Pertanika J. Sci. & Technol. 34 (3): 1685 - 1702 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Studying the Effects of Knowledge Distillation and the Fragility of Lightweight BERT-based Models

Allistair Nallie Konsil[1]*, Stephanie Chua[1], Jacey Lynn Minoi[1], Md Mizanur Rahman[2], Gerraint Gillan[2], Rafazila Ramli[2], Rasitasam Safii[3], Ahmad Sofian bin Shminan[4], and Lee Jun Choi[4]

[1]Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia
[2]Department of Community Medicine and Public Health, Faculty of Medicine and Health Sciences, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia
[3]Kulliyyah of Medicine, International Islamic University Malaysia, Jalan Sultan Ahmad Shah, Bandar Indera Mahkota, 25200 Kuantan, Pahang, Malaysia
[4]Faculty of Cognitive Sciences and Human Development, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia

ABSTRACT

Large language models (LLMs) such as BERT (Bidirectional Encoder Representations from Transformers) have caused strides of progress in the field of natural language processing (NLP), with lightweight variants such as DistilBERT, MobileBERT and TinyBERT being developed to lower the resource requirements to deploy the models in real-world settings. However, while past research has investigated improving non-distilled models on a variety of benchmarks by changing their architecture, there are limited studies that explore how the lightweight variants may perform in the same conditions. This study applies modifications and ensemble techniques on lightweight BERT models in extractive question answering (QA) to address this gap. The experiments were conducted using three datasets: SQuAD, AdversarialQA, and a newly curated Sexual and Reproductive Health QA (SRHQA) dataset consisting of 1000 samples. From the results, it was shown that applying the same modifications that would enhance the base BERT models to the lightweight variants generally caused a 0.14 - 5.20% F1 decrease in performance, with marginal exceptions observed in specific dataset-model combinations.

ARTICLE INFO

Article history:
Received: 08 December 2025
Accepted: 06 June 2026
Published: 19 June 2026

DOI: https://doi.org/10.47836/pjst.34.3.17

E-mail addresses:
allistair99@gmail.com (Allistair Nallie Konsil)
chlstephanie@unimas.my (Stephanie Chua)
jacey@unimas.my (Jacey Lynn Minoi)
rmmizanur@unimas.my (Md Mizanur Rahman)
22010275@siswa.unimas.my (Gerraint Gillan)
22010339@siswa.unimas.my (Rafazila Ramli)
razitasham@iium.edu.my (Rasitasam Safii)
sasofian@unimas.my (Ahmad Sofian bin Shminan)
cljun@unimas.my (Lee Jun Choi)
* Corresponding author

e-ISSN: 2231-8526

Ensembling, on the other hand, showed improvements across all the datasets, ranging from 2.19 - 19.46% F1 over the BERT baseline. The results of the study highlight the sensitivity of the lightweight models, their trade-offs with efficiency, and that an ensemble is a valid approach to utilising the lightweight models without architectural modifications.

Keywords: BERT, ensemble, lightweight models, question answering, robustness

INTRODUCTION

Transformer-based large language models such as BERT have been shown to perform quite well compared to previous architectures in several NLP tasks, with key advantages being their ability to process more tokens in parallel and being able to produce high-quality embeddings. However, one of the drawbacks of the base BERT model is that it is computationally intensive, making it difficult to deploy where resources are limited, such as edge devices. This led to the development of lightweight variants like DistilBERT, MobileBERT and TinyBERT, which apply several techniques to reduce their overall parameter count.

While these compressed models have their benefits of being faster and having a lower barrier for use, there has been some research that shows they are also more fragile. Their performance varies when they are trained in architectural or when retrained in more specialised domains. This makes them less reliable and brings into question their ability for use in more high-stakes domains such as sexual and reproductive health (SRH), where misinformation can have dire consequences.

Ensembling shows promise in enhancing the stability of LLMs by aggregating multiple predictions, which in turn reduces variance and the effects of outliers. However, this technique is often utilised for its ability to edge out extra performance on benchmarks using large non-distilled models, with limited usage for lightweight models in more niche areas like extractive question answering on a domain. With these gaps in mind, this study investigates how lightweight BERT-based models perform on a mix of general and domain-specific datasets for extractive question answering after applying architectural modifications, while also looking into how ensembling can be utilised to improve the performance using the modified models.

Related Works

Transformer Model

A large part of research done in NLP before 2017 utilised neural networks such as recurrent neural networks (RNNs) or long short-term memory networks (LSTMs), which are sequential in nature, with a major restriction being that they would become exponentially more expensive as the sequence length grew.

The introduction of the Transformer architecture (Vaswani et al., 2017) was a paradigm shift because it showed that neural networks can be trained to process tokens simultaneously instead of just sequentially. They utilised self-attention to create high-quality embeddings that translated into improved performance on benchmarks. The design choices also allowed for faster training and iteration on longer texts; other than that, they used positional encoding, which helped preserve the word order, while residual connections reduced the problem of vanishing gradients and improved stability. All these design choices contributed to the Transformer architecture achieving state-of-the-art results on a variety of NLP tasks, such as machine translation and sentiment analysis, surpassing previous architectures at the time.

The Transformer architecture, which was originally used for machine translation, consisted of an encoder and decoder component. However, researchers realised you can utilise just the encoder or decoder for specialised tasks, for example, a decoder-only architecture such as GPT (generative pretrained transformer) has been shown to excel at auto-regressive sequence modelling and is effective at open-ended text generation. Other than that, BERT (Devlin et al., 2019), which is an encoder-only model pretrained on two pre-training tasks, masked language modelling and next sentence prediction, has been shown to produce high-quality embeddings that significantly improve the performance on downstream tasks. However, due to their large size, these models are computationally expensive.

Lightweight BERT Models

With the success of large pretrained transformer models and the demand for a more practically deployable solution that can be utilised in resource-constrained conditions, this led to research branching into increasing efficiency. DistilBERT Sanh et al. (2019) was one of the earliest studies in reducing the parameter count of the base BERT model. This was done by reducing the number of encoder layers and using knowledge distillation to make the remaining layers mimic the teacher model; they were able to retain 95% of BERT's performance with a reduced parameter count. DistilBERT showcased that transferring knowledge from a large teacher model to a smaller student model is a viable strategy for reducing the total parameter count and the computational cost of inference.

MobileBERT Sun et al. (2020) was a study aiming to develop a model that could be used on edge devices. They utilised architectural modifications such as bottleneck structures and thin layers to compress the model, retaining performance. MobileBERT also applies knowledge distillation to distil the intermediate representations and predictions from the teacher BERT.

TinyBERT Jiao et al. (2020) focus on knowledge distillation, where they employ a two-stage distillation process, first by doing a general distillation, then a task-specific

distillation. For both general and task-specific distillation, they investigated passing the generated embeddings, attention and hidden layers from the teacher to the student while still retaining performance even with drastically reducing the number of parameters. The three models utilised knowledge distillation to reduce the parameter count while retaining their performance, allowing them to be used in contexts where a full-sized BERT model may be too computationally taxing. However, there is still the question of whether these models can really be utilised in areas where accuracy is critical, although they may be efficient and could be deployable with weak robustness and sensitivity to modifications, the trade-off may not be worth it.

Modifications to BERT

After the introduction of BERT, extensive usage showed that the generated embeddings were able to perform quite well in downstream tasks. This motivated researchers to try to create variants of the architecture that may edge out more performance. One of the most common and easily implemented modifications is the addition of layers utilising other neural network architectures, such as BiLSTM or convoluted neural networks, with the aim of adding better capture of sequential dependencies. For example, Dhar and Chauhan (2023) added BiLSTM layers into DistilBERT and showed that it could improve the performance of classification on web content datasets. Similarly, Fang et al. (2025) applied a BERT-BiLSTM model for detecting malicious comments and noted that they were able to get improved performance compared to the baseline BERT model.

Although these studies show that modifications as simple as adding another layer can improve performance, some studies have reported that for distilled models, these modifications may cause worsening performance. Diddee et al. (2022) and Lorenzoni et al. (2024) reported in their papers that some modifications to architecture or hyperparameters can disturb the optimised weights of lightweight models, which points to a trade-off between model performance and robustness in distilled models. For this study, fragility is defined as the tendency for performance to degrade when the models are subjected to changes in hyperparameters, architecture or subject domain.

Ensemble Learning

Ensemble learning is a technique that integrates the predictions of multiple models to improve the accuracy and robustness. It has been used extensively in NLP due to the technique being able to reduce variance and make up for the weakness of individual models. Traditional ensemble strategies such as bagging Breiman (1996) and boosting Schapire (1990) demonstrate how training either multiple independently sampled models or sequentially refined learners can yield stronger generalisation, while stacking Wolpert (1992) introduces a meta-learning framework that combines model outputs

through an additional predictive layer. Averaging is the selected ensemble method due to several reasons. Unlike the previously mentioned methods, averaging does not require ensemble-specific training, such as training a meta learner; this helps isolate the effects of modifications from improvements introduced during training.

Ensembles in Question Answering

The ensemble technique is effective in a variety of NLP tasks, including extractive question answering. For instance, Lee et al. (2019) combined BioBERT with BiLSTM-Attention layers for biomedical natural language inference, using majority voting to aggregate predictions from multiple models. Zhou et al. (2019) used weighted averaging with BERT and BiDAF models to achieve top results on SQuAD 2.0. Pranesh et al. (2020) introduced QuesBELM, a stacking ensemble of BERT-based models for Google's Natural Questions dataset, demonstrating the benefits of combining models of varying architectures.

In social media health text classification, Dang et al. (2020) applied ensembles of BERT-Large, Bio+Clinical BERT, and BERT-Base, achieving higher F1 scores through averaging and cross-validation. Pearce et al. (2021) added BiLSTM layers into BERT embeddings, forming a layered ensemble that improved extractive QA performance across NewsQA and CovidQA. These studies showed the potential and effectiveness of using ensembles for improving the performance and stability on a variety of tasks and datasets.

Despite extensive work on ensemble learning with large BERT models, there is limited research focusing on using lightweight BERT in ensembles. The robustness of distilled models when applying architectural modifications also remains underexplored, especially in extractive QA tasks. This study addresses these gaps by examining the sensitivity of lightweight BERT models to modifications and evaluating whether averaging ensembles can enhance robustness. The findings aim to highlight the effects of applying naïve modifications to optimised distilled models without further knowledge distillation.

MATERIALS AND METHODS

Data Sources and Collection

The study uses three datasets, all of which are in English, representing both general-purpose and domain-specific contexts for extractive question answering tasks. The first dataset is the Stanford Question Answering Dataset (SQuAD). SQuAD is a widely used benchmark dataset for extractive QA. It contains over 100,000 question-answer pairs based on Wikipedia articles with extractive QA requiring the model to determine the answer span of a question within a given context. The large sample size and comprehensive range of general topics make it a suitable dataset for setting the baseline model performance to compare with existing studies. The second dataset is AdversarialQA, which extends the QA challenge by introducing adversarial examples designed to test the robustness of QA systems.

During the development of the dataset, a human-in-the-loop approach was used, where a comparison of the F1 score from human and model predictions was done. Only questions where the base model achieves an F1 score below a threshold of 40% are retained. Adding AdversarialQA to the study allows us to evaluate how lightweight BERT variants handle complex contexts. Lastly, the third dataset is the Sexual and reproductive health question answering (SRHQA) dataset, which was custom-curated for this research. It focuses on the topic of sexual and reproductive health, an area where accuracy is critical due to the consequences of misinformation. Sources were selected through a systematic review; paragraphs of text were extracted from the sources and were then proofread and verified by professionals from two universities in Malaysia. The dataset contains approximately 1,000 annotated samples, which consist of context, question and answer pairs. The SRHQA dataset is used to evaluate how well models can perform under domain shift.

The SRHQA dataset was constructed using publicly available educational and medical texts sourced from verified educational and medical publications. No human subjects were recruited or involved in data collection; all source material was drawn from existing published documents. Accordingly, formal IRB or ethics board approval was not required under standard institutional research guidelines. The dataset has been made publicly available on HuggingFace under a CC-BY 4.0 license to support reproducibility and future research in domain-specific QA.

To further characterise the dataset, Table 1 provides aggregate corpus statistics. The dataset comprises 64 passages and 1,010 QA pairs. Average context length is 379.6 words per passage (range: 100-507 words), and average answer span length is 7.3 words (range: 1-57 words). Domain coverage spans sexual health and STIs, anatomy and puberty, relationships and emotions, menstruation and hygiene, mental and emotional wellbeing, reproduction and pregnancy, abuse, consent and safety, cancer and medical conditions, and gender and sexual identity.

Data Preprocessing

The domain-specific dataset was formatted to follow the structure of SQuAD to keep it consistent with the other two datasets. The annotation tool Haystack was utilised for labelling and formatting the dataset for the extractive question answering task.

Table 1
SRHQA dataset aggregate statistics

Statistic	Value
Total passages	64
Total QA pairs	1,010
Average Context Length	379.6 (range: 100-507)
Average Answer Span Length	7.3 (range: 1-57)

Additional steps were taken to clean the dataset, such as removing unnecessary characters and adjusting the token length of the text to fit within the model context window.

The dataset was then partitioned into two subsets, one for training and another for validation, with a ratio of 8 to 2. The training subset was used during the model and ensemble training process, enabling the models to learn the relationships between questions and their corresponding answers, optimise performance, and assess generalisation capabilities. The validation subset, on the other hand, was reserved for evaluating the model's performance on held-out validation data.

Model Configuration

In this study, four pre-trained models were selected for extractive question answering. These include BERT (csarron/bert-base-uncased-squad-v1), DistilBERT (distilbert/distilbert-base-uncased-distilled-squad), MobileBERT (csarron/mobilebert-uncased-squad-v1), and TinyBERT (deepset/tinybert-6l-768d-squad2), all of which were obtained from the Hugging Face model repository. Each selected model was pretrained on the SQuAD dataset since it provides a large sample for training, allowing the models to get familiar with the extractive QA task while also setting a baseline. Input tokenisation was performed using the respective tokeniser for each model. TinyBERT, however, utilises the BERT base tokeniser due to its architectural compatibility. The maximum input sequence length was set to 512 tokens for all models to ensure a consistent and sufficient context window.

It should be noted that the TinyBERT checkpoint used (deepset/tinybert-6l-768d-squad2) was pre-trained on SQuAD v2, while the remaining models were pre-trained on SQuAD v1. This versioning difference was retained as it reflects the best publicly available TinyBERT checkpoint for extractive QA at the time of experimentation. SQuAD v2 introduces unanswerable questions in addition to the answerable questions present in SQuAD v1; however, since evaluation in this study focuses exclusively on answerable span prediction across all datasets, the practical impact of this checkpoint difference is expected to be minimal. The relative performance patterns of TinyBERT remain consistent with the other lightweight models across AdversarialQA and SRHQA, suggesting that the versioning difference does not materially affect the study's conclusions.

To evaluate the effects of architectural modification, a bidirectional LSTM (BiLSTM) layer was added to each model variant. The BiLSTM layer was placed between the transformer encoder and the final classification layer. To make sure that the BiLSTM layer is compatible with the encoder output, the hidden size is set to match the dimensions. The modified models were fine-tuned using the same hyperparameters as the base model to isolate the impact of the architectural change.

During training, the models were trained for three epochs using a learning rate of 3e-5 and the Adam optimiser with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon = 1e\text{-}8$. The learning rate followed a linear scheduling strategy. The training batch size was set to 6, while evaluation

was performed with a batch size of 60. A fixed random seed of 42 was used throughout all experiments. This was a deliberate methodological choice to ensure full deterministic reproducibility. The hyperparameters of the training and evaluation setup are shown in Table 3. All experiments were conducted on Google Colab using a single T4 GPU with approximately 16 GB of VRAM, an Intel Xeon CPU and 13 GB of RAM. The versions used for the software are listed below in Table 2.

Transfer Learning Setup

Each of the pretrained models was fine-tuned on the AdversarialQA and SRHQA datasets to get the performance before and after being adapted to each dataset. The pretrained weights were loaded from the Hugging Face repository during initialisation and then fine-tuned on the dataset through single-stage training. During fine-tuning, variants of each base model were created, which included the addition of a BiLSTM layer. This was done to investigate the sensitivity of lightweight models to architectural changes. Apart from this modification, all models followed the same training pipeline and hyperparameter settings that were specified in the model configuration section. After fine-tuning, model performance was assessed, with evaluation done on the held-out validation split of each respective dataset.

Table 2
Software environment

Component	Version
Python	3.10
PyTorch	2.0.1+cu118
HuggingFace Transformers	4.35.0
CUDA	11.8
datasets (HF)	2.14.0

Table 3
Hyperparameters

Parameter	Value
Batch size (Training)	6
Batch size (Evaluation)	60
Epochs	3
Random seed	42
Learning rate	3e-5
Optimiser	Adam
Adam β_1	0.9
Adam β_2	0.999
Adam ε	1e-8
Maximum Context Length	512 tokens

Ensemble Implementation

The averaging ensemble is generally composed of fine-tuned variants of the base lightweight models, DistilBERT, MobileBERT and TinyBERT. These models were chosen to evaluate the robustness and whether using an ensemble can mitigate any performance drops. Variants of these models, fine-tuned on specific datasets, were grouped accordingly. For instance, to evaluate the ensemble on the AdversarialQA dataset, the corresponding fine-tuned variants of the three lightweight models were selected. Each base model was trained independently to ensure that any performance improvements are from the ensemble instead of shared training procedures, while also accommodating the hardware setup.

Predictions from the three models were aggregated using a probability averaging strategy. This was done by adapting the question answering pipeline from Hugging Face to accept multiple models simultaneously. During inference, each model produces a probability distribution over the potential answer tokens. Z-score normalisation is then applied to these distributions to make the magnitude of predictions consistent across models to prevent any one model from dominating the final prediction. The final ensemble prediction was obtained by averaging the normalised probabilities token-wise. This averaged probability distribution was then passed to the existing post-processing function, which maps the probabilities to the final answer span. Model diversity was primarily derived from differences in distillation levels and architectural variations among the lightweight models.

Evaluation Metrics

Model performance was evaluated using metrics used in the SQuAD paper for the extractive question answering task, which are F1 and EM, which focus on partial and exact correctness of predictions. F1 is used to measure the overlap between the prediction and the ground truth; it acts as a measure for soft correctness and is calculated by the harmonic mean of precision and recall, as shown in Equation 1. The Exact Match (EM) is a binary metric that evaluates whether the model can predict the same answer as the ground truth. Together, these metrics provide a comprehensive assessment of both approximate and exact model performance.

$$F_1\ Score = \frac{2\ \times Precision \times Recall}{Precision + Recall} \quad [1]$$

RESULTS AND DISCUSSION

Individual Model Performance

The baseline evaluation of lightweight BERT models revealed that all three variants, DistilBERT, MobileBERT, and TinyBERT, performed well on g eneral-purpose datasets but exhibited varying degrees of sensitivity when the addition of a BiLSTM layer was applied, as shown in Table 4. On the SQuAD dataset, DistilBERT achieved an F1 score of 86.56% and an Exact Match (EM) of 78.94%, while MobileBERT and TinyBERT recorded F1 scores of 89.94% and 88.91%, and EM scores of 82.5% and 81.49%, respectively. These results align with previous studies reporting competitive performance for lightweight variants in standard QA tasks. However, with the additional layer, a consistent decrease in performance is shown, ranging from 0.13 to 0.31% for all lightweight models. These decreases aren't seen in the non-distilled model, which has an increase of 0.16% F1 and 0.41% EM. The ensemble also shows an increase of 2.9% F1 and 3.8 EM.

The AdversarialQA dataset was designed to confuse models by including adversarial samples that scored a low F1 score during the curation of the dataset. Due to this, a lower overall performance is to be expected, as seen by the F1 and EM scores of all models in Table 5. Following fine-tuning on the AdversarialQA training set, DistilBERT achieved 23.2% EM and 32.79% F1, while MobileBERT and TinyBERT scored 29.6% EM / 39.53% F1 and 27.5% EM / 37.85% F1 for the base models, respectively. These results demonstrate the model's sensitivity to context perturbations and the limitations of modified distilled architectures in handling adversarial samples. However, the use of an ensemble was able to help in reducing the weakness of individual models, resulting in the highest F1 and EM out of the evaluated models.

In the SRHQA dataset, which contains domain-specific terminology, baseline models displayed a further decrease in performance, as shown in Table 6. MobileBERT reached 73.94% F1 and 57.23% EM, and TinyBERT recorded 62.57% F1 and 42.08% EM.

Table 4
SQuAD dataset

Model	F_1	EM	F1 Difference	EM Difference
DistilBERT	86.56	78.94	-	-
DistilBERT + BiLSTM	86.64	78.81	0.08	-0.13
MobileBERT	89.94	82.5	-	-
MobileBERT + BiLSTM	89.63	82.19	-0.31	-0.31
TinyBERT	88.91	81.49	-	-
TinyBERT + BiLSTM	88.77	81.28	-0.14	-0.21
BERT	88.11	80.48	-	-
BERT + BiLSTM	88.27	80.89	0.16	0.41
Ensemble	91.01	84.28	2.9	3.8

Table 5
AdversarialQA dataset

Model	F_1	EM	F1 Difference	EM Difference
DistilBERT	32.79	23.20	-	-
DistilBERT + BiLSTM	32.49	22.20	-0.30	-1.00
MobileBERT	39.53	29.60	-	-
MobileBERT + BiLSTM	39.84	28.90	0.31	-0.70
TinyBERT	37.85	27.50	-	-
TinyBERT + BiLSTM	37.64	27.20	-0.21	-0.30
BERT	38.08	27.00	-	-
BERT + BiLSTM	38.33	27.10	0.25	0.10
Ensemble	40.27	30.00	2.19	3.00

Table 6
SRHQA dataset

Model	F_1	EM	F1 Difference	EM Difference
DistilBERT	59.25	40.2	-	-
DistilBERT + BiLSTM	60.08	40.4	0.83	0.2
MobileBERT	79.14	64.26	-	-
MobileBERT + BiLSTM	73.94	57.23	-5.2	-7.03
TinyBERT	66.23	44.26	-	-
TinyBERT + BiLSTM	62.57	42.08	-3.66	-2.18
BERT	65.86	42.48	-	-
BERT + BiLSTM	68.61	44.26	2.75	1.78
Ensemble	85.32	67.23	19.46	24.75

DistilBERT, when modified, on the other hand, achieved 60.08% F1 and 40.4% EM, an increase from the baseline performance. This result is considered an outlier instead of a pattern since the same performance gains are not seen on other datasets. The reduced performance highlights the challenge of transferring lightweight models to specialised domains and the importance of assessing model robustness beyond standard benchmarks.

Impact of Architectural Modifications

To evaluate how architectural modifications affect the performance of lightweight transformer models, a BiLSTM layer was added between the BERT encoder and the final output head of each base model. The non-distilled model showed the expected outcome of having a slight performance boost, which is consistent with the results seen from other studies. However, for the lightweight models, a consistent decrease in performance is seen across all three datasets.

This suggests that the internal representations acquired during distillation are highly optimised, and additional structural modifications disrupt the alignment learned from the teacher models. In contrast, the base BERT model generally showed stable or improved performance under similar modifications, which further supports that lightweight models are less robust compared to the base model.

From the results, there is a trade-off in efficiency and robustness seen in how the lightweight models adapt to the addition of a BiLSTM layer in their architecture and in how the base BERT model performs. Generally, we see a decrease in performance. This can be attributed to the smaller parameter count of the lightweight models and the learned representations from the teacher already making use of the limited parameters. With the addition of the BiLSTM layer, the model needs to relearn representations that can utilise the characteristics of the new layer. For lightweight models, this is difficult; however, for the non-distilled BERT model, we can see that it can adapt by inferring from the increased performance across the three datasets. Practically, when working with compact models, architectural modifications need to be carefully considered and planned.

These results also show lightweight models having limited ability for generalising across different datasets and domains. For example, when evaluating on SQuAD, a general domain dataset and SRHQA, a domain-specific dataset, both of which follow the same formatting, they perform at different levels of accuracy. This is reflected in lower F1 and EM scores on SRHQA relative to SQuAD. The limited ability to generalise can be attributed to the training process. Since these models are obtained through knowledge distillation, they inherently have a narrower representation space than their teacher models. This reduces flexibility when exposed to domains with distinct linguistic structures. The AdversarialQA dataset was excluded from this generalisation analysis, as its intentionally difficult questions yielded consistently low F1 scores (<40%), making direct comparison inappropriate.

The general pattern of weakening performance in lightweight models following the addition of a BiLSTM layer can be attributed to the limited internal representations produced by knowledge distillation. Because knowledge distillation compresses the teacher's learned representations into a smaller parameter space, the student models carry less internal redundancy, making them more sensitive to downstream changes (Bai et al., 2024). When a BiLSTM layer is added after the transformer encoder, it introduces additional sequential processing that requires the model to project its compressed representations into a new feature space. For full-sized BERT, which retains broader representational capacity and higher internal redundancy, such structural additions are feasible without disturbing the underlying representations. The contrast between BERT's stable or improved performance and the lightweight models' weakening performance supports the hypothesis that the fragility of distilled models is specifically a consequence of the limitations introduced during the distillation process itself (Diddee et al., 2022; Lorenzoni et al., 2024).

An exception to this general trend is observed with DistilBERT on the SRHQA dataset, where the addition of a BiLSTM layer produced a marginal improvement (F1: 59.25% → 60.08%, EM: 40.20% → 40.40%). This improvement does not replicate across other datasets. DistilBERT degrades under BiLSTM modification on AdversarialQA (-0.30 F1) and shows no meaningful improvement on SQuAD (+0.08 F1). The non-replicating nature of this result identifies it as context-specific rather than structural. This study employed three datasets spanning general-purpose, adversarial, and domain-specific contexts precisely to test whether the fragility pattern holds across conditions. The directional consistency of degradation across datasets, 7 of 9 lightweight model-dataset combinations show degradation or no meaningful improvement, distinguishes the general finding from random variation. An isolated positive deviation that does not replicate across the same model on other datasets cannot be attributed to architectural advantage and does not undermine the overall pattern.

The BERT + BiLSTM configuration on AdversarialQA yields F1 of 38.33% and EM of 27.10%, representing an improvement of +0.25 F1 and +0.10 EM over the BERT baseline (F1 38.08%, EM 27.00%). All models in Table 5 were fine-tuned on the AdversarialQA training set before evaluation. Among lightweight models, DistilBERT and TinyBERT show the expected degradation under BiLSTM modification (-0.30 and -0.21 F1, respectively), while MobileBERT shows a marginal positive deviation (+0.31 F1). Consistent with the DistilBERT SRHQA case, this deviation does not replicate across datasets — MobileBERT degrades under BiLSTM modification on both SQuAD (-0.31 F1) and SRHQA (-5.20 F1). The cross-dataset consistency of MobileBERT's degradation on two of three datasets, and the absence of replication of the positive deviation, confirms that the AdversarialQA result is context-specific and does not constitute evidence of a structural architectural benefit. This behaviour aligns with Tran and Kretchmar (2024), who showed that models trained on SQuAD-style datasets do not inherently generalise to adversarial inputs and often fail under misleading or contradictory question formulations.

Ensemble Performance

Table 7 presents the comparison between the performance of the BERT baseline and the averaging ensemble. For the SQuAD dataset, the ensemble was able to score an F1 of 91.01 and EM of 84.28, improving by 2.9 and 3.8 points from the BERT baseline, respectively. A similar trend is observed on the AdversarialQA dataset, where the ensemble obtains an F1 score of 40.27 and an EM score of 30.00, which is a substantial improvement of 2.19 in F1 and 3.00 in EM over the baseline model. The results on SRHQA show the largest improvements with the ensemble scoring an F1 of 85.32 and EM of 67.23, which improve on the results of the base BERT model by 19.46 and 24.75, respectively.

Table 7
Ensemble performance

Dataset	Model	F_1	EM	F1 Difference	EM Difference
SQuAD	BERT	88.11	80.48	-	-
	Ensemble	91.01	84.28	2.9	3.8
AdversarialQA	BERT	38.08	27.00	-	-
	Ensemble	40.27	30	2.19	3.00
SRHQA	BERT	65.86	42.48	-	-
	Ensemble	85.32	67.23	19.46	24.75

In general, a trend appears when comparing the results across all three datasets, which is that the use of an ensemble to aggregate the predictions often resulted in better performance than the base model. This is especially evident when the sub model performs well, as seen by MobileBERT's results on the SRHQA dataset, which in turn contributes largely to the overall ensemble performance. These improvements show that a simple method, such as averaging, can stabilise and mitigate weaknesses of individual models, providing a practical approach for improving lightweight models without resorting to architectural-level changes.

Across all experiments, the results indicate that lightweight models struggle to accommodate architectural changes. Adjustments such as adding extra layers frequently led to declines in performance when compared to their unmodified counterparts. While some isolated cases showed gains in F1 or EM, these improvements were inconsistent across datasets and typically involved trade-offs elsewhere. In contrast, the base BERT model demonstrated stable and consistent gains when the same BiLSTM layer was introduced, suggesting that larger, non-distilled architectures are more resilient to structural modifications. Meanwhile, the averaging ensemble approach yielded reliable improvements across all datasets without requiring any architectural changes, reinforcing ensembling as an effective and robust strategy for enhancing lightweight model performance.

CONCLUSION

The results of the experiments highlight that the lightweight transformer models are sensitive to naïve modifications, as seen through the addition of a BiLSTM layer, which consistently resulted in degraded performance and through the model's inability to retain performance when undergoing domain shift. Using averaging, an improvement in F1 and EM was seen across all datasets, with the highest increase from the SRHQA dataset. Overall, the results indicate that ensembles can stabilise the predictions of individual models and that the method of distillation provides enough diversity to make averaging viable.

Theoretical and Practical Contributions

The study provides empirical evidence on the fragility of lightweight models when architectural changes are applied. Previous studies have explored applying ensembling

and modifications to BERT; however, there are limited studies that look into applying them to lightweight models. By recording the results of how these techniques affect the distilled models, this research contributes to the knowledge of trade-offs between efficiency, robustness and model complexity.

The results of this study provide practical considerations when deploying lightweight models on edge devices, such as whether transfer learning is appropriate, whether redistilling is required and whether modifications are a viable idea. Although the lightweight models are more appropriate for edge or mobile devices because of their lower requirements, usage requires careful deliberation. Ensembling is a reliable way of using multiple models to improve overall performance without increasing the architectural complexity of each model. Domain shift requires careful consideration and ample preparation, especially for the data to sufficiently adapt the general-purpose model to a specialised domain.

Limitations and Future Directions

The results of the experiments highlight the fragility of distilled BERT models when architectural modifications are applied. To address the fragility, future research should look into utilising the existing distilled weights instead of trying to retrain the model. For example, techniques like Low-Rank Adaptation (LoRA) could be an area to pursue since it minimally disrupts the original weights.

Dataset size was also observed to be a factor in the model performance on the respective validation set, with models trained on the larger dataset, SQuAD, often showing better generalisation than those trained on the much smaller SRHQA dataset. Future research could investigate methods for enhancing the generalisation of these lightweight models in low-resource settings.

The use of an ensemble showed a general improvement in performance over the base BERT model; the strategy that was used was simple averaging. From the base results, it was shown that MobileBERT often outperformed the other lightweight models. A weighted averaging configuration with MobileBERT having a higher weight may be more effective. Other than this, research could investigate how to specifically address the weakened robustness of lightweight post-distillation, possibly through a secondary distillation phase like in TinyBERT or an improved knowledge distillation pipeline. Future work should also consider input-level robustness evaluation through controlled perturbations such as typo injection, synonym replacement, and context shuffling, to quantify model sensitivity beyond a single adversarial dataset.

Overall, these directions emphasise the need to balance model efficiency with robustness, highlighting that performance alone is insufficient when lightweight models are highly sensitive to architectural changes. Continued research along these lines could enhance the practical utility and adaptability of distilled models for real-world applications.

In conclusion, while lightweight transformer models offer better efficiency, they also show notable sensitivity when modifications are added to their architecture and domain shifted. Using averaging ensembles, the predictions of multiple models can be stabilised and the performance enhanced, especially in domain-specific contexts. These findings contribute to both theoretical insights and practical guidance for the development and deployment of reliable, efficient NLP systems. By highlighting the trade-offs between computational efficiency, robustness, and accuracy, this research informs best practices for leveraging distilled BERT models in real-world extractive question answering applications.

ACKNOWLEDGEMENT

This work was supported by the Vice-Chancellor High Impact Research Grant (UNI/F08/VC-HIRG/85491/P05-02/2022). The authors would like to acknowledge Universiti Malaysia Sarawak (UNIMAS) for the support provided towards this publication.

REFERENCES

Bai, G., Chai, Z., Ling, C., Wang, S., Lu, J., Zhang, N., Shi, T., Yu, Z., Zhu, M., Zhang, Y., Song, X., Yang, C., Cheng, Y., & Zhao, L. (2024). Beyond efficiency: A systematic survey of resource-efficient large language models. *arXiv*. https://doi.org/10.48550/arXiv.2401.00625

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123-140. https://doi.org/10.1007/BF00058655

Dang, H., Lee, K., Henry, S., & Uzuner, Ö. (2020). Ensemble BERT for classifying medication-mentioning tweets. In G. Gonzalez-Hernandez, A. Z. Klein, I. Flores, D. Weissenbacher, A. Magge, K. O'Connor, A. Sarker, A.-L. Minard, E. Tutubalina, Z. Miftahutdinov, & I. Alimova (Eds.), *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task* (pp. 37-41). Association for Computational Linguistics. https://aclanthology.org/2020.smm4h-1.5/

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1, pp. 4171-4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Dhar, A., & Chauhan, I. (2023). A multi-category content classification approach using DistilBERT transformer and BiLSTM. In *2023 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)* (pp. 727-732). IEEE. https://doi.org/10.1109/ICCCIS60361.2023.10425205

Diddee, H., Dandapat, S., Choudhury, M., Ganu, T., & Bali, K. (2022). Too brittle to touch: Comparing the stability of quantisation and distillation towards developing low-resource MT models. In P. Koehn, L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussà, C. Federmann, M. Fishel, A. Fraser, M. Freitag, Y. Graham, R. Grundkiewicz, P. Guzman, B. Haddow, M. Huck, A. Jimeno Yepes, T. Kocmi, A.

Martins, M. Morishita, C. Monz, M. Nagata, T. Nakazawa, M. Negri, A. Névéol, M. Neves, M. Popel, M. Turchi, & M. Zampieri (Eds.), *Proceedings of the Seventh Conference on Machine Translation (WMT)* (pp. 870-885). Association for Computational Linguistics. https://aclanthology.org/2022.wmt-1.80/

Fang, Z., Zhang, H., He, J., Qi, Z., & Zheng, H. (2025). Semantic and contextual modelling for malicious comment detection with BERT-BiLSTM. In *2025 4th International Symposium on Computer Applications and Information Technology (ISCAIT)* (pp. 1867-1871). IEEE. https://doi.org/10.1109/ISCAIT64916.2025.11010257

Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., & Liu, Q. (2020). TinyBERT: Distilling BERT for natural language understanding. In T. Cohn, Y. He, & Y. Liu (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 4163-4174). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.372

Lee, L.-H., Lu, Y., Chen, P.-H., Lee, P.-L., & Shyu, K.-K. (2019). NCUEE at MEDIQA 2019: Medical text inference using an ensemble BERT-BiLSTM-attention model. In D. Demner-Fushman, K. B. Cohen, S. Ananiadou, & J. Tsujii (Eds.), *Proceedings of the 18th BioNLP Workshop and Shared Task* (pp. 528-532). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-5058

Lorenzoni, G., Portugal, I., Alencar, P., & Cowan, D. (2024). Exploring variability in fine-tuned models for text classification with DistilBERT. *arXiv*. https://doi.org/10.48550/arXiv.2501.00241

Pearce, K., Zhan, T., Komanduri, A., & Zhan, J. (2021). A comparative study of transformer-based language models on extractive question answering. *arXiv*. https://doi.org/10.48550/arXiv.2110.03142

Pranesh, R. R., Shekhar, A., & Pallavi, S. (2020). QuesBELM: A BERT-based ensemble language model for natural questions. In *2020 5th International Conference on Computing, Communication and Security (ICCCS)* (pp. 1-5). IEEE. https://doi.org/10.1109/ICCCS49678.2020.9277176

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv*. https://doi.org/10.48550/arXiv.1910.01108

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning, 5*(2), 197-227. https://doi.org/10.1007/BF00116037

Sun, Z., Yu, H., Song, X., Liu, R., Yang, Y., & Zhou, D. (2020). MobileBERT: A compact task-agnostic BERT for resource-limited devices. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 2158-2170). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.195

Tran, S. Q., & Kretchmar, M. (2024). Towards robust extractive question answering models: Rethinking the training methodology. In *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 2222-2236). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.129

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *arXiv*. https://doi.org/10.48550/arXiv.1706.03762

Wolpert, D. H. (1992). Stacked generalisation. *Neural Networks, 5*(2), 241-259. https://doi.org/10.1016/S0893-6080(05)80023-1

SUPPLEMENTARY DATA

Code and Data Availability

Training and evaluation code can be found at https://github.com/AllistairNK/SRHQAExperiments. The SRHQA dataset can be found at https://huggingface.co/datasets/allistair99/SRH1K. The dataset is released under the CC-BY 4.0 license.