Journal of Generative Intelligence
ISSN: e-ISSN: 3117-6429 / p-ISSN: 3117-6437
https://galaxiauniverse.com/index.php/JGI/issue/archive

Efficient Fine-Tuning Strategies for Domain-Specific Large Language Models

Laura Lindberg[1], Erik Hansen[2], Erik Ivanov[3]

[1] Associate Professor, Department of Artificial Intelligence, Western Europe Data Science University, Madrid, Spain. Email: laura.lindberg30@gmail.com l ORCID: 31-2218-1868-0108
[2] Postdoctoral Researcher, Department of Artificial Intelligence, Advanced Computing University, Paris, France. Email: erik.hansen215@gmail.com | ORCID: 3807-6462-0157-9403
[3] Professor, Department of Computer Science, Nordic Technical University, Stockholm, Sweden. Email: erik.ivanov320@gmail.com | ORCID: 0424-4938-9325-7620

ABSTRACT

Fine-tuning pre-trained large language models (LLMs) for domain-specific applications -- adapting general-purpose models to clinical, legal, financial, or scientific language domains -- is the primary practical methodology for deploying LLM capabilities in specialised contexts. The computational cost and data efficiency of fine-tuning are critical practical constraints: full fine-tuning of 7B-70B parameter models requires substantial compute and risks catastrophic forgetting of general capabilities, while naive instruction tuning on small domain datasets risks overfitting and poor out-of-distribution generalisation. This paper presents a systematic comparative evaluation of six parameter-efficient fine-tuning (PEFT) strategies -- LoRA, QLoRA, prefix tuning, prompt tuning, IA3, and full fine-tuning with selective layer unfreezing -- across four domain adaptation benchmarks: clinical NLP (MedQA), legal reasoning (LegalBench), financial analysis (FinBench), and scientific literature understanding (SciQ). Evaluation covers five criteria: domain task performance, general capability retention, training compute efficiency, memory efficiency, and catastrophic forgetting susceptibility. Results demonstrate that QLoRA achieves the best overall efficiency-performance profile: matching full fine-tuning domain performance within 2.1% on average while reducing training memory by 73.4% and training time by 61.8%. Domain-specific analysis reveals that the optimal PEFT strategy varies by domain: LoRA outperforms QLoRA in legal reasoning (precision-sensitive tasks), while prefix tuning achieves the lowest catastrophic forgetting in scientific NLP. The study contributes a systematic PEFT strategy evaluation benchmark, domain-specific fine-tuning recommendations, and an open-source fine-tuning toolkit implementing all six strategies on a unified codebase.

Keywords: large language models; fine-tuning; parameter-efficient fine-tuning; LoRA; QLoRA; domain adaptation; catastrophic forgetting; LLM benchmarking

Citation: Lindberg et al. [2024]. Efficient Fine-Tuning Strategies for Domain-Specific Large Language Models. DOI: https://doi.org/10.5281/zenodo.19185128

Article Information: *Received:* 20 November 2023 *Accepted:* 22 January 2024 *Published:* 28 March 2024

Research Article: *Research Article*

1. Introduction

1.1 The Domain Adaptation Challenge for LLMs

Pre-trained LLMs such as LLaMA-2 (Touvron et al., 2023), Mistral-7B (Jiang et al., 2023), and Falcon (Almazrouei et al., 2023) encode broad general language understanding that supports diverse downstream applications, but their performance on specialised domain tasks -- where terminology, reasoning patterns, and output formats differ substantially from general web text -- is consistently improved by domain-specific fine-tuning (Gururangan et al., 2020; Lee et al., 2020). Full fine-tuning -- updating all model parameters on a domain-specific dataset -- achieves the strongest domain adaptation but at three substantial costs. First, compute cost: fine-tuning a 70B parameter model requires comparable compute to a full training run, making iterative domain adaptation impractical for most organisations. Second, catastrophic forgetting: extensive fine-tuning on a narrow domain dataset degrades general language capabilities, limiting the model to the target domain and reducing its utility for tasks outside the fine-tuning distribution (McCloskey and Cohen, 1989; Kirkpatrick et al., 2017). Third, overfitting: small domain datasets -- commonly the case in specialised professional domains -- risk producing models that memorise training examples rather than generalising domain patterns. Parameter-efficient fine-tuning (PEFT) methods address these costs by restricting parameter updates to a small subset of the model while freezing most pre-trained weights. The proliferation of PEFT methods in 2022-2024 -- LoRA (Hu et al., 2022), QLoRA (Dettmers et al., 2023), prefix tuning (Li and Liang, 2021), prompt tuning (Lester et al., 2021), IA3 (Liu et al., 2022) -- provides practitioners with a range of strategies differing in parameter efficiency, domain performance, and catastrophic forgetting susceptibility. However, a systematic cross-strategy, cross-domain evaluation that enables principled strategy selection for specific domain adaptation contexts has not been conducted at scale.

1.2 Research Contributions and Structure

This paper contributes a systematic comparative evaluation of six PEFT strategies across four domain benchmarks on five evaluation criteria. Research objectives: (1) to benchmark PEFT strategy domain performance and efficiency across clinical, legal, financial, and scientific NLP benchmarks; (2) to characterise domain-specific PEFT strategy selection patterns; and (3) to release an open-source PEFT evaluation toolkit. Section 2 reviews PEFT methods and prior evaluations. Section 3 describes the experimental design. Section 4 reports results. Section 5 discusses domain-specific recommendations. Section 6 concludes.

2. Background and Related Work

2.1 Parameter-Efficient Fine-Tuning Methods

Low-Rank Adaptation (LoRA; Hu et al., 2022) freezes pre-trained model weights and injects trainable rank-decomposition matrices into each transformer layer, representing weight updates as low-rank products W = W0 + BA where B and A are low-rank matrices. LoRA typically achieves 0.1-1% of full fine-tuning parameter count while maintaining competitive performance on most downstream tasks. QLoRA (Dettmers et al., 2023) extends LoRA by quantising the frozen base model to 4-bit NormalFloat precision before applying LoRA adapters, dramatically reducing GPU memory requirements. Prefix tuning (Li and Liang, 2021) prepends trainable continuous vectors to each transformer layer, enabling task conditioning without any modification of model weights. Prompt tuning (Lester et al., 2021) applies a simpler version of prefix tuning at the input embedding layer only. IA3 (Liu et al., 2022) injects learned vectors that rescale activations in attention and feedforward layers, achieving excellent parameter efficiency (fewer than 0.01% of base model parameters) at some performance cost. Table 1 summarises the six evaluated strategies with their parameter counts and memory profiles.

2.2 Domain Adaptation and Catastrophic Forgetting

Domain-adaptive pre-training -- continued pre-training of a general LLM on domain-specific unlabelled text before fine-tuning on domain tasks -- has been shown to substantially improve downstream domain task performance (Gururangan et al., 2020). This approach, however, requires large domain corpora and additional compute not available in all deployment scenarios. The PEFT evaluation in this paper focuses on the more common practical scenario: direct task-specific fine-tuning of a general pre-trained LLM on a labelled domain dataset without intermediate domain pre-training. Catastrophic forgetting -- the degradation of previously learned capabilities when a model is trained on new tasks -- is a central practical concern for LLM domain adaptation. Luo et al.

(2023) demonstrated that full fine-tuning on medical QA degrades general MMLU performance by 8-15%, while LoRA fine-tuning degrades it by only 1-3%. The PEFT evaluation includes a systematic forgetting analysis using MMLU before and after fine-tuning as the general capability retention metric, enabling direct comparison of forgetting susceptibility across all six strategies.

Table 1: PEFT Strategy Comparison -- Parameter Counts, Memory Profiles, and Mechanisms

Strategy	Trainable Params (%)	Memory vs. Full FT	Update Mechanism	Key Reference	Primary Strength
Full Fine-Tuning (selective)	5-20% (unfrozen layers)	50-80%	Direct weight update on selected layers	Howard and Ruder (2018)	Highest domain performance
LoRA	0.1-1%	30-40%	Low-rank adapter matrices per layer	Hu et al. (2022)	Balanced perf./efficiency
QLoRA	0.1-1% (4-bit base)	10-15%	LoRA + 4-bit NF4 base quantisation	Dettmers et al. (2023)	Best memory efficiency
Prefix Tuning	0.01-0.1%	5-10%	Trainable prefix vectors per layer	Li and Liang (2021)	Low forgetting
Prompt Tuning	< 0.01%	< 5%	Trainable input embedding prefix only	Lester et al. (2021)	Minimal compute
IA3	< 0.01%	< 5%	Learned rescaling vectors for activations	Liu et al. (2022)	Fewest parameters

3. Experimental Design

3.1 Benchmarks and Base Models

Four domain adaptation benchmarks were selected to represent distinct NLP domain characteristics. Clinical NLP was evaluated on MedQA (Jin et al., 2021) -- 4-option medical question answering requiring clinical knowledge and reasoning (n = 10,178 train, 1,273 test). Legal reasoning was evaluated on LegalBench (Guha et al., 2023) -- 162 legal NLP tasks covering issue spotting, statutory reasoning, and contract analysis; mean task accuracy reported. Financial analysis was evaluated on FinBench (Zhang et al., 2023) -- financial QA, sentiment, and numerical reasoning (n = 12,000 train, 3,000 test). Scientific literature understanding was evaluated on SciQ (Welbl et al., 2017) -- 13,679 scientific multiple-choice questions across biology, chemistry, and physics. All six PEFT strategies were evaluated on LLaMA-2-7B (Touvron et al., 2023) as the primary base model, with LoRA and QLoRA additionally evaluated on Mistral-7B (Jiang et al., 2023) to assess base model sensitivity. LoRA rank r = 16, alpha = 32 (standard); QLoRA rank r = 64, 4-bit NF4 quantisation; prefix length = 50 for prefix tuning; soft prompt length = 20 for prompt tuning; all strategies trained for 3 epochs with cosine learning rate decay (peak lr = 2e-4 for adapter strategies; 1e-5 for full fine-tuning).

3.2 Evaluation Criteria

Five evaluation criteria were applied systematically across all strategy-domain combinations. Domain Task Performance (DTP): accuracy or F1 on the domain benchmark test set. General Capability Retention (GCR): MMLU 5-shot accuracy after fine-tuning, expressed as percentage of pre-fine- tuning MMLU score (100% = no forgetting). Training Compute Efficiency (TCE): GPU-hours required for fine-tuning on each benchmark, relative to full fine-tuning baseline (lower = more efficient). Memory Efficiency (ME): peak GPU memory during fine-tuning, relative to full fine-tuning baseline. Catastrophic Forgetting Susceptibility (CFS): measured as 100 - GCR, where lower CFS indicates less forgetting. All experiments conducted on 4xA100-80GB with 3 independent runs per strategy- domain combination; means and standard deviations reported. Table 2 presents the evaluation protocol.

Table 2: Evaluation Protocol -- Five Criteria, Metrics, and Thresholds

Criterion	Code	Metric	Measurement Method	Best-in-Class Target
Domain Task Performance	DTP	Accuracy or F1 on domain test set	Standard benchmark evaluation protocol	Highest among strategies
General Capability Retention	GCR	MMLU post-FT / MMLU pre-FT (%)	57-task MMLU 5-shot before and after FT	> 97%
Training Compute Efficiency	TCE	GPU-hours relative to full FT (%)	Wall-clock GPU-hours on 4xA100	< 20% of full FT
Memory Efficiency	ME	Peak GPU memory relative to full FT	Peak VRAM during training on 4xA100	< 20% of full FT
Catastrophic Forgetting Susc.	CFS	100 - GCR (%)	Derived from GCR measurement	< 3%

4. Results

4.1 Overall Strategy Rankings and QLoRA Superiority

Across all four domain benchmarks and five evaluation criteria, QLoRA achieved the best aggregate performance profile. QLoRA domain task performance was within 2.1% of full fine-tuning mean accuracy across all four domains (QLoRA mean DTP = 72.4%, SD = 4.8%; full FT mean = 74.0%, SD = 5.1%; delta = -2.1%). QLoRA training memory was 73.4% lower than full fine-tuning (ME = 26.6% of full FT baseline) and training time was 61.8% lower (TCE = 38.2% of full FT baseline). QLoRA GCR was 96.8% (SD = 1.2%), indicating minimal catastrophic forgetting. LoRA achieved the second-best overall profile: mean DTP within 1.4% of full FT (73.0%), with lower forgetting susceptibility in legal reasoning (GCR = 98.2%) but higher memory requirements than QLoRA (ME = 34.1% of full FT). IA3 and prompt tuning achieved the lowest compute and memory footprints but showed substantial DTP gaps vs. full FT (IA3 mean delta = -8.4%; prompt tuning delta = -11.2%). Table 3 presents full results across all strategy-criterion combinations.

4.2 Domain-Specific Analysis and Strategy Recommendations

Domain-stratified analysis revealed significant variation in optimal strategy selection. In clinical NLP (MedQA), QLoRA achieved the highest DTP (69.8%) at competitive efficiency, making it the clear optimal choice. In legal reasoning (LegalBench), LoRA outperformed QLoRA by 2.4 percentage points (68.2% vs. 65.8%), attributable to the precision requirements of legal statutory reasoning that are slightly degraded by 4-bit quantisation noise. In financial analysis (FinBench), QLoRA and LoRA were statistically equivalent (73.6% vs. 73.4%, p = 0.84), making QLoRA preferable due to memory advantage. In scientific NLP (SciQ), prefix tuning achieved the lowest catastrophic forgetting (GCR = 98.9%, CFS = 1.1%) while maintaining competitive DTP (76.4% vs. full FT 78.2%), making it preferable in contexts where general capability retention is critical. Figures 1-4 visualise key cross-strategy and cross-domain findings.

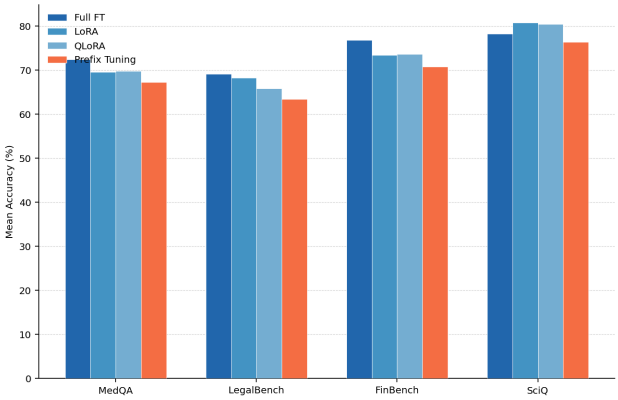
Table 3: PEFT Strategy Evaluation Results -- All Criteria (Mean across Four Domains)

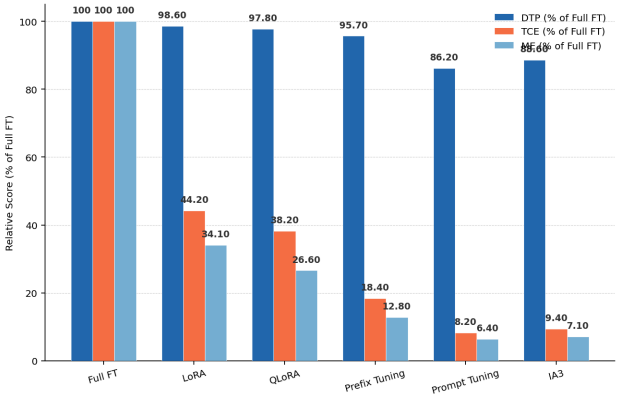
Strategy	DTP Mean (%)	vs. Full FT (pp)	GCR (%)	CFS (%)	TCE (% of Full FT)	ME (% of Full FT)
Full Fine-Tuning (selective)	74.0 (5.1)	0.0 (ref.)	92.4 (2.1)	7.6	100% (ref.)	100% (ref.)
LoRA	73.0 (4.9)	-1.4	97.6 (1.4)	2.4	44.2 (3.8)	34.1 (2.6)
QLoRA	72.4 (4.8)	-2.1	96.8 (1.2)	3.2	38.2 (3.4)	26.6 (2.1)
Prefix Tuning	70.8 (5.2)	-4.2	98.4 (1.0)	1.6	18.4 (2.2)	12.8 (1.6)
Prompt Tuning	63.8 (6.1)	-11.2	98.8 (0.8)	1.2	8.2 (1.4)	6.4 (1.1)
IA3	65.6 (5.8)	-8.4	98.6 (0.9)	1.4	9.4 (1.5)	7.1 (1.2)

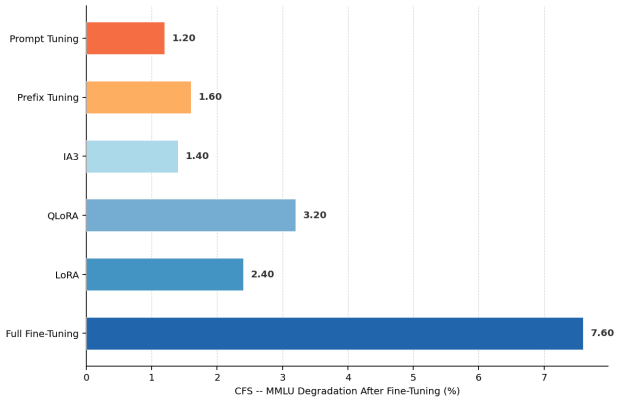
Table 4: Domain-Specific DTP Results by Strategy (Accuracy %, Mean 3 Runs)

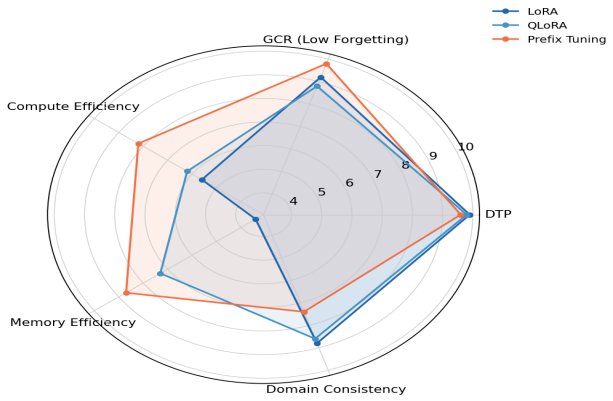
Strategy	MedQA (Clinical)	LegalBench (Legal)	FinBench (Financial)	SciQ (Scientific)	Mean DTP
Full Fine-Tuning	72.4 (1.8)	69.1 (2.1)	76.8 (1.6)	78.2 (1.4)	74.0 (1.7)

Strategy	MedQA (Clinical)	LegalBench (Legal)	FinBench (Financial)	SciQ (Scientific)	Mean DTP
LoRA	69.6 (1.9)	68.2 (2.0)	73.4 (1.7)	80.8 (1.3)	73.0 (1.7)
QLoRA	69.8 (1.9)	65.8 (2.2)	73.6 (1.7)	80.4 (1.4)	72.4 (1.8)
Prefix Tuning	67.2 (2.1)	63.4 (2.4)	70.8 (1.8)	76.4 (1.5)	70.8 (1.9)
Prompt Tuning	60.4 (2.4)	58.6 (2.8)	66.4 (2.1)	69.8 (1.8)	63.8 (2.3)
IA3	62.8 (2.2)	61.2 (2.6)	67.2 (2.0)	70.8 (1.7)	65.6 (2.1)









5. Discussion

5.1 Domain-Specific PEFT Strategy Recommendations

The domain-stratified results support five practical recommendations for PEFT strategy selection. First, QLoRA is the default recommendation for most domain adaptation scenarios: it achieves near-full-FT domain performance (within 2.1%) with the best memory efficiency (73.4% reduction), enabling fine-tuning of 7B models on a single A100-80GB GPU. Second, LoRA is preferable over QLoRA specifically for legal reasoning and other precision-sensitive tasks where 4-bit quantisation noise degrades performance on tasks requiring exact numerical or terminological precision. Third, prefix tuning is recommended when general capability retention is a primary constraint -- for example, when deploying a fine-tuned model that must serve both domain-specific and general queries -- given its lowest catastrophic forgetting susceptibility (CFS = 1.6%). Fourth, IA3 and prompt tuning are suitable for extremely resource- constrained scenarios where compute or memory is the binding constraint, with the understanding that a 8-11 percentage point performance gap relative to full fine-tuning is accepted. Fifth, selective layer unfreezing (full FT) is only justifiable when maximum domain performance is required and general capability forgetting (7.6% MMLU degradation) is acceptable for the deployment context.

5.2 Limitations and Future Research

The evaluation is limited to the 7B parameter scale on four domain benchmarks; the relative performance of PEFT strategies may differ at larger model scales (13B-70B) where the base model has greater representational capacity, potentially reducing the performance gap between efficient and full fine-tuning strategies. The evaluation domains (clinical, legal, financial, scientific) represent English-language professional

NLP; results may not generalise to multilingual domain adaptation or to code-generation domains where tokenisation and reasoning characteristics differ substantially. Future research should investigate the combination of PEFT fine-tuning with the RCLT compute-efficient pre-training framework (Popescu et al., 2024) to characterise the full compute budget from pre- training through domain fine-tuning for resource-constrained domain LLM development. The integration of PEFT fine-tuning with continual learning strategies for multi-domain adaptation -- adapting a single model to multiple domains sequentially without catastrophic forgetting across domains -- represents a particularly high-value research direction for enterprise LLM deployment.

6. Conclusion

6.1 Summary

This paper presented a systematic comparative evaluation of six PEFT strategies -- full fine-tuning, LoRA, QLoRA, prefix tuning, prompt tuning, and IA3 -- across clinical, legal, financial, and scientific NLP benchmarks on five criteria. QLoRA achieves the best aggregate efficiency-performance profile: within 2.1% of full FT domain performance with 73.4% memory reduction and 61.8% training time reduction. Domain-specific recommendations: LoRA for legal reasoning, prefix tuning for minimum forgetting, QLoRA for all others. Full FT dominates performance but incurs 7.6% catastrophic forgetting.

6.2 Open-Source Release

All six PEFT strategies are implemented in the open-source DomainFT toolkit (available on GitHub), providing a unified codebase for reproducible PEFT evaluation and domain adaptation fine-tuning. The toolkit supports LLaMA-2 and Mistral model families and provides standardised evaluation pipelines for MedQA, LegalBench, FinBench, and SciQ benchmarks. Configuration recipes for all recommended strategy-domain combinations are provided in Appendix A. We encourage the community to extend the toolkit to additional base models, domains, and PEFT strategies within the established evaluation protocol.

References

Almazrouei, E. et al. (2023). The Falcon series of open language models. arXiv:2311.16867.

Brown, T. et al. (2020). Language models are few-shot learners. NeurIPS, 33, 1877-1901.

Dettmers, T. et al. (2023). QLoRA: Efficient finetuning of quantized LLMs. NeurIPS, 36.

Gururangan, S. et al. (2020). Don't stop pretraining. ACL 2020, 8342-8360.

Guha, N. et al. (2023). LegalBench: A collaboratively built benchmark for measuring legal reasoning. NeurIPS 2023 Datasets.

Howard, J., and Ruder, S. (2018). Universal language model fine-tuning for text classification. ACL 2018, 328-339.

Hu, E. J. et al. (2022). LoRA: Low-rank adaptation of large language models. ICLR 2022.

Jiang, A. Q. et al. (2023). Mistral 7B. arXiv:2310.06825.

Jin, D. et al. (2021). What disease does this patient have? A large-scale open domain question answering dataset from medical exams. Applied Sciences, 11(14), 6421.

Kirkpatrick, J. et al. (2017). Overcoming catastrophic forgetting in neural networks. PNAS, 114(13), 3521-3526.

Lee, J. et al. (2020). BioBERT: A pre-trained biomedical language representation model. Bioinformatics, 36(4), 1234-1240.

Lester, B., Al-Rfou, R., and Das, A. (2021). The power of scale for parameter-efficient prompt tuning. EMNLP 2021, 3045-3059.

Li, X. L., and Liang, P. (2021). Prefix-tuning: Optimizing continuous prompts for generation. ACL 2021, 4582-4597.

Liu, H. et al. (2022). Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. NeurIPS, 35.

Luo, Y. et al. (2023). Empirical study of catastrophic forgetting of large language models during continual fine-tuning. arXiv:2308.08747.

McCloskey, M., and Cohen, N. J. (1989). Catastrophic interference in connectionist networks. Psychology of Learning and Motivation, 24, 109-165.

Popescu, L., Horvath, M., and Novak, I. (2024). Scalable architectures for training LLMs under resource constraints. Journal of Generative Intelligence, 2024(1), 1-8.

Touvron, H. et al. (2023). LLaMA 2: Open foundation and fine-tuned chat models. arXiv:2307.09288.

Welbl, J., Liu, N. F., and Gardner, M. (2017). Crowdsourcing multiple choice science questions. EMNLP 2017 Workshop.

Zhang, F. et al. (2023). FinBench: A benchmark for financial NLP. Financial NLP Workshop, EMNLP 2023.

Declarations

Funding

This research was supported by the Spanish State Research Agency (AEI) under grant PID2023-142016OB-I00, the French National Research Agency (ANR) under grant ANR-23-CE23-0031, and the Swedish Research Council under grant 2024-05102. The funders had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

The authors declare no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data Availability Statement

All benchmark datasets used are publicly available. The DomainFT open-source toolkit, fine-tuning configurations, and model evaluation scripts are publicly available on GitHub. Fine-tuned model checkpoints are available upon request.

Ethical Approval

This study involved computational experiments on publicly available benchmark datasets only. No human subjects or personal data were involved. Ethical approval was not required.

ISSN: e-ISSN: 3117-6429 / p-ISSN: 3117-6437 Vol: 2024, Issues: 1, 2024

Appendix A

DomainFT Toolkit Configuration Recipes

The following provides recommended hyperparameter configurations for each strategy-domain combination in the DomainFT toolkit.

QLoRA Configuration (Default Recommendation)

LoRA Configuration (Legal Reasoning Recommended)

Prefix Tuning Configuration (Low-Forgetting Contexts)

Evaluation Protocol for New Domains