Training Free Adaptive Text Classification through Aggregated Large Language Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

Class-incremental classification problems typically requires continual learning of the underlying algorithm to adapt to new classes. While current-generation large language models (LLMs) can have excellent few shot performance on several tasks, many tasks still require retraining to account for distribution shifts at either the inputs or task level. Continual learning techniques could be applied to LLMs, yet this requires retraining multiple task- and distribution-specific LLMs versions. Additionally, for specific applications like medical applications, maintaining compliance with regularity standards becomes challenging as models evolves, requiring transparency and accountability. Besides the costs and complexity of retraining multiple models, continually learned models are also prone to performance degradation due to data drifts and catastrophic forgetting. We overcome these challenges by introducing a semantic search-based method that simultaneously uses multiple LLM vectorizers/encoders and prompts without requiring any fine tuning. We depict that our proposed method has performance comparable to that of LLM fine tuning for clinical (MR and CT protocoling) datasets. In this approach, instead of being restricted to fine-tuning a single LLM, multiple foundation models(LLMs)/vectorizers will be leveraged simultaneously, maximizing their capability without incurring extra expenses for retraining or fine tuning, meaning that their individual potentials will be aggregated to determine the final outcomes. Our method could be utilized for continual learning environments, eliminating the need for retraining and adapts dynamically to incoming data, ensuring continuous updating. This approach uses the diverse perspectives and strengths provided by different LLMs and prompts, enhancing the robustness and comprehensiveness of the responses. By aggregation of different foundation models without the need for fine-tuning, this method demonstrates encouraging accuracy and reliability for medical and non-medical datasets, as multiple LLMs/prompts can highlight various aspects of the same issue, mitigating the biases and limitations that may arise from using a single prompt or model.

1 Introduction

Text classification is the process of assigning categories to text based on its content, a fundamental task in natural language processing (NLP). Recent Large Language Models (LLMs) have significantly enhanced text classification capabilitiesSun et al. [2023]. LLMs have shown the ability for in-context learning (ICL)Thoppilan et al. [2022], Ouyang et al. [2022] but despite their success, language models using ICL still lag behind fine-tuned models in text classification. This is because they struggle with complex language tasks such as understanding clauses and ironyZhang et al. [2022], Kojima et al. [2022], and they are limited to use only a small portion of the training data, reducing their effectiveness.

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

Continual learning, also known as lifelong learning, allows a model to continuously acquire and adapt to added information without erasing previous knowledgeYang et al. [2024]. This capability is essential for scenarios where the model must remain up to date with new data and evolving patterns Huang et al. [2021], Wang et al. [2023], Jang et al. [2021]. For continual learning, catastrophic forgetting is an important obstacle and is especially evident in text classification, where semantic nuances and evolving patterns worsen the issue. Addressing this challenge is crucial for deploying robust and adaptive NLP systems in real-world applications. Another key challenge of continual learning is balancing plasticity (the ability to learn added information) and stability (the ability to retain previously learned information). Moreover, adapting model architecture to handle new classes at inference time is also a significant challenge. On other hand, some specific applications like medical applications are subject to regularity requirements, and these regulations mandate strict data privacy and security measures. Continual learning models must therefore be designed to comply with these regulations, which means that models need to be retrained with regularity constrained in mind, adding an extra layer of complexity.

In this paper, we propose a method for Class-incremental classification in an online environment using LLMs, without the need for re-training. In this method, instead of fine-tuning LLMs, we utilize different LLMs in their pre-trained form, partition the data into various fields, create different prompts from them, and apply these prompts or different fields with different LLMs for their encoding version. As a result, we will obtain a bag of information-rich encoded data, where each LLM contributes its own unique vision and perspective to the encoded version. It is like leveraging multiple experts, each with their own viewpoints, working together to complete the classification task. Two different methods are then introduced to draw a conclusion from this diverse and information-rich bag of data. The first method, called Aggregated Decision Making in Multi-Index System (MIS)3.1.1, we explore a methodology where multiple data sources (processed by different foundation models), contribute to a decision-making process. Each data source provides its own set of recommendation and to determine the optimal choice, these individual outputs are aggregated using statistical methods. This approach ensures a comprehensive evaluation by synthesizing diverse inputs, ultimately leading to a more balanced decision. In the Integrated Embedding Analysis (IEA) for Enhanced Decision-Making section 3.1.2, we utilized another methodology for optimizing selection process using multiple data sources. This approach involves concatenating embeddings generated by each vectorizer/LLMs/prompts/data fields and applying Principal Component Analysis (PCA) to reduce their dimensionality. This approach was tested on two datasets, a clinical MR and CT dataset where inputs from the primary physician are used for patient protocol. For both cases, using the proposed approach demonstrated strong performance compared to fine-tuning LLMs. considering that the computational resources required for model updates and maintenance would be drastically reduced since the models will not be fine-tuned on an online environment. Also, in regulated industries like healthcare, the need to constantly manage and re-validate models due to continual learning updates would be eliminated. On the other hand, users and stakeholders would have greater trust in models that consistently perform well without the risk of degrading over time. The proposed method can utilize different LLMs and prompts, which offers significant benefits, including enhanced diversity of thought and robustness, as it combines the unique strength and perspectives of each LLMs. Additionally, by aggregating the outputs of different LLMs and/or prompts, the model increases overall accuracy and reliability, which ensures a more balanced and comprehensive understanding, reducing biases and limitations inherent in individual models.

2 Related Work

Continual learning in text classification is addressed through various innovative methods aimed at mitigating catastrophic forgetting and handling data imbalance. Continual pre-training refreshes LLMs with new data periodically to ensure their relevance and effectiveness. Recent studies focus on integrating Large Language Models (LLMs) into continuous learning frameworks for text classification. This involves evolving methodologies to improve how LLMs process new information while preserving previously acquired knowledge. Research by Ke et al. Ke et al. [2023] underscores the importance of this approach in maintaining LLM accuracy across different domains and tasks. Moreover. in-context learning has transformed text classification by using pre-trained knowledge through prompt-based queries, minimizing the need for extensive training. Innovations by Schick and Schütze Schick and Schütze [2019] and Han et al. Han et al. [2023] have enhanced the precision of these models in real-world applications.

Moreover, Huang et al. (2021) Huang et al. [2021] introduced an information disentanglement based regularization to maintain task-specific information distinct, thus allowing the model to perform well on a sequence of text classification tasks without interference. Jang et al. (2021) Jang et al. [2021] tackled data imbalance in continual learning by segmenting the data distribution into exclusive subsets, ensuring effective focus on underrepresented classes. Wang et al. (2023) Wang et al. [2023] propose a lightweight snapshot-based approach using knowledge distillation, which enables the integration of new and old knowledge without extensive resource requirements. Ermis et al. (2022) adapt transformers with adapters for continual learning, demonstrating their potential beyond typical use cases in text tasks Ermis et al. [2022]. Lastly, Pasunuru et al. (2021) explored continual few-shot learning, addressing the rapid adaptation to new tasks with limited data Pasunuru et al. [2021]. All these studies train their models on continual learning settings, demonstrating various strategies to enhance learning retention and adaptability over sequential tasks.

Prompt-based methods could also be used for complex text classification problems. Sun et al. (2023) Sun et al. [2023] introduced progressive prompting for complex text classification tasks. This method uses a step-by-step prompting process to first identify simple clues and then perform deeper reasoning to make final decisions. Recent research also explores progressive prompting and retrieval-augmented methods to enhance continual learning. For instance, Liu et al. (2024) Junhua et al. [2024] introduced Linguistic-Adaptive Retrieval-Augmented Language Models (LARA) to improve multi-turn intent classification, demonstrating significant advancements in handling complex conversational contexts.

3 Methodology

In the process of making optimal selections from multiple data sources, different effective strategies can be employed. In this section, two different methods, "Aggregated Decision Making in Multi-Index System" and "Integrated Embedding Analysis for Enhanced Decision-Making" are discussed in detail below.

3.1 MIS: Aggregated Decision Making in Multi-Index System

3.1.1 Creating historical indexing structures

In text classification tasks, each input field can contribute valuable information and the integration of various fields, either individually or through combinations using different prompts, enhances the classification processes. As described before, the core idea of this method is to retrieve labels for the current query using historical data. One of the paper's main contributions is to develop a method for retrieving information about a query from individual data fields, in addition to a combination of data fields within prompt(s) without fine-tuning or retraining. Different prompts, whether applied to single fields or combinations of fields, can yield useful insights, and improve the overall performance of the model. Moreover, each data segment (individual fields or prompts) can have its own set of indexing structures using a variety of encoders/LLMs. Indeed, different vectorizer(s) (LLM, TFIDF, etc.) can be applied to the same data filed or prompts, and consequently, different LLMs may develop their own distinct understanding of that data filed/prompts. Therefore, for each of data points and/or prompts, utilizing their corresponding vectorizer(s) (LLM, TFIDF, etc.), we will synthesize their vectorized version and integrate them to their corresponding indexing structure. A dimensionality reduction method (such as PCA) may be considered as a preliminary step before integrating the vectors into their corresponding indexing structure. This process will be applied to all available historical data, resulting in multiple instances of indexing structures, each containing the most up-to-date version of the data as previously described Fig. 1. In the context of inference time, and given that data arrives incrementally, it is crucial to note that indexing structures will be updated with every query from a user. Employing various LLMs, akin to using multiple judges or referees, allows for gathering insights from each model without a cap on the number of 'judges' used. Unlike continual learning, this can be achieved without re-training or dedication a specific instance for each model for every user. Fig. 1. illustrates the concept of creating series of indexing tables for each vectorizer/encoder/LLM and data field/prompts.

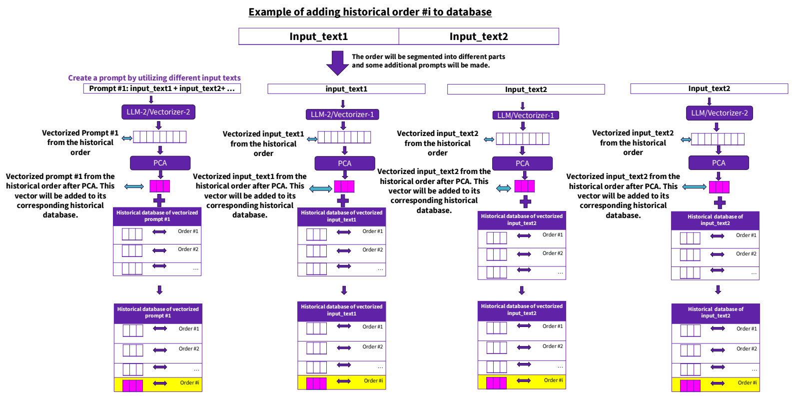

Figure 1: Aggregated Decision Making in Multi-Index System, creating historical indexing structures. This method leverages the strength of multiple LLMs to handle a diverse set of prompts and fields and each indexing table provides specific insights related to the query.

3.1.2 Inferencing Time

When a query is received, any datapoint in the historical dataset, which corresponds to a label, can be considered as a potential candidate for the query. For each instance of a query, distinct sets of data segments and prompts are generated in a way identical to how they are prepared for historical data, Fig. 2. The method previously described is then consistently applied to each set of data segments and prompts of the query, using their corresponding vectorizer(s), resulting in the creation of corresponding vectorized versions of them. For each vectorized query, a similarity search is implemented against the data in their corresponding indexing structure (Using FAISS in this paper, Douze et al. [2024]), and the historical datapoints are ranked based on their similarity scores, which can be computed using the dot product, or alternatively distance metrics such as Euclidean distance. As the results, each table presents a unique ranking of candidates along with their scores. In the subsequent step we must aggregate these results to identify the historical datapoint that achieves the highest rank and highest score. A variety of methods including Borda count and Bayesian techniques, can be considered for aggregating results from different indexing tables. These methods can be customized for each problem based on its specific nature, demonstrating the proposed method's power and flexibility.

A viable method for selecting the optimal candidate from multiple indexing structures involves computing a weighted average of scores across all the indexing structures. As mentioned above, for each query, each datapoint from the historical dataset is considered as a candidate. Each candidate's score is aggregated, assigning a score of zero to any candidate not listed among top candidates of an indexing structure. To narrow down the candidate's pool, each indexing structure sets a score threshold and candidates scoring below this threshold are not considered as top candidates. Subsequently the weighted average scores of all top ranked candidates are calculated. The candidate with the highest aggregated score is selected, and a specific attribute from this candidate is utilized for the query. For the results presented in this paper, the weight assigned to each index is calculated as the inverse of the variance of the similarity scores for the data. Using the inverse of the variance as weights for the weights assigned to each indexing table is chosen because it inherently prioritizes more consistent indexing tables. By doing so, we reduce the influence of those with high variability, who might introduce more noise and less reliable assessments. After calculating the average score for each candidate in the historical data using all the scores provided by every indexing table, different methods could be used to choose the winner. For example, the class corresponding to the historical data with the highest score could be considered as the suggested class for the query. Additionally, top N candidates can be used and by clustering, the class that appears the most could be considered as the winner. Moreover, different methods such as using mean and standard deviations, percentiles, or the elbow method, etc. can be utilized. These methods can vary based on the nature of the problem and whether that dataset is balanced or imbalanced.

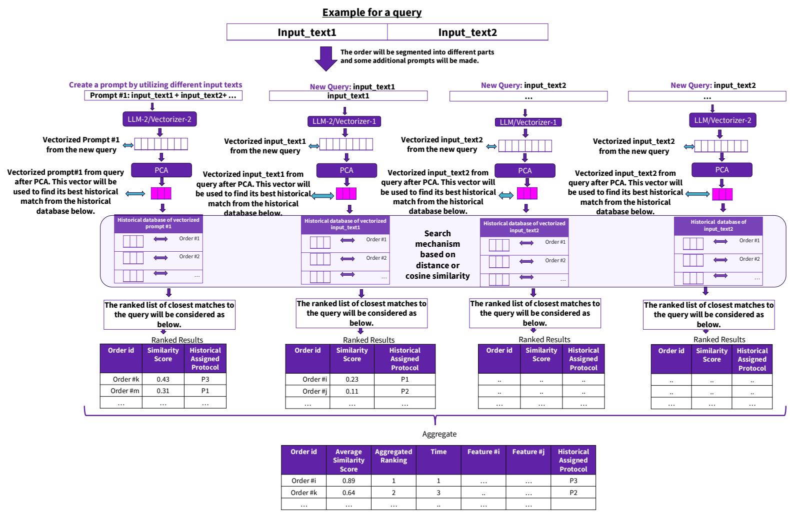

Figure 2: Aggregated Decision Making in Multi-Index System, Inferencing Time

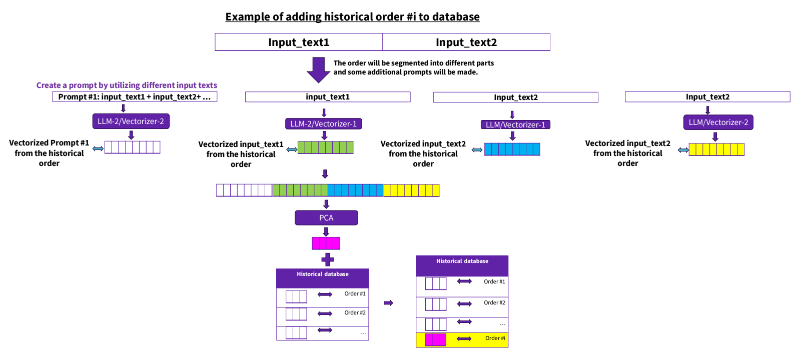

Figure 3: IEA: Integrated Embedding Analysis for Enhanced Decision-Making

3.2 IEA: Integrated Embedding Analysis for Enhanced Decision-Making

In this method, different data segments and prompts along with tier corresponding embeddings are generated using a variety of LLMs/vectorizers/prompts, similar to the approach described in 3.1.1. Then, each set of embeddings is normalized to ensure uniform scale across different sources. These normalized embeddings are then concatenated, forming a single, comprehensive embedding. This concatenated embedding is then subjected to PCA to reduce dimensionality while preserving the most critical information. The PCA transformed data is subsequently stored in a unified index table. The decision-making process involves selecting the class of the option that is closest to a given query or clustering the results and finding the most common class in top selections.

4 Results

It should be mentioned that we used the FAISS library for efficient similarity search in our indexing and searching mechanisms Douze et al. [2024].

4.1 Magnetic Resonance Imaging (MRI) Protocoling Dataset

We employed a dataset collected from a major site in the (BLIND). This dataset includes Magnetic Resonance Imaging (MRI), which each entry consists of two key inputs, “Reason for Exam” and “Suggested Procedure”. The aim of classification is to classify these inputs to suggest an appropriate “Protocol” for scan. Intelligent protocoling for scans is crucial in modern radiology as it significantly reduces the workload of radiologists and improves efficiency. It minimizes the need for manual intervention, allowing radiologist to focus on interpreting results rather than setting up scans, plus it enhances patient care by reducing wait times and the potential for human error, leading to more timely and precise diagnosis. This dataset consist of 1000 historical samples (22 Protocols), and the results are calculated for 2379 queries (assuming it is an online environment and the queries are coming one after each other, but the historical data sized is kept as 1000).

Reason for Exam(input)	Suggested procedure (input)	Protocol (label: output)
re-eval HCC for possible future tx. H/o HCV c/b HCC s/p TACE x4	Mr Abdomen w and wo iv contrast	Abdomen Liver Gadavist
Hx of Bosniak 2F left renal cyst. Surveillance imaging.	Mr Abdomen w and wo iv contrast	Abdomen Renal Renal Adrenal
prostate cancer surveillance, multiparametric MRI for biopsy planning, 3D lab 3D recons for treatment planning.	Mr Pelvis w and wo iv contrast	Pelvis Male Pelvis Prostate Local

Table 1: Examples from MR protocoling dataset illustrating three fields: reason for exam, procedure, and protocol(label)

Three different LLMs (bioMistral Labrak et al. [2024], BioGPTLuo et al. [2022], GatorTronYang et al. [2022]) are used, all downloaded from HuggingFace websiteWolf et al. [2019], and the selected prompts are (Two different input fields are: reason for exam and Procedure):

1-"An patient has been brought to our hospital, accompanied by observations *reason for exam* and *Procedure* from another physician. What protocol for a scan would you suggest for it, considering these characteristics?", and 2-"With the arrival of an patient at our hospital, accompanied by *reason for exam* and *Procedure* from another expert, we're looking for a fitting protocol for a scan. Can you provide your recommendation?".

In fact, the prompts generated from the combination of two inputs, reason for exam and Procedure, will be utilized to predict the corresponding protocol to scan(label). For this dataset, both MIS and IEA methods are used. For MIS, the process explained in 3.1.1 is implemented on 1000 historical data samples, each embedded vector from LLMs is treated independently, being added directly to the indexing tables. For each query (2379 cases), following what is discussed in 3.1.2, the scores of every historical data is calculated, the inverse of the variance of scores for all the historical data (1000 scores for each index) has been used to weigh each index for the weighted average, weighted average is applied and then the class associated with the candidate having the highest score is selected as the wining class. On the other hand, in IEA, embeddings from different LLMs are concatenated, followed by the application of PCA to reduce dimensionality to 256 on this concatenated set for all historical data as well as each query, and the class associated with the candidate from the historical data having the highest score is selected.

As can be seen in Fig. 4, for Prompt 2, the performance of both MIS and IEA is acceptable when compared to different Large Language Models (LLMs). Notably, integrating different LLMs generally yields better results than integrating embeddings with Principal Component Analysis (PCA), primarily because the former maintains more distinct information from each model. In the case of combining GatorTron and BioGPT, both methods—MIS and IEA—perform similarly, a combination of both BioGPT and GatorTron performs better than each individually. However, when considering the combination of GatorTron, BioGPT, with bioMistral, it seems combining bioMistral with each, makes the results worse. This suggests that addingbioMistral does not enhance, and may even detract from, the performance achieved by just GatorTron or BioGPT. However, for MIS, integrating GatorTron, bioMistral and BioGPT performs the best.The results generally demonstrate that aggregating outputs from multiple Large Language Models (LLMs) can improve the F1 score. Additionally, employing individual LLMs with various configurations of prompts also leads to improvements in the F1 score, depicted in Fig. 4, suggesting that different prompts can be likened to reshaping the input into various forms, thereby enabling the method to make more informed decisions. In Fig. 4 it can be seen that using prompt 1 in addition to prompt 2 improved the results for GatorTron, reinforcing the idea of having more prompt, and also emphasizing the power of the proposed method for the online domain.

As depicted in Fig. 4, without the need for retraining, including more prompts improved the results; however, it may sometime worsen the results. This anomaly may be attributed to the specific nature of the some prompts possibly introducing complexities or nuances not well-handled by the particular LLM's training regimen. Additionally, the integration of not suitable prompts could lead to an increase in response ambiguity or a decrease in coherence, challenging the specific LLM's ability to generate pertinent and cohesive outputs. This underscores the importance of careful prompt selection and customization in leveraging the full capabilities of pre-trained LMMs. Indeed, the balance between redundancy and novelty in the input prompts can influence model output. Redundant information might reinforce certain responses, but excessive redundancy could limit the model's generative capabilities. Novel information, while potentially enhancing the model's response, could also introduce uncertainties if it falls outside the model's training experience.

Moreover, it is noteworthy that in some instances, adding a LLM can actually deteriorate the results, a factor that must be considered in the aggregation strategy. The optimal number of LLMs to employ depends on the problem's specifics. Not every configuration of LLMs yields better outcomes when combined; certain configurations outperform the aggregated approach. This indicates the importance of the specific type of LLMs used, a phenomenon observed with both methods (MIS, IEA). This observation can be attributed to several factors. First, there is the issue of redundancy; additional LLMs may introduce overlapping information that does not contribute new insights but merely repeats existing data. This redundancy can worsen the unique contributions of each individual LLM, leading to a plateau or even a decrease in performance. Secondly, when multiple models (such as LLM X, LLM Y, and LLM Z) are aggregated, their individual biases and errors can interact in complex ways. This interaction can lead to unexpected behavior in the aggregated output, where the compounded biases or errors reverses the advantages each model brings individually. For instance, while LLM X and LLM Y may complement each other's strengths, the addition of LLM Z might introduce conflicting approaches or assumptions that disrupt the synergy between LLM X and LLM Y. This can be observed in Fig. 4 where BioGPT or GatorTron alone perform better compared to their combination with bioMistral.

4.2 Computed Tomography (CT) Protocoling Dataset

Another dataset, CT protocling, was also used for this paper. Only BioGPT and GatorTron were used based on the insights from the previous section, as these two models were found to be effective in improving the results. This dataset consists of 1000 historical records with 23 labels, and 13939 records are treated as queries over the time. Below are examples from the dataset along with their various fields.

Reason for Exam(input)	Suggested procedure (input)	Protocol (label: output)
pmh metastatic breast ca and ho sbo. pw tachycardia. concern for sbo	ct abdomen pelvis w iv contrast	abdomen and pelvis
etoh cirrhosis, hcc screening.	ct abdomen liver w iv contrast triphasic	triphasic liver
history of breast cancer, recent unintentional weight loss and back pain	ct chest abdomen pelvis w iv contrast	chest abdomen pelvis

Table 2: Examples from the CT protocoling dataset illustrating three fields: reason for exam, procedure, and protocol(label)

The same prompts and process outlined in the previous section for MIS and IEA was applied to this dataset. A similar trend was observed, where using a combination of prompts and LLMs improved the results as depicted in Fig. 5. Interestingly, specific prompts performed better with certain LLM, for example, the first prompt performed better for BioGPT, while the second prompt performed more effectively with GatorTron. Consequently, the best performance was achieved by combining (BioGPT, the first prompt) and (GatorTron, the 2nd prompt). Overall, considering the size of the historical data (1000 samples) and the number of queries (13939), the method demonstrated strong

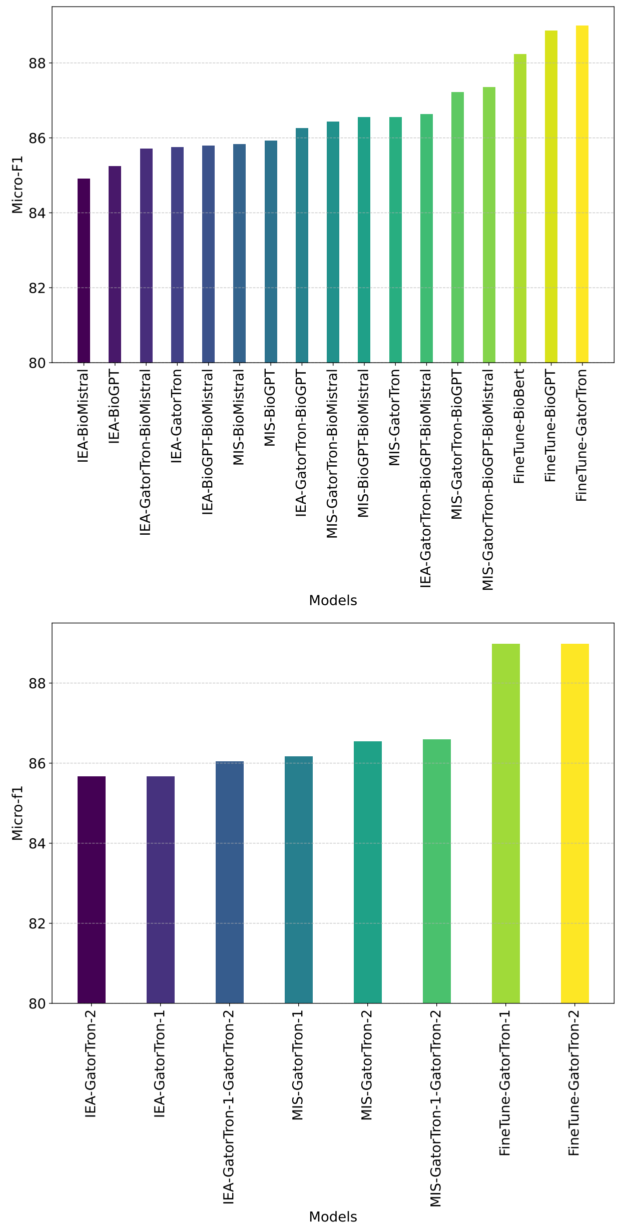

Figure 4: (a) The results for the MR clinical dataset using IEA(Integrated Embedding Analysis) and MIS(Multi-Index System), for the first prompt. (b) The results for the MR clinical dataset using IEA(Integrated Embedding Analysis) and MIS(Multi-Index System), combination of different prompts (Prompt 1 and Prompt 2).

performance based on the F1-score in Fig. 5, indicating its potential for effectively addressing the classification task.

5 Conclusion

In this paper, a method has been developed for classification tasks to implement foundation models including LLMs, that eliminates the need for retraining or fine-tuning and could be effectively used in online environments. In this method, a diverse and comprehensive set of information about a query is created by segmenting the data, forming prompts, and implementing different LLMs on them. It simplifies the aggregation of different LLMs while maintaining high accuracy and performance. This

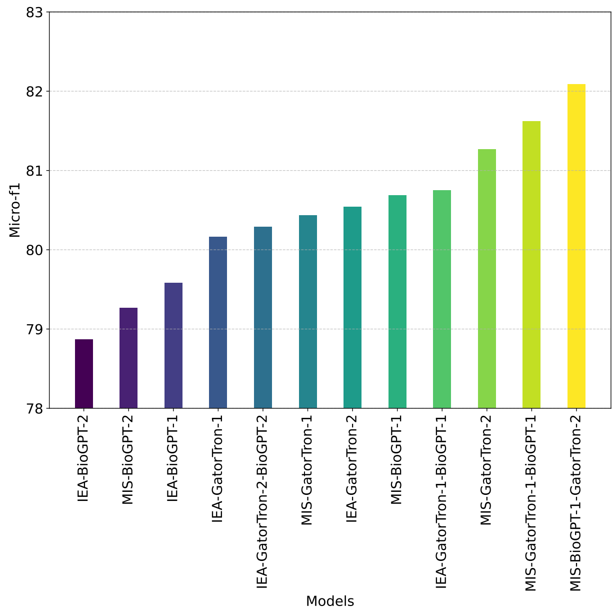

Figure 5: The results for the CT clinical dataset using IEA(Integrated Embedding Analysis) and MIS(Multi-Index System) for a combination combination of different prompts (Prompt 1 and Prompt 2) and BioGPT and GatorTron.

approach avoids complexities of continual learning (such as frequent retraining of LLMs, catastrophic forgetting, regularity approvals), reducing computational costs significantly, and is easy to implement and deploy, even in the medical domains with their own specific challenges.

It should be mentioned that when using concatenated encoders followed by PCA for searching, the combined feature vectors from different encoders can provide a richer and more comprehensive representation of the data. This method works well when the dataset benefits from a diverse set of features, as the concatenation captures various aspects of the data, enhancing the overall search accuracy. PCA then reduces the dimensionality, retaining the most informative components, which helps in making the search more efficient while preserving the essential characteristics of the data. This approach could be particularly more useful for datasets with complex features, where a single representation might not suffice. On the other hand, using multiple indices with an aggregation method (like MIS) can be more effective for datasets where different indices can specialize in capturing distinct characteristics of the data, like the clinical dataset in Section. 4.1. In this scenario, each index can focus on a specific feature subset or aspect of the data and then MIS helps in aggregating the results, thus reducing the impact of outliers or noise. MIS could be more beneficial for datasets with varied or noisy features, as the ensemble of indices can provide a more stable and reliable search result through consensus.

The results shown in this paper indicate that both MIS and IEA methods perform well for text classification without requiring additional training for both clinical datasets. This presents a valuable opportunity to replace continuous learning with foundation models by using pretrained only models. None of the models need fine-tuning during evaluation, eliminating the need for customization for each customer or site while the performance remains comparable and acceptable compared to fine-tuning. Additionally, in this method, the historical data stays up-to-date with incoming queries, as new data are classified and feedback with specified true labels is received by human in the loop in an online environment. This approach offers significant flexibility in choosing different combinations of LLMs, prompts, and datasets without incurring excessive computational costs, and allows for the aggregation of different models. Additionally, while our approach focused on simpler methodologies for aggregated decision making, alternative techniques, such as graph-based methods, could potentially offer more sophisticated aggregation and are worth exploring in future studies. However, there remain significant areas for further research. Key among these is identifying the optimal methods for aggregating results from multiple LLMs. Determining which LLM or prompt should be included or excluded in the aggregation process is crucial for enhancing performance and efficiency.

References

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. 2024.

Beyza Ermis, Giovanni Zappella, Martin Wistuba, Aditya Rawal, and Cédric Archambeau. Continual learning with transformers for image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3774–3781, 2022.

Xiaochuang Han, Daniel Simig, Todor Mihaylov, Yulia Tsvetkov, Asli Celikyilmaz, and Tianlu Wang. Understanding in-context learning via supportive pretraining data. *arXiv preprint arXiv:2306.15091*, 2023.

Yufan Huang, Yanzhe Zhang, Jiaao Chen, Xuezhi Wang, and Diyi Yang. Continual learning for text classification with information disentanglement based regularization. *arXiv preprint arXiv:2104.05489*, 2021.

Joel Jang, Yoonjeon Kim, Kyoungho Choi, and Sungho Suh. Sequential targeting: A continual learning approach for data imbalance in text classification. *Expert Systems with Applications*, 179: 115067, 2021.

Liu Junhua, Tan Yong Keat, and Fu Bin. Lara: Linguistic-adaptive retrieval-augmented llms for multi-turn intent classification. *arXiv preprint arXiv:2403.16504*, 2024.

Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, and Bing Liu. Continual pre-training of language models. *arXiv preprint arXiv:2302.03241*, 2023.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213, 2022.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. Biomistral: A collection of open-source pretrained large language models for medical domains, 2024.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6), 09 2022. ISSN 1477-4054. doi: 10.1093/bib/bbac409. URL `https://doi.org/10.1093/bib/bbac409`. bbac409.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Ramakanth Pasunuru, Veselin Stoyanov, and Mohit Bansal. Continual few-shot learning for text classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5688–5702, 2021.

Timo Schick and Hinrich Schütze. Learning semantic representations for novel words: Leveraging both form and context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6965–6973, 2019.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. Text classification via large language models. *arXiv preprint arXiv:2305.08377*, 2023.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022.

Jue Wang, Dajie Dong, Lidan Shou, Ke Chen, and Gang Chen. Effective continual learning for text classification with lightweight snapshots. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 10122–10130, 2023.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B Costa, Mona G Flores, Ying Zhang, Tanja Magoc, Christopher A Harle, Gloria Lipori, Duane A Mitchell, William R Hogan, Elizabeth A Shenkman, Jiang Bian, and Yonghui Wu. A large language model for electronic health records. *npj Digital Medicine*, 5(1):194, 2022.

Yutao Yang, Jie Zhou, Xuanwen Ding, Tianyu Huai, Shunyu Liu, Qin Chen, Liang He, and Yuan Xie. Recent advances of foundation language models-based continual learning: A survey. *arXiv preprint arXiv:2405.18653*, 2024.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*, 2022.