The Journal of Education in Perioperative Medicine

Special Article

Fine-Tuning Large Language Models to Enhance Programmatic Assessment in Graduate Medical Education

Gregory J. Booth, MD
Thomas Hauert, DO
Mike Mynes, MD
John Hodgson, MD
Elizabeth Slama, MD
Ashton Goldman, MD
Jeffrey Moore, DO

Introduction

Narrative feedback is very important to guide trainee growth in a competency-based medical education framework.[1,2] However, organizing large amounts of unstructured text into a format that is suitable for a specific task can be labor intensive. Natural language processing (NLP) can help overcome this challenge. NLP is a collection of artificial intelligence (AI) techniques in which computer systems extract meaning from (ie, discriminative or traditional AI) or produce (ie, generative AI) written or spoken human language. NLP is increasingly used in medical education, including applications such as personalizing learning experiences,[3] identifying gender bias in trainee assessments,[4] and enhancing program assessment of trainee competence.[5]

Our group previously published an NLP technique that classifies narrative comments on anesthesiology trainee performance assessments into the Accreditation Council for Graduate Medical Education (ACGME) subcompetencies to enhance programmatic assessment.[5] We integrated this innovation into our residency program, which substantially reduces the time required by our Clinical Competency Committee to organize narrative feedback on each resident, guides competency-based trainee self-reflections, and facilitates semiannual faculty feedback. Although programs can design assessment forms to guide comments toward specific subcompetencies, an assessment that lists all 23 anesthesiology subcompetencies would be cumbersome and likely result in limited assessment submissions. Our program's approach was to guide comments toward a limited number of competencies that are either seldom discussed or difficult for the NLP technique to identify (eg, Systems-Based Practice), and provide ample space for free-text narratives. Although the innovation performed well, it used an NLP technique that is no longer regarded as state-of-the-art.

The NLP algorithm used in our innovation was called FastText,[6] which falls into the broader category of NLP techniques called "bag of words" (BoW). BoW approaches do not consider word order in their predictions; it is as if all the words for a given input are dumped into a "bag" and are used in aggregate to make a prediction. Deep learning NLP algorithms, including large language models (LLMs) such as ChatGPT,[7] are currently regarded as the state-of-the-art approach in many NLP applications. Most deep learning LLMs incorporate information about word sequence and context and therefore outperform many BoW approaches. The aim of this study was to explore applications of contemporary deep learning LLMs in medical education. We hypothesized that deep learning LLMs can more accurately identify ACGME subcompetencies in trainee assessments than our previous BoW approach.

Methods

This study was approved by the Naval Medical Center Portsmouth Institutional Review Board in compliance with all applicable federal regulations governing the protection of human subjects. The specific NLP task was multiclass classification in which deep learning models learned to predict which ACGME subcompetency was best reflected in narrative clinical performance evaluations on anesthesiology trainees.

Data Source and Preprocessing

Techniques for curating the dataset were described previously.[5] Briefly, narrative assessments on residents across the four US military anesthesiology training programs between July 1, 2019, and June 30, 2021, were collected. Expert reviewers (ie, anesthesiology program directors and associate program directors) assigned each sentence with 1 of the 23 anesthesiology ACGME subcompetencies from Milestones 2.0 that best described it or N for not useful or D for demographic (eg, "this is an evaluation for a day on labor and delivery"). Although assessments predated Milestones 2.0, which were implemented on July 1, 2021, we elected to use Milestone 2.0 subcompetencies for model use with ongoing applications.

continued on next page

continued from previous page

Data from 3 programs were pooled to build the training dataset ($N = 10\,218$ comments), and data from the fourth program were used as the validation dataset ($N = 2255$ comments). Each comment represented a separate narrative by a staff anesthesiologist on a resident's performance, labeled with the ACGME subcompetency that best applied. Subcompetencies represented by <1% of the training data were combined into the most similar subcompetency, yielding 16 possible categories (final categories and distribution of comments are presented in the authors' prior work[5]). For example, all three Systems-Based Practice subcompetencies were combined into 1 category.

NLP Algorithm Basics

Despite many model architectures for NLP classification, there is a general flow that algorithms follow (Figure 1). Algorithms work on numbers, not words. Therefore, NLP algorithms first convert text to numbers in a process called encoding (sometimes called "embedding" depending on the specific approach). The goal of encoding is to assign numbers to text inputs that capture semantic meaning; thus, linguistic information from written or spoken language is encoded into abstracted mathematical relationships. In original BoW algorithms, a given input for a model was always assigned the same numeric encoding despite the context in which it was used. Advanced BoW algorithms, such as FastText, modified input encodings based on a few surrounding words, which allowed models to account for limited context related to how a word was used.

Improved encoding techniques were the crux of many advances in LLM performance over the past several years. With the advent of embedding, a specialized encoding technique used in models such as FastText and Word2Vec,[8] natural language could be represented by rich vectors of continuous data, boosting performance on many NLP tasks. Advances in embedding (eg, Transformer models[9]) now allow algorithms to consider long word sequences and distant context around an input, thereby empowering NLP models to better understand intricacies of human language. With these embedding techniques, a given input may have different numeric embeddings depending on how it is used and thus enhance performance on downstream tasks. This approach is more akin to human conversations; a word or phrase could carry very different meaning depending on the context and flow of a conversation.

Approach to NLP Model Training

In this investigation, deep learning was performed using the Bidirectional Encoder Representations from Transformers (BERT) family of models,[10] a deep learning neural network architecture. BERT is an LLM that embeds linguistic relationships from text learned from training on more than 1 billion words from over 10 000 books and thousands of Wikipedia pages.

One limitation of training and deploying LLM systems is model size. As LLMs grow in complexity, they tend to perform better at the expense of higher computational requirements. This concept is important to consider in our application (ie, residency assessments) because more complex models increase costs and may present challenges related to data privacy. Because trainee assessments may include sensitive data, the ideal system would operate in a contained environment with limited risk of data spillage. Although some AI companies may allow users to train models with users' data, we aimed to retain control over our models and the data used to train them. Further, AI models can be implemented on mobile devices. Typical mobile applications are on the order of 50 to a few hundred megabytes (MB), which is much smaller than most contemporary LLMs, which can be several gigabytes.

We explored performance of the original BERT-base model fine-tuned on our dataset in addition to 4 smaller-sized BERT models to investigate the tradeoff between model complexity and accuracy: BERT-base (436 MB), BERT-medium (167 MB), BERT-small (116 MB), BERT-mini (45 MB), and BERT-tiny (18 MB). A custom training pipeline was created in Python 3.10 (Wilmington, DE) with the Transformers Python library.[11] All models were trained using the following hyperparameters: learning rate = 2×10^{-5}, batch size = 16, and 3 epochs.

Performance Metrics

The primary performance metric was F1. F1 is a measure of accuracy that is calculated from recall (ie, sensitivity) and precision (ie, positive predictive value). We calculated F1 with 95% confidence intervals (CIs) on the validation dataset. F1 ranges from 0 to 1 (1 represents perfect performance). There is no consistent interpretation of what F1 score constitutes satisfactory performance. Interpretation of F1 depends on the context of its application. For example, some systems are designed to optimize recall, whereas others preferentially seek high precision. The utility of F1 in the present investigation is to provide a metric commonly used in classification performance assessment that is consistent across all comparisons. Further, we sought a performance metric that could represent global model performance. Because the data are class imbalanced (ie, unequal distribution of subcompetencies), our F1 implementation assessed performance of the models while treating all classes as equally important so that model performance on rare category labels were not overshadowed by the highly represented categories.

Additionally, we calculated area under the receiver operating characteristic curve (AUC) values with 95% CIs for each ACGME category to allow direct comparison of model performance to our original BoW approach. CIs were calculated using 250 bootstrap samples. AUC in this investigation represents the probability that the model scores a randomly selected narrative that belongs to a particular subcompetency as more likely to belong to that subcompetency than a randomly selected narrative belonging to a different subcompetency. AUC is generally interpreted as poor (0.50 to 0.59), fair (0.60 to 0.69), good (0.70 to 0.79), excellent (0.80 to 0.89), and outstanding (0.90 and above).

Sensitivity Analysis

Transformer models learn how to represent human language from the text used to train them. Therefore, an LLM may be skilled at interpreting 1 language but not another. Similarly, an LLM trained on a broad collection of topics may not perform well in a specific domain, such as medicine. To explore whether performance on our

continued on next page

continued from previous page

medical education application could be boosted by using an LLM more familiar with medical language, we also fine-tuned a model called SciBERT.[12] SciBERT has the same architecture as BERT-base but was trained using more than 1 million full-text biomedical publications, thereby learning how to represent English language in the scientific domain.

Results

Model Performance

There were no statistically significant differences in F1 between the model used in our prior investigation (FastText) and BERT-base, BERT-medium, BERT-small, or BERT-mini (Figure 2). BERT-tiny performed worse. On sensitivity analysis, SciBERT performance was not significantly better than FastText or BERT-base.

BERT-mini was the smallest model with equivalent performance to FastText based on F1 (Figure 2). BERT-mini was 94% smaller (45 MB versus 784 MB). Considering AUC per ACGME category for BERT-mini versus FastText, performance was equivalent on all of the 16 categories except Patient Care 7 (Situational Awareness and Crisis Management) and Systems-Based Practice (Table 1; AUC per ACGME category for FastText were published in the authors' prior work[5]).

BERT-mini performance on each category ranged from poor (the worst was Systems-Based Practice, AUC of 0.5) to outstanding (Patient Care 5, Airway Management, AUC of 0.9). Nine of the 16 categories demonstrated performance of good or better (AUC of at least 0.70). Four of the 16 categories demonstrated fair performance, and 3 demonstrated poor performance.

Discussion

In this investigation, deep learning LLMs learned how to interpret anesthesiology trainee narrative evaluations. The LLMs did not perform more accurately than the authors' prior work,[5] which used an older NLP technique that has limited ability to account for word order or context (FastText). However, comparable accuracy was achieved using an LLM that was 94% smaller than the model used in the prior work. This improvement in computational efficiency advances our understanding of the optimal approach to integrating NLP technologies into medical education applications.

BERT-mini demonstrated good or excellent performance, identifying most ACGME competencies discussed in trainee assessments, similar to FastText. Areas of poor or fair performance were also similar, such as Systems-Based Practice. We were surprised that the larger LLMs used in this investigation did not outperform FastText. We posited in our prior publication that one limitation may be related to poor inter-rater agreement for some competencies,[5] which limits an LLM's ability to learn consistent patterns, regardless of the LLM size or complexity. Another reason could be that most narrative comments were only 1 to 2 sentences, so the transformer architectures were unable to leverage broad linguistic contexts. Another limitation was that we used pretrained models. Intermediate training, or additional pretraining, can improve domain-specific performance. For example, BioBERT took the pretrained BERT model and modified its weights using a large collection of biomedical texts, which enhanced performance on several biomedical-specific tasks.[13] In our sensitivity analysis, SciBERT did not improve performance, but it is possible that intermediate training on BERT using a large corpus of feedback language or anesthesiology-specific language could improve performance for our task.

Although we explored performance metrics of the LLMs, we did not measure utility such as quantifying value added. Our program uses the FastText implementation described in this investigation and has realized substantive value in several areas, including reduced time and improved consistency by the Clinical Competency Committee while mapping assessment comments to competencies and facilitating ACGME requirements related to feedback and evaluation. We use the model to guide a competency-based trainee self-reflection exercise, and we provide faculty with semiannual feedback on the ACGME competencies discussed in the narratives that they write. We have found particular value in the system's ability to identify comments as "not useful" and provide feedback on which competencies their comments address in comparison to the average core faculty member during our targeted faculty feedback to help them better align their comments with ACGME language. Our findings in this investigation suggest that we can realize the same value with a smaller model, which potentially could be integrated into mobile applications.

Transformer-based LLMs are extremely powerful NLP tools and currently demonstrate state-of-the-art performance on many tasks. However, there is a paucity of evidence on technological barriers to deploying LLMs in medical education applications. In the authors' experience, computational requirements for hosting and running NLP models can be quite costly and technically challenging. For example, many LLMs require expensive graphics processing units to operate. These graphics processing units may only be available to institutions through subscription services. Therefore, using an LLM for medical education purposes could require moving data to the model for fine-tuning it to perform a specific task and/or using it to make predictions or generate outputs. Both of those situations raise concerns for data privacy. Smaller LLMs may mitigate these concerns because they may be able to operate within an institution's computing framework or even on a mobile device, assuming that they are still performant. Aside from privacy, such implementations may be more accessible by applications for faster and simpler use.

Deploying LLMs on end-user devices is an active area of research in the broader AI community. For example, Google (Mountain View, CA) released a compact version of their powerful transformer LLMs called Gemini Nano in late 2023, which was optimized for use on mobile android devices.[14] Gemini Nano is a distilled version of their larger Gemini model, similar to the smaller versions of BERT used in the present work. Recent research from Apple (Cupertino, CA) highlights advances in LLM memory usage on iOS to accommodate LLM computational requirements and speed LLM performance on their devices.[15] Our work complements these efforts in the AI community by providing evidence on methods to improve

continued on next page

continued from previous page

NLP efficiency in the medical education domain.

Conclusion

NLP has the potential to transform trainee feedback through several important mechanisms. We showed how transformer-based LLMs can automatically organize unstructured narrative assessments for program and/or trainee review. We showed that improvements in NLP model architecture can dramatically reduce computational resources without sacrificing performance. Our work is important for advancing the integration of LLMs into graduate medical education workflows and is aligned with advances in the broader AI community on how best to deploy models that are performant yet small enough for hosting on end-user devices or locally at an institution, thereby potentially improving speed and mitigating concerns related to data privacy.

Acknowledgments

We would like to acknowledge the following who helped develop the training dataset used in this study: Andy Cronin, MD, Angela McElrath, MD, Kyle Cyr, MD, Charles Sibley, MD, Martin Ismawan, MD, Alyssa Zuehl, MD, James Slotto, MD, Maureen Higgs, MD, and Matt Haldeman, MD.

References

1. Kogan JR, Hatala R, Hauer KE, Holmboe E. Guidelines: the do's, don'ts and don't knows of direct observation of clinical skills in medical education. *Perspect Med Educ.* 2017;6(5):286-305.
2. Edgar LM, McLean S, Hogan SO, Hamstra S, Holmboe ES. *The Milestones Guidebook*. Chicago, IL: ACGME; 2020.
3. Feng S, Shen Y. ChatGPT and the future of medical education. *Acad Med*. 2023;98(8):867-8.
4. Heath JK, Weissman GE, Clancy CB, et al. Assessment of gender-based linguistic differences in physician trainee evaluations of medical faculty using automated text mining. *JAMA Netw Open*. 2019;2(5):e193520.
5. Booth GJ, Ross B, Cronin WA, et al. Competency-based assessments: leveraging artificial intelligence to predict subcompetency content. *Acad Med*. 2023;98(4):497-504.
6. Joulin A, Grave E, Bojanowski P, et al. Bag of tricks for efficient text classification. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Volume 2, Short Papers*. Stroudsburg, PA: Association for Computational Linguistics; 2017:427-31.
7. OpenAI. ChatGPT (March 14 version) (Large language model). https://chat.openai.com/chat. Accessed November 10, 2023.
8. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. https://doi.org/10.48550/arXiv.1301.3781. Published January 16, 2013.
9. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: *Advances in Neural Information Processing Systems. 31st Conference on Neural Information Processing Systems*. Red Hook, NY: Curran Associates, Inc; 2017:6000-10.
10. Devlin J, Chang M, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding. https://doi.org/10.48550/arXiv.1810.04805. Published October 11, 2018.
11. Wolf T, Debut L, Sanh V, et al. Transformers: state-of-the-art natural language processing. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Stroudsburg, PA: Association for Computational Linguistics; 2020:38-45.
12. Beltagy I, Lo K, Cohan A. SciBERT: a pretrained language model for scientific text. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. Stroudsburg, PA: Association for Computational Linguistics; 2019:3615-20.
13. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. 2020;36(4):1234-40.
14. Anil R, Borgeaud S, Wu Y, et al. Gemini: a family of highly capable multimodal models. https://doi.org/10.48550/arXiv.2312.11805. Published December 19, 2023.
15. Alizadeh K, Mirzadeh I, Belenko D, et al. LLM in a flash: efficient large language model inference with limited memory. https://doi.org/10.48550/arXiv.2312.11514. Published December 12, 2023.

continued on next page

continued from previous page

The following authors are in both the Department of Anesthesiology, Uniformed Services University, Bethesda, MD, and Department of Anesthesiology and Pain Medicine, Naval Medical Center Portsmouth, Portsmouth, VA: ***Gregory J. Booth*** *is an Associate Professor at Uniformed Services University and Program Director, Anesthesiology Residency at Naval Medical Center Portsmouth;* ***Mike Mynes*** *and* ***Elizabeth Slama*** *are Assistant Professors at Uniformed Services University and Staff Anesthesiologists at Naval Medical Center Portsmouth;* ***Jeffrey Moore*** *is an Assistant Professor at Uniformed Services University and Program Director, Pain Medicine Fellowship, and Associate Designated Institutional Official at Naval Medical Center Portsmouth.* ***Thomas Hauert*** *is an Anesthesiology Resident Physician at Naval Medical Center Portsmouth, Portsmouth, VA.* ***Ashton Goldman*** *is an Associate Professor at Uniformed Services University, Bethesda, MD, and a Staff Orthopedic Surgeon at the Department of Orthopedic Surgery and Sports Medicine at Naval Medical Center Portsmouth, Portsmouth, VA.* ***John Hodgson*** *is an Associate Professor and Program Director, Anesthesiology Residency at University of South Florida, Tampa, FL.*

Corresponding author: *Gregory J. Booth, Uniformed Services University, 4301 Jones Bridge Road, Bethesda, MD 20814.*

Email address: *Gregory J. Booth: gjbooth2@gmail.com*

Disclaimer: *The views expressed in this article reflect the results of research conducted by the authors and do not necessarily reflect the official policy or position of the Department of the Navy, Department of Defense, or the US Government. We are military service members. This work was prepared as part of our official duties. Title 17 U.S.C. 105 provides that "Copyright protection under this title is not available for any work of the United States Government." Title 17 U.S.C. 101 defines a US Government work as a work prepared by a military service member or employee of the US Government as part of that person's official duties.*

Disclosures: *The authors have no competing interests to disclose.*

Funding: *The authors received no funding for this work.*

Abstract

Background: Natural language processing is a collection of techniques designed to empower computer systems to comprehend and/or produce human language. The purpose of this investigation was to train several large language models (LLMs) to explore the tradeoff between model complexity and performance while classifying narrative feedback on trainees into the Accreditation Council for Graduate Medical Education subcompetencies. We hypothesized that classification accuracy would increase with model complexity.

Methods: The authors fine-tuned several transformer-based LLMs (Bidirectional Encoder Representations from Transformers [BERT]-base, BERT-medium, BERT-small, BERT-mini, BERT-tiny, and SciBERT) to predict Accreditation Council for Graduate Medical Education subcompetencies on a curated dataset of 10 218 feedback comments. Performance was compared with the authors' previous work, which trained a FastText model on the same dataset. Performance metrics included F1 score for global model performance and area under the receiver operating characteristic curve for each competency.

Results: No models were superior to FastText. Only BERT-tiny performed worse than FastText. The smallest model with comparable performance to FastText, BERT-mini, was 94% smaller. Area under the receiver operating characteristic curve for each competency was similar on BERT-mini and FastText with the exceptions of Patient Care 7 (Situational Awareness and Crisis Management) and Systems-Based Practice.

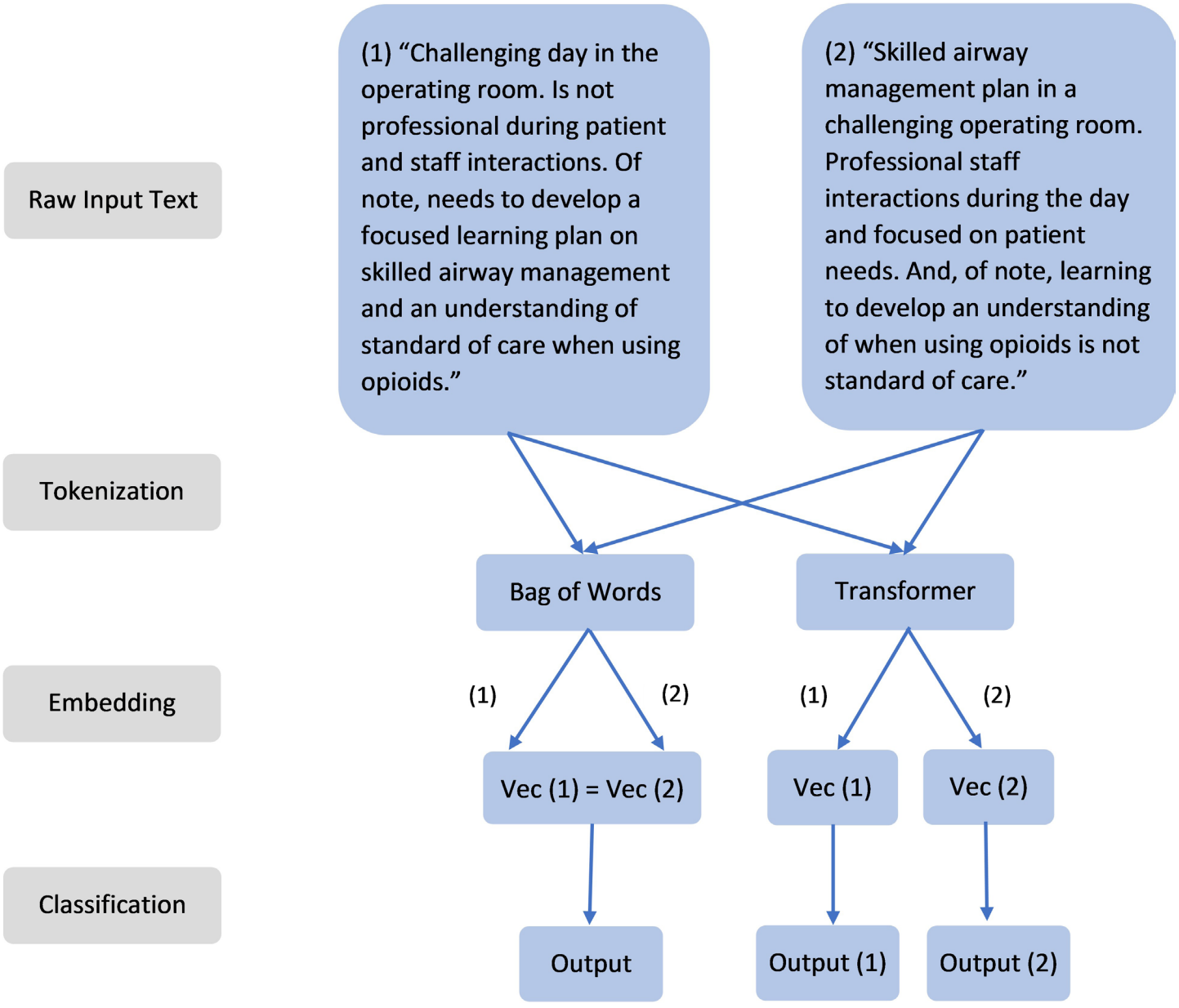
Discussion: Transformer-based LLMs were fine-tuned to understand anesthesiology graduate medical education language. Complex LLMs did not outperform FastText. However, equivalent performance was achieved with a model that was 94% smaller, which may allow model deployment on personal devices to enhance speed and data privacy. This work advances our understanding of best practices when integrating LLMs into graduate medical education.

Keywords: Natural language processing, artificial intelligence, anesthesiology, graduate medical education, large language model

continued from previous page

Figures

Figure 1. *Example of text classification using bag of words (BoW) and transformer large language models (LLMs). In this example, both inputs have identical words, but word order and context convey much different meanings. First, raw input text is tokenized, or split, into chunks of characters/words. Then, a BoW or transformer LLM embeds, or encodes, the tokens into numeric representations. BoW does not account for word order, unlike transformers. Therefore, BoW creates identical embedding vectors for both inputs (Vec 1, the vector representation for input (1) is identical to Vec 2, the vector representation for input (2)) because each has the same set of words. However, the transformer creates distinct vectors for each input, which capture different semantics. In this example, the final layer on the algorithm is classification. BoW produces the same prediction for both inputs, whereas the transformer produces distinct predictions. The final layer on the transformer LLM could be replaced by a different fine-tuned layer, such as one for question answering. With a different fine-tuned layer, the tokenization and embedding process would be the same, but the end-task would solve a different problem.*



continued on next page

continued from previous page

Figures *continued*

Figure 2. *Model performance and size. Performance is quantified by F1 score with 95% confidence intervals (CIs). Model size is quantified in megabytes. All Bidirectional Encoder Representations from Transformers (BERT) models are substantially smaller than FastText. With the exception of BERT-tiny, F1 scores for all BERT models are equivalent to FastText (ie, the 95% CIs overlap for all models except BERT-tiny). The BERT-tiny F1 score is significantly lower.*

continued on next page

continued from previous page

Table

Table 1. *Model Performance Measured by Area Under the Receiver Operating Characteristic Curve* [a]

	SciBERT	BERT-base	BERT-medium	BERT-small	BERT-mini	BERT-tiny
Model Size	442 MG	436 MB	167 MB	116 MB	45 MB	18 MB
AUC						
PC1	0.82 (0.78, 0.87)	0.78 (0.72, 0.83)	0.77 (0.72, 0.82)	0.76 (0.71, 0.81)	0.77 (0.72, 0.82)	0.62 (0.58, 0.67)
PC2	0.78 (0.75, 0.82)	0.75 (0.71, 0.79)	0.73 (0.69, 0.77)	0.74 (0.71, 0.78)	0.72 (0.69, 0.76)	0.64 (0.61, 0.68)
PC3	0.79 (0.73, 0.87)	0.73 (0.66, 0.80)	0.73 (0.66, 0.80)	0.71 (0.66, 0.77)	0.67 (0.62, 0.73)	0.51 (0.50, 0.53)
PC4	0.67 (0.63, 0.70)	0.68 (0.64, 0.72)	0.67 (0.64, 0.71)	0.69 (0.66, 0.73)	0.70 (0.67, 0.74)	0.57 (0.54, 0.59)
PC5	0.90 (0.86, 0.94)	0.92 (0.88, 0.96)	0.93 (0.90, 0.96)	0.93 (0.89, 0.96)	0.90 (0.87, 0.94)	0.89 (0.85, 0.93)
PC7	0.65 (0.61, 0.69)	0.64 (0.59, 0.68)	0.59 (0.56, 0.64)	0.58 (0.54, 0.62)	0.52 (0.50, 0.54)	0.53 (0.50, 0.55)
PC8	0.67 (0.60, 0.74)	0.61 (0.55, 0.69)	0.61 (0.56, 0.69)	0.76 (0.69, 0.85)	0.63 (0.58, 0.71)	0.50 (0.50, 0.50)
PC10	0.89 (0.85, 0.93)	0.91 (0.87, 0.94)	0.89 (0.84, 0.93)	0.88 (0.83, 0.92)	0.86 (0.82, 0.91)	0.81 (0.76, 0.87)
MK1	0.89 (0.86, 0.92)	0.91 (0.88, 0.93)	0.88 (0.85, 0.91)	0.88 (0.85, 0.91)	0.87 (0.84, 0.90)	0.86 (0.83, 0.89)
MK2	0.59 (0.55, 0.64)	0.60 (0.56, 0.65)	0.59 (0.55, 0.64)	0.59 (0.54, 0.63)	0.54 (0.51, 0.58)	0.51 (0.50, 0.52)
P	0.65 (0.63, 0.68)	0.64 (0.62, 0.67)	0.63 (0.61, 0.65)	0.65 (0.62, 0.68)	0.64 (0.61, 0.67)	0.61 (0.58, 0.63)
ICS	0.82 (0.78, 0.86)	0.84 (0.79, 0.88)	0.85 (0.82, 0.89)	0.82 (0.77, 0.86)	0.81 (0.77, 0.85)	0.73 (0.69, 0.78)
PBLI	0.70 (0.67, 0.74)	0.75 (0.71, 0.79)	0.70 (0.67, 0.75)	0.69 (0.66, 0.73)	0.66 (0.63, 0.70)	0.51 (0.50, 0.53)
SBP	0.57 (0.53, 0.61)	0.52 (0.50, 0.54)	0.50 (0.50, 0.50)	0.50 (0.50, 0.52)	0.50 (0.50, 0.50)	0.50 (0.50, 0.50)
D	0.85 (0.82, 0.88)	0.86 (0.83, 0.89)	0.86 (0.83, 0.88)	0.86 (0.83, 0.89)	0.85 (0.82, 0.88)	0.84 (0.81, 0.87)
N	0.86 (0.85, 0.88)	0.87 (0.85, 0.89)	0.87 (0.85, 0.89)	0.87 (0.85, 0.89)	0.87 (0.85, 0.89)	0.85 (0.84, 0.87)

Abbreviations: AUC, area under the receiver operating characteristic curve; BERT, Bidirectional Encoder Representations from Transformers; D, demographic; ICS, interpersonal and communication skills; MB, megabytes; MK, medical knowledge; P, professionalism; PBLI, practice-based learning and improvement; PC, patient care; SBP, systems-based practice; N, not useful.

[a] Data are presented as AUC (95% confidence interval).