Generated by autopeerreview.com

Manuscript review report

TPTT: Transforming Pretrained Transformer into Titans

Issues found	Review Date	Review Time	Recommendations
	2025-06-24 11:26:18	881.92 s	69

Overall Summary

The manuscript needs refinement to improve clarity, conciseness, and adherence to academic writing standards. Key areas for attention include title optimization, structural reorganization, and language polishing. Figures must be properly cited, and acronym definitions should be consistent.

Section Review Summary

Section	Status	Issues
Title	✖	1
Title Page	✖	3
Abstract	✖	1
Structure	✖	5
Main Headings	✖	2
Language	✖	49
Acronyms	✖	3

Section	Status	Issues
Figures and Tables	✗	5
Referencing	✓	
Funding	✗	

Title Recommendations (1 issue)

The title should be more descriptive and highlight key aspects of the work.

Location: Title

Title Reviewer @ autopeerreview.com

```
Original: TPTT: Transforming Pretrained Transformer into Titans
```

```
Suggestion: TPTT: Efficiently Transforming Pretrained Transformers with
Linearized Attention
```

Explication: More descriptive, replacing the less specific "Titans" with "Linearized Attention" which is a key component, and adding the word "efficiently" to highlight an important property.

Title Page Recommendations (3 issues)

The title page needs complete author information, including affiliations and correct email format, and proper formatting.

Location: Header

Title Page Reviewer @ autopeerreview.com

```
Original: \begin{document}
```

```
Suggestion: \maketitle\n\begin{abstract}\nRecent advances in large language
models (LLMs) have led to remarkable progress in natural language processing,
but their computational and memory demands remain a significant challenge,
particularly for long-context inference. We introduce TPTT (Transforming
Pretrained Transformer into Titans), a novel framework for enhancing
pretrained Transformer models with efficient linearized attention mechanisms
and advanced memory management. TPTT employs techniques such as Memory as Gate
(MaG) and mixed linearized attention (LiZA). It is fully compatible with the
Hugging Face Transformers library, enabling seamless adaptation of any causal
LLM through parameter-efficient fine-tuning (LoRA) without full retraining. We
show the effectiveness of TPTT on the MMLU benchmark with models of
approximately 1 billion parameters, observing substantial improvements in both
efficiency and accuracy. For instance, Titans-Llama-3.2-1B achieves a 20%
increase in Exact Match (EM) over its baseline. Statistical analyses and
comparisons with recent state-of-the-art methods confirm the practical
scalability and robustness of TPTT. The source code is available at
https://github.com/fabienfrfr/tptt and the Python package at
https://pypi.org/project/tptt/.\n\end{abstract}\n\n\textit{Keywords:} Large
Language Models; Transformers; Attention Mechanism; Memory Management;
Parameter-Efficient Fine-Tuning\n\n\begin{document}
```

Explication: Added `\maketitle` command to display the title, author, and date. Included the abstract and keywords within the header for proper formatting and visibility.

Location: Authors

Title Page Reviewer @ autopeerreview.com

```
Original: \thanks{\texttt{fabien.furfaro@gmail.com}}
```

```
Suggestion: \thanks{\texttt{fabien.furfaro@\{institution\}.com}}
```

Explication: Suggest using an institutional email address instead of a personal one. Replace {institution} with the appropriate domain name.

Location: Authors

Title Page Reviewer @ autopeerreview.com

```
Original: \author{\large Fabien
Furfaro\thanks{\texttt{fabien.furfaro@gmail.com}}}
```

```
Suggestion: \author{\large Fabien Furfaro\thanks{Corresponding author:
\texttt{fabien.furfaro@\{institution\}.com}} \newline \textit{\{Department\},
\{Institution\}, \{City\}, \{Country\}}}
```

Explication: Explicitly specify the corresponding author and include the author's affiliation details (Department, Institution, City, Country). Replace {institution} and the bracketed placeholders with actual values. Use `\newline` to separate the author and affiliation information.

Abstract Recommendations (1 issue)

Consider using a structured abstract to improve readability and information accessibility.

Location: Abstract

Abstract Reviewer @ autopeerreview.com

```
Original: Abstract
```

Explication: The abstract is unstructured (does not contain subheadings). This is common in many journals, but structured abstracts can improve readability and information retrieval.

Structure Recommendations (5 issues)

The manuscript's structure needs reorganization, especially regarding the introduction's length, section naming conventions, and separation of results and discussion.

Location: Introduction

Structure Reviewer @ autopeerreview.com

```
Original: Introduction
```

Explication: The introduction is very short (~138 words) compared to the abstract (~1084 words). Expand the introduction to provide more context, motivation, and a clearer outline of the paper's contributions.

Location: Methodology

Structure Reviewer @ autopeerreview.com

```
Original: Methodology
```

Explication: Consider renaming "Methodology" to "Method" for conciseness, especially if it aligns with common practices in the research area.

Location: Training Procedure

Structure Reviewer @ autopeerreview.com

```
Original: Training Procedure
```

Explication: Integrate the "Training Procedure" section into the "Methodology" section, as training details are typically considered part of the experimental method.

Location: Experiments and Results

Structure Reviewer @ autopeerreview.com

```
Original: Experiments and Results
```

Explication: Consider renaming "Experiments and Results" to just "Results" or perhaps "Experimental Results." The experimental setup should be described in a separate "Experiment" or "Experimental Setup" section, either as a subsection of "Methods" or preceding "Results.

Location: Discussion and Conclusion

Structure Reviewer @ autopeerreview.com

```
Original: Discussion and Conclusion
```

Explication: Separate the "Discussion and Conclusion" section into two distinct sections: "Discussion" and "Conclusion." This allows for a more focused interpretation of results and a clearer summary of the paper's key findings and implications.

Main Headings Recommendations (2 issues)

Main headings should be concise and accurately reflect the content of the sections.

Location: Dynamic Memory as Gate Scheduling (LiZA MaG Callback)

Main Headings Reviewer @ autopeerreview.com

```
Original: Dynamic Memory as Gate Scheduling (LiZA MaG Callback)
```

```
Suggestion: Dynamic MaG Scheduling (LiZA Callback)
```

Explication: Shortened the title, using the acronym MaG and removing the redundant "Memory as Gate". "Callback" is sufficient without repeating "MaG".

Location: Evaluation Metrics and Benchmark Results

Main Headings Reviewer @ autopeerreview.com

```
Original: Evaluation Metrics and Benchmark Results
```

```
Suggestion: Benchmark Results
```

Explication: The section focuses primarily on benchmark results, and evaluation metrics are introduced within that context. Removing "Evaluation Metrics" makes the title more concise and aligned with the primary content.

Language Recommendations (49 issues)

The manuscript needs language revisions to enhance clarity, conciseness, and maintain a consistent academic tone through sentence structure and word choice.

Location: Not specified

Language Reviewer @ autopeerreview.com

Original: Recent advances in large language models (LLMs) have led to remarkable progress in natural language processing, but their computational and memory demands remain a significant challenge, particularly for long-context inference.

> Suggestion: Large language models (LLMs) have enabled advances in natural language processing; however, their computational and memory demands remain a challenge, particularly for long-context inference.

Explication: Reduced sentence length and improved flow while maintaining academic tone.

Location: Introduction

Language Reviewer @ autopeerreview.com

Original: However, the quadratic computational complexity and substantial memory requirements of the standard self-attention mechanism remain a significant bottleneck, particularly for long-context inference.

> Suggestion: However, the quadratic computational complexity and memory requirements of self-attention limit long-context inference.

Explication: Shortened sentence by removing redundancy and focusing on the core issue.

Location: Introduction

Language Reviewer @ autopeerreview.com

Original: Efficient attention mechanisms~\cite{katharopoulos2020transformers, yang2024parallelizing} have been proposed to reduce the complexity of self-attention from quadratic to linear or near-linear, making it more tractable for long sequences.

> Suggestion: Efficient attention mechanisms~\cite{katharopoulos2020transformers, yang2024parallelizing} reduce the complexity of self-attention to linear or near-linear, improving tractability for long sequences.

Explication: Reduced wordiness and improved clarity by using a more direct construction.

Location: Introduction

Language Reviewer @ autopeerreview.com

Original: Recurrent architectures and internal memory mechanisms~\cite{mercat2024linearizing, behrouz2024titans} have also been developed to enhance the model's ability to capture long-range dependencies, drawing inspiration from cognitive memory in biological systems rather than hardware memory.

> Suggestion: Recurrent architectures and internal memory mechanisms~\cite{mercat2024linearizing, behrouz2024titans} enhance a model's ability to capture long-range dependencies, drawing inspiration from cognitive memory rather than hardware memory.

Explication: Improved conciseness and clarity by removing unnecessary phrasing ("have also been developed" and "in biological systems").

Location: Related Work

Language Reviewer @ autopeerreview.com

Original: Despite these advances, most existing methods require significant architectural modifications or training from scratch, which limits their applicability to already pretrained models.

> Suggestion: Most existing methods require significant architectural modifications or training from scratch, limiting their applicability to pretrained models.

Explication: Removed "Despite these advances," as the contrast is implied. Also shortened "already pretrained" to "pretrained".

Location: Related Work

Language Reviewer @ autopeerreview.com

Original: For example, solutions like FlashAttention and Mamba focus on efficient architectures, while others such as LoLCat and Liger convert standard attention to linearized forms; notably, Liger exploits similar properties to those leveraged in this work, but does not rely on explicit linearization injection.

> Suggestion: For example, FlashAttention and Mamba focus on efficient architectures, while LoLCat and Liger convert standard attention to linearized forms. Liger exploits similar properties to those in this work, but does not rely on explicit linearization injection.

Explication: Shortened the sentence for clarity and flow. Removed redundant "solutions like" and rephrased the "notably" clause.

Location: Related Work

Language Reviewer @ autopeerreview.com

Original: In this work, we introduce TPTT (Transforming Pretrained Transformer into Titans), a framework that transforms pretrained transformer models into efficient and scalable architectures by incorporating linearized attention mechanisms and advanced internal memory augmentation.

> Suggestion: We introduce TPTT (Transforming Pretrained Transformer into Titans), a framework that enhances pretrained transformer models with linearized attention mechanisms and internal memory augmentation.

Explication: Streamlined the sentence by removing "efficient and scalable architectures by incorporating" and using "enhances with".

Location: Related Work *Language Reviewer @ autopeerreview.com*

Original: The goal is to unlock the potential of already trained models by equipping them with memory-augmented capabilities through lightweight adaptation.

> Suggestion: TPTT unlocks the potential of pretrained models by providing memory-augmented capabilities through lightweight adaptation.

Explication: Strengthened the sentence using active voice and replacing "already trained" with "pretrained.

Location: Linearized Attention Mechanisms *Language Reviewer @ autopeerreview.com*

Original: Standard self-attention in transformers computes pairwise interactions between all tokens, resulting in quadratic complexity with respect to sequence length.

> Suggestion: Self-attention computes pairwise interactions between tokens, resulting in quadratic complexity with sequence length.

Explication: Removed redundant "standard self-attention in transformers" and shortened "with respect to" for conciseness.

Location: Linearized Attention Mechanisms *Language Reviewer @ autopeerreview.com*

Original: To address this, linearized attention mechanisms approximate the softmax attention using linear operations, typically by projecting queries and keys through a feature map \u03c6.

> Suggestion: Linearized attention mechanisms approximate softmax attention using linear operations, projecting queries and keys through a feature map \u03c6.

Explication: Shortened the sentence by removing "To address this, " and "typically by".

Location: Linearized Attention Mechanisms

Language Reviewer @ autopeerreview.com

Original: This reduces computational and memory costs, enabling efficient processing of long sequences while maintaining modeling power.

> Suggestion: This reduces computational cost and memory usage, enabling efficient processing of long sequences while maintaining performance.

Explication: Replaced "computational and memory costs" with "computational cost and memory usage" for clarity and conciseness. Changed "modeling power" to "performance" for better terminology.

Location: Linearized Attention Mechanisms

Language Reviewer @ autopeerreview.com

Original: where \ud83d\udc03_t, \ud83d\udc03_i, and \ud83d\udc03_i are the query, key, and value vectors at positions t and i, respectively, and \u03b2_i is a gating vector (or scalar) modulating the keys and values.

> Suggestion: where \ud83d\udc03_t, \ud83d\udc03_i, and \ud83d\udc03_i represent the query, key, and value vectors at positions t and i, respectively, and \u03b2_i is a gating vector (or scalar) modulating the keys and values.

Explication: Replaced "are" with "represent" for better accuracy.

Location: Memory as Gate (MaG)

Language Reviewer @ autopeerreview.com

Original: To further enhance long-range dependency modeling, we introduce an internal memory augmentation mechanism, Memory as Gate (MaG), inspired by the Titans architecture.

> Suggestion: To enhance long-range dependency modeling, we introduce Memory as Gate (MaG), an internal memory augmentation mechanism inspired by the Titans architecture.

Explication: Slight restructuring to improve flow and conciseness.

Location: Memory as Gate (MaG)

Language Reviewer @ autopeerreview.com

Original: Unlike hardware memory, this mechanism enables the model to store and recall contextual information over extended sequences, analogous to cognitive memory.

> Suggestion: Unlike hardware memory, MaG enables the model to store and recall contextual information over long sequences, similar to cognitive memory.

Explication: Clarified subject by replacing "this mechanism" with "MaG". Changed "extended" to "long" for better readability. Shortened "analogous to" to "similar to".

Location: Memory as Gate (MaG)

Language Reviewer @ autopeerreview.com

Original: This allows the model to leverage both the efficiency of linear attention and the expressivity of softmax attention, and can be seen as a form of memory-augmented gating.

> Suggestion: This allows the model to leverage the efficiency of linear attention and the expressiveness of softmax attention, functioning as memory-augmented gating.

Explication: Corrected "expressivity" to "expressiveness" for better academic tone. Replaced "can be seen as a form of" with "functioning as" for conciseness.

Location: Parallel Delta Rule Modeling

Language Reviewer @ autopeerreview.com

Original: In this work, the feature mapping function of linear attention is approximated by DeltaNet.

> Suggestion: We approximate the feature mapping function of linear attention using DeltaNet.

Explication: Changed to active voice for clarity and conciseness.

Location: Parallel Delta Rule Modeling

Language Reviewer @ autopeerreview.com

Original: The recurrent update of internal memory states is formulated in two ways, following:

> Suggestion: The recurrent update of internal memory states follows one of two formulations:

Explication: Improved clarity and conciseness.

Location: Parallel Delta Rule Modeling

Language Reviewer @ autopeerreview.com

Original: The final state of one chunk becomes the initial state of the next, supporting efficient memory usage for long sequences.

> Suggestion: The final state of a chunk initializes the next, supporting efficient memory usage for long sequences.

Explication: Improved flow and conciseness.

Location: Parallel Delta Rule Modeling

Language Reviewer @ autopeerreview.com

Original: The output \ud83d\udc03_lin is then computed from these memory states.

> Suggestion: We then compute the output \ud83d\udc03_lin from these memory states.

Explication: Changed to active voice for clarity.

Location: Integration with Pretrained Models *Language Reviewer @ autopeerreview.com*

Original: Our approach injects linearized attention and memory augmentation modules into pretrained transformer models. The process involves:

> Suggestion: TPTT injects linearized attention and memory augmentation modules into pretrained transformer models, as follows:

Explication: Improved clarity and specifies TPTT instead of "Our approach". "The process involves:" was replaced with "as follows:" for conciseness.

Location: Integration with Pretrained Models *Language Reviewer @ autopeerreview.com*

Original: 1. Identification of Target Modules: Key attention layers to be modified are identified using tools such as get_tptt_model.

> Suggestion: 1. Identify Target Modules: Use tools like `get_tptt_model` to identify key attention layers for modification.

Explication: Made the sentence more direct and active, and used code formatting for `get_tptt_model`.

Location: Integration with Pretrained Models *Language Reviewer @ autopeerreview.com*

Original: 2. Modification of Attention Layers: These layers are replaced or extended with the proposed LiZAttention module, which implements both linear and softmax attention with linear projection weigth sharing and MaG.

> Suggestion: 2. Modify Attention Layers: Replace or extend target layers with the LiZAttention module, which implements linear and softmax attention with linear projection weight sharing and MaG.

Explication: Made the sentence more direct and active. Corrected spelling of "weigth" to "weight".

Location: Integration with Pretrained Models *Language Reviewer @ autopeerreview.com*

Original: 3. Training and Fine-Tuning: The modified model is fine-tuned using parameter-efficient methods such as LoRA, ensuring optimal adaptation to the new mechanisms without full retraining.

Suggestion: 3. Fine-Tune: Fine-tune the modified model using parameter-efficient methods like LoRA for adaptation, avoiding full retraining.

Explication: Reduced wordiness and made the sentence more direct.

Location: Integration with Pretrained Models *Language Reviewer @ autopeerreview.com*

Original: This procedure enables the transformation of any causal pretrained LLM into a memory-augmented, efficient architecture with minimal overhead, that without any new layers.

Suggestion: This procedure transforms any causal pretrained LLM into a memory-augmented, efficient architecture with minimal overhead, and without new layers.

Explication: Improved conciseness and grammar.

Location: LiZAttention Module *Language Reviewer @ autopeerreview.com*

Original: The \texttt{LiZAttention} module is a core component of the TPTT architecture, designed to synergistically combine linearized attention and standard (softmax) attention mechanisms.

Suggestion: The \texttt{LiZAttention} module is a core component of TPTT, synergistically combining linearized attention and standard (softmax) attention mechanisms.

Explication: Improved conciseness.

Location: LiZAttention Module *Language Reviewer @ autopeerreview.com*

Original: This hybrid approach leverages the computational efficiency of linear attention while retaining the expressivity of vanilla attention.

Suggestion: This hybrid approach leverages the efficiency of linear attention while retaining the expressiveness of vanilla attention.

Explication: Replaced "expressivity" with "expressiveness" for better academic tone.

Location: LiZAttention Module *Language Reviewer @ autopeerreview.com*

Original: To support long-context inference, \texttt{LiZAttention} maintains a cache of intermediate states and implements a recurrent information mechanism for efficient internal memory management.

Suggestion: To support long-context inference, \texttt{LiZAttention} caches intermediate states and uses a recurrent mechanism for internal memory management.

Explication: Shortened sentence for conciseness.

Location: Efficient Internal Memory Management *Language Reviewer @ autopeerreview.com*

Original: The cache of intermediate states maintained by \texttt{LiZAttention} enables a recurrent information, efficiently supporting long-context inference without excessive computational overhead.

Suggestion: The \texttt{LiZAttention} cache of intermediate states enables recurrent information processing, supporting long-context inference efficiently.

Explication: Improved clarity and conciseness. Removed "excessive computational overhead" as it's implied by "efficiently".

Location: Efficient Internal Memory Management *Language Reviewer @ autopeerreview.com*

Original: This approach allows the model to scale to longer sequences, leveraging both local and global context.

Suggestion: This allows the model to scale to longer sequences, leveraging local and global context.

Explication: Removed redundant "approach".

Location: Parameter-Efficient Fine-Tuning with LoRA *Language Reviewer @ autopeerreview.com*

Original: This approach reduces the number of trainable parameters and memory requirements, while maintaining performance comparable to full fine-tuning.

Suggestion: LoRA reduces trainable parameters and memory requirements, while maintaining performance comparable to full fine-tuning.

Explication: Clarified subject by replacing "This approach" with "LoRA".

Location: Dynamic Memory as Gate Scheduling (LiZA MaG Callback) *Language Reviewer @ autopeerreview.com*

Original: A important component of the training process is the LiZA MaG callback, which dynamically adjusts the Memory as Gate (MaG) weighting parameter during training.

Suggestion: An important component of the training process is the LiZA MaG callback, which dynamically adjusts the Memory as Gate (MaG) weighting parameter during training.

Explication: Corrected "A important" to "An important.

Location: Dynamic Memory as Gate Scheduling (LiZA MaG Callback)

Language Reviewer @ autopeerreview.com

Original: The MaG weight is initialized at 0.01 and linearly increased to 0.5 over the first 100 steps, facilitating a smooth transition from reliance on vanilla (softmax) attention to linearized attention.

Suggestion: The MaG weight is initialized at 0.01 and linearly increased to 0.5 over the first 100 steps, enabling a smooth transition from vanilla (softmax) attention to linearized attention.

Explication: Replaced "facilitating" with "enabling" for conciseness.

Location: Dynamic Memory as Gate Scheduling (LiZA MaG Callback)

Language Reviewer @ autopeerreview.com

Original: This schedule allows the model to effectively balance the two attention mechanisms, optimizing performance throughout training.

Suggestion: This schedule effectively balances the attention mechanisms, optimizing performance during training.

Explication: Improved conciseness.

Location: Dynamic Memory as Gate Scheduling (LiZA MaG Callback)

Language Reviewer @ autopeerreview.com

Original: The callback is integrated directly into the training loop, ensuring adaptive control of the MaG parameter and enhancing the model's adaptability and efficiency.

Suggestion: The callback integrates into the training loop, ensuring adaptive control of the MaG parameter and enhancing the model's adaptability and efficiency.

Explication: Improved conciseness by removing "directly".

Location: Experimental Setup *Language Reviewer @ autopeerreview.com*

Original: Mixed precision training and gradient clipping at 1.0 were employed to optimize computational efficiency and stability.

> Suggestion: We used mixed precision training and gradient clipping at 1.0 to optimize computational efficiency and stability.

Explication: Changed to active voice.

Location: Training Results *Language Reviewer @ autopeerreview.com*

Original: The results indicate consistent and efficient learning across architectures, by low final loss values and stable gradient norms.

> Suggestion: The results indicate consistent and efficient learning across architectures, evidenced by low final loss values and stable gradient norms.

Explication: Replaced "by" with "evidenced by" for greater precision.

Location: Training Results *Language Reviewer @ autopeerreview.com*

Original: The use of Low-Rank Adaptation (LoRA) and the Memory as Gate (MaG) mechanism shows effective in optimizing training dynamics and convergence.

> Suggestion: Low-Rank Adaptation (LoRA) and the Memory as Gate (MaG) mechanism are effective in optimizing training dynamics and convergence.

Explication: Simplified the sentence structure for clarity.

Location: Evaluation Metrics and Benchmark Results *Language Reviewer @ autopeerreview.com*

Original: These metrics respectively measure strict correctness, partial overlap, and quasi-exactness between model outputs and ground truth answers, providing a nuanced view of model performance~\cite{hendrycks2020measuring}

> Suggestion: These metrics measure strict correctness, partial overlap, and quasi-exactness between model outputs and ground truth answers, providing a nuanced view of performance~\cite{hendrycks2020measuring}.

Explication: Removed "respectively" for conciseness. Also, changed "model performance" to "performance.

Location: Evaluation Metrics and Benchmark Results *Language Reviewer @ autopeerreview.com*

Original: TPTT models, and especially Titans-Llama-3.2-1B, consistently outperform their base counterparts in EM, with better PEM and PQEM scores.

Suggestion: TPTT models, especially Titans-Llama-3.2-1B, consistently outperform their base counterparts in EM, and achieve better PEM and PQEM scores.

Explication: Improved sentence structure.

Location: Evaluation Metrics and Benchmark Results *Language Reviewer @ autopeerreview.com*

Original: This shows the benefit of integrating linearized attention and internal memory mechanisms for complex language understanding tasks.

Suggestion: This highlights the benefits of integrating linearized attention and internal memory mechanisms for complex language understanding.

Explication: Shortened sentence for conciseness.

Location: Discussion and Comparison *Language Reviewer @ autopeerreview.com*

Original: Compared to recent state-of-the-art methods such as Mamba~\cite{gu2023mamba}, LoLCat~\cite{zhang2024lolcats}, and Liger~\cite{lan2025liger}, TPTT stands out by enabling the transformation of existing pretrained models without full retraining, while maintaining good benchmark performance.

Suggestion: Compared to methods like Mamba~\cite{gu2023mamba}, LoLCat~\cite{zhang2024lolcats}, and Liger~\cite{lan2025liger}, TPTT transforms existing pretrained models without full retraining, while maintaining benchmark performance.

Explication: Improved conciseness by removing "recent state-of-the-art" and simplifying the sentence structure.

Location: Discussion and Comparison *Language Reviewer @ autopeerreview.com*

Original: The observed improvements in EM and better PEM/PQEM scores highlight the effectiveness of TPTT's linearized attention and memory augmentation for efficient and robust LLM adaptation.

Suggestion: The EM improvements and better PEM/PQEM scores highlight the effectiveness of TPTT's linearized attention and memory augmentation for efficient and robust LLM adaptation.

Explication: Reduced wordiness and improved flow.

Location: Discussion and Conclusion

Language Reviewer @ autopeerreview.com

Original: In this paper, we have introduced TPTT, a novel framework for enhancing pretrained Transformer models by integrating efficient linearized attention mechanisms and internal memory augmentation.

Suggestion: We introduced TPTT, a framework for enhancing pretrained Transformer models with linearized attention mechanisms and internal memory augmentation.

Explication: Improved conciseness and flow.

Location: Discussion and Conclusion

Language Reviewer @ autopeerreview.com

Original: Experimental results on the MMLU benchmark shows significant improvements in both efficiency and accuracy, with robust statistical analyses and favorable comparisons to state-of-the-art methods.

Suggestion: MMLU benchmark results demonstrate significant improvements in efficiency and accuracy, with robust statistical analyses and favorable comparisons to other methods.

Explication: Improved clarity, active voice, and conciseness.

Location: Discussion and Conclusion

Language Reviewer @ autopeerreview.com

Original: The use of LoRA allows for efficient and flexible fine-tuning, enabling rapid adaptation to new tasks and domains.

Suggestion: LoRA enables efficient and flexible fine-tuning, allowing rapid adaptation to new tasks and domains.

Explication: Improved conciseness.

Location: Discussion and Conclusion

Language Reviewer @ autopeerreview.com

Original: While our results are promising, broader validation on additional benchmarks and real-world scenarios is needed to fully assess the generalizability and robustness of the approach.

Suggestion: Broader validation on benchmarks and real-world scenarios is needed to fully assess the generalizability and robustness of TPTT.

Explication: Removed "While our results are promising" for conciseness. Clarified subject by specifying TPTT.

Location: Discussion and Conclusion *Language Reviewer @ autopeerreview.com*

Original: Future work will focus on optimizing the integration process, exploring more sophisticated internal memory mechanisms, and extending the evaluation to larger models and a wider range of benchmarks.

> Suggestion: Future work will optimize the integration process, explore sophisticated internal memory mechanisms, and extend evaluations to larger models and more benchmarks.

Explication: Improved clarity and conciseness by removing "more" and "a wider range of.

Location: Discussion and Conclusion *Language Reviewer @ autopeerreview.com*

Original: In summary, TPTT provides a practical, scalable, and effective library for upgrading pretrained Transformers, with strong empirical results and promising implications for the future of efficient language modeling.

> Suggestion: TPTT is a practical, scalable, and effective library for upgrading pretrained Transformers, with strong results and promising implications for efficient language modeling.

Explication: Shortened the sentence.

Location: Related Work *Language And Acronyms Reviewer @ autopeerreview.com*

Original: Our approach is fully compatible with existing frameworks and enables rapid adaptation of any causal LLM to long-context tasks via parameter-efficient fine-tuning with LoRA, without requiring full retraining.

> Suggestion: TPTT is compatible with existing frameworks, enabling rapid adaptation of causal large language models (LLMs) to long-context tasks via parameter-efficient fine-tuning (LoRA), without full retraining.

Explication: The acronym 'LLM' was not defined on its first use in the text. This suggestion adds the full definition, "large language models". Also, "Our approach" was replaced with "TPTT" and the text was made more concise for clarity.

Acronyms Recommendations (3 issues)

Ensure all acronyms are defined upon their first use and are not redundantly defined later in the text.

Location: Abstract

Acronyms Reviewer @ autopeerreview.com

Original: We show the effectiveness of TPTT on the MMLU benchmark with models of approximately 1 billion parameters, observing substantial improvements in both efficiency and accuracy.

Suggestion: We show the effectiveness of TPTT on the Massive Multitask Language Understanding (MMLU) benchmark with models of approximately 1 billion parameters, observing substantial improvements in both efficiency and accuracy.

Explication: The acronym "MMLU" was not defined in the abstract. Added definition at first mention.

Location: Discussion and Conclusion

Acronyms Reviewer @ autopeerreview.com

Original: Our approach leverages parameter-efficient fine-tuning (LoRA) to enable the rapid adaptation of large language models (LLMs) to long-context tasks, without the need for full retraining.

Suggestion: Our approach leverages parameter-efficient fine-tuning (LoRA) to enable the rapid adaptation of LLMs to long-context tasks, without the need for full retraining.

Explication: Acronym 'LLMs' was already defined as 'large language models' earlier in the body (in 'Introduction'). Remove this redundant definition.

Location:

Acronyms Reviewer @ autopeerreview.com

Original: Our approach is fully compatible with existing frameworks and enables rapid adaptation of any causal LLM to long-context tasks via parameter-efficient fine-tuning with LoRA, without requiring full retraining.

Suggestion: Our approach is fully compatible with existing frameworks and enables rapid adaptation of any causal large language model (LLM) to long-context tasks via parameter-efficient fine-tuning with LoRA, without requiring full retraining.

Explication: The acronym 'LLM' was used without its full form being defined. This recommendation defines 'LLM' as 'large language model' upon its first usage.

Figures and Tables Recommendations (5 issues)

Figure and table captions require more context and clear summaries of key results. All figures should be explicitly cited within the text.

Location: fig:approach_overview

Figures and Tables Reviewer @ autopeerreview.com

Original: Overview of the TPTT architecture.

Suggestion: Overview of the TPTT architecture, demonstrating the integration of linearized attention with memory management for efficient processing and output generation. The left panel shows the decoder-only architecture with parallel linear (LiZAttention) and vanilla attention. The right panel illustrates the linearized attention mechanism, including shared weights for query (Q), key (K), value (V), and output (O) projections, state memory (S) management, and the Memory as Gate (MaG) weighting mechanism using Delta Rule and AdaptativeAvgPool1D.

Explication: The original caption lacked a clear, concise summary of the key result. The suggested caption provides a more structured description by first summarizing the architecture's overall function and then detailing the components shown in the left and right panels. Also reworded to improve flow and readability.

Location: tab:training-metrics

Figures and Tables Reviewer @ autopeerreview.com

Original: Training performance metrics for TPTT models.

Suggestion: Training performance metrics, including loss and accuracy, during the training phase of the TPTT models, demonstrating learning efficiency and convergence.

Explication: The suggested caption provides more context by specifying the types of metrics and indicating the purpose of showing these metrics (learning efficiency and convergence). This makes the table more self-contained.

Location: tab:mmlu-results

Figures and Tables Reviewer @ autopeerreview.com

Original: MMLU benchmark results (one-shot) with statistical analysis. Each pair groups a Titans model and its base counterpart.

Suggestion: MMLU benchmark results (one-shot) with statistical analysis comparing the performance of Titans models against their base counterparts. Statistical significance is indicated where applicable.

Explication: The revised caption explicitly states that the comparison between Titans and base models is the focus, and adds that statistical significance is indicated in the table.

Location: Not specified

Figures and Tables Reviewer @ autopeerreview.com

```
Original: Overview of the TPTT architecture. On the left, the diagram
illustrates a decoder-only architecture where linear attention is injected in
parallel of vanilla attention (LiZAttention). On the right, the detailed
architecture of the linearized attention mechanism is depicted, highlighting
the shared weights for query (Q), key (K), value (V), and output (O)
projections. It also shows the management of the state memory (S) and the
combination of outputs through the Memory as Gate (MaG) weighting mechanism.
The diagram emphasizes the integration of linearized attention mechanisms and
advanced memory management techniques, such as Delta Rule and
AdaptativeAvgPool1D, contributing to processing and output generation.
```

Explication: Figure 1 is present but is never cited in the text.

Location: Not specified

Figures and Tables Reviewer @ autopeerreview.com

```
Original: Algorithm 1: LiZAttention Forward Pass
```

Explication: Figure 2 is present but is never cited in the text.