# BioWE-LLM

## Final Project Report

**Demo video link:** https://drive.google.com/file/d/1yhcl69R1aFlywOX3ZYm5EnBUBbmfAke-/view

Our project was motivated by the desire to improve the accuracy of question-and-answer (Q&A) tasks, as these are critical to providing timely and effective healthcare. We aimed to create a pre-trained language model that is fine-tuned on biomedical literature to surpass current AI technologies and achieve higher benchmarks in Q&A accuracy. Our primary focus was to develop a reliable and accessible model within the biomedical domain that can benefit healthcare professionals by providing more accurate and efficient information retrieval. By achieving higher benchmarks in Q&A accuracy, we can enhance diagnostic precision and provide more targeted treatments, leading to improved patient care. Accurate Q&A systems also allow for quicker access to relevant information, saving time for healthcare professionals and improving information retrieval. Additionally, this can lead to easier and faster uncovering of patterns and insights, which can further research and educational advancements. By facilitating more informed decision-making, our language model can contribute to better healthcare outcomes and improved patient satisfaction. Moreover, patients can benefit from a more accurate and accessible Q&A system, as this can provide them with clearer information about their conditions and treatments. This can improve their health literacy, help them make more informed decisions about their healthcare, and enhance their overall experience with the healthcare system. Ultimately, our language model can have a substantial impact on public health policies and interventions on a global scale. By optimizing resources, it can lead to better healthcare outcomes, reduce healthcare costs, and improve the overall quality of healthcare.

### ➢ Related work

We found the BioGPT paper to be highly relevant to our project and it served as a great inspiration for us. Published by Microsoft in September 2024, the paper introduces a language model that outperforms previous natural language processing models in the biomedical field. It was pre-trained on vast amounts of biomedical literature, making it highly accurate when generating and processing biomedical text. This inspired us to create our own model, called BioWE-LLM, by combining two existing models. We believe that this model represents an advancement towards the application of artificial intelligence in healthcare, which is a crucial field that can benefit greatly from AI technology. After thoroughly studying the BioGPT paper, we realized that our model could potentially assist in research, diagnostics, education, and other areas of the biomedical domain.

### ➢ Overall architecture/workflow and the main functional components

Below is an overview of our model workflow and the key functional components that we use in our architecture. Our models of choice are BioBERT and BioELECTRA, and we use PubMedQA as our dataset. We analyze the results of our Question-Answering process and also explain the benefits of using weighted Ensemble methods in our approach.

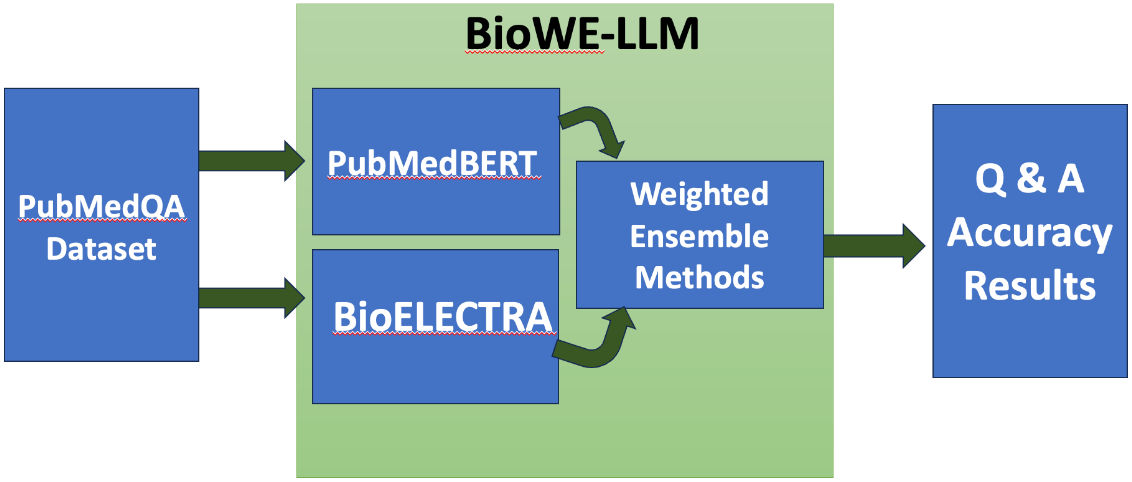


**Figure 1:** BioWE-LLM Workflow

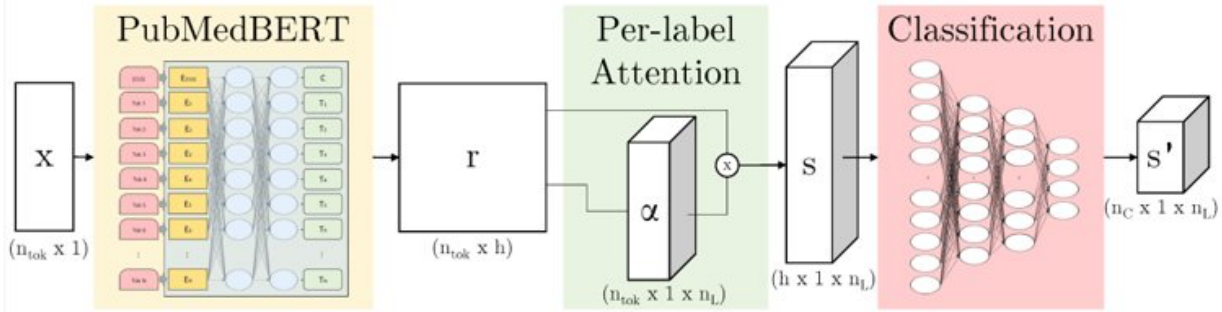


**Figure 2:** overall architecture of PubMedBERT

The diagram illustrates the overall structure of PubMedBERT. The input text, represented by X, is first tokenized into words. Each token is then embedded and processed through PubMedBERT via E layers. The resulting output, r, is then passed through a pre-label attention mechanism. This mechanism helps to identify the relevant parts of the text for each label. The output of the attention mechanism is a summary vector, S, which is then passed to the classification layer. The classification layer predicts the output, S', by classifying the input.

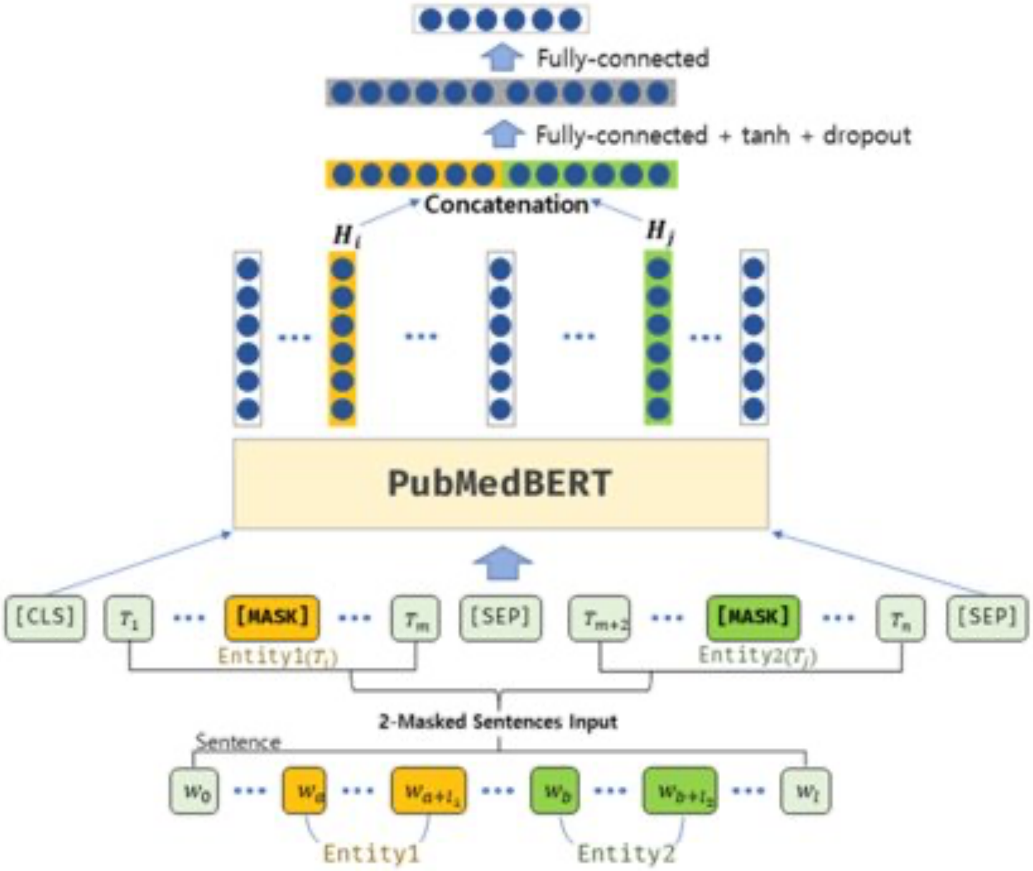


**Figure 3:** Masked tokens of input text in PubMedBERT

In pre-training, a technique called masked tokenization is used to help the model learn the context and meaning of words in the biomedical domain. Context tokenization is used to help the model learn the context and meaning of words. This involves masking certain parts of the text and requiring the model to predict these masked tokens. The hidden layers are then concatenated and processed through a fully connected layer, followed by an activation function and dropout for regularization. Finally, another fully connected layer is added to predict the original value of the masked tokens.

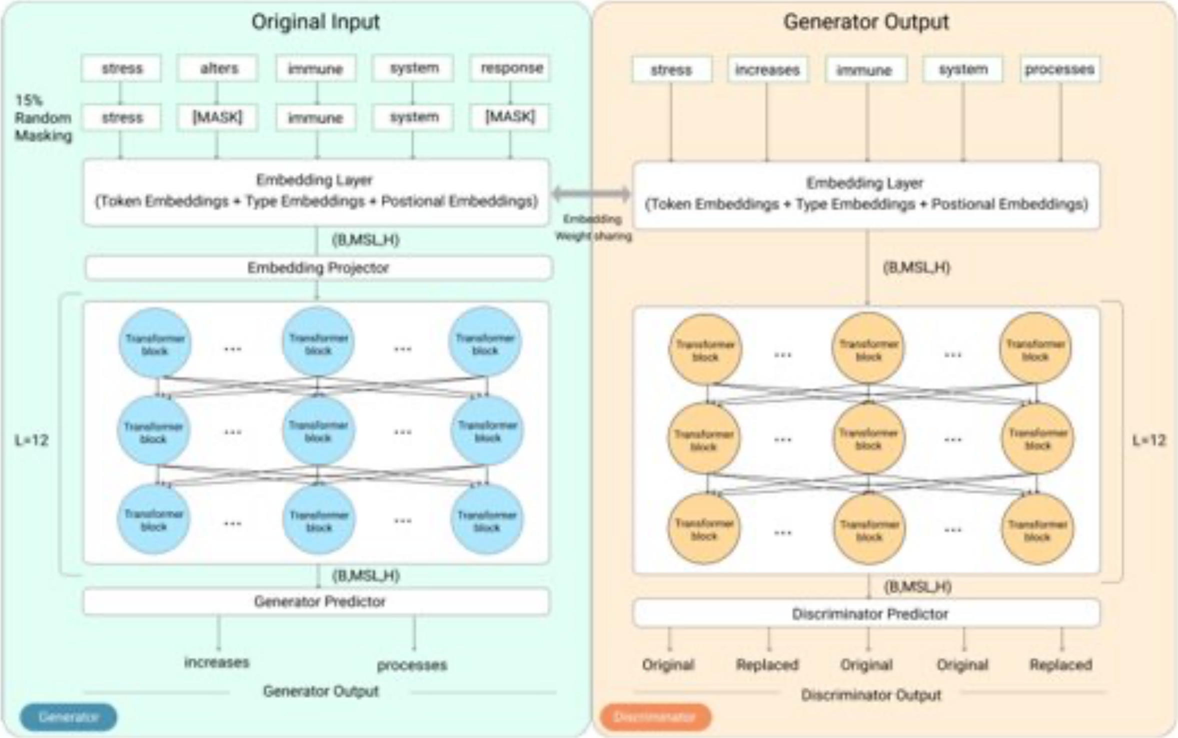


**Figure 4:** Overview of ELECTRA-Base model Pretraining. Output shapes are mentioned in parenthesis after each block. (B=Batch Size, MSL=Maximum Sequence Length, H=Hidden size)

The above figure shows the replaced token detection pre-training method used in ELECTRA-based models, which BioELECTRA was inspired by. This approach has two main components: the generator and the discriminator. During the pre-training phase, the generator tokenizes the input text and randomly selects 15% of the tokens to mask. It then tries to fill in the blanks by predicting replacements for the masked tokens. For example, it may suggest "increases" instead of "alters" or "processes" instead of "response".

The output of the generator is then used as the input of the discriminator which determines whether each token is authentic or a substitution made by the generator. This model is used in BioELECTRA to learn more efficiently because it can distinguish between correct and incorrect tokens, instead of just predicting masked words like traditional BERT models. Both the generator and discriminator have the same transformer-based architecture, even though they have different roles.

### Weighted Ensemble Methods

We utilized Weighted Ensemble Methods for our model, which is a type of ensemble learning technique that combines multiple models to make predictions. Each model in the ensemble is assigned a weight, indicating its importance in the final prediction. The weights can be learned or pre-determined based on certain criteria. In our project, we assigned more weight to bioELECTRA than to pubmedBERT since we knew that bioELECTRA outperforms pubmedBERT (according to the bioGPT paper).

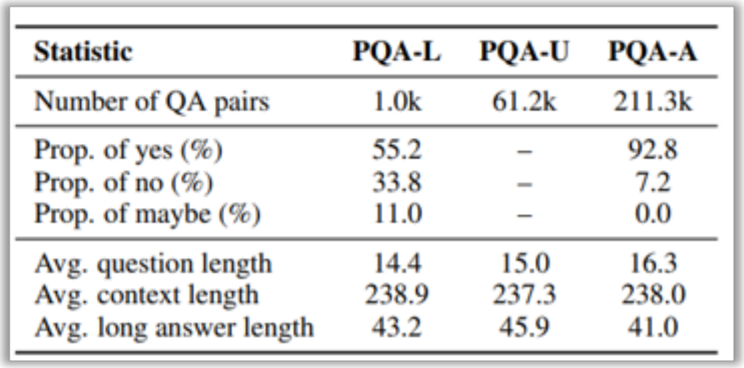
## Advantages of Weighted Ensemble Methods

Weighted ensemble methods offer several advantages over single models. By combining the predictions of multiple models, weighted ensemble methods can often achieve higher accuracy compared to using a single model. The ensemble can effectively leverage the strengths of individual models and mitigate their weaknesses, resulting in more reliable predictions. An ensemble of models is typically more robust to noisy or biased data compared to a single model. By aggregating the predictions from multiple models, the ensemble can reduce the impact of outliers and errors in individual models, leading to more robust and stable predictions. Weighted ensemble methods can improve the generalization capability of models. Each model in the ensemble might capture different aspects of the data or learn different patterns, thus covering a larger portion of the feature space. This diversification helps the ensemble perform well on unseen data and reduces overfitting. Ensemble methods can compensate for the bias inherent in any individual model. By combining models with different biases, the ensemble can minimize the impact of individual biases and tend to provide more balanced and accurate predictions. Weighted ensemble methods allow for the combination of various models, enabling the utilization of different learning algorithms or approaches. This flexibility allows the ensemble to adapt to different data types, problems, or domains, making it a more versatile and applicable solution. Depending on the approach used for weighting, some weighted ensemble methods provide transparency in terms of understanding the contributions of individual models. By analyzing the weights assigned to each model, it is possible to gain insights into the importance or reliability of each constituent model, aiding in model selection and interpretation.

## Dataset

The PubMedQA dataset is divided into three subsets: labeled, unlabeled, and artificially generated. They are referred to as PQA-L (labeled), PQA-U (unlabeled), and PQA-A (artificial), respectively. The architecture of the PubMedQA dataset is illustrated in the figure below.

Moreover, as shown in the table below, PQA-L and PQA-A contain questions that can be answered with yes, no, or maybe, with varying question length, context length, and long answer length. In this specific project, we only used the PQA-L, which is the labeled subset of the PubMedQA dataset for fine-tuning our model.



| Statistic | PQA-L | PQA-U | PQA-A |
|---|---|---|---|
| Number of QA pairs | 1.0k | 61.2k | 211.3k |
| Prop. of yes (%) | 55.2 | – | 92.8 |
| Prop. of no (%) | 33.8 | – | 7.2 |
| Prop. of maybe (%) | 11.0 | – | 0.0 |
| Avg. question length | 14.4 | 15.0 | 16.3 |
| Avg. context length | 238.9 | 237.3 | 238.0 |
| Avg. long answer length | 43.2 | 45.9 | 41.0 |

**Table:** PubMedQA dataset statistics

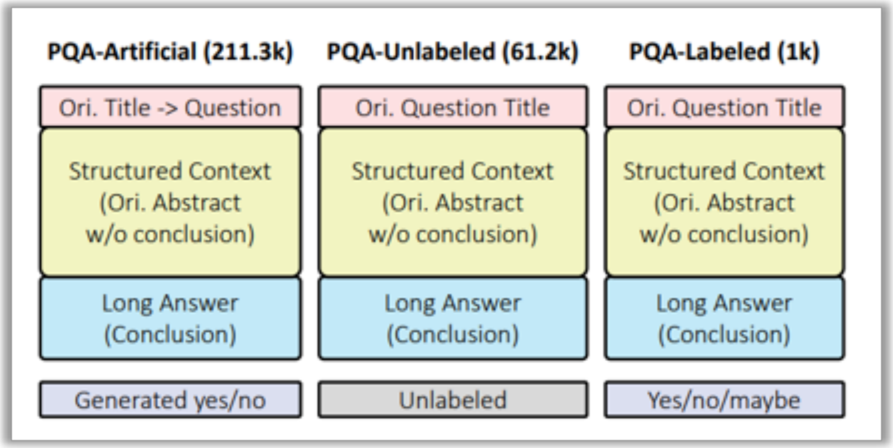


**Figure:** Architecture of PubMedQA dataset PubMedQA is split into three subsets, PQA-A (rtificial), PQA-U (nlabeled), and PQA-L (abeled)

**Metrics:**

Accuracy

**Evaluation results and analysis**

Fig. 6 and Fig. 7 show training and validation accuracy and loss for BioWE-LLM framework on PubMedQA dataset, respectively, which are the best result among ten runs. To validate the effectiveness of the BioWE-LLM framework, we employed PubMedBERT, BioELECTRA, and BioGPT models as a baseline models for comparison.

Table 2 shows the experimental results of this project. According to this table, the BioWE-LLM framework achieves an average accuracy of 79.8% outperforming PubMedBERT, BioELECTRA, and BioGPT models.

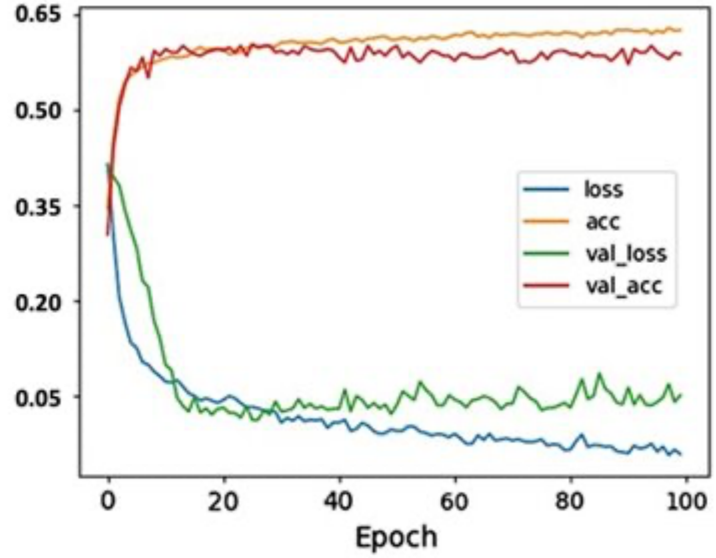


**Figure:** Accuracy and loss for training and validation of BioELECTRA

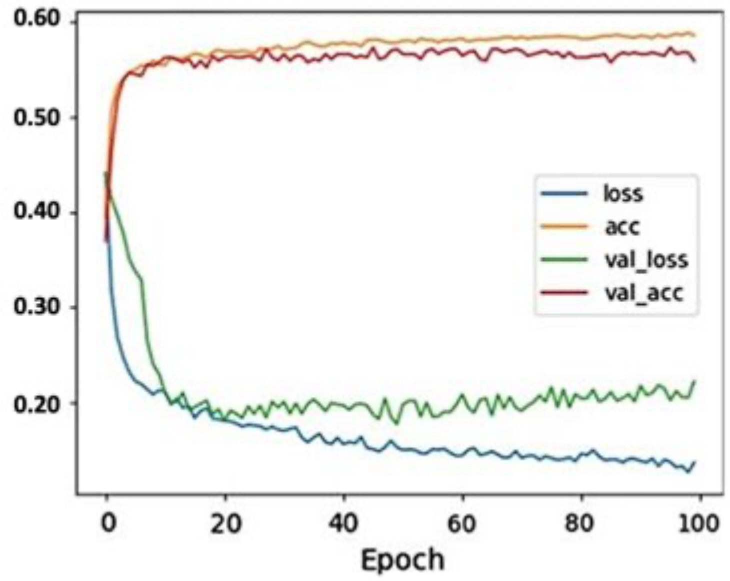


**Figure:** Accuracy and loss for training and validation of PubMedBERT

**Table:** Experimental result and comparison with baseline models

| Model | Accuracy |
|---|---|
| PubMedBERT | 56.8 |
| BioELECTRA | 63.2 |
| BioGPT | 78.2 |
| **BioWE-LLM (ours)** | **79.8** |

**Experimental analysis for three scenarios (easy, average, and challenging cases)**

1. **Easy scenario & Average scenario**
   In our analysis of the PubMedQA dataset with our LLM architecture, we focused on what can be classified as *average cases* in the context of question complexity.
   We acknowledge that our current dataset does not include *easy cases*, which are characterized by questions with straightforward, directly stated answers. The absence of these simpler queries in our training data means our model has been optimized for a more complex level of inquiry, typical of *average cases*. The dataset is typically involved questions where answers are not directly stated but can be inferred from the text. They require a moderate level of comprehension and the ability to integrate information from different parts of a biomedical paper. Our model, trained on this subset of the PubMedQA dataset, is tailored to handle such inquiries, where the answers are not explicitly presented but can be deduced from the given text.
2. **Challenging scenario**
   However, for *challenging cases* that require deep contextual understanding, multi-faceted reasoning, or synthesis from various sections or multiple papers, our current model might face limitations due to its training primarily on labeled data. Including the unlabeled portion of the PubMedQA dataset in future work could enhance the model's capabilities in these complex scenarios. This expansion would allow the model to learn from a broader range of texts and questions, potentially improving its ability to handle intricate biomedical queries that demand a more nuanced understanding of the subject matter.

**Contributions of the project**

The weighted ensemble of PubMedBERT and BioElectra models can offer significant contributions to the field of biomedical text mining and natural language processing. Here are several areas where this ensemble can make a difference:

1.Improved performance: Combining the strengths of both PubMedBERT and BioElectra models can lead to improved performance on various biomedical text mining tasks. PubMedBERT, a variant of BERT pre-trained on PubMed abstracts, has shown excellent performance on text classification, named entity recognition, and relation extraction in the biomedical domain. BioElectra, on the other hand, is specifically trained on biomedical literature and can capture domain-specific knowledge effectively. The weighted ensemble can leverage the complementary characteristics of both models to achieve even higher performance.

2. Domain-specific understanding: Biomedical literature contains complex scientific and medical concepts. The BioElectra model is trained to understand and represent these domain-specific terms and concepts effectively. By incorporating BioElectra into the weighted ensemble, it can provide a better understanding of domain-specific terms, medical abbreviations, and relation extraction in the biomedical context.

3. Transfer learning: Transfer learning is a powerful technique that allows models to learn from large-scale pre-training tasks and generalize well to specific downstream tasks. Both PubMedBERT and BioElectra models benefit from transfer learning, where they are pre-trained on large biomedical text corpora. The weighted ensemble can leverage the pre-trained knowledge from both models to perform well on a wide range of biomedical text mining tasks without extensive task-specific training.

4. Model interpretability: One challenge in the field of deep learning is model interpretability. While deep neural networks are black boxes, ensemble models can provide better interpretability by aggregating the predictions of multiple models. The weighted ensemble of PubMedBERT and BioElectra can enhance interpretability by leveraging the collective intelligence of both models.

5. Robustness and generalization: Including multiple models in an ensemble can improve robustness and generalization. Each model can have its own biases and limitations, but by combining their predictions, the ensemble can mitigate these individual weaknesses. The weighted ensemble can produce more reliable and generalized predictions by considering multiple perspectives.

6. Continuous improvement and adaptability: The field of natural language processing is rapidly evolving. New models, architectures, and techniques constantly emerge. The weighted ensemble framework can be easily adapted to incorporate newer models, allowing for continuous improvements and leveraging the latest advancements in the field.

Overall, the weighted ensemble of PubMedBERT and BioElectra models can make significant contributions to biomedical text mining by improving performance, providing domain-specific understanding, leveraging transfer learning, enhancing interpretability, improving robustness and generalization, and allowing for continuous improvement and adaptability**.**