraction with deep neural network model", *Neurocomputing*, 383, 113–121, 2020. https://doi.org/10.1016/j.neucom.2019.11.083.

[3] J. Devlin, M. W. Chang, K. Lee, K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", *Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Minneapolis, MN, USA, 02-07 June 2019, 4171-4186.

[4] Y. Guan, J. Whitehill, "Transformer-Encoder Trees for Efficient Multilingual Machine Translation and Speech Translation", *arXiv*, 2025. https://doi.org/10.48550/arXiv.2509.17930.

[5] J. Ainslie, J. Lee-Thorp, M. de Jong, Y. Zemlyanskiy, F. Lebrón, S. Sanghai, "GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints", *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*, Singapore, Singapore, 06-10 December 2023, 4895-4901. https://doi.org/10.18653/v1/2023.emnlp-main.298.

[6] N. Kalchbrenner, P. Blunsom, "Recurrent continuous translation models", *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, Seattle, Washington, USA, 18-21 October 2013, 1700-1709. https://doi.org/10.18653/v1/d13-1176.

[7] I. Sutskever, O. Vinyals, Q. V. Le, "Sequence to sequence learning with neural networks", *Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS 2014)*, Montreal, Canada, 08-13 December 2014.

[8] D. Bahdanau, K. Cho, Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate", *arXiv*, 2014. https://doi.org/10.48550/arXiv.1409.0473.

[9] D. Friedman, A. Wettig, D. Chen, "Learning Transformer Programs", *arXiv*, 2023. https://doi.org/10.48550/arXiv.2306.01128.

[10] P. Michel, O. Levy, G. Neubig, "Are sixteen heads really better than one?", *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NIPS 2019)*, Vancouver, Canada, 08-14 December 2019.

[11] E. Voita, D. Talbot, F. Moiseev, R. Sennrich, I. Titov, "Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned", *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, 28 July-02 August 2019, 5797-5808. https://doi.org/10.18653/v1/P19-1580.

[12] L. Meng, M. Goodwin, A. Yazidi, P. Engelstad, "A Manifold Representation of the Key in Vision Transformers", *Proceedings of the 41st Computer Graphics International Conference (CGI Annual)*, Geneve, Switzerland, 01-05 July 2024, 311-322. https://doi.org/10.1007/978-3-031-82021-2_22.

[13] Y. Lu, X. Zhou, W. He, J. Zhao, T. Ji, T. Gui, Q. Zhang, X. J. Huang, "LongHeads: Multi-Head Attention is Secretly a Long Context Processor", *arXiv*, 2024. https://doi.org/10.48550/arXiv.2402.10685.

[14] P. Jin, B. Zhu, L. Yuan, S. C. Yan, "MoH: Multi-Head Attention as Mixture-of-Head Attention", *arXiv*, 2025. https://doi.org/10.48550/arXiv.2410.11842.

[15] T. Ji, B. Guo, Y. Wu, Q. P. Guo, L. X. Shen, Z. Chen, X. P. Qiu, Q. Zhang, T. Gui, "Towards Economical Inference: Enabling DeepSeek's Multi-Head Latent Attention in Any Transformer-based LLMs", *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL Annual)*, Vienna, Austria, 27 July-01 August 2025, 33313–33328. https://doi.org/10.18653/v1/2025.acl-long.1597.

[16] B. İşlek, R. Katırcı, H. Celik, "Enhancing Question Answering Systems Through Optimal Hyperparameter Tuning in GRU", *8th International Artificial Intelligence and Data Processing Symposium (IDAP)*, Malatya, Türkiye, September, 21-22 September 2024. https://doi.org/10.1109/IDAP64064.2024.10710732.

[17] G. A. Lujan-Moreno, P. R. Howard, O. G. Rojas, D. C. Montgomery, "Design of experiments and response surface methodology to tune machine learning hyperparameters, with a random forest case-study", *Expert Systems with Applications*, 109, 195–205, 2018. https://doi.org/10.1016/j.eswa.2018.05.024.

[18] W. Pannakkong, K. Thiwa-anont, K. Singthong, P. Parthanadee, J. Buddhakulsomsiri, "Hyperparameter Tuning of Machine Learning Algorithms Using Response Surface Methodology: A Case Study of ANN, SVM , and DBN", *Mathematical Problems in Engineering*, 2022, 8513719, 2022. https://doi.org/10.1155/2022/8513719.

[19] S. Islam, H. Elmekki, A. Elsebai, J. Bentahar, N. Drawel, G. Rjoub, W. Pedrycz, "A comprehensive survey on applications of transformers for deep learning tasks", *Expert Systems with Applications*, 241, 122666, 2024. https://doi.org/10.1016/j.eswa.2023.122666.

[20] H. Çelik, R. Katırcı, B. İşlek, "Effect Of Parameters On Performance In Question-Answer Model With Simple Rnn Deep Learning Method", *International Conference on Scientific and Innovation Research-III*, Sivas, Türkiye, 03-05 May 2024. https://doi.org/10.5281/zenodo.11320381.

[21] R. Katırcı, H. Çelik, "Learning Rate Sensitivity In Transformer Models: A Case Study In Neural Machine Translation", *International Conference on Scientific and Innovation Research-IV*, Sivas, Türkiye, 30-31 May 2025. https://doi.org/10.5281/zenodo.15769066.

[22] N. Shazeer, M. Stern, "Adafactor: Adaptive learning rates with sublinear memory cost", *Proceedings of 35th International Conference on Machine Learning (ICML)*, Stockholm, Sweden, 10-15 July 2018.

[23] L. Y. Liu, H. M. Jiang, P. C. He, W. Z. Chen, X. D. Liu, J. F. Gao, J. W. Han, "On the Variance of the Adaptive Learning Rate and Beyond", *The Eighth International Conference on Learning Representations (ICLR 2020)*, Virtual Only Conference, 26 April-1 May 2020. https://doi.org/10.48550/arXiv.1908.03265.

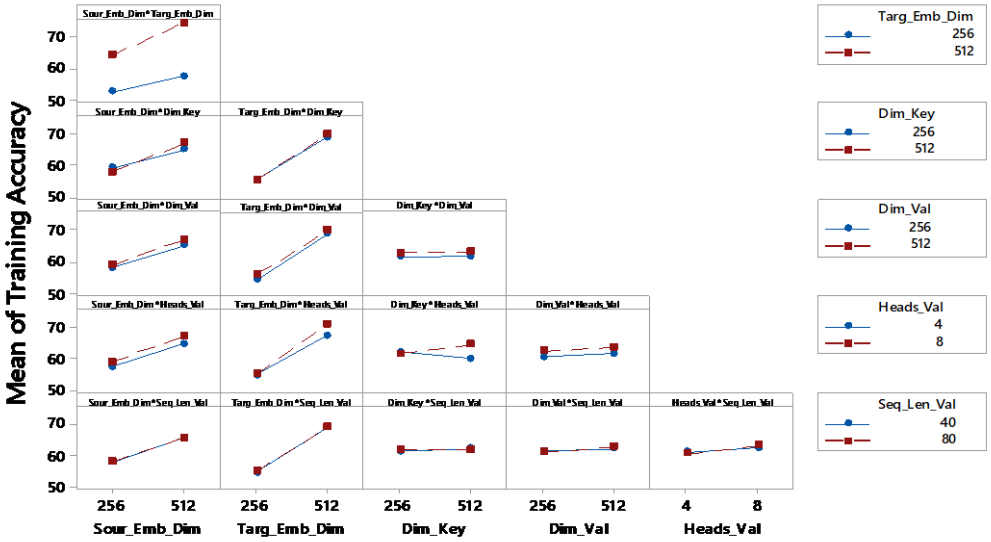
[24] K. Shen, J. L. Guo, X. Tan, S. L. Tang, R. Wang, J. Bian, "A Study on ReLU and Softmax in Transformer", *arXiv*, 2023. https://doi.org/10.48550/arXiv.2302.06461.

[25] H. S. Fang, J. U. Lee, N. S. Moosavi, I. Gurevych, "Transformers with Learnable Activation Functions", *Proceeding of the 17th Conference of the European-Chapter of the Association-for-Computational-Linguistics (EACL)*, Dubrovnik, Croatia, 02-06 May 2023, 2337-2353. https://doi.org/10.18653/v1/2023.findings-eacl.181.

APPENDIX

This supplementary section provides detailed training and validation accuracy plots for all three language pairs (DE–TR, EN–TR and FR–TR). Specifically, interaction plots are presented to demonstrate the combined effects of six core transformer hyperparameters—source embedding dimension, target embedding dimension, key dimension, value dimension, number of attention heads and sequence length—are presented to demonstrate how these architectural parameters jointly influence model performance. These figures support the main findings presented in Section III by providing additional insights into the learning dynamics and convergence behavior observed during both training and validation phases. As shown in Fig. A1–A3, the interactions between model parameters exhibit consistent trends across language configurations, highlighting nonlinear effects that underscore the importance of proper hyperparameter tuning for achieving optimal translation accuracy.

(a)



(b)

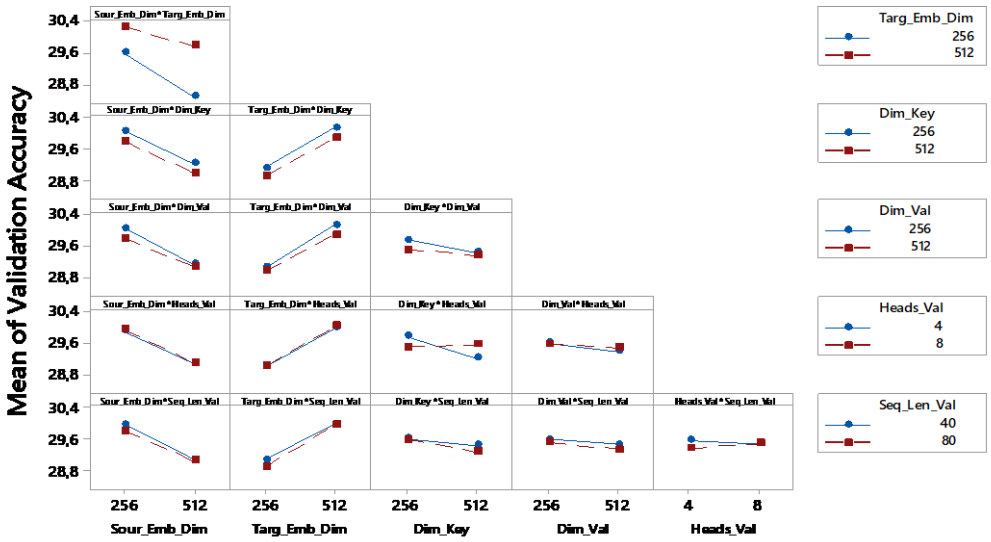


Fig. A1. DE-TR interaction plots for: (a) Training accuracy, (b) Validation accuracy.

(a)

(b)

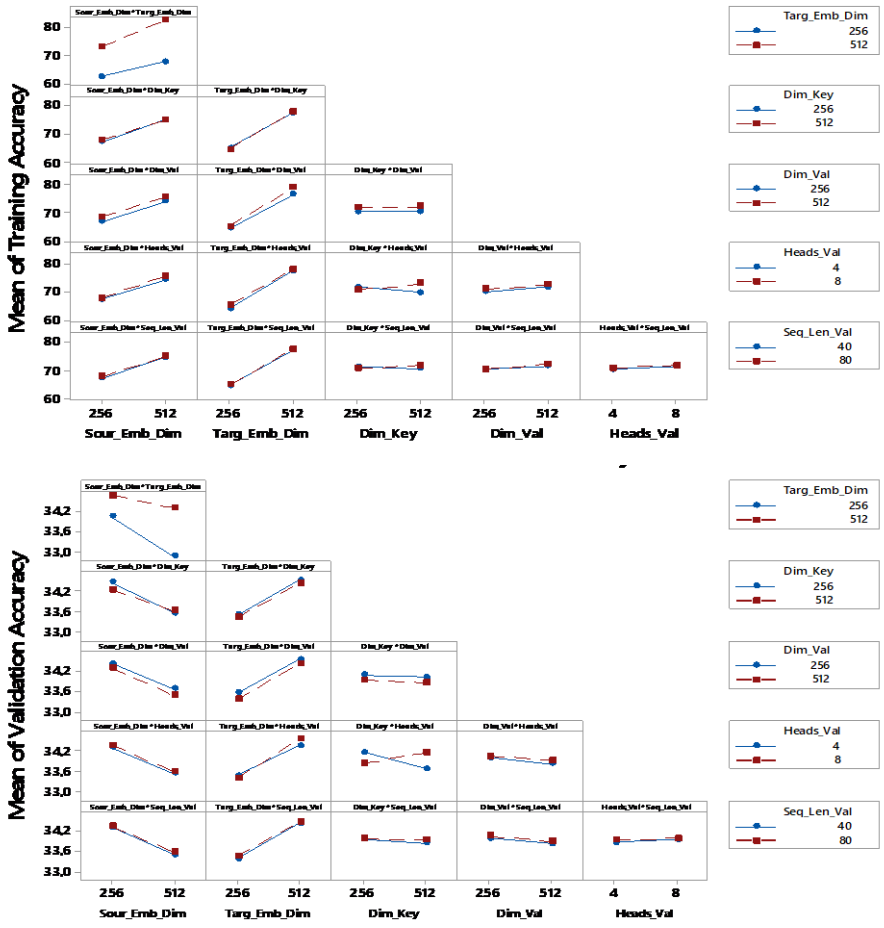


Fig. A2. EN-TR interaction plots for: (a) Training accuracy, (b) Validation accuracy.

(a)

(b)

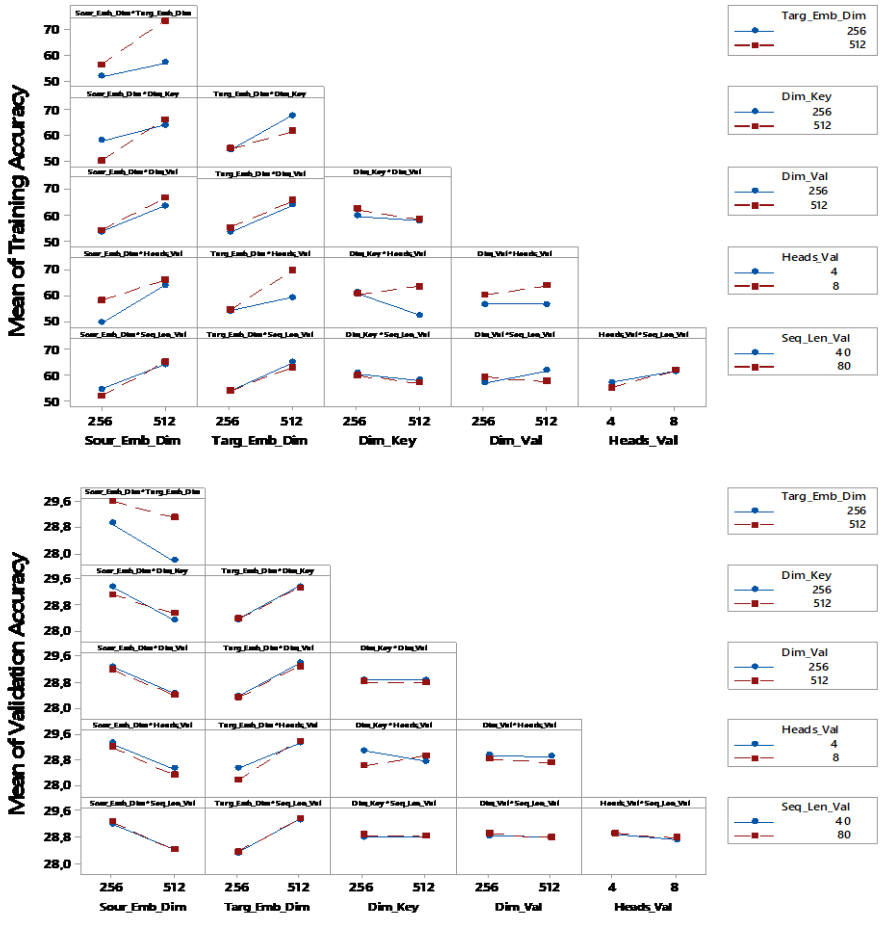


Fig. A3. FR-TR interaction plots for: (a) Training accuracy, (b) Validation accuracy.