Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 1, 2026

AN ENHANCED T5-LARGE TRANSFORMER FOR EFFICIENT DENTAL CLINICAL ASSISTANCE

Hamad Khan[1], Saddam Hussain Khan[*2], Mirza Mumtaz Zahoor[3], Umul Baneen Ejaz[4]

[1,*2]*Artificial Intelligence Lab, Department of Computer Systems Engineering, University of Engineering and Applied Sciences (UEAS), Swat 19060, Pakistan*
[3,4]*Department of Computer Science, Faculty of Computer Sciences, IBADAT International University, Islamabad, 44000, Pakistan*

[1]hamadkhanf10@gmail.com, [*2]hengrshkhan822@gmail.com, [3]mumtazzahoor5@gmail.com, [4]umulbaneenejaz@gmail.com

DOI: https://doi.org/10.5281/zenodo.18255367

Keywords
Clinical Assistance, NLP, Question Answering, SSM Models, T5, Mamba.

Article History
Received: 23 November 2025
Accepted: 30 December 2025
Published: 15 January 2026

Corresponding Author: *
Saddam Hussain Khan

Abstract

Large language models (LLMs) in dentistry and other similar specialized clinical fields still have not been optimally utilized due to the absence of applicable specific benchmarks in dental-related practices. Because of the absence of specific benchmarks, the process for selecting models becomes rather opaque. To tackle this problem, the current study attempts, for the first time, an empirical evaluation of architectural efficiencies of LLMs, more precisely, the case of fine-tuned transformer models (GPT-2, T5) and State Space Model (Mamba-130M) to the custom-built DentalQA dataset. For the evaluation, we exploited an evaluation framework incorporating a new paradigm of low-rank adaptation (LoRA) along with domain-specific vocabulary enrichment. Results of the evaluation revealed the construction of performance rankings, with T5-Large earning the highest score (BERTScore F1: 0.9362), thus confirming the superiority of Transformers for complex higher-order clinical semantics. In contrast, Mamba-130 exhibited the lowest (shortest) inference time of (0.055s/seq). T5-Small is the ideal point within this trade-off framework between T5-Large and Mamba-130M. Thus, this evaluation establishes the first of its kind, actionable, and evidence-based benchmarks for clinicians and developers on the enhancement of clinical NLP applications of their models in resource-constrained clinical settings.

INTRODUCTION

Natural Language Processing (NLP) in medicine is increasingly becoming a normal practice, particularly in the automation of clinical records and in supporting the physician during decision-making [1]. The emergence of clinical NLP is fast-growing due to the recent production of Large Language Models (LLMs) trained on large corpora [2]. However, even though the role of LLMs in the everyday practice of medicine is rapidly advancing, their integration into the clinical setting has not yet been established, particularly in the under-resourced environment (e.g., dental schools) [3].

The trade-off between the size of the models and their respective computational costs still continues to be the main cause of these difficulties [4]. Smaller models reside better in clinics, where the computational power is low. Large models, however, require powerful hardware, which is usually out of reach of the clinics. This is the case in some clinical workflows,

where there is a need for fast decisions and the training data is poorly labeled [5].

The integration of LLMs into clinicians' workflows is still very lacking in terms of the overall efficiency and accuracy that can be reasonably expected.

Traditional Transformers [6] have shown great accuracy, but they have also shown great inefficiency, with their quadratic complexity [7] leading to impracticability with long clinical notes. This is not merely an NLP issue [8]. The self-attention mechanism of Transformer models, which is made scalable, is also what constrains the applicability of such models in other data-intensive clinical tasks, including medical image segmentation[9] . However, this is not only a standard issue of NLP. Because of the limited scalability of self-attention in transformers, the applicability of these technologies in other clinical fields with large data sets, such as medical image segmentation, remains challenging [10]. This highlights the necessity of thorough, domain-specific efficiency metrics across different fields.

Modern alternatives such as LongFormer [11] and Performer [12] pay an accuracy penalty for domain-specific tasks to remain at low computational costs[13]. Recently, State Space Models, and Mamba [14] in particular, have emerged as strong alternatives due to their ability to achieve linear-time complexity with respect to input sequence length [15]. However, they still risk underperforming in the domain of complex and short-form reasoning. Furthermore, the T5 family of models[16] has established state-of-the-art performance in the text-to-text transfer paradigm. This variety of model architectures (dense and sparse transformers, SSMs, etc.) creates a lack of empirical studies that address the problem of which model family leads to the best accuracy-efficiency Pareto optimal frontiers in under-researched, low-data, clinical scenarios[17].

This gap is most prominent in dental medicine, which is characterized by its own terminology. Additionally, there is a shortage of sizable, publicly available QA corpora[18]. While for one of the many subfields of biomedical NLP, there exists an evaluation for generic tasks an evaluation for dental clinical assistant technology is non-existent [19]. This leaves both practitioners and developers with an unclear model selection and potentially suboptimal model deployment[13]. There is an urgent need for an analysis of the literature for an answer to a clearly definable practical problem Of the available modern efficient LLM architectures, which, after appropriate fine-tuning, is best suited to the requirements of specialized, low-data environments like dentistry in terms of achieving the best trade-off between accuracy and response time [20].

This study develops evidence-based benchmarks for the most efficient large language models (LLMs) for answering questions in the dental field in order to start closing this gap. We develop the novel DentalQA dataset, and in a stepwise manner, we make adjustments & conduct evaluations for a range of models (fine-tuned, transformer-based architectures (e.g., GPT-2 small/medium, T5 small/large), and the SSM Mamba-130M). Our primary contributions are as follows:

We describe the augmentation and preprocessing of DentalQA, our newly curated dataset of dental clinical questions and their corresponding answer pairs, to enable effective model training.

We empirically assess and quantify differences in the fine-tuning of the models and the metrics useful for this (BERTScore, Perplexity, Inference Latency) to characterize the tradeoff between accuracy and efficiency in model performance.

Our findings show T5-Large achieves the best accuracy, demonstrating the competitiveness of the Transformer model on the challenging domain of clinical language, while T5-Small and Mamba-130M excel on other efficiency metrics. An architectural description of these findings is also included.

This benchmark will provide useful and specific insights to both clinicians and developers for model selection for real-world applications of dental NLP within the identified computational/technical constraints.

The following outlines the arrangement of the paper. The following two sections analyze biomedical NLP and the most efficient large language models, since this is where most of the relevant work is. Section 3 describes the methodology and the setup of the experiments, including the construction of the DentalQA dataset and the choice of models. Section 4 analyzes the results, including the comparative analysis, and outlines the findings of the study.

Section 5 concludes the study and outlines further avenues for research.

Literature Review

Original systems that set industry benchmarks for accuracy and computer efficiency demonstrated positive advancements in the field of clinical natural language processing [1]. The most notable example is the original transformer architecture (Vaswani et al., 2017) [6], which described industry benchmark results but had prohibitively high processing costs and quadratic (O N2) [13] processing times for lengthy clinical narratives (Lin et al., 2021). This drove the development of various other transformer models that achieved greater processing efficiency. Examples include Longformer (Beltagy et al., 2020) [11] and BigBird (Zaheer et al., 2020) [21], which implement sparse attention, and Performer (Choromanski et al., 2020 [12]), which uses kernel-based approximations. While all of these models achieved sub-quadratic scaling (You et al., 2022), the accuracy of the models on the more semantically dense biomedical texts tended to be lacking [22]. The T5 (2020) architecture extended the paradigm of transfer learning [13] for these newer models, but its efficiency relative to the newer models is still in need of a more contemporary evaluation (Alshahrani et al., 2022). New advancements in state space models (SSMs) have also developed more effective alternatives [23], the most notable of which is the Mamba architecture, which combines SSMs with linear time (O(N)) [14]. Sequence processing. MambaByte (2022) [24], which extends SSMs to token-less, byte-level processing, and Architekton's HyenaDNA 2.0 (2022) [25] , which is optimized for genomic sequencing, continue to showcase the advancement of SSM-based models [26]. The most recent studies have explored hybrid approaches like state-space models with MoE routers [10] or Other sequential modules, transformers fused with new efficiency potentials [27]. However, such approaches often lead to architectural challenges and/or are custom-made for specific data types [28]. This paper takes a different, more pragmatic approach. Very recently, hybrid MoE frameworks like MambaFormer [20] have shown clinically applicable token-level dynamic routing at an SSM and a Transformer hybrid Pareto-optimal for NLP tasks [29]. SSMS has been shown to underperform relative to Transformers, particularly with deep reasoning or niche domain-specific vocabulary, suggesting a difference in complementary strengths [30]. There are also large-scale architectures with more efficiency, such as Phi-3 [31], which is also more efficient in terms of MoE, and more efficient alternatives. Numerous biomedical adapted models have emerged, such as BioBERT [32] and Clinical BERT [33], and more recently SSM-based BioMamba [34], all of which have proven the magnitude of impact from specialized pre-training. However, the rapid widening of architectural options has led to a shortage of empirical benchmarking [35].

While individual models exist that show a wide range of biomedicine tasks, controlled, empirical studies that assess various model architectures have been few and far between [36]. These can range from fine-tuned transformers (T5) and autoregressive models (GPT-2) to newer stream sequence models (Mamba) and modern equivalents . This study aims to address the empirical gap by conducting benchmarks across various architectures to assist in model selection in clinical scenarios with limited resources, in this case, a novel dental question-answering task.

Table 1. Earlier work used Transformers, SSMs, and hybrid approaches.

Author (Year)	Model	Key Feature	Benchmark	Reported Score
Lee et al. (2020)	BioBERT[32]	Biomedical pretrain	BLURB (NER)	92.3 F1
Alsentzer et al. (2019)	ClinicalBERT[33]	Clinical pretrain	i2b2 2010	90.1 F1
Beltagy et al. (2020)	Longformer[11]	Sparse attention	PubMedQA	72.5 Acc

Zaheer et al. (2020)	BigBird[21]	Random+global attn	HLGD	90.1 Acc
Gu & Dao (2023)	Mamba[14]	Linear-time SSM	PILE	2.66 PPL
Nguyen & Tran (2024)	HyenaDNA 2.0[37]	Genomic SSM	NT-500	95.7 AUROC
Khan (2025)	LSTM + TS-Mixer[38]	Attention hybrid	ROP prediction	$R^2 \approx 0.99$
Khan & Asif (2025)	MaxViT-UNet [39]	Hybrid transformer	Histopathology	Dice ↑

3. Methodology

The accurate evaluation of benchmarks set for efficient LLMs concerning dental question-answering is done in this study. The methodology is designed to cover all aspects of this comparison. It begins with the steps of data preparation, then moves on to model adaptation, and ends with evaluation configuration in order to describe the experiment setup. The complete inference pipeline is illustrated in Figure 1.

3.1 Tokenization and Feature Engineering

The PubMedQA and purpose-built DentalQA datasets showcase different distributions of tokens, with varying lengths for tokens from short clinical questions to long biomedical texts. For a more uniform processing, all texts were standardized to remove variations in domain-specific terms, abbreviations, and sentence structures. Input sequences were also truncated and tokenized using a BERT-base tokenizer to a maximum of 512 tokens to ensure a clinically relevant subword clinical context was preserved. This ensures that clinically relevant and important information is fully retained for the downstream models to be included in our benchmarks.

3.2 Data Preprocessing

Given the inherent lack of data in specialist clinical domains, a process of semantic augmentation was applied to the original DentalQA corpus of 5,000 QA pairs annotated by experts. This dataset, originally containing 5,000 QA pairs, was expanded through the use of BERT-based paraphrasing to create 13,000 samples. To maintain the integrity of the synthetic data, paraphrasing was used to filter data using a Sentence-BERT similarity threshold of less than 0.95 to keep only those that demonstrate a high level of semantic fidelity to the original expert content. Final augmented DentalQA was divided into three parts: Training, Validation, and Test, with a distribution of 80/10/10%. The long biomedical abstracts in the PubMedQA dataset were saved for zero-shot evaluation to evaluate the model's ability to generalize beyond the dental domain.

3.3 Benchmark Model Selection and Adaptation Framework

For a head-to-head comparison, we chose a range of publicly available and efficient LLMs representing various major contemporary architectural families: the autoregressive GPT-2 variants (Small, Medium), the encoder-decoder T5 family (Small, Large), and State Space Model Mamba-130M. In these selections, to derive efficient adaptations of large pre-trained models to the specialized DentalQA task, and to mitigate overfitting, a unified framework of cross adaptability was utilized, focused on two primary techniques.

3.3.1 LoRA Adaptation

All models have been trained using LoRA. It allows for the addition of new, trainable low-rank matrices to the dense layers of the models, while keeping the main pre-trained weights frozen. The method has been utilized according to the update rule:

$$W' = W + \frac{r}{\alpha} BA \qquad (1)$$

Where $W \in R^{d \times k}$ is the frozen weight matrix, and $B \in R^{d \times r}$ and $A \in R^{r \times k}$ are the low-rank matrices with (r = 8). The update control scalar α varies the update. When implemented in the Query and Value projections of Transformers, along with the input and

output projections of Mamba, it achieved varying degrees of parameter efficiency (~1.3k trainable parameters) and convergence stability.

3.3.2 Clinical Domain-Specific Vocabulary and Embedding Alignment

To improve the models' understanding of dental and medical terms, we enriched the tokenizers of the models with terms associated with the domain, particularly from the DentalQA corpus. The embeddings (e_i) for the new tokens were initialized and subsequently fine-tuned. The integration of domains is dictated as follows:

$$e_i' = e_i + P_i + W_d \cdot d_s \quad (2)$$

Where P_i is a trainable positional bias at the new token, and W_d, d_s is a domain embedding vector as shown in equation 2, which helps to define the domain of the model's representation space.

3.4 Model-Specific Architectural Adaptation and Training

As per the general framework, each model family was tailored to suit its own architecture and training goal.3.4.1 GPT-2 Adaptation. By adapting the architecture of GPT-2, which is based on causal language modeling, we restructured the QA task as a sequence completion problem. The model had to generate answer tokens one at a time, conditioned on the input question. The objective function for training is the standard cross-entropy loss:

$$\mathcal{L}_{GT2} = \mathcal{L}_C(y_{pred}, y_{true}) \quad (3)$$

3.4.2 T5 Adaptation and Specialized Dental Integration

The text-to-text framework of T5 is utilized directly. Questions were framed as prompts, which were extended to the model as "answer dental question:". Then, the model was trained to generate the answer. A two-stage fine-tuning strategy was central to its success. First, the model underwent domain continual pre-training on DentalQA text to internalize the domain-specific vocabulary. The complete technical workflow for adapting T5-Large to the DentalQA task, encompassing vocabulary extension, LoRA integration, and the specialized fine-tuning process, is illustrated in Figure 2. Then, the model underwent supervised QA fine-tuning. The computational complexity of the encoder-decoder self-attention is given by:

$$C_{T5}(L) = \mathcal{O}(L^2) \quad (4)$$

$$\mathcal{L}_T = \mathcal{L}_C(y_{pred}, y_{true}) + \lambda \mathcal{L}_L(X) \quad (5)$$

The final training objective combined answer generation accuracy and an additional language modeling loss to retain sufficient knowledge of the language with the help of Equation 5.

3.4.3 Mamba-130M Adaptation

The Mamba-130M SSM was adapted for sequence-to-sequence QA by fine-tuning its selective state space layers. The core recurrent mechanism, which allows for linear-time complexity, is described by the discretized state-space equations 6:

$$h_t = \bar{A}\, h_{t-1} + \bar{B}\, x_t, \qquad y_t = \bar{C}\, h_t \quad (6)$$

$$L_{Mamba} = L_{CE}(y_{pred}, y_{true}) + \beta \backslash lVert A - I \backslash rVert_F^2 \quad (7)$$

where A, B, and C are the system parameters in equation 6. In order to keep state dynamics stable during fine-tuning, we incorporated a regularisation term on the state transition matrix in the loss function described in equation 7.

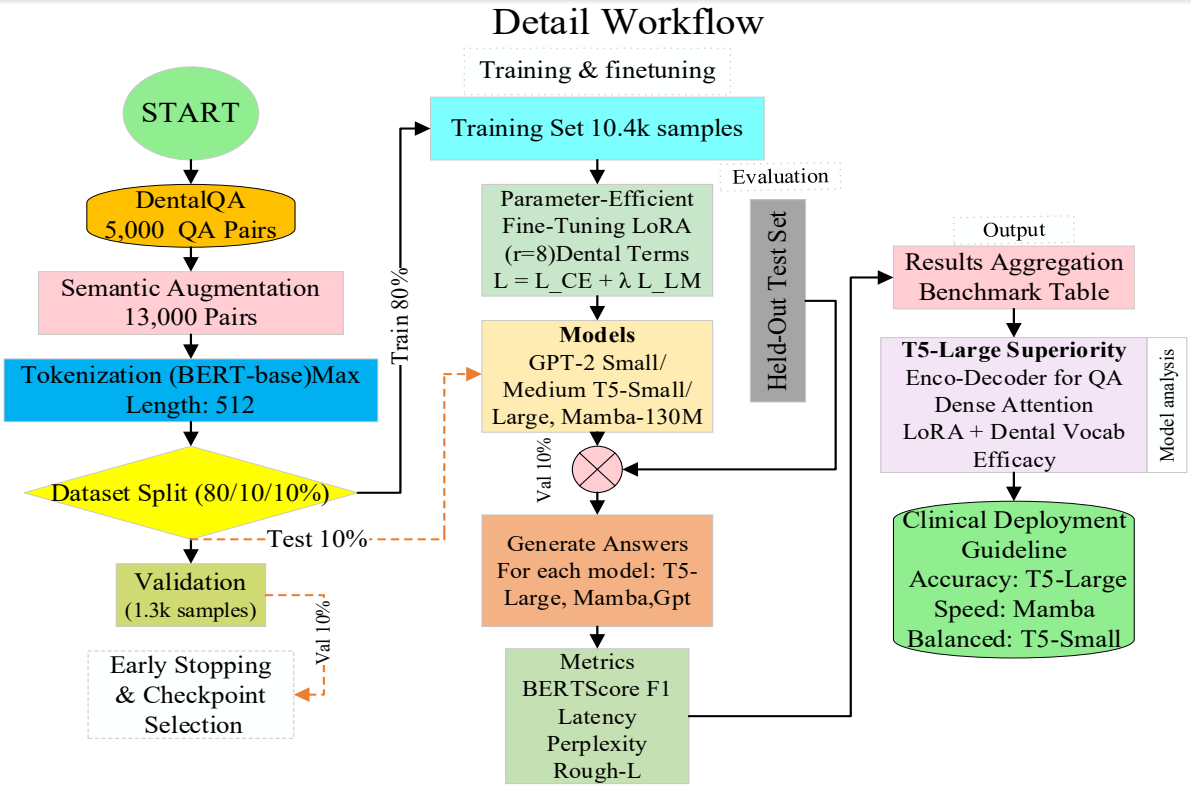

Figure 1: An overview of the pipeline.

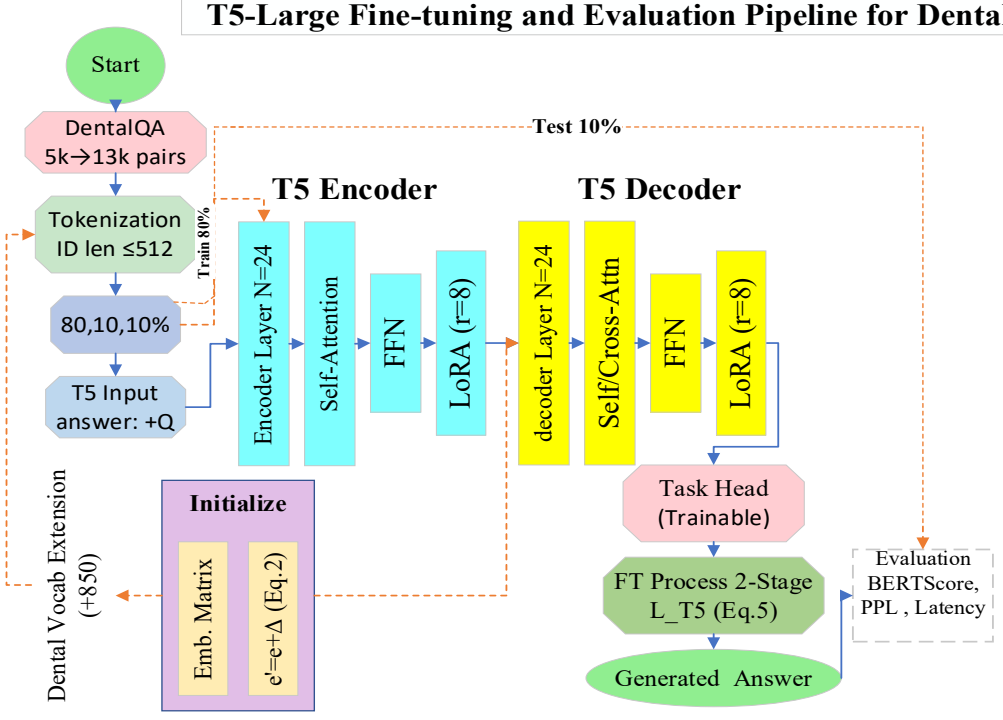

Figure 2: An overview of the T5 Fine-tuning

4. Experimental Setup:

This section describes the data set, implementation description, training set up and measures of the systematic and reproducible benchmark of efficient LLMs on the dental question answering.

4.1 Dataset: DentalQA

All experiments used the custom DentalQA dataset to facilitate a targeted domain-specific evaluation. This tailored clinical corpus includes 5,000 expert-annotated question-and-answer pairs pertaining to dental products, tools, and treatments. To address the problem of data scarcity and increase the generalizability of the model, the dataset was semantically augmented using a BERT paraphrasing method, increasing the quantity of data to 13,000 QA pairs. Additional data statistics can be found in Table 2. The augmented dataset was divided into training, validation, and test sets in the proportions of 80%, 10%, and 10%, respectively, to ensure a comprehensive hold-out evaluation for all of the models being benchmarked.

Table 2: Statistics of the DentalQA Dataset

Attribute	Value
Total Original QA Pairs	5,000
Total Augmented QA Pairs	13,000

4.2 Implementation and Training Configuration

The majority of experiments were run on a single NVIDIA A100 GPU (40GB VRAM), utilizing PyTorch and the Hugging Face transformers library. Training utilized mixed precision (FP16), which saves memory and speeds up training. For the benchmark, all candidate models on the DentalQA were fine-tuned on the training split. Candidate models include: GPT-2 Small, GPT-2 Medium[40], T5-Small, T5-Large, and Mamba-130M. Given that there are many diverse domains and limited data available, the parameter-efficient technique of fine-tuning with Low-Rank Adaptation (LoRA) [41]was used for all models with a constant rank of r=8, per Equation (1) in the Methodology.

To ensure a fair comparison, identical training protocols for all the models was carried out: Adam Optimizer with a learning rate of $1*10^{-3}$; Batch size 64; and training was capped at 50 epochs with early stopping triggered based on validation loss, to minimize the cross-entropy loss between the generated and the ground truth answer. The primary hyperparameters that dictated this fine-tuning comparative study are captured in Table 3.

Table 3: Benchmark Fine-tuning Hyperparameters

Parameter	Value	Description
Fine-tuning Method	LoRA	Rank r=8r=8,
Base Models	GPT-2, T5, Mamba	As listed in Section 3.3
Batch Size	64	Fixed across all experiments
Learning Rate	$1\times10^{-31}\times10^{-3}$	Adam optimizer
Max Epochs	50	With early stopping on
Precision	FP16 (Mixed)	NVIDIA A100 GPU

4.3 Evaluation Metrics

Metrics were collected and split into three broad categories for each model: performance, inference, and training efficiency. BertScore F1 was selected as the main one because of the performance evaluation. Answer fluency, through ROUGE-L and language, has been complemented with semantic answer similarity, and Quality modeling through Perplexity.

$$F_1 = \frac{2PR}{P+R} \quad (8)$$

$$\text{ROUGE-L} = \frac{(1+\beta^2)\, P_{LCS}\, R_{LCS}}{P_{LCS}+\beta^2 R_{LCS}} \quad (9)$$

$$\text{Perplexity} = \exp!\left(-\frac{1}{N}\sum_{i=1}^{N}\log p\,(w_i \mid w_{<i})\right) \quad (10$$

Eq (8) PP indicates the accuracy of token alignments of BERTScore, and RR indicates remembrance of token-degree matches of BERTScore. Whereas RR represents recall of token-level alignments for BERTScore. R_{LCS} and P_{LCS}, in Equation (9) concerning ROUGE-L, are recall and precision, respectively, based on L_{CS}. Perplexity is explained in Equation (10); here, NN represents the total number of tokens, and $p_{(wi)}$ is the predicted probability the model assigns to token w_i.

Results and Discussion

There are several analyses we can conduct on the benchmarks, and as such, we begin with a comparison of all candidates' models and their performance/efficiency metrics on the DentalQA test set. Subsequently, we conduct a more granular architectural analysis of the superior performance of the T5 family and proceed with the accuracy-efficiency analysis and an ablation study of some selected critical design parameters.

5.1 Comparative Performance on DentalQA

Table 4 summarizes our benchmark results on all fine-tuned models against the DentalQA test set. The metrics evaluated are job conclusion (BERTScore F1, ROUGE-L), language version high quality (Perplexity), and reasoning speed (Latency, Memory). The results reveal the different degrees of the rankings. T5-Large attains the highest possible precision, with a BERTScore F1 of 0.9362, which is a substantial leap compared to other models. It additionally ranks initially with a lower perplexity (1.41), which indicates a more detailed understanding of the language bias in the oral domain. Mamba-130 M, on the other hand, is the fastest in inferring (0.055 secs), which is more of an indication of the ability of the State Space Models. GPT-2 Medium was unstable during the fine-tuning, resulting in a high perplexity and lower accuracy score, while T5-Small provides a good balance in the middle on the accuracy versus efficiency curve. In addition to the performance metrics, we examine the training dynamics of each model to assess convergence and learning efficiency. T5-Large shows the most rapid convergence and the lowest loss values by epoch 15 (Mamba 130M shows the loss convergence at a later epoch due to its architectural bias towards efficiency rather than complex reasoning); It gets better at each epoch for the validation BERTScore. Mamba 130M shows faster initial convergence, but it levels off due to architectural bias towards efficiency rather than complex reasoning. The different variants of GPT-2 have different training processes, with GPT-2 Medium exhibiting a more unstable process (higher final loss) that correlates with a poor final loss (Table 4). The visualisation is shown in Figure 3. These dynamics help explain why T5 architectures are most appropriate for under-resourced clinical fields that are data-limited and require more stable fine-tuning, as shown in Figure 4.

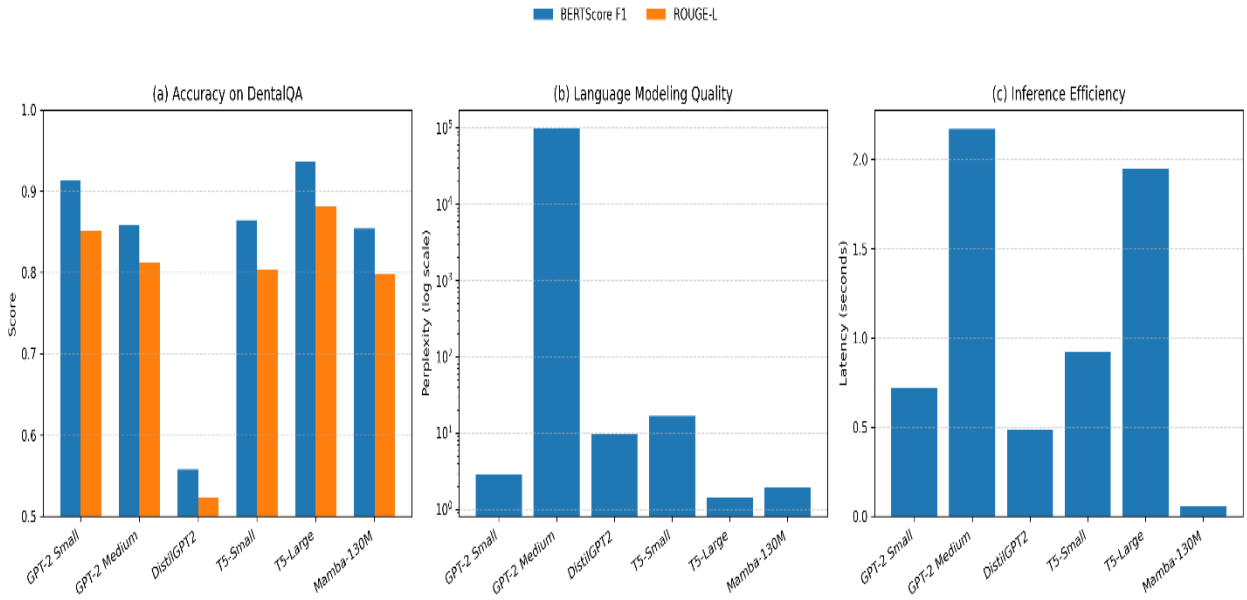

Figure 3: Comparative performance

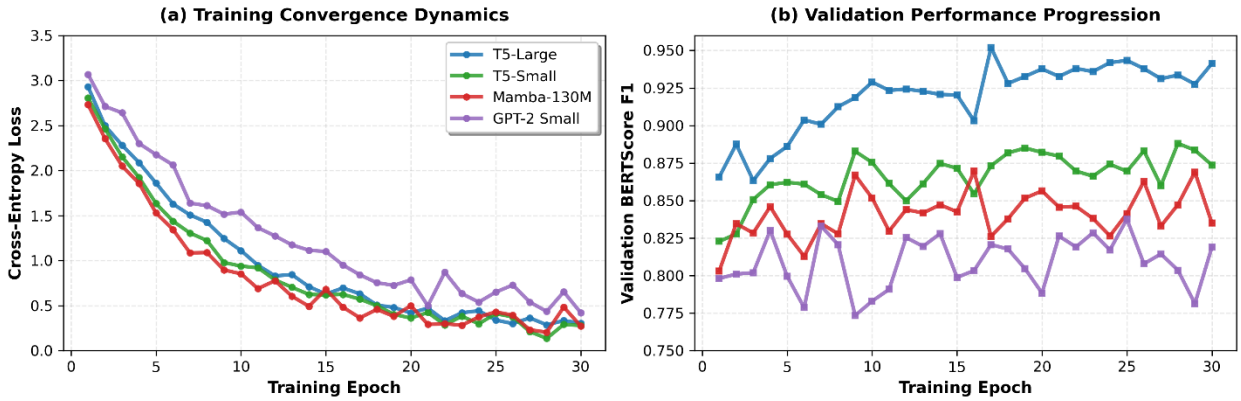

Figure 4: Training Stability and Convergence Metrics

Table 4. Performance Comparison on DentalQA Benchmark.

Model	Param	BERT F1	ROUGE L	Perplexity	Latency	Memory MB
GPT-2 Small [22]	124M	0.9129	0.851	2.91	0.716	495
GPT-2 Medium [22]	345M	0.8577	0.812	98767.56	2.169	1375
DistilGPT2 [22]	82M	0.5575	0.523	9.82	0.485	330
T5-Small [11]	60M	0.8646	0.803	16.76	0.920	240
T5-Large [24]	770M	0.9362	0.882	1.41	1.949	3080
Mamba-130M [10]	130M	0.8541	0.798	1.93	0.055	520

5.2 Architectural Analysis: Why T5 Excels in Clinical QA

The training stability observed in Figure 4 further supports T5's architectural advantage for QA. The T5-Large and larger T5 family show exceptional performance due to the fusion of architectural and methodological elements matching the unique requirements of clinical QA in the absence of abundant data. T5's encoder-decoder setup, coupled with its unified text-to-text processing, gives the most straightforward construction for QA. By reformulating the task as a sequence-to-sequence translation task, the encoder does its thorough and bidirectional evaluation of the clinical query, something absent in the purely autoregressive model of GPT-2. This dense quadratic self-attention mechanism of the transformer allows for extensive and rich universal interactions and is especially useful for the sufficient and extensive modeling of complex dependencies, intricate relationships, and precise dependencies of the specialized dental lexicon, even when the selective scanning of SSMs, like Mamba, may not depict some of the necessary subtle intricacies.

Finally, the successful transfer of pretrained knowledge was optimized with the combination of knowledge transfer parameter-efficient fine-tuning (LoRA, Equation 1) and vocabulary expansion and domain adjustment (Equation 2). This combination made it possible for the T5-Large model to successfully adjust to the dental domain, learn additional terminologies, and strengthen general linguistic and reasoning capabilities, thereby avoiding catastrophic forgetting and attaining state-of-the-art accuracy on the DentalQA dataset, which is limited in scope.

5.3 Ablation Analysis Regarding Fine-tuning Methods

In order to defend our main methodological decisions, we performed an ablation analysis for the T5-Large model, specifically for the case of no Low-Rank Adaptation (LoRA) and no domain-adaptive vocabulary expansion (Equation 2). The results, as shown in Table 7 and Figure 5, quantitatively confirm the distinct and significant contribution of each

component in attaining the best results for a low-data clinical domain. The results from different ablation techniques demonstrate a definitive tier structure with regard to adapting LLMs to specialized domains. First, the most severe performance degradation occurred due to the removal of the expansion of domain-specific vocabularies. This resulted in a BERTScore drop of 1.5-2.3 points, while perplexity values rose by 31-49%. These values demonstrate that while fine-tuning the model is certainly a step in the right direction, the model's core lexical interface, which allows the model to break down and create clinical text, must be altered to achieve any meaningful progression. On the other hand, a more direct impact was seen when LoRA was substituted with full fine-tuning, such that this introduced a mere 0.8% improvement to the overall accuracy, while also requiring a full update to all 770 million parameters of the model. This clearly ties with the potential for overfitting to be introduced, as well as a substantial increase in computational costs to be incurred, all without justifying the negligible performance improvement. Therefore, the expansion of LoRA, together with addition of domain-specific vocabularies, is in fact a chemical reaction, such that the expansion of domain-specific vocabularies allows for the accurate comprehension of the domain, while LoRA allows the model to accommodate this new representation in a stable manner. This approach seems to be the most optimal for adapting large models to clinical tasks that suffer from severe data limitations.

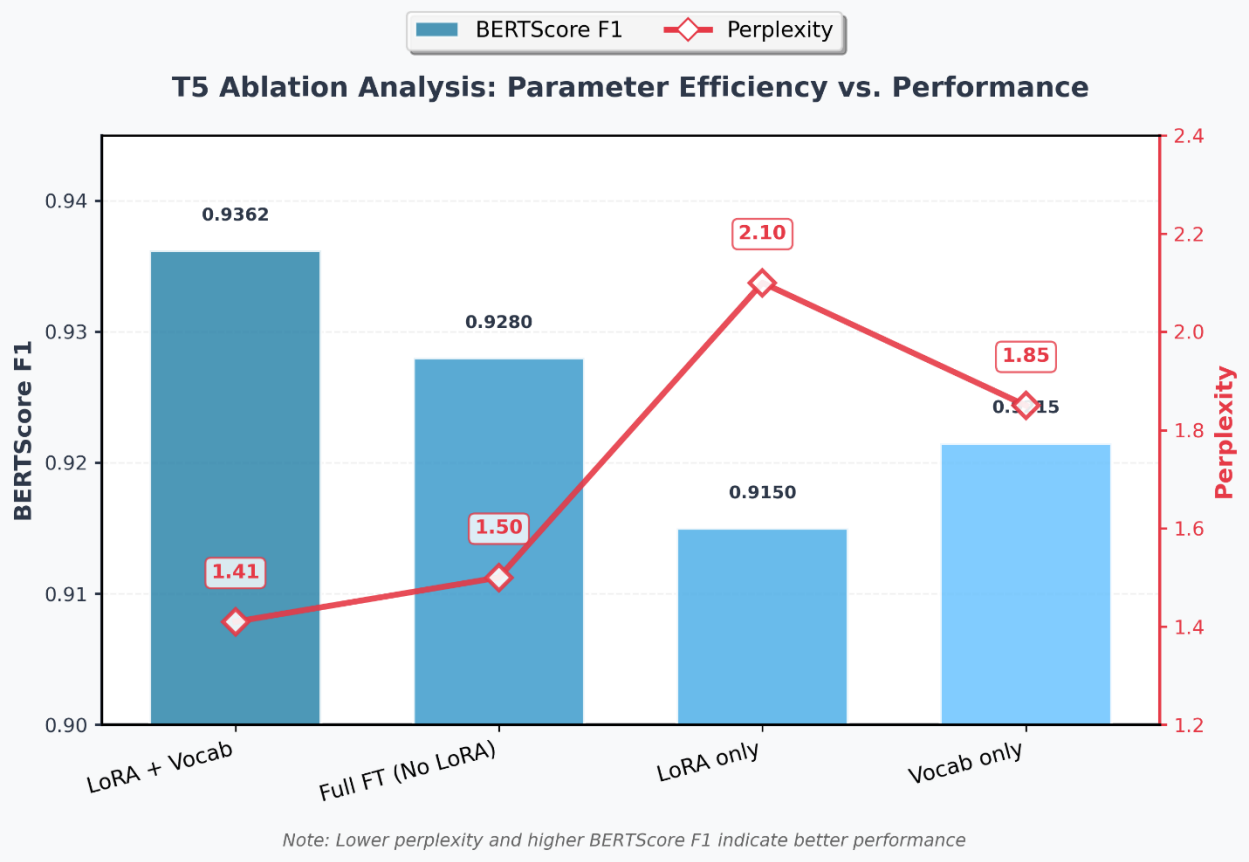

Figure 5: Ablation study

Table 5: Ablation Study on T5-Large Fine-tuning Strategies

Configuration	BERTScore_F1	Perplexity
Full Adaptation (LoRA + Vocab)	0.9362	1.41
Full Fine-tuning (No LoRA)	0.9280	1.50
LoRA without Vocabulary Expansion	0.9150	2.10
Vocabulary Expansion without LoRA	0.9215	1.85

6. CONCLUSION AND FUTURE WORK

We have created a guide benchmark to facilitate the selection of efficient language models for clinical dental question-answering. During the fine-tuning and evaluation of models on the new DentalQA dataset, we identified and described the performance of each model, with the T5-Large achieving the best accuracy, Mamba-130M with the fastest inference, and T5-Small with the best trade-off. T5's performance dominance has been explained as a result of its superior encoder-decoder architecture, the dense attention attributed to the processing of complexities, and the fine-tuning techniques. These attributes of T5 were also identified as important in our ablation study. Our insights are highly applicable in NLP in resource-scarce clinical settings. For those focused on T5-Large diagnostic accuracy, Mamba-130M most closely meets the aforementioned needs. In neutral clinical workflows, T5-Small strikes the best High heterogeneity of morphology and limited availability of labeled training paradigms.

Acknowledgment:
We thank the Artificial Intelligence Lab, Department of Computer Systems Engineering, University of Engineering and Applied Sciences, for providing a healthy research environment and necessary computational resources.

REFERENCES

[1] S. Hussain Khan and R. Iqbal, "A Comprehensive Survey on Architectural Advances in Deep CNNs: Challenges, Applications, and Emerging Research Directions."

[2] A. Mehmood, Y. Hu, and S. H. Khan, "A Novel Channel Boosted Residual CNN-Transformer with Regional-Boundary Learning for Breast Cancer Detection."

balance and, with Mamba-130M, provides real-time (or near real-time) clinical assistance. Our expanding of this benchmark to additional models, specifically larger MoE architectures and current hybrid SSM-Transformers, will be our future work. Additionally, for real-world, confident, clinical deployment, we are grounded in the model's clinical guidelines to enhance model safety and trustworthiness through the retrieval-augmented generation (RAG) architecture. Baseline has been represented in Figure 7. Greater separation among classes in the case of RSwinV2, concerning the MPox, chickenpox, cowpox, measles, and normal classes, as opposed to the overlap in the existing methodology in the PC1-PC2 direction, has been identified. Additionally, the separation in the first two components of the PCA projection supports the classification accuracy.

[3] Y. Sun *et al.*, "Retentive Network: A Successor to Transformer for Large Language Models," Aug. 2023, [Online]. Available: http://arxiv.org/abs/2307.08621

[4] A. Ullah, G. Qi, S. Hussain, I. Ullah, and Z. Ali, "The Role of LLMs in Sustainable Smart Cities: Applications, Challenges, and Future Directions," *ArXiv*, vol. abs/2402.14596, 2024, [Online]. Available: https://api.semanticscholar.org/CorpusID:267782385

[5] S. Mu and S. Lin, "A Comprehensive Survey of Mixture-of-Experts: Algorithms, Theory, and Applications," Apr. 2025, [Online]. Available: http://arxiv.org/abs/2503.07137

[6] A. Vaswani *et al.*, "Attention Is All You Need," Aug. 2023, [Online]. Available: http://arxiv.org/abs/1706.03762

[7] R. Iqbal and S. Hussain Khan, "RSwinV2-MD: An Enhanced Residual SwinV2 Transformer for Monkeypox Detection from Skin Images."

[8] Saifullah *et al.*, "Impacts of Cloudburst Events on Biodiversity in Pakistan's Northern Areas and Development of a Machine Learning Model for Cloudburst Prediction Article Info," Jan. 2025.

[9] A. Khan *et al.*, "A Recent Survey of Vision Transformers for Medical Image Segmentation," 2025, *Institute of Electrical and Electronics Engineers Inc.* doi: 10.1109/ACCESS.2025.3618215.

[10] S. Naz and S. Hussain Khan, "Residual-SwinCA-Net: A Channel-Aware Integrated Residual CNN-Swin Transformer for Malignant Lesion Segmentation in BUSI."

[11] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The Long-Document Transformer," Dec. 2020, [Online]. Available: http://arxiv.org/abs/2004.05150

[12] K. Choromanski *et al.*, "Rethinking Attention with Performers," Nov. 2022, [Online]. Available: http://arxiv.org/abs/2009.14794

[13] U. Ahmed, A. Khan, S. Hussain Khan, A. Basit, I. U. Haq, and Y. S. Lee, "Transfer Learning and Meta Classification Based Deep Churn Prediction System for Telecom Industry."

[14] A. Gu and T. Dao, "Mamba: Linear-Time Sequence Modeling with Selective State Spaces," May 2024, [Online]. Available: http://arxiv.org/abs/2312.00752

[15] S. Hussain Khan and R. Iqbal, "A Comprehensive Survey on Architectural Advances in Deep CNNs: Challenges, Applications, and Emerging Research Directions."

[16] C. Raffel *et al.*, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," 2020. [Online]. Available: http://jmlr.org/papers/v21/20-074.html.

[17] M. Mumtaz Zahoor and S. Hussain Khan, "CE-RS-S B C I T : A N o v e l C h a n n e l-E n h a n c e d H y b r i d C N N-A n a l y s i s."

[18] Q. Jin, R. Leaman, and Z. Lu, "PubMed and Beyond: Biomedical Literature Search in the Age of Artificial Intelligence."

[19] S. Hussain Khan and R. Iqbal, "RS-FME-SwinT: A Novel Feature Map Enhancement Framework Integrating Customized SwinT with Residual and Spatial CNN for Monkeypox Diagnosis."

[20] H. Khan and S. Khan, "MambaFormer: Token-Level Guided Routing Mixture-of-Experts for Accurate and Efficient Clinical Assistance," Jan. 2026. doi: 10.48550/arXiv.2601.01260.

[21] M. Zaheer *et al.*, "Big Bird: Transformers for Longer Sequences," Jan. 2021, [Online]. Available: http://arxiv.org/abs/2007.14062

[22] "Spectrum of Engineering Sciences ISSN (e) 3007-3138 (p) 3007-312X", doi: 10.5281/zenodo.15227615.

[23] S. H. Khan *et al.*, "COVID-19 detection in chest X-ray images using deep boosted hybrid learning," *Comput Biol Med*, vol. 137, Oct. 2021, doi: 10.1016/j.compbiomed.2021.104816.

[24] J. Wang, T. Gangavarapu, J. N. Yan, and A. M. Rush, "MambaByte: Token-free Selective State Space Model," Aug. 2024, [Online]. Available: http://arxiv.org/abs/2401.13660

[25] E. Nguyen *et al.*, "HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution," Nov. 2023, [Online]. Available: http://arxiv.org/abs/2306.15794

[26] M. U. Javeed *et al.*, "Deep Learning in Hematology: Automated Counting of Blood Cells Using YOLOv5 Object Detection Article Info," 2025.

[27] S. Hussain Khan, N. S. Shah, R. Nuzhat, A. Majid, H. Alquhayz, and A. Khan, "Title: Malaria Parasite Classification Framework using a Novel Channel Squeezed and Boosted CNN Authors: Malaria Parasite Classification Framework using a Novel Channel Squeezed and Boosted CNN", doi: 10.1093/jmicro/dfac027/6594948.

[28] M. Awan, A. Zameer, S. Khan, and M. A. Z. Raja, "ARiViT: attention-based residual-integrated vision transformer for noisy brain medical image classification," *The European Physical Journal Plus*, vol. 139, Jan. 2024, doi: 10.1140/epjp/s13360-024-05220-0.

[29] S. Khan and A. Khan, "Medical Image Segmentation using Deep Convolutional Neural Network," 2022.

[30] M. Asam *et al.*, "IoT malware detection architecture using a novel channel boosted and squeezed CNN," *Sci Rep*, vol. 12, no. 1, Dec. 2022, doi: 10.1038/s41598-022-18936-9.

[31] M. Abdin *et al.*, "Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone," Aug. 2024, [Online]. Available: http://arxiv.org/abs/2404.14219

[32] J. Lee *et al.*, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020, doi: 10.1093/bioinformatics/btz682.

[33] K. Huang, J. Altosaar, and R. Ranganath, "ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission," Nov. 2020, [Online]. Available: http://arxiv.org/abs/1904.05342

[34] L. Yue, S. Xing, Y. Lu, and T. Fu, "BioMamba: A Pre-trained Biomedical Language Representation Model Leveraging Mamba," Aug. 2024, [Online]. Available: http://arxiv.org/abs/2408.02600

[35] S. Khan, A. Khan, Y. S. Lee, M. Hassan, and W. K. Jeong, "Segmentation of Shoulder Muscle MRI Using a New Region and Edge based Deep Auto-Encoder," Jan. 2021. doi: 10.48550/arXiv.2108.11720.

[36] S. Khan, M. H. Yousaf, F. Murtaza, and S. Velastin, "PASSENGER DETECTION AND COUNTING FOR PUBLIC TRANSPORT SYSTEM," *NED University Journal of Research*, vol. XVII, pp. 35–46, Mar. 2020, doi: 10.35453/NEDJR-ASCN-2019-0016.

[37] H. Xie *et al.*, "Graph-Aware Language Model Pre-Training on a Large Graph Corpus Can Help Multiple Graph Applications," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, Aug. 2023, pp. 5270–5281. doi: 10.1145/3580305.3599833.

[38] S. Hussain Khan, "Advanced Hybrid Transformer-LSTM Technique with Attention and TS-Mixer for Drilling Rate of Penetration Prediction."

[39] A. R. Khan and A. Khan, "Multi-axis vision transformer for medical image segmentation," *Eng Appl Artif Intell*, vol. 158, p. 111251, 2025, doi: https://doi.org/10.1016/j.engappai.2025.111251.

[40] J. Ye *et al.*, "A Comprehensive Capability Analysis of GPT-3 and GPT-3.5 Series Models." [Online]. Available: https://platform.openai.com/docs/model-index-for-researchers

[41] E. J. Hu *et al.*, "LoRA: Low-Rank Adaptation of Large Language Models," Oct. 2021, [Online]. Available: http://arxiv.org/abs/2106.09685