Addressing Few-Shot LLM Classification Instability Through Explanation-Augmented Distillation

William Muntean[1] **and Joe Betts**[1]
[1]National Council of State Boards of Nursing (NCSBN)
Chicago, IL
Correspondence: wmuntean@ncsbn.org

Abstract

Large language models (LLMs) are increasingly adopted for educational assessment despite evidence that specialized models achieve superior performance. This study compares few-shot in-context learning with explanation-augmented knowledge distillation for exam question classification using medical education data. Few-shot learning exhibited substantial performance instability, with accuracy varying up to 14 percentage points based on example selection, while knowledge distillation provided consistent 70.1% accuracy after proper hyperparameter optimization. Though neither LLM approach matched specialized BERT performance (80.5%), knowledge distillation eliminated the reliability issues plaguing few-shot methods, offering organizations a stable solution for leveraging existing LLM infrastructure in operational assessment applications.

1 Introduction

Large language models (LLMs) have gained widespread adoption across educational assessment applications, driven by their versatility and the appeal of unified infrastructure that can handle multiple tasks without maintaining separate specialized models. However, this adoption occurs despite evidence that task-specific approaches often achieve superior performance. Bucher and Martini (2024) demonstrated that fine-tuned smaller models, including BERT-based classifiers, significantly outperform both zero-shot and few-shot LLM approaches in text classification tasks. This performance gap raises important questions about how organizations already invested in LLM infrastructure can most effectively leverage these capabilities for reliable educational assessment applications, even when accepting that peak performance may require specialized alternatives.

Few-shot in-context learning represents the most straightforward approach to LLM-based classification, requiring no model training while promising reasonable performance through carefully selected examples. However, recent research has revealed substantial instability in few-shot classification performance, with accuracy varying significantly based on example selection, ordering, and prompt construction choices (Nguyen and Wong 2023; Alves et al. 2023; Wan et al. 2023). This variability extends beyond minor fluctuations, with identical examples presented in different orders producing measurably different classification outcomes. For operational assessment systems requiring consistent and reliable performance, such instability undermines the practical utility of few-shot approaches, even when average performance might be acceptable. The sensitivity to configuration choices introduces an additional layer of complexity that conflicts with the apparent simplicity that makes few-shot learning initially attractive.

Knowledge distillation offers a promising solution for organizations committed to LLM-based approaches, enabling the transfer of reasoning capabilities from large models to smaller, more efficient counterparts while maintaining performance consistency. Unlike few-shot learning, knowledge distillation produces stable models that do not depend on carefully curated examples at inference time. Explanation-augmented distillation extends this approach by incorporating the reasoning patterns and decision processes of teacher models, potentially capturing more nuanced classification strategies than traditional output-only distillation methods (Xu et al., 2024). While this approach may not achieve the peak performance of specialized classifiers, it represents an optimization strategy for organizations seeking to maximize the reliability and efficiency of LLM-based classification within existing infrastructure constraints. This study evaluates whether explanation-augmented knowledge distillation can provide the consistency and computational efficiency needed for operational deploy-

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 2: Works in Progress, pages 197–203

ment while delivering competitive performance relative to unstable few-shot alternatives.

1.1 Knowledge Distillation in Natural Language Processing

Knowledge distillation has emerged as a powerful technique for transferring capabilities from large, computationally expensive models to smaller, more efficient alternatives while maintaining competitive performance. Originally developed for computer vision applications (Hinton et al., 2015), the approach has been successfully adapted to natural language processing tasks, where the computational demands of large language models create significant deployment challenges. Traditional knowledge distillation focuses on matching output distributions between teacher and student models, enabling smaller models to approximate the decision boundaries learned by their larger counterparts (Gou et al., 2021).

Recent advances in explanation-augmented knowledge distillation extend beyond output matching to incorporate the reasoning processes of teacher models. This approach leverages the natural language generation capabilities of large language models to produce detailed explanations alongside predictions, creating richer training signals for student models (DeepSeek-AI et al., 2025). By learning to replicate both the decisions and reasoning patterns of teacher models, student models may achieve better generalization and more robust performance across diverse inputs. However, the effectiveness of explanation-augmented distillation for classification tasks in educational domains remains underexplored.

1.2 Few-Shot Learning Instability

While few-shot in-context learning offers apparent simplicity for LLM deployment, mounting evidence reveals significant performance instability across different configuration choices. Studies have documented substantial variance in classification accuracy based on example selection, with different sets of representative examples producing measurably different results even when controlling for example quality and domain coverage (Nguyen and Wong, 2023). This instability extends to example ordering effects, where identical examples presented in different sequences can alter model predictions.

The sensitivity of few-shot learning to prompt construction choices poses particular challenges for operational deployment in educational assessment. Beyond random variation, systematic biases may emerge when examples exhibit consistent characteristics that do not represent the full complexity of the classification task (Tjuatja et al., 2024). These findings suggest that the apparent simplicity of few-shot learning may be misleading, as achieving reliable performance requires careful curation and validation of example sets—a process that may be as complex as traditional model training approaches.

1.3 Exam Question Classification

Educational assessment systems rely heavily on accurate classification of exam questions into predefined content domains to ensure proper test construction, maintain content validity, and support diagnostic feedback (Kane, 2006). This classification task involves mapping individual questions to taxonomic categories that reflect the knowledge, skills, or competencies being assessed. In medical education, for example, questions must be aligned with clinical domains, procedural categories, or competency frameworks to ensure comprehensive coverage of required learning outcomes (Bridge et al., 2003).

Traditional approaches to question classification have relied on manual expert review or rule-based systems, but the scale of modern item banks and the complexity of question content have motivated automated classification methods. Recent advances in data-driven approaches have also extended to optimizing assessment items themselves, including systematic methods for refining item options (Muntean et al., 2025). While specialized models like fine-tuned BERT classifiers have demonstrated superior performance for this task (Bucher and Martini, 2024), many educational organizations seek to benefit from existing LLM infrastructure for question classification as part of broader assessment workflows.

The stakes for classification accuracy in educational assessment are particularly high, as misclassified questions can compromise test validity, lead to content imbalances, and undermine the reliability of score interpretations (Messick, 1995). This context demands not only reasonable classification performance but also consistent and predictable behavior across diverse question types and content areas.

1.4 Research Questions

This study investigates the effectiveness of explanation-augmented knowledge distillation for exam question classification compared to few-shot in-context learning approaches. Specifically, we address the following research questions:

RQ1 How does few-shot in-context learning performance vary when examples are systematically selected based on question difficulty (easy vs. difficult vs. mixed examples within each content domain)?

RQ2 Can explanation-augmented knowledge distillation produce student models that achieve competitive classification accuracy compared to few-shot learning approaches while maintaining greater performance consistency?

RQ3 How sensitive is explanation-augmented knowledge distillation to hyperparameter choices, and what configurations optimize the trade-off between performance and training efficiency?

2 Methods

2.1 Dataset

We utilized a subset of a medical examination item bank containing 6,839 multiple-choice questions labeled according to eight high-level test blueprint domains (National Council of State Boards of Nursing, 2023). Items were randomly selected from questions that had passed all statistical pretest criteria. The dataset was partitioned using stratified sampling to maintain domain proportions: 4,103 questions (60%) for training, 1,368 questions (20%) for validation, and 1,368 questions (20%) for testing. Question difficulty was determined using population-calibrated item difficulty values, with difficulty distributions roughly equivalent across the eight content domains. The classification task involved mapping individual question to their corresponding content domains based on the medical knowledge and competencies being assessed.

To establish performance benchmarks, we implemented a BERT-based classification model using the all-MiniLM-L6-v2 sentence transformer (Wang et al., 2020). We fine-tuned the model using contrastive learning with 8,000 question pairs (1,000 pairs per domain, consisting of 500 positive and 500 negative pairs), representing the specialized classification method that has been shown to outperform LLM-based approaches in similar text classification tasks.

2.2 Few-Shot Learning Experiments

We conducted few-shot learning experiments primarily using GPT-OSS-20B (OpenAI et al., 2025), with pilot studies on GPT-OSS-120B, LLaMA 4 Maverick (Meta AI, 2025), and Claude Sonnet 3.7 (Anthropic, 2025) to validate that instability patterns generalize across different large language models. All models were accessed through Databricks environment endpoints. To systematically investigate the impact of example difficulty on few-shot performance, we created three experimental conditions based on population-calibrated difficulty values. For easy examples, we selected the 25 easiest items and randomly divided them into 5 sets of 5 items per domain. For difficult examples, we selected the 25 most difficult items and applied the same division strategy. For mixed examples, we randomly selected 5 items per domain, repeated 5 times. Each condition resulted in 5 replications of 40 few-shot examples (5 examples from each of the 8 domains), enabling assessment of both systematic bias effects and random variation. All few-shot examples were drawn from the training set to prevent data leakage.

The prompt structure consisted of task instructions, content domain definitions, few-shot examples with their classifications, and repeated instructions with output format specifications. All models were required to follow structured output formatting to ensure consistent response parsing. Each replication was evaluated on the complete test set to quantify performance variability across different example selections.

2.3 Knowledge Distillation Experiments

We implemented explanation-augmented knowledge distillation using LLaMA 3.1 405B as the teacher model and LLaMA 3.1 8B as the student model. For all 4,103 training examples, we prompted the teacher model to generate detailed rationales explaining why each question belonged to its specified domain and why alternative domains were less appropriate. This process created question-explanation-classification triplets that enabled the student model to learn both the reasoning patterns and classification decisions of the teacher model.

The student model underwent full parameter fine-tuning on these explanation-augmented sequences

using Databricks parameter sweep functionality. We conducted systematic hyperparameter optimization across a 2×2 experimental design with learning rates of 1×10^{-6} versus 1×10^{-7} and training epochs of 1 versus 2. This design enabled assessment of hyperparameter sensitivity while maintaining computational feasibility for the full parameter fine-tuning approach.

2.4 Evaluation

We evaluated all approaches using overall accuracy and weighted F_1-score on the held-out test set (1,368 questions) to ensure unbiased performance assessment. For few-shot learning approaches, we measured performance consistency by calculating the standard deviation and range (maximum - minimum accuracy) across the 5 replications within each difficulty condition. This analysis quantifies both the magnitude and variability of performance instability across different example selections, enabling direct comparison with the consistent performance of knowledge distillation approaches.

3 Results

The experimental results demonstrate clear performance differences between approaches and reveal significant instability in few-shot learning methods. The BERT baseline achieved the highest overall performance with 80.5% accuracy and 80.4% weighted F_1-score, confirming prior findings that specialized models outperform LLM-based approaches for text classification tasks. However, the comparison between few-shot learning and knowledge distillation reveals important insights about the viability of LLM-based classification methods.

3.1 Few-Shot Learning Performance and Instability

Few-shot learning performance varied substantially based on example difficulty, with counterintuitive results regarding the relationship between example difficulty and classification accuracy. Models performed best when provided with difficult examples (62.7% accuracy, 60.7% F_1), followed by random examples (56.0% accuracy, 54.0% F_1), and worst with easy examples (52.8% accuracy, 51.4% F_1). This unexpected finding suggests that difficult questions may provide richer contextual information or more distinctive features that help models distinguish between content domains.

More critically, few-shot learning exhibited substantial performance instability across different example selections within each difficulty condition. The difficult examples condition showed the highest variability, with accuracy ranging from 56.8% to 71.0% (standard deviation = 6.1%) and F_1 scores ranging from 53.9% to 68.9% (standard deviation = 6.7%). Easy examples demonstrated moderate instability with accuracy ranging from 47.3% to 58.2% (standard deviation = 4.4%), while random examples showed the most consistent performance with accuracy ranging from 51.7% to 60.7% (standard deviation = 3.9%). Despite this relative consistency, even the random condition exhibited meaningful performance variation that could impact operational deployment reliability.

The instability patterns were consistent across multiple large language models tested in pilot studies, including GPT-OSS-120B, LLaMA 4 Maverick, and Claude Sonnet 3.7, indicating that few-shot learning instability represents a general phenomenon rather than model-specific behavior. This cross-model consistency strengthens the evidence that example selection significantly impacts few-shot classification performance regardless of the underlying architecture.

3.2 Knowledge Distillation Performance and Stability

Explanation-augmented knowledge distillation results revealed extreme sensitivity to hyperparameter selection, with learning rate choice proving critical for successful model training. The optimal configuration using learning rate 1×10^{-6} and 2 training epochs achieved 70.1% accuracy and 70.4% weighted F_1-score, representing competitive performance relative to few-shot learning approaches while completely eliminating the instability associated with example selection.

Hyperparameter analysis revealed dramatic performance differences based on learning rate selection. Models trained with learning rate 1×10^{-6} substantially outperformed those trained with 1×10^{-7}, likely due to catastrophic forgetting effects at the extremely low learning rate that prevented adequate adaptation to the classification task. The 1×10^{-7} learning rate produced poor performance regardless of epoch count (44.0% accuracy with 1 epoch, 47.0% accuracy with 2 epochs), while the 1×10^{-6} learning rate enabled effective learning (60.8% accuracy with 1 epoch, 70.1% accuracy with 2 epochs). The improvement from 1 to 2 epochs at the higher learning rate suggests that additional training time benefits explanation-

augmented distillation when appropriate learning rates are used.

Knowledge distillation demonstrated complete stability across evaluation runs, producing identical performance metrics without the variability that characterized few-shot approaches. This stability emerges from the fundamental difference in methodology: rather than relying on a small set of potentially biased examples at inference time, the distilled model learns from the complete training dataset during fine-tuning. The student model internalizes classification patterns through exposure to the full range of question types and difficulty levels, eliminating dependence on the specific characteristics of a limited example set. Once trained, the distilled model produces consistent predictions without requiring carefully curated examples, removing the primary source of instability that affects few-shot learning.

3.3 Comparative Analysis

The knowledge distillation approach addresses the primary limitation of few-shot learning by providing consistent performance without dependence on carefully curated examples. While the best knowledge distillation configuration (70.1% accuracy) did not exceed the maximum few-shot performance (71.0% accuracy with difficult examples), it achieved performance within the range of few-shot results while eliminating the risk of poor performance due to unfavorable example selection. Notably, the knowledge distillation performance exceeded the minimum few-shot performance across all difficulty conditions and matched the performance of the best few-shot condition (difficult examples) while avoiding the substantial variability that makes few-shot approaches unreliable.

The comparison reveals a fundamental trade-off between peak performance potential and performance consistency. Few-shot learning offers the possibility of higher performance when examples are carefully selected, but carries substantial risk of poor performance with different example choices. Knowledge distillation provides predictable performance that falls within the middle-to-upper range of few-shot results, representing a viable solution for organizations requiring reliable classification performance from LLM-based approaches. The elimination of example-dependent variability makes knowledge distillation particularly suitable for operational deployment where consistent performance is more valuable than occasional peak performance.

4 Discussion

This study provides empirical evidence that explanation-augmented knowledge distillation offers a viable solution to the instability problems that plague few-shot in-context learning for educational question classification. While neither LLM-based approach achieves the performance of specialized BERT classifiers, the findings reveal important practical considerations for organizations committed to leveraging existing LLM infrastructure for assessment applications.

The substantial performance variability observed in few-shot learning—with accuracy ranges exceeding 14 percentage points in some conditions—represents a significant barrier to operational deployment. This instability extends beyond random variation to include systematic biases based on example characteristics, as demonstrated by the counterintuitive finding that difficult examples produced better classification performance than easy examples. This result suggests that few-shot learning may be sensitive to the cognitive complexity and feature richness of selected examples in ways that are difficult to predict or control. The consistency of these instability patterns across multiple large language models indicates that the problem is fundamental to the few-shot learning paradigm rather than specific to particular architectures. For educational assessment applications, this variability is particularly concerning as unreliable classification performance can compromise test validity and undermine confidence in automated systems. The observed variability means that identical classification tasks could produce different results depending solely on example selection choices, creating potential fairness and consistency issues in high-stakes assessment environments.

Explanation-augmented knowledge distillation addresses these limitations by fundamentally changing the relationship between examples and model performance. Rather than depending on a small set of potentially biased examples at inference time, the distilled model learns from comprehensive exposure to the full training dataset, internalizing classification patterns that remain consistent across evaluations. This methodological difference eliminates the primary source of instability in few-shot approaches while maintaining competitive performance levels. The extreme sen-

sitivity to hyperparameter selection observed in our distillation experiments, particularly the dramatic performance differences between learning rates, highlights the importance of systematic optimization rather than relying on conventional parameter choices. The poor performance at learning rate 1×10^{-7} likely reflects catastrophic forgetting, where the extremely conservative learning rate prevented adequate adaptation to the classification task. However, once properly configured, the distilled model produces stable and reliable performance without the variability that characterizes few-shot approaches.

The trade-off between peak performance and consistency revealed in our results reflects broader considerations in educational technology deployment. While few-shot learning may occasionally achieve higher performance with optimal example selection, the risk of poor performance with suboptimal examples may be unacceptable in assessment contexts where consistent behavior is essential. Knowledge distillation provides a middle path that sacrifices some performance potential for greater reliability and predictability, making it particularly suitable for operational assessment applications where consistency is paramount.

Several limitations should be considered when interpreting these results. Our evaluation focused on a single domain (medical education) and classification task, and generalization to other educational contexts requires further investigation. The hyperparameter space explored for knowledge distillation was limited, and more comprehensive optimization might yield improved performance. Future research should investigate the effectiveness of explanation-augmented distillation across diverse educational domains, examine different distillation methods, and analyze the quality and utility of generated explanations. Additionally, research into methods for automatically selecting optimal few-shot examples or reducing example dependency could address some of the limitations identified in few-shot approaches.

This study demonstrates that explanation-augmented knowledge distillation provides a practical solution to the instability problems inherent in few-shot learning approaches for educational question classification. The elimination of example-dependent variability, combined with competitive performance levels, makes knowledge distillation particularly suitable for operational assessment applications where consistency and reliability are paramount. These findings contribute to the growing understanding of how to effectively deploy large language models in educational contexts while managing their inherent limitations and operational constraints, offering organizations a viable path to leverage existing LLM infrastructure reliably and consistently.

References

Duarte Alves, Nuno Guerreiro, João Alves, José Pombal, Ricardo Rei, José de Souza, Pierre Colombo, and Andre Martins. 2023. Steering large language models for machine translation with finetuning and in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11127–11148, Singapore. Association for Computational Linguistics.

Anthropic. 2025. Claude 3.7 sonnet [large language model]. https://www.anthropic.com/news/claude-3-7-sonnet. Anthropic PBC.

Patrick D Bridge, Joseph Musial, Robert Frank, Thomas Roe, and Shlomo Sawilowsky. 2003. Measurement practices: methods for developing content-valid student examinations. *Medical teacher*, 25(4):414–421.

Martin Juan José Bucher and Marco Martini. 2024. Fine-tuned 'small' llms (still) significantly outperform zero-shot generative ai models in text classification. *Preprint*, arXiv:2406.08660.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *Preprint*, arXiv:1503.02531.

Michael Kane. 2006. Content-related validity evidence in test development. *Handbook of test development*, 1:131–153.

Samuel Messick. 1995. Standards of validity and the validity of standards in performance asessment. *Educational measurement: Issues and practice*, 14(4):5–8.

Meta AI. 2025. Llama 4 maverick (17b-128e) [large language model]. https://huggingface.co/meta-llama/

Llama-4-Maverick-17B-128E-Instruct. Meta Platforms, Inc.

William Muntean, Joe Betts, Zhuoran Wang, and Hao Jia. 2025. Comparing data-driven methods for removing options in assessment items. *Journal of Educational Measurement*. Online First.

National Council of State Boards of Nursing. 2023. Next generation nclex: Nclex-rn test plan. https://www.ncsbn.org/public-files/2023_RN_Test%20Plan_English_FINAL.pdf. National Council of State Boards of Nursing, Chicago, IL. Accessed September 2025.

Tai Nguyen and Eric Wong. 2023. In-context example selection with influences. *Preprint*, arXiv:2302.11042.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, and 108 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Lindia Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkwar, and Graham Neubig. 2024. Do LLMs exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026.

Xingchen Wan, Ruoxi Sun, Hanjun Dai, Sercan O. Arik, and Tomas Pfister. 2023. Better zero-shot reasoning with self-adaptive prompting. *Preprint*, arXiv:2305.14106.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Preprint*, arXiv:2002.10957.

Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. 2024. A survey on knowledge distillation of large language models. *Preprint*, arXiv:2402.13116.