machine learning & knowledge extraction

MDPI

Brief Report

Small or Large? Zero-Shot or Finetuned? Guiding Language Model Choice for Specialized Applications in Healthcare

Lovedeep Gondara [1,2,*], Jonathan Simkin [1], Graham Sayle [3], Shebnum Devji [1], Gregory Arbour [3] and Raymond Ng [3]

[1] British Columbia Cancer Registry, Provincial Health Services Authority, Vancouver, BC V6H 4C1, Canada; jonathan.simkin@bccancer.bc.ca (J.S.)
[2] School of Population and Public Health, University of British Columbia, Vancouver, BC V5Z 4E8, Canada
[3] The Data Science Institute, University of British Columbia, Vancouver, BC V5Z 4E8, Canada
* Correspondence: lovedeep.gondara@ubc.ca

Abstract

Objectives: To guide language model (LM) selection by comparing finetuning vs. zero-shot use, generic pretraining vs. domain-adjacent vs. further domain-specific pretraining, and bidirectional language models (BiLMs) such as BERT vs. unidirectional LMs (LLMs) for clinical classification. **Materials and Methods:** We evaluated BiLMs (RoBERTa, Pathology-BERT, Gatortron) and LLM (Mistral nemo instruct 12B) on three British Columbia Cancer Registry (BCCR) pathology classification tasks varying in difficulty/data size. We assessed zero-shot vs. finetuned BiLMs, zero-shot LLM, and further BCCR-specific pretraining using macro-average F1 scores. **Results:** Finetuned BiLMs outperformed zero-shot BiLMs and zero-shot LLM. The zero-shot LLM outperformed zero-shot BiLMs but was consistently outperformed by finetuned BiLMs. Domain-adjacent BiLMs generally outperformed generic BiLMs after finetuning. Further domain-specific pretraining boosted complex/low-data task performance, with otherwise modest gains. **Conclusions:** For specialized classification, finetuning BiLMs is crucial, often surpassing zero-shot LLMs. Domain-adjacent pretrained models are recommended. Further domain-specific pretraining provides significant performance boosts, especially for complex/low-data scenarios. BiLMs remain relevant, offering strong performance/resource balance for targeted clinical tasks.

Keywords: small language models; large language models; AI in healthcare

Received: 25 July 2025
Revised: 14 October 2025
Accepted: 15 October 2025
Published: 17 October 2025

Citation: Gondara, L.; Simkin, J.; Sayle, G.; Devji, S.; Arbour, G.; Ng, R. Small or Large? Zero-Shot or Finetuned? Guiding Language Model Choice for Specialized Applications in Healthcare. *Mach. Learn. Knowl. Extr.* **2025**, *7*, 121. https://doi.org/10.3390/make7040121

1. Introduction

The rapid advancement and proliferation of language models (LMs) have revolutionized the field of Natural Language Processing (NLP) [1]. From compact, BERT [2] type, bidirectional small language models (BiLMs) to massive large language models (LLMs) like GPT-4, practitioners now have more choices than ever. This proliferation extends to healthcare NLP, where various LMs have demonstrated superior performance over traditional approaches [3,4]. This raises a critical question: in the era of language models, which LM should be used for a given task? The decision involves several considerations.

1. Do we need to finetune a model or can we use it in its zero-shot capacity?
2. Are domain-adjacent pretrained models better than generic pretrained models?
3. Is further pretraining on domain-specific data helpful?
4. With the rise of LLMs, are BiLMs like BERT still relevant?

This paper delves into these crucial decision points and explores the various trade-offs. For empirical evidence, we use the electronic pathology data from the British Columbia Cancer Registry (BCCR) and study various scenarios that help us answer the questions above.

Scenario (a): Easy problem, binary classification, large training data. For this scenario, we use the reportability classification problem [5], where the training data consists of 40,000 labeled pathology reports with the goal being to classify the reports into reportable or non-reportable tumors (based on the guidelines set by governing bodies). The test data for this scenario consists of 20,400 pathology reports unseen by the models during the training phase.

Scenario (b): Medium-hard problem, multi-class classification, limited training data. For this scenario, we use the tumor group classification problem [6], where the training data consists of 16,000 pathology reports with the goal being to classify the pathology reports into one of nineteen tumor groups, where most tumor groups have less than 1000 samples in the training data. The test data for this scenario consists of 2058 pathology reports.

Scenario (c): Hard problem, multi-class classification, small training data. This scenario is based on classifying histology from pathology reports for leukemia, where we have the six most common histology codes and only 1000 pathology reports as the training data and 447 reports are used as the test data.

These scenarios were selected to represent a spectrum of common challenges in clinical NLP, varying in task complexity (binary vs. multi-class) and data availability (large vs. limited vs. small).

2. Models and Metrics

For empirical evaluation, we use several off-the-shelf BiLMs: a strong general-purpose model like RoBERTa (125M parameters) [7], and models specifically pretrained on broader clinical text (pathologyBERT (108M parameters) [8], and Gatortron (345M parameters) [9]). For our LLM, we use Mistral nemo instruct (12B parameters) [10] selected as a representative high-performance open-source LLM that can fit on a commodity GPU. We further pretrain RoBERTa and Gatortron on 1M pathology reports from BCCR using Masked Language Modeling (MLM) and denote these domain-specific pretrained models as BCCRoBERTa and BCCRTron. For performance reporting, we use the macro-averaged F1 scores. All experiments were performed on a desktop machine, with Intel Xeon processor, 32 GB RAM and NVIDIA A5000 GPU.

2.1. Off-the-Shelf Models: Finetuning vs. Zero-Shot

A fundamental decision when deploying a pretrained language model concerns the adaptation strategy. Should we invest resources in finetuning the model on a task-specific dataset, or can the model be used effectively "out-of-the-box" using zero-shot approaches? Large Language Models (LLMs), particularly those with billions of parameters, have demonstrated remarkable zero-shot and few-shot capabilities [9]. By providing task instructions and, optionally, a few examples directly within the input prompt, these models can often perform surprisingly well on tasks for which they weren't explicitly trained. This approach is advantageous as it requires minimal or no task-specific labeled data and can significantly speed up development cycles. However, zero-shot performance can be variable and highly sensitive to prompt phrasing, and may not reach the peak accuracy achievable through dedicated training for complex or highly specialized tasks. Finetuning, alternatively, involves updating the model's weights using a labeled dataset specific to the target task [11]. This typically leads to higher performance on the specific task, better adaptation to domain-specific nuances, and potentially more reliable outputs. The trade-offs include the need for labeled data and the computational cost of the training process.

For empirical evaluation using our scenarios, we observe that for BiLMs, finetuning consistently outperforms the zero-shot approaches where zero-shot performance ranges from 0.34 to 0.40 for scenario (a), 0.01 for scenario (b), and 0.02–0.13 for scenario (c). After finetuning, the performance increased to 0.95–0.97 for scenario (a), 0.78–0.85 for scenario (b), and 0.61–0.78 for scenario (c). For LLM, in its zero-shot capacity, it reached the performance of 0.76 for scenario (a), 0.54 for scenario (b), and 0.65 for scenario (c). LLM zero-shot performance was evaluated using task-specific instructions provided via prompting. To mirror resource constraints encountered in the real-world applications, we do not finetune the LLM. We provide detailed results in Table 1. This aligns with expectations, as BiLMs lack the zero-shot capabilities of modern LLMs, particularly for out-of-domain tasks.

Table 1. Detailed results for all models and all scenarios.

Models	Task 1 (Reportability Classification)	Task 2 (Tumor Group Classification)	Task 3 (Histology Classification)
RoBERTa—zeroshot	0.34	0.01	0.02
PathologyBERT—zeroshot	0.40	0.01	0.04
Gatortron—zeroshot	0.34	0.01	0.13
Mistral—zeroshot	0.76	0.54	0.65
RoBERTa—finetuned	0.96	0.78	0.61
PathologyBERT—finetuned	0.95	0.81	0.60
Gatortron—finetuned	0.97	0.85	0.78
BCCRoBERTa—finetuned	0.97	0.84	0.71
BCCRTron—finetuned	0.97	0.85	0.89

2.2. The Value of Further Domain-Specific Pretraining

The question of whether further pretraining on domain-specific data is beneficial often arises, especially when base models are already trained on broad or domain-adjacent datasets. For instance, if a model like Gatortron or pathologyBERT has been pretrained on a large corpus of clinical notes, is there value in continuing the pretraining process specifically on the data from one's own institution before finetuning for a downstream task? The argument for further pretraining rests on the potential for the model to learn domain-specific vocabulary, jargon, syntax, and subtle semantic relationships that might be underrepresented or absent in the initial pretraining corpus [12]. Further pretraining on specific data allows the model to adjust its internal representations to better reflect this unique linguistic distribution, potentially leading to improved performance on downstream tasks like information extraction or classification within that narrow domain.

However, this process requires access to a substantial amount of unlabeled domain-specific text and incurs significant computational costs. The benefits may also diminish if the initial pretraining data was already highly relevant or if the downstream task dataset is large enough for finetuning to effectively adapt the model. Therefore, the decision to pursue further pretraining requires a careful cost–benefit analysis, weighing the potential performance gains against the required resources and the degree of divergence between the existing model's knowledge and the target domain's specific characteristics.

Using BCCRoBERTa and BCCRTron, for scenarios (a) and (b), we observe that finetuned BCCRoBERTa outperforms finetuned RoBERTa (0.97 vs. 0.96 and 0.84 vs. 0.78 respectively), whereas BCCRTron shows no improvement compared to Gatortron. This suggests that for these tasks (a and b) with relatively larger finetuning datasets, Gatortron's initial large-scale clinical pretraining may have already captured sufficient domain-relevant

features, while the general-purpose RoBERTa benefited more from pathology-specific adaptation. For scenario (c), we observe large differences when using further pretrained models (0.71 vs. 0.61 for BCCRoBERTa vs. RoBERTa and 0.89 vs. 0.78 for BCCRTron vs. Gatortron).

2.3. Revisiting the Questions

With the help of the information provided above, we conclude the paper by revisiting the questions and providing our recommendations based on our experience and empirical evidence, while acknowledging the limitation that we have only tested a small set of models.

1. Do we need to finetune a model or can we use it in its zero-shot capacity?

Our evaluation clearly shows that finetuned models outperform their zero-shot counterparts, especially for BiLMs, as they have limited zero-shot capabilities. LLMs, as anticipated perform better in zero-shot capacity than their BiLM counterparts, but are outperformed by finetuned BiLMs.

Our recommendation: BiLMs should be finetuned on the task-specific dataset. A finetuned BiLM outperforms a zero-shot LLM for many well-defined tasks, especially where the models are intended for use in a specialized domain (cancer pathology reports in our case).

2. Are domain-adjacent pretrained models better than general pretrained models?

We compare the finetuned version of general pretrained models (RoBERTa) with domain-adjacent pretrained models (PathologyBERT and Gatortron), and we observe that the domain-adjacent models, when finetuned, often outperform general pretrained models, especially for complex tasks where the data availability for finetuning is scarce.

Our recommendation: Echoing the recommendations from [13], where possible, finetune a domain-adjacent pretrained model compared to a general model.

3. Is further pretraining on domain-specific data helpful?

While further pretraining on domain-specific data offers small gains for simple tasks, it improves the downstream model performance for complex tasks and/or when the finetuning dataset is small (such as histology classification in our case).

Our recommendation: If the data and compute resources are available, it is worthwhile to further pretrain BiLMs for an extra performance boost.

4. With the rise of LLMs, are BiLMs like BERT still relevant?

We have shown that for specialized tasks, BiLMs still endure and outperform zero-shot LLMs. While LLMs excel at generation, complex reasoning, and few-shot learning across a vast range of topics, BiLMs often provide sufficient or even superior performance for many well-defined NLP tasks, such as text classification, named entity recognition, etc., particularly when task-specific training data is available for finetuning. Further, for resource-constrained environments, which are commonplace, BiLMs offer a compelling balance of performance and efficiency. Further, when it comes to explainability, BiLMs are more amenable to explainability techniques focused on input token importance, making them a good fit for clinical use cases [14,15].

Our recommendation: If the task is well-defined (example: classification) and the data is available for finetuning, BiLMs should be the first choice.

Author Contributions: L.G., J.S., and G.S. conceived and designed the study. S.D. is the SME and helped with data preparation. G.A. and R.N. contributed to the interpretation of the results. L.G. and J.S. drafted the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The paper uses real-world data with sensitive patient information; therefore, it cannot be made publicly available. For access to the data and code, please contact Lovedeep Gondara (lovedeep.gondara@ubc.ca).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Qin, L.; Chen, Q.; Feng, X.; Wu, Y.; Zhang, Y.; Li, Y.; Li, M.; Che, W.; Yu, P.S. Large language models meet NLP: A survey. *arXiv* **2024**, arXiv:2405.12819. [CrossRef]
2. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; Volume 1 (long and short papers), pp. 4171–4186. [CrossRef]
3. Bedi, S.; Liu, Y.; Orr-Ewing, L.; Dash, D.; Koyejo, S.; Callahan, A.; Fries, J.A.; Wornow, M.; Swaminathan, A.; Lehmann, L.S.; et al. A systematic review of testing and evaluation of healthcare applications of large language models (LLMs). *medRxiv* **2024**. [CrossRef]
4. Wang, D.; Zhang, S. Large language models in medical and healthcare fields: Applications, advances, and challenges. *Artif. Intell. Rev.* **2024**, *57*, 299. [CrossRef]
5. Gondara, L.; Simkin, J.; Arbour, G.; Devji, S.; Ng, R. Classifying Tumor Reportability Status From Unstructured Electronic Pathology Reports Using Language Models in a Population-Based Cancer Registry Setting. *JCO Clin. Cancer Inform.* **2024**, *8*, e2400110. [CrossRef] [PubMed]
6. Gondara, L.; Simkin, J.; Devji, S.; Arbour, G.; Ng, R. ELM: Ensemble of Language Models for Predicting Tumor Group from Pathology Reports. *arXiv* **2025**, arXiv:2503.21800. [CrossRef]
7. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized BERT pretraining approach. *arXiv* **2019**, arXiv:1907.11692. [CrossRef]
8. Santos, T.; Tariq, A.; Das, S.; Vayalpati, K.; Smith, G.H.; Trivedi, H.; Banerjee, I. PathologyBERT-pre-trained vs. a new transformer language model for pathology domain. In Proceedings of the AMIA Annual Symposium Proceedings, San Francisco, CA, USA, 29 April 2023; Volume 2022, p. 962.
9. Yang, X.; Chen, A.; PourNejatian, N.; Shin, H.C.; Smith, K.E.; Parisien, C.; Compas, C.; Martin, C.; Flores, M.G.; Zhang, Y.; et al. Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records. *arXiv* **2022**, arXiv:2203.03540. [CrossRef]
10. Jiang, A.Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D.S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. Mistral 7B. *arXiv* **2023**, arXiv:2310.06825. [CrossRef]
11. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
12. Howard, J.; Ruder, S. Universal language model fine-tuning for text classification. *arXiv* **2018**, arXiv:1801.06146. [CrossRef]
13. Gururangan, S.; Marasović, A.; Swayamdipta, S.; Lo, K.; Beltagy, I.; Downey, D.; Smith, N.A. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv* **2020**, arXiv:2004.10964. [CrossRef]
14. Kraišniković, C.; Harb, R.; Plass, M.; Al Zoughbi, W.; Holzinger, A.; Müller, H. Fine-tuning language model embeddings to reveal domain knowledge: An explainable artificial intelligence perspective on medical decision making. *Eng. Appl. Artif. Intell.* **2025**, *139*, 109561. [CrossRef]
15. Kokalj, E.; Škrlj, B.; Lavrač, N.; Pollak, S.; Robnik-Šikonja, M. BERT meets shapley: Extending SHAP explanations to transformer-based classifiers. In Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation, Online, 19 April 2021; pp. 16–21.