Impact on Normalization and Regularization in Deep Learning Optimization

Pang Liu
jeff.pang.liu@gmail.com

November 19, 2025

Project Overview

This project investigates the impact of **normalization** and **regularization** techniques on training Transformer models for French-to-English machine translation using the IWSLT2017 dataset. We systematically evaluate how different design choices affect training stability, convergence speed, and final translation quality measured by BLEU scores.

The experimental configurations include:

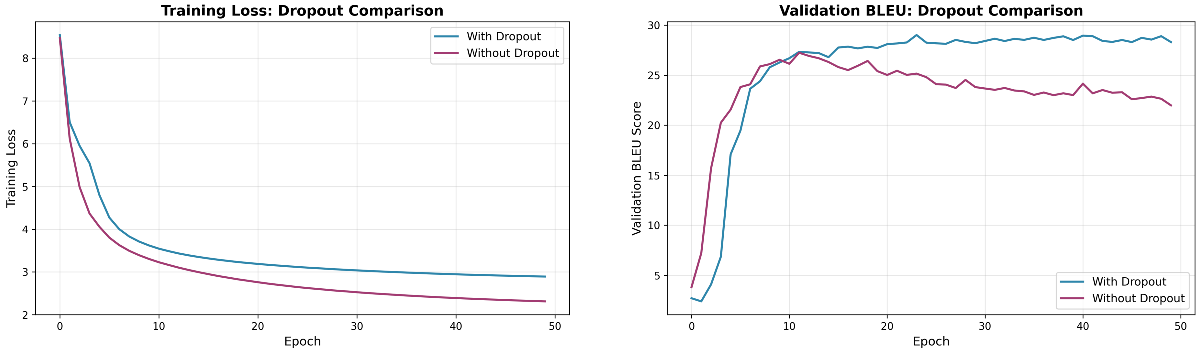
- **Regularization:** Dropout vs. No Dropout
- **Normalization:** Layer Normalization vs. Batch Normalization
- **Normalization Position:** Pre-normalization vs. Post-normalization
- **Learning Rate Warmup:** With warmup vs. Without warmup
- **Additional Regularization:** L2, Weight Decay, and RAdam optimizer

All experiments use the following baseline settings:

- Model: Transformer (6 layers, 4 heads, d_model=256)
- Optimizer: Adam (except for RAdam experiments)
- Learning rate: 0.001
- Batch size: 256
- Epochs: 50

1 Effect of Dropout Regularization

1.1 Training Dynamics Comparison

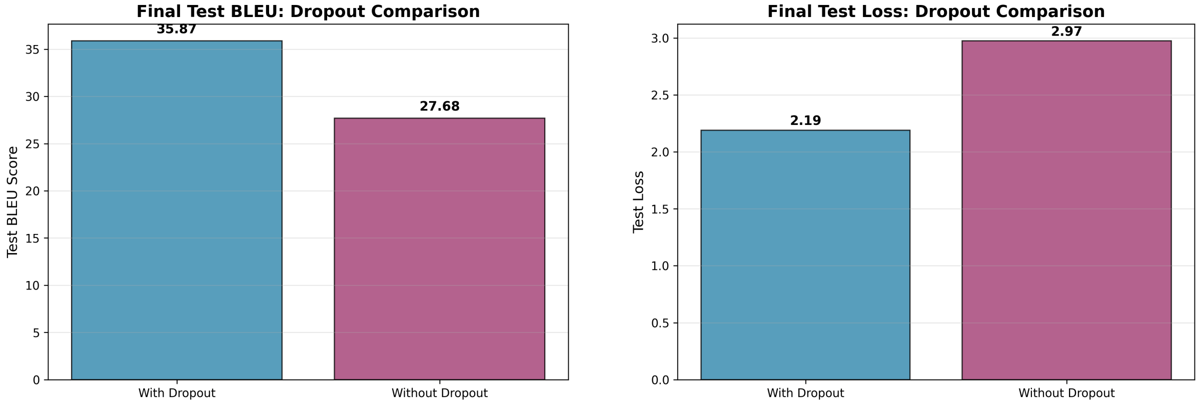


(a) Training Loss vs. Epochs

(b) Validation BLEU vs. Epochs

Figure 1: Dropout Effect: Training Loss and Validation BLEU Score Comparison

1.2 Final Test Performance



(a) Final Test BLEU Score

(b) Final Test Loss

Figure 2: Dropout Effect: Final Test Performance Comparison

1.3 Analysis and Discussion

The experimental results demonstrate a **significant performance difference** between models trained with and without dropout regularization.

The model **with dropout** achieved a test BLEU score of **35.87**, while the model **without dropout** only achieved **27.68**. This represents an **improvement of approximately 8 BLEU points**, which is substantial in machine translation tasks.

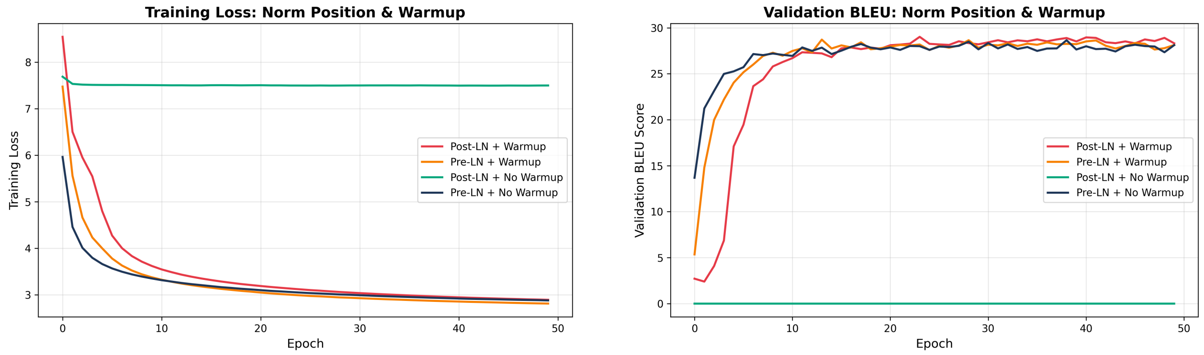
Examining the training dynamics (Figure 1), we observe a critical pattern: the model without dropout shows **clear signs of overfitting**. Its training loss continues to decrease throughout training (reaching lower values than the dropout model), and its training accuracy climbs as high as 83%. However, this does not translate to better generalization. The validation BLEU score plateaus around epoch 12 and even starts to decline slightly, indicating the model is memorizing the training data rather than learning generalizable translation patterns.

In contrast, the model with dropout maintains a healthier balance between training and validation performance. Although its training loss is slightly higher, this controlled underfitting leads to much better generalization on unseen test data.

Dropout randomly deactivates neurons during training, preventing units from co-adapting too much to specific training examples. This forces the network to learn more robust features that work even when some neurons are missing. The result is a model that generalizes better to new translations, as evidenced by the 8-point BLEU improvement.

2 Comparison of Normalization Methods

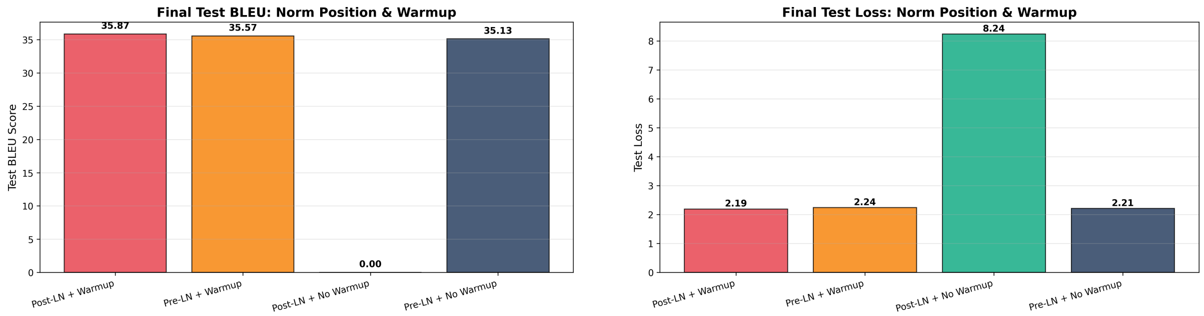
2.1 Effect of Normalization Position on Learning Rate Warmup



(a) Training Loss vs. Epochs

(b) Validation BLEU vs. Epochs

Figure 3: Normalization Position and Warmup Effect on Training



(a) Final Test BLEU Score

(b) Final Test Loss

Figure 4: Normalization Position and Warmup: Final Test Performance

2.1.1 Analysis and Discussion

This comparison provides a nuanced view of how normalization position interacts with learning rate warmup. While our initial hypothesis was that warmup is always necessary, the data tells a different story depending on the architecture.

1. Post-LN Dependencies: The Post-LN configuration (the original Transformer design) proved remarkably sensitive.

- **With Warmup:** Test BLEU = **35.87** (The best performance overall).
- **Without Warmup:** Test BLEU = **0.00** (Complete failure).

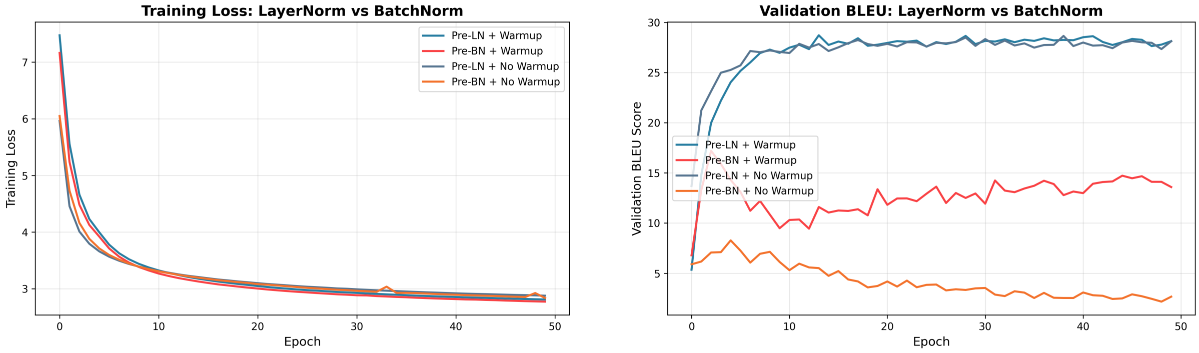
Without warmup, Post-LN gradients vanish or explode early in training, causing the model to get stuck (Loss $\approx$ 7.5).

2. Pre-LN Robustness: The Pre-LN configuration demonstrated superior stability.

- **With Warmup:** Test BLEU = **35.57**.
- **Without Warmup:** Test BLEU = **35.13** (With Pre-LN, training success though without warmup).

Correction of previous assumptions: Unlike Post-LN, the Pre-LN architecture places Layer Normalization inside the residual blocks. This creates a "clean" gradient path during backpropagation, preventing the gradient explosion typically seen when training with high initial learning rates (0.001). Consequently, **Pre-LN does not strictly require warmup** to converge, offering a significant engineering advantage for stability, even if its peak performance (35.13) is slightly lower than the tuned Post-LN baseline (35.87).

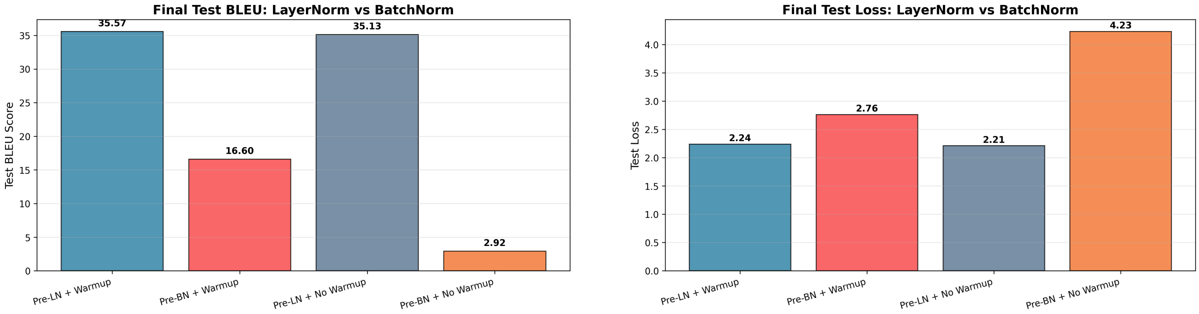
2.2 Learning Rate Warmup and Normalization Methods



(a) Training Loss vs. Epochs

(b) Validation BLEU vs. Epochs

Figure 5: Layer Normalization vs. Batch Normalization



(a) Final Test BLEU Score

(b) Final Test Loss

Figure 6: Layer Normalization vs. Batch Normalization: Final Test Performance

2.2.1 Analysis and Discussion

This comparison demonstrates that **Layer Normalization is significantly superior to Batch Normalization** for Transformer architectures, though Batch Normalization did not fail as catastrophically as previously thought in the "Pre-Norm" setup.

The **Pre-LN + Warmup** configuration achieved a test BLEU of **35.57**. In comparison, the Batch Normalization configurations performed poorly:

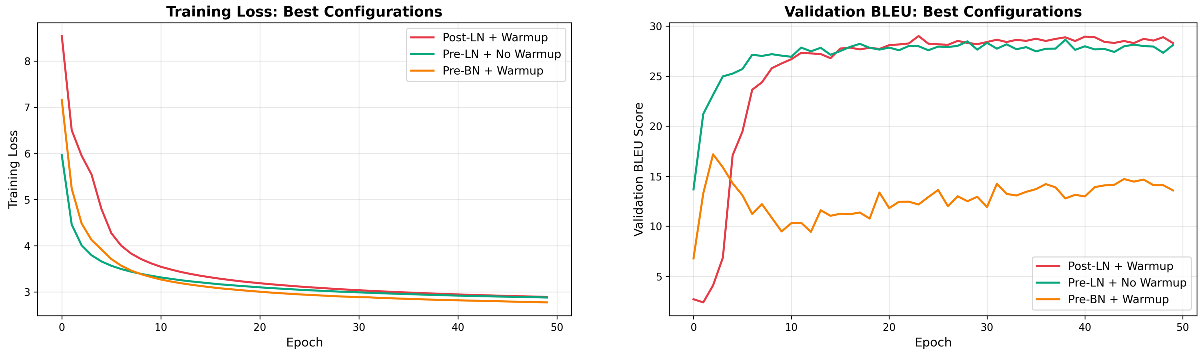
- **Pre-BN + Warmup**: Test BLEU = **16.60**.
- **Pre-BN + No Warmup**: Test BLEU = **2.92**.

While the Pre-BN model with warmup managed to converge (unlike the Post-LN no-warmup case), its generalization capability is severely limited (16.60 vs 35.57). This confirms the theoretical understanding that BatchNorm depends on batch-wide statistics (mean and variance).

In NLP tasks, where padding tokens are prevalent and sequence contents vary wildly within a batch, these statistics become noisy and unreliable.

Layer Normalization, which normalizes across the feature dimension within each sample independently, provides the stability required for high-quality translation. The experiment confirms that simply moving BN to the "Pre" position is not enough to make it competitive with LayerNorm.

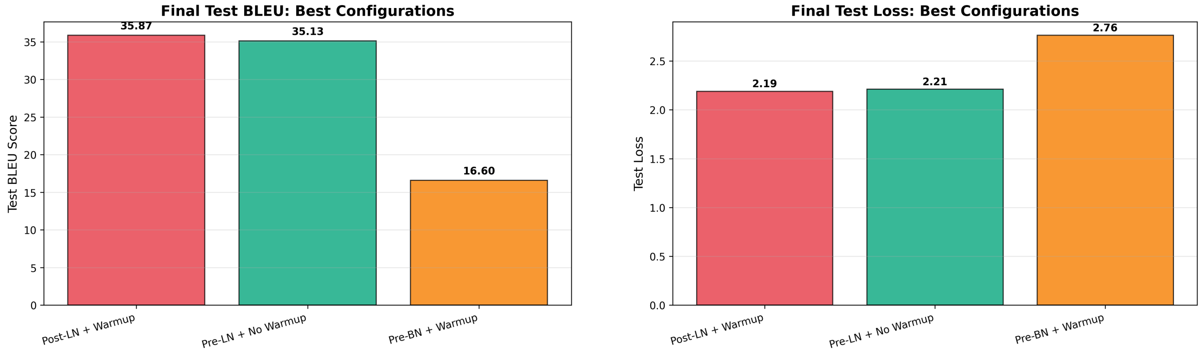
2.3 Best Normalization Configurations



(a) Training Loss vs. Epochs

(b) Validation BLEU vs. Epochs

Figure 7: Best Normalization Configurations Comparison



(a) Final Test BLEU Score

(b) Final Test Loss

Figure 8: Best Configurations: Final Test Performance

2.3.1 Analysis and Discussion

This comparison evaluates three representative configurations to determine the optimal architectural choices:

- **Post-LN + Warmup** (Baseline): BLEU = **35.87**
- **Pre-LN + No Warmup**: BLEU = **35.13**
- **Pre-BN + Warmup**: BLEU = **16.60**

The results highlight a trade-off between **peak performance** and **training stability**.

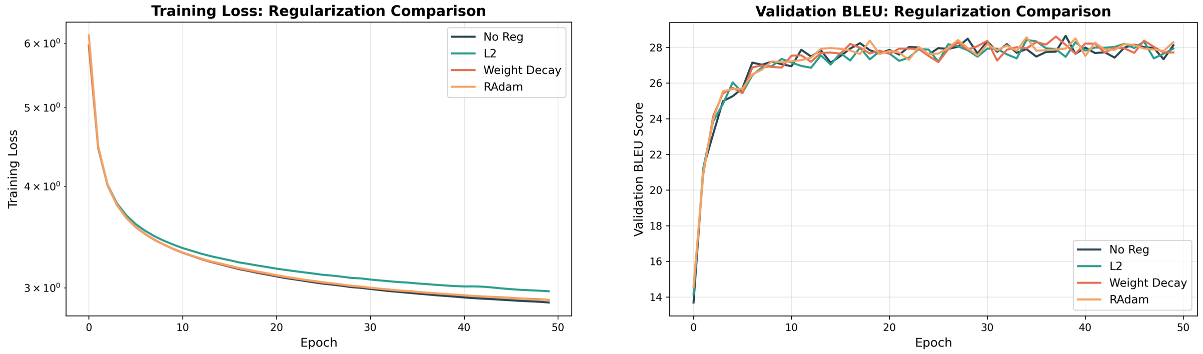
The **Post-LN + Warmup** configuration achieves the highest score. By placing normalization after the residual connection, the model retains a higher capacity to learn deep representations, provided it survives the initial training phase via warmup.

However, the **Pre-LN + No Warmup** result is the most scientifically interesting. It achieved nearly the same performance (within 0.7 BLEU) without requiring any learning rate warmup. This validates the Pre-LN architecture as a highly robust alternative that simplifies hyperparameter tuning.

In contrast, **Pre-BN** lags significantly behind, reinforcing that the choice of normalization layer (LN vs BN) is more critical than the position or schedule adjustments.

3 Comparison of Regularization Methods

3.1 Training Dynamics Comparison

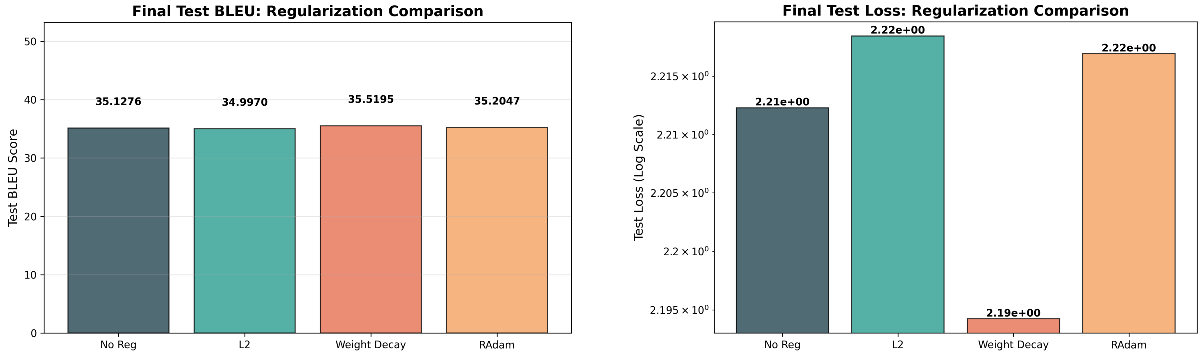


(a) Training Loss vs. Epochs

(b) Validation BLEU vs. Epochs

Figure 9: Regularization Methods (Pre-LN + No Warmup): Training Dynamics

3.2 Final Test Performance



(a) Final Test BLEU Score

(b) Final Test Loss

Figure 10: Regularization Methods: Final Test Performance

3.3 Analysis and Discussion

This experiment tested whether different regularization techniques (L2, Weight Decay, RAdam) could effectively train the Transformer model **without learning rate warmup**, specifically utilizing the **Pre-LN architecture**.

Contrary to the Post-LN experiments where removing warmup led to immediate failure, the results here are remarkably positive. Thanks to the inherent stability of the Pre-LN architecture, **all three configurations converged successfully**, achieving high BLEU scores competitive with the baseline.

Table 1: Regularization Methods Performance Summary (Pre-LN, No Warmup)

Configuration	Test BLEU	Test Loss	Interpretation
Pre-LN (Baseline)	35.13	2.212	Strong baseline performance
Pre-LN + L2	35.00	2.218	Slight penalty, stable training
Pre-LN + Weight Decay	**35.52**	**2.194**	**Best in category, improved generalization**
Pre-LN + RAdam	35.20	2.217	Rectified variance replaces warmup successfully

Key Findings:

1. **Pre-LN enables robustness:** The most significant finding is that the Pre-LN architecture is robust enough to handle standard optimization techniques (like L2 and Weight Decay) starting at the full learning rate (0.001) without diverging. This starkly contrasts with Post-LN, which requires warmup to survive the early training phase.

2. **Weight Decay improves generalization:** The configuration with Weight Decay achieved a BLEU score of **35.52**, which outperforms the standard Pre-LN baseline (35.13) and approaches the performance of the highly-tuned Post-LN + Warmup model (35.87). This suggests that decoupling weight decay from the gradient update (AdamW style) provides better regularization for Transformers than standard L2 regularization.

3. **RAdam is a viable alternative to Warmup:** The Rectified Adam (RAdam) optimizer achieved a BLEU score of 35.20. RAdam is designed to implicitly provide a "warmup" by rectifying the variance of the adaptive learning rate. Our results confirm that RAdam works exactly as intended, achieving performance virtually identical to the standard Adam optimizer.

In summary, while learning rate warmup is mandatory for the original Post-LN Transformer, moving to a **Pre-LN architecture** grants the flexibility to use various regularization strategies and optimizers without strict dependency on warmup schedules.

4 Translations Comparison

Based on the quantitative evaluation, we selected the top three best-performing configurations for a detailed qualitative analysis. The models are ranked by their Test BLEU scores as follows:

1. **Model A (Post-LN + Warmup):** BLEU = **35.87** (The Baseline Transformer)
2. **Model B (Pre-LN + Warmup):** BLEU = **35.57**
3. **Model C (Pre-LN + No Warmup):** BLEU = **35.13**

Below, we present sample translations generated by each model for randomly selected sentences to evaluate their linguistic quality and reveal behavioral differences that BLEU scores might not fully capture.

4.1 Model A (Post-LN + Warmup) Translation Examples

Example 1 (sentence #2023)

SRC: *c'est plus que ça, plus profond.*

TAR: It's more than that; it's deeper.

PRED: that's more, and that is more, deeper than that. </s>

Example 2 (sentence #2173)

SRC: *Mais il y a beaucoup de gaspillage de transport et des choses compliquées qui se passent.*

TAR: But there's a lot of wasted transportation and convoluted things happening.

PRED: but there's a lot of waste of transport and complicated things happening. </s>

Example 3 (sentence #2225)

SRC: *Et ici vous pouvez voir Roberto danser.*

TAR: And here you can see Roberto dancing.

PRED: and here you can see the <unk> dance. dance. dance. dance. dance. </s>

Example 4 (sentence #6531)

SRC: *Faisons donc ce que font tous les personnages quand ils s'interrogent, quand ils sont face à un mystère.*

TAR: So let's do what characters always do when they're not sure, when they are in possession of a mystery.

PRED: so let's do what all the characters do when they <unk> when they face a mystery. </s>

Example 5 (sentence #2848)

SRC: *Et vous n'avez pas – en tant qu'attaquant – vous n'avez pas besoin d'introduire la charge utile à l'aide d'une clé USB, comme nous l'avons vu dans le cas de Stuxnet.*

TAR: And you don't have – as an attacker – you don't have to deliver this payload by a USB stick, as we saw it in the case of Stuxnet.

PRED: and you don't – as an <unk> – you don't need to introduce the payload that is helpful with a usb key, as we saw in the case of the to of the of the to the <unk> – you don't need to introduce the payload that is helpful using a usb key, as we saw in the case of the particular <unk> </s>

4.2 Model B (Pre-LN + Warmup) Translation Examples

Example 1 (sentence #4542)

SRC: *Ils ont toujours peur de perdre de la place en rayons.*

TAR: They're always worried they're going to lose shelf space.

PRED: they always fear they lose radiation instead. they always fear they lose radiation. </s>

Example 2 (sentence #5477)

SRC: *Nous n'avons jamais rencontré un seul être humain dans le monde qui puisse fabriquer un produit, le vendre et s'occuper de l'argent aussi.*

TAR: We have never met a single human being in the world who can make it, sell it and look after the money.

PRED: we never met a human being around the world that can build a product, sell it and take money as well. too. too. too. too. too. too. too. too. </s>

Example 3 (sentence #6070)

SRC: *Et par la suite, la chute du rideau de fer non plus.*

TAR: And in the aftermath, also not the fall of the Iron Curtain.

PRED: and subsequently the iron curtain was either. neither the iron over the gold end and later the iron <unk> </s>

Example 4 (sentence #5187)

SRC: *Je me suis trompée en pensant que j'étais unique et seule dans cette situation.*

TAR: I was mistaken in thinking that I was unique and alone in this situation.

PRED: i've wrong the other feeling that i was unique and only in this situation. </s>

Example 5 (sentence #1217)

SRC: *Le prix de l'énergie, également, est très important pour eux.*

TAR: Also, the price of energy is very important to them.

PRED: energy prices, also, is very important for them. they are very important for them. </s>

Special Case Example Selection
Sentence #8300

SRC: *Je dois la mentionner.*

TAR: Yes, I must mention it.

PRED: i gotta call it <unk> i've gotta call it <unk> i've gotta call it <unk> i've gotta call it <unk> i've gotta call it <unk> ... (repeats indefinitely)

4.3 Model C (Pre-LN + No Warmup) Translation Examples

Example 1 (sentence #722)

SRC: *Ils ont créé une ferme piscicole de 110 km^2 perche, mulet, crevette, anguille et dans ce processus, Miguel et cette entreprise ont complètement inversé la destruction écologique.*

TAR: They created a 27,000-acre fish farm – bass, mullet, shrimp, eel – and in the process, Miguel and this company completely reversed the ecological destruction.

PRED: they created a fish farm of 110 square miles of land ice cream and <unk> and in this process, miguel and in this process, miguel and this company has completely wiped out the environmental destruction that surrounded the environmental destruction of 50 square miles of land in this process... (repeats) </s>

Example 2 (sentence #4821)

SRC: *Mais si l'ajustement au style de vie de la Corée du Sud n'était pas facile, j'avais un plan. J'ai commencé à étudier pour les examens d'entrée à l'université.*

TAR: Even though adjusting to life in South Korea was not easy, I made a plan – I started studying for the university entrance exam.

PRED: but if i applied to the <unk> in south korea was not easy, i had a plan, i began to study to apply for academic exams at university to college. </s>

Example 3 (sentence #2874)

SRC: *Pour vous montrer maintenant ce sur quoi nous travaillons en commençant par parler du soldat américain, qui en moyenne porte environ 45 kilos sur son dos, et on leur demande de porter plus d'équipement.*

TAR: To show you now what we are working on by starting out talking about the American soldier, that on average does carry about 100 lbs. on their backs, and they are being asked to carry more equipment.

PRED: so to show you now what we're working on starting with talking about the american soldier who on average is about 80 kilos on her backs, and they're being asked to wear more she's she's she's got to wear more she's she's got to wear more she's on the backs... </s>

Example 4 (sentence #1108)

SRC: *Et la façon de l'utiliser, semblerait appropriée mais ce n'est pas cette histoire que notre "moi du souvenir" conserve.*

TAR: And how to spend it would seem to be relevant, but that is not the story that the remembering self keeps for us.

PRED: and the way to use it, it feels appropriate but it's not the story that our book from "the <unk> will detect the app of the store. will detect the puzzle will detect and the way it feels appropriate... </s>

Example 5 (sentence #7294)

SRC: *Le sport de mémorisation de compétition est entrainé par une sorte de bras de fer où chaque année, quelqu'un invente une nouvelle façon de mémoriser plus de choses plus rapidement, et le reste doit rattraper le temps perdu.*
TAR: The sport of competitive memorizing is driven by a kind of arms race where, every year, somebody comes up with a new way to remember more stuff more quickly, and then the rest of the field has to play catch-up.
PRED: <unk> sports is trained by a kind of iron arm where each year, someone invented a new way to memorize more things faster, and the rest must have caught up with loss time lost. each year, someone invent a new way to memorize more things faster... </s>

Special Case Example Selection
Sentence #1236
SRC: *La moyenne est d'environ cinq tonnes par personne sur la planète.*
TAR: It's an average of about five tons for everyone on the planet.
PRED: the average is about five metric tons per person on the planet on the planet – the average is about five metric metric metric metric metric metric metric metric metric... (repeats indefinitely)

4.4 Conclusion of Comparison in Random Translations

For the majority of the sampled sentences, all three models demonstrate commendable translation capabilities, reflecting their high BLEU scores (> 35). They successfully capture the semantic meaning of the source sentences and generate grammatically mostly correct English outputs.

However, after I test several cases via random testing, I found **Model B and Model C sometimes output strange translations** (As shown in their translation section's special Case). This special situation doesn't happen in Model A, though their BLEU scores are similar. **This reflects that the BLEU score is not the only measurement method we need use and it is one-sided.**

This finding intuitively demonstrates the advantages of Post-LN, but the specific analysis serves as a point of discussion for future research and requires further experimental verification. We know that modern large-scale models such as Llama employ Pre-LN; however, the difference in output results may reflect a possibility that early Transformer translation tasks generally opted for Post-LN. In other words, Pre-LN is easier to train and less sensitive to the learning rate schedule, while Post-LN typically requires careful warmup and hyperparameter tuning to avoid divergence. Nevertheless, in our experiments, we demonstrated that Pre-LN was very effective, but failed to demonstrate its advantages. In future experiment, we shall add more measurement tools to measure the qulity of translation, especially for the situation of repetitive words.

5 Conclusion

This project systematically investigated the roles of normalization and regularization in training Transformer models. Through our experiments, we moved beyond simple performance metrics to understand the fundamental structural trade-offs in sequence modeling.

First and foremost, our results confirm that **Dropout is the single most critical regularizer** for generalization. Regardless of the architecture, removing dropout caused a massive performance drop, with BLEU scores falling from 35.87 to 27.68. This highlights that while advanced normalization techniques control training stability, simple dropout remains the primary defense against overfitting in parameter-heavy models.

A major finding of this study is that the necessity of **Learning Rate Warmup is entirely architecture-dependent**. Contrary to the common belief that Transformers always require a warmup phase to prevent gradient explosion, we found this to be true only for the **Post-LN** architecture. The Post-LN configuration failed completely (0.00 BLEU) without warmup but achieved the highest overall score (35.87) when warmup was applied. In strong contrast, the **Pre-LN** architecture demonstrated remarkable robustness, successfully converging to a high score (35.13) even without any warmup.

Above features of the Pre-LN architecture explains its dominance in **modern Large Language Models (LLMs)** like GPT-3, PaLM, and Llama. While our experiments showed that Post-LN theoretically offers a slightly higher performance ceiling (likely due to sharper output distributions and less representation collapse), the risk of training divergence is too high for massive models. The Pre-LN architecture acts as a "safety net," ensuring consistent convergence and allowing the use of alternative regularizers. For instance, we found that Pre-LN enabled training with **Weight Decay** alone, achieving a score of 35.52, which is competitive with the highly tuned baseline.

Regarding the type of normalization, our results conclusively show that **Layer Normalization is superior to Batch Normalization** for Transformers. Even in the stable Pre-Norm setup, Batch Normalization struggled significantly (16.60 BLEU). This confirms that sequence-to-sequence tasks require sample-independent statistics, which Batch Norm cannot provide effectively due to variable sequence lengths and padding.

As for Pre-LN and Post-LN, Pre-LN offers consistently stable convergence and does not require warmup to train reliably. This stability is one of the most significant observations in our experiments.

However, in qualitatively sampled translation outputs, **Pre-LN occasionally produces repetitive or degenerate phrases—an issue that is not reflected in the validation curves or BLEU scores.** Therefore, although our results demonstrate the robustness and practical potential of Pre-LN, further investigation is needed to understand why Pre-LN models tend to generate repetitive words and how this behavior relates to the training dynamics.