Training Dynamics of Large Language Models Under Domain Adaptation in Clinical Domains

Anonymous ACL submission

Abstract

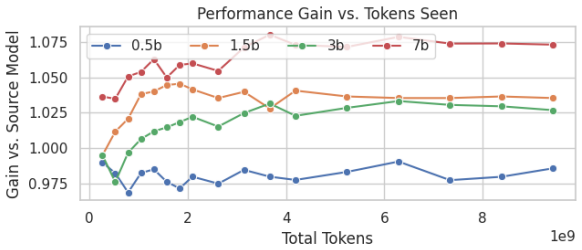
While scaling laws for general-purpose language models are well-documented, the empirical trajectories governing their specialized adaptation to high-stakes clinical environments remain poorly understood. In this study, we systematically characterize the adaptation trajectories of the *Qwen2.5* and *Qwen3* model families within the German medical domain through continuous pre-training and model merging. We evaluate performance through a dual-metric lens, combining objective knowledge benchmarks via multiple-choice question answering with an assessment of generative proficiency through automated preference rankings. Our results reveal a critical asynchronous scaling behavior between factual recall and generative proficiency; while models achieve rapid stylistic alignment within the first 2B to 3B tokens, objective knowledge acquisition scales more gradually and exhibits significantly shallower improvement curves. Furthermore, our results demonstrate that once a sufficient architectural capacity is reached, domain-specific training allows models to bridge the performance gap to significantly larger generalist counterparts, challenging the assumption that raw parameter scale is the primary determinant of domain proficiency. These findings demonstrate that domain proficiency is not a monolithic acquisition process but a series of decoupled trajectories, providing an empirical blueprint for compute-optimal specialization in resource-constrained clinical environments.

1 Introduction

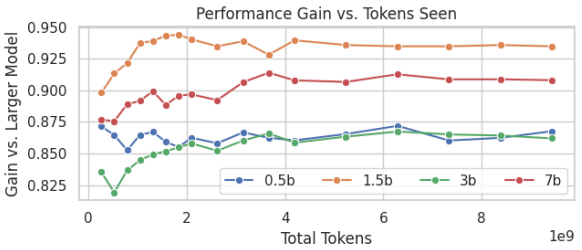
Large Language Models (LLMs) have rapidly become central to innovation in data-driven fields, with healthcare representing a particularly dynamic and challenging area for their deployment (Naveed et al., 2025; Meyer et al., 2024). These models are increasingly leveraged to interpret complex clinical information and support a range of medical tasks, offering the potential to enhance both the efficiency and quality of care. However, translating the broad capabilities of LLMs into reliable, domain-specific performance is a multifaceted challenge. Specialized domains, such as medicine, demand not only expertise of technical terminology and structured documentation, but also robust adaptation to local linguistic and operational contexts. Furthermore, the need for secure, efficient, and compliant AI solutions underscores the importance of developing parameter-efficient architectures that can be tailored to diverse regulatory and hardware environments (Belcak et al., 2025).

To address these limitations, domain adaptation through continuous pre-training (*CPT*) has emerged as a standard methodology for internalizing specialized knowledge (Shi et al., 2025). Yet, current research in this area remains limited by two critical gaps. First, existing studies typically treat domain adaptation as a static, "before-and-after" task, overlooking the nuanced evolution of model performance as a function of computational investment (Chen et al., 2025; Christophe et al., 2024). While recent work has begun to explore the scaling behavior of *CPT*, these investigations have focused almost exclusively on factual knowledge benchmarks. Second, there is a significant evaluation gap: the majority of domain-adaptation literature relies on objective Multiple Choice Question Answering (*MCQA*) benchmarks to measure success (Guo et al., 2025; Singhal et al., 2025). This narrow focus fails to account for generative utility, which refers to the model's ability to produce coherent, instruction-following, and professionally-toned medical prose. This capability is arguably the more critical metric for practical clinical application.

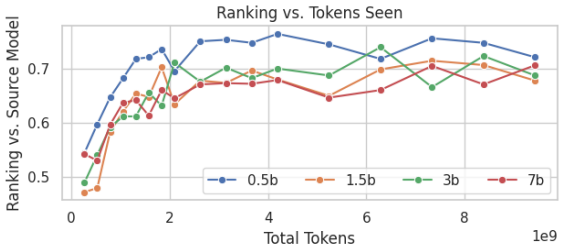
This paper bridges these gaps by providing a systematic, comparative analysis of two state-of-the-art model families, *Qwen2.5* and *Qwen3*, across parameter scales ranging from $0.5B$ to $8B$. By



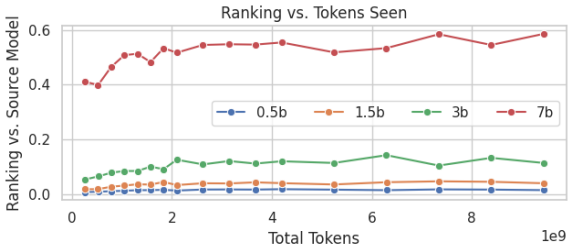
(a) *MCQA*: against source model



(b) *MCQA*: against next larger model



(c) Generative rankings: against source model



(d) Generative rankings: against next larger model

Figure 1: Scaling trajectories for the *Qwen2.5* family across domain-specific token exposure. The top row tracks factual knowledge (*MCQA*) and the bottom row assesses generative utility via Davidson preference rankings. These plots illustrate adaptation dynamics relative to the original source models and relative parity achieved against the next-tier general-purpose variants (e.g., $7B$ vs. $14B$).

tracking the adaptation process across billions of tokens and utilizing weight-merging techniques to preserve conversational alignment, we move beyond factual recall to investigate the compute-performance frontier of generative proficiency. Our work provides a dual-metric analysis: we utilize traditional *MCQA* benchmarks to track knowledge injection and an open-ended "LLM-as-a-judge" framework to approximate real-world usability.

Our findings offer three primary contributions:

1. We investigate the temporal dynamics of factual knowledge injection and generative quality, identifying a rapid initial alignment phase followed by convergence points that are heavily dictated by the base models.

2. We demonstrate that domain-specific compute can bridge the performance gap between scales, identifying a capacity threshold where specialized models outperform generalist counterparts with twice the parameter count.

3. We reveal that factual recall and generative proficiency follow decoupled evolutionary paths, with linguistic alignment occurring faster and more steeply than the gradual injection of factual knowledge.

By investigating these performance trajectories and efficiency trade-offs, this study provides actionable insights for researchers and practitioners aiming to transform resource-efficient models into viable clinical tools without sacrificing their foundational instruction-following strengths.

The remainder of this paper is organized as follows: Section 2 reviews related work in domain adaptation. Sections 3 and 4 describe the methodology, adaptation pipeline, and experimental configuration. Our findings are reported in Section 5 and discussed in Section 6. Finally, Section 7 provides concluding remarks and directions for future research.

2 Related Work

This section surveys related work on adapting large language models to specialized domains.

2.1 Model Scale and Specialization

Recent studies have highlighted the interplay between model scale, domain specialization, and computational constraints. Junior et al. (2025) systematically studied Portuguese *Qwen2.5* models and found that while larger models retain more knowledge, the marginal benefits of domain specialization decrease with size. Under compute-limited settings, smaller specialized models can outperform

general-purpose counterparts, achieving lower perplexity with fewer training steps. Using *MCQA* benchmarks, the authors show that specialized models reach optimal performance more quickly, exhibit higher sample efficiency, and are less prone to catastrophic forgetting than their general counterparts. These findings underscore the potential efficiency gains of targeted domain adaptation, particularly in resource-constrained environments. However, their evaluation is limited to *MCQA* tasks, and the effects across model scales and broader adaptation metrics remain largely unexplored.

Building on this work, our study directly examines how model scale affects domain adaptation dynamics in the medical domain, going beyond single-task *MCQA* benchmarks. We systematically evaluate adaptation trajectories across multiple model sizes, considering both generative proficiency and factual knowledge, revealing divergent learning dynamics and scale-efficiency crossovers that were not captured in prior work.

2.2 Domain Adaptation of LLMs

Continual pre-training (CPT) has emerged as a common approach for adapting general-purpose LLMs to specialized domains. Its effectiveness depends heavily on corpus quality and alignment with the model's prior knowledge. Indiscriminate pre-training can harm performance (Öncel et al., 2024), while targeted CPT combined with careful model merging preserves general reasoning capabilities and enhances domain fluency (Niyogi and Bhattacharya, 2024; Siriwardhana et al., 2024; Yang et al., 2024b). Complementary work emphasizes balancing general and domain-specific knowledge. Colombo et al. (2024) show that incorporating a small proportion of general data during domain adaptation preserves broad capabilities, and Gu et al. (2024) formalize this trade-off via the Critical Mixture Ratio. Together, these studies demonstrate that domain specialization is not solely determined by model scale, but also by careful data selection, training design, and compute allocation.

2.3 Merging Large Language Models

Resource-efficient adaptation strategies, such as model merging, offer a practical alternative to full-scale fine-tuning. Surveys and recent experiments (Yang et al., 2024a; Nobari et al., 2025) demonstrate that model merging can yield substantial performance gains with reduced computational cost. Larger models generally facilitate more successful merging, while different merging strategies perform comparably at scale (Yadav et al., 2024a). In the medical domain, Spherical Linear Interpolation (SLERP) (Shoemake, 1985; Goddard et al., 2024) has been shown to achieve strong empirical results, highlighting the potential of these techniques for creating high-performing, compute-efficient LLMs.

3 Methodology

This paper systematically investigates the relationship between computational investment and the efficacy of domain adaptation in LLMs. We continuously track model performance throughout the training process, rather than relying on single-point evaluations.

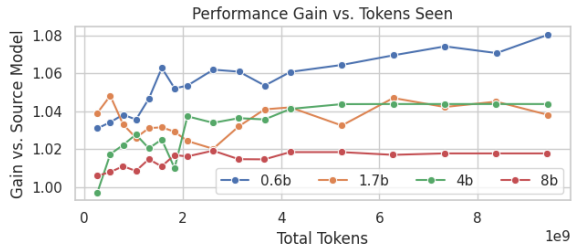
3.1 Domain Adaptation Pipeline

The adaptation methodology follows a structured, multi-stage pipeline proposed by Lu et al. (2025). This pipeline is designed to infuse specialized domain knowledge while preserving the instruction-following capabilities inherent in the base models. The process begins with a foundation model that has already undergone general-purpose instruction tuning, which serves as the base for all subsequent domain-specific refinement.

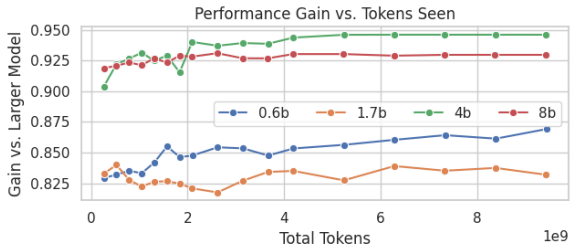
The first phase of the pipeline involves CPT on a curated, domain-specific corpus. During this stage, the model is exposed to raw, unlabeled text using a standard causal language modeling objective. To prevent catastrophic forgetting and preserve instruction-following capabilities, we merge the continuously pretrained model with the original base. Following related literature (Lu et al., 2025; Labrak et al., 2024), we use Spherical Linear Interpolation (SLERP), which accounts for the high-dimensional spherical geometry of the weight space.

3.2 Evaluation

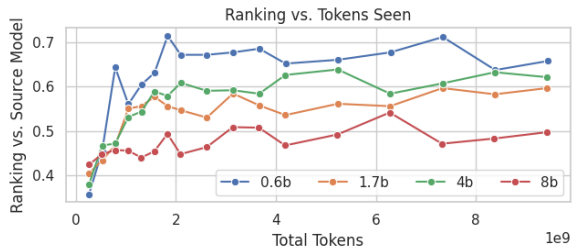
We evaluate multiple checkpoints of the adapted models along two complementary axes: objective knowledge, assessed through Knowledge-Based Benchmarks, and generative proficiency, assessed via a preference-ranking approach. This dual evaluation captures not only the acquisition of medical facts but also the model's ability to apply that knowledge coherently and in an instruction-compliant manner.



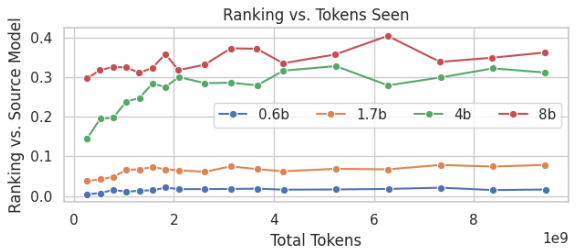
(a) *MCQA*: against source model



(b) *MCQA*: against next larger model



(c) Generative rankings: against source model



(d) Generative rankings: against next larger model

Figure 2: Scaling trajectories for the *Qwen3* family across domain-specific token exposure, analogous to Figure 1. The top row tracks factual knowledge while the bottom row assesses generative utility, comparing adapted variants against their original source models and the next-tier general-purpose baselines (e.g., $8B$ vs. $14B$).

3.2.1 Knowledge-Based Benchmarks

Objective performance is measured through *MCQA* using standardized benchmarks relevant to the target domain. These benchmarks provide a stable and reproducible metric for assessing the model's internal "world knowledge" independent of generative formatting or linguistic style. We report one-shot accuracy across all saved checkpoints to observe how factual accuracy scales with training compute.

To isolate the impact of continuous pre-training, we focus on the Relative Performance Gain. This metric is defined as the ratio of the adapted checkpoint's performance $Acc(C_i)$ to that of the selected baseline model $Acc(M_{bl})$, typically the original instruction-tuned source model:

$$G_{rel}(C_i) = \frac{Acc(C_i)}{Acc(M_{bl})} \quad (1)$$

3.2.2 Preference Ranking

To assess the practical application of domain knowledge we follow existing literature by evaluating generative proficiency using a "LLM-as-a-Judge" approach (Dubois et al., 2025; Zheng et al., 2023; Li et al., 2024; Sun et al., 2024; Zhang et al., 2024). For each domain-specific prompt, both a selected training checkpoint and a baseline model, usually the non-adapted instruction-tuned source model, generate free-text responses. A high-capacity judge model then evaluates these responses in a head-to-head format, assigning a win, loss, or tie to each response pair.

The resulting pairwise comparison data is then aggregated into global rankings using the *Davidson* model (Davidson, 1970), which serves as an extension of the *Bradley-Terry* model (Bradley and Terry, 1952) to account for ties in preference. This statistical approach allows us to estimate the latent strength parameters for each checkpoint on a linear scale, providing a normalized measure of generative performance across the training trajectory.

4 Experimental Setup

This section describes the technical implementation of our domain adaptation pipeline.

4.1 Domain-Specific Corpus

The continuous pre-training phase utilizes a specialized subset of the *FineWeb-2* dataset (Penedo et al., 2025). To ensure high domain relevance, we employ a filtered version of the corpus specifically tailored for the german medical domain. The filtering process utilizes a hybrid classification pipeline that combines classification models with *LLM* reasoning to extract medical discourse (see Appendix C for further details). The filtered dataset consists of $7.3M$ unique documents, which amount to $9.5B$ tokens with respect to the *Qwen2.5* tokenizer.

4.2 Model Selection

To investigate the impact of compute across different architectural iterations and scales, we select two model families from the *Qwen* series spanning $0.5B$ to $8B$ parameters, to investigate scaling trends in resource constrained environments. All selected models are instruction-tuned versions of the base weights, serving as the starting point for our adaptation-and-merge pipeline.

4.3 Evaluation Setup

Our objective evaluation focuses on the German medical domain. We utilize a subset of the MMLU (Massive Multitask Language Understanding) benchmarks (Hendrycks et al., 2021). Specifically, we select subtasks that demonstrate the highest correctness ratios for the medical field as reported by Gema et al. (2025), ensuring the evaluation is grounded in high-signal data. These include *Clinical Knowledge*, *Anatomy* and *College Medicine*. Additionally, we include a machine-translated German version of the *MedQA* benchmark (Jin et al., 2021). We use the Eval-Harness framework (Gao et al., 2024) for all *MCQA* evaluations.

For open-ended evaluation, we utilize a translated version of the *MedInstruct* dataset (Zhang et al., 2023) as the source of prompts, consisting of 217 entries. Responses are generated using *vLLM* (Kwon et al., 2023) and evaluated using a head-to-head comparison against the selected baseline. We employ *GPT-4.1-mini* as our Judge model. The exact prompt template used by the Judge model is presented in Figure 4 in the Appendix.

5 Findings

The following sections analyze the evolution of model proficiency across the domain adaptation process. All reported metrics correspond to the merged model checkpoints generated via the protocol outlined in Section 3.1.

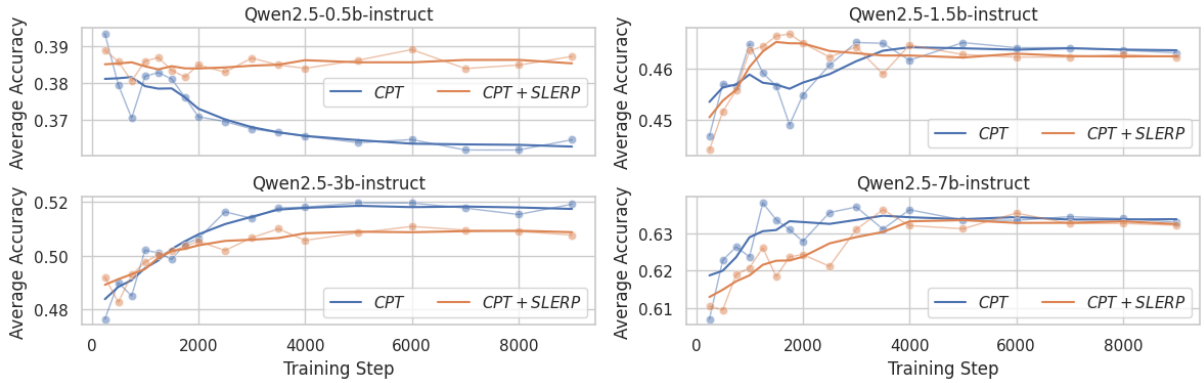
5.1 Knowledge Acquisition in the Medical Domain

The impact of continuous pre-training on factual recall follows consistent scaling patterns across both *Qwen* families. (Figures 3a, 3b in the Appendix). While absolute accuracy remains primarily a function of model scale, domain adaptation can narrow the knowledge gap. Adapted smaller models consistently approach the performance of larger, non-adapted counterparts, indicating that domain-specific compute can partially compensate for raw parameter count.

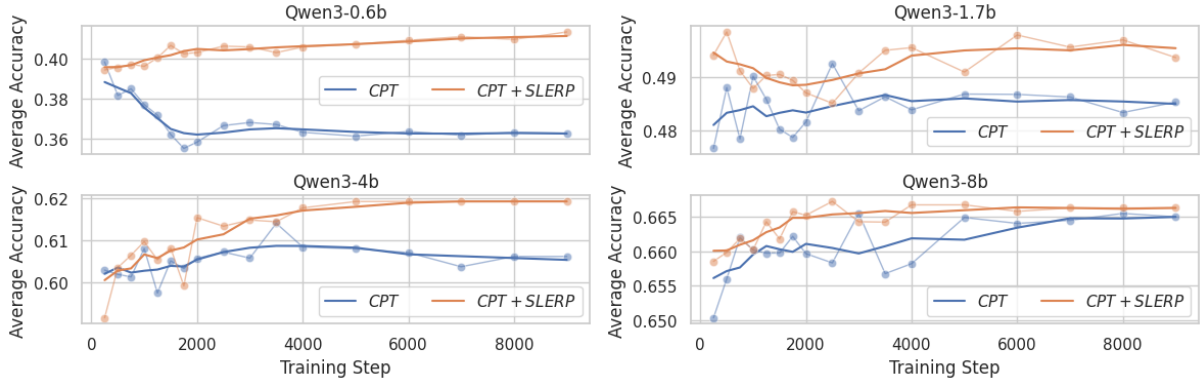
As illustrated in Figures 1a and 2a, models achieve the most significant relative gains within the first $4B$ tokens, followed by a convergence plateau. However, a more nuanced picture emerges when these trajectories are normalized by cumulative compute (Appendix Figures 5a and 6a). This visualization reveals a divergence between token-wise and compute-wise efficiency. While models across all scales exhibit similar relative gains per seen token, smaller models demonstrate superior compute efficiency by reaching their respective performance ceilings at a fraction of the total compute investment. In contrast, larger architectures require a significantly higher computational budget to achieve comparable relative improvements, though they ultimately remain necessary for extending the absolute scaling frontier. The most notable discrepancy between the two families involves the largest parameter tiers in each family. The *Qwen2.5*-$7B$ model shows substantial relative gains, supporting the idea that a certain level of architectural capacity is required to successfully internalize and leverage specialized medical compute. Conversely, the *Qwen3* $8B$ variant exhibits a conservative improvement of only 2%, being notably outperformed in relative gain by its own $0.6B$ counterpart. We hypothesize that this stems from pre-training saturation, specifically, given the release timelines, the *Qwen3* models were likely exposed to the *FineWeb-2* corpus during their initial base training. This alignment suggests a "catastrophic overtraining" effect (Springer et al., 2025), where models trained significantly beyond their compute-optimal point on a specific distribution develop a rigidity that increases resistance to further fine-tuning on similar data distributions.

5.2 Generative Strength and Preference Ranking

The open-ended generative evaluation reveals trends that diverge from objective *MCQA* findings, highlighting a clear distinction between linguistic alignment and factual recall. Across both families, continuous pre-training yields a robust increase in generative proficiency. As shown in the preference ranking trajectories in Figures 1c and 2c, all models exhibit a rapid ascent against their respective source models, typically converging after only $2B$ tokens.



(a) Qwen2.5 Model Family



(b) Qwen3 Model Family

Figure 3: Absolute performance trajectories on medical *MCQA* benchmarks. Each plot illustrates the accuracy progression for the raw continued pre-trained weights (CPT) and the resulting model after weight-merging with the source model ($CPT + SLERP$).

While token-wise scaling is uniform, the families diverge in their convergence thresholds. The *Qwen2.5* family shows stronger overall gains, converging between ranking scores of 0.65 and 0.70. In contrast, *Qwen3* models plateau lower, at scores between 0.55 and 0.65, with the exception of the $8B$ variant showing virtually no improvement, stagnating near 0.5. This disparity reinforces the pre-training saturation hypothesis mentioned above.

A notable finding is the disproportionate gain in the smallest parameter tiers. Despite limited capacity, the $0.5B$ and $0.6B$ variants are significant outliers, demonstrating the most dramatic increases in ranking scores. However, the link between stylistic and factual adaptation varies. In the *Qwen2.5* family, the $0.5B$ model's ranking improved despite stagnant *MCQA* scores, whereas the *Qwen3* $0.6B$ model showed holistic improvement across both metrics. This suggests that while small models consistently benefit from stylistic alignment, factual internalization remains highly sensitive to the base model's specific architecture or initial training state.

The "Win-Tie-Loss" distributions visualized in Figures 7 and 8 in the Appendix provide further mechanical insight. For the smallest variants, "Ties" dominate, suggesting the judge model often finds the adapted and source models equally proficient. We attribute this to the constraints of weight interpolation in low-parameter regimes. The limited weight space often results in merged checkpoints that remain structurally similar to the baseline (Yadav et al., 2024b; Lu et al., 2025). As model size increases, the "Tie" category shrinks, suggesting that larger architectures possess the capacity to develop more distinct, high-quality features that allow a judge to clearly differentiate the adapted model from its initialization.

5.3 Cross-Scale Performance Gap

To evaluate the practical utility of domain-specific compute, we analyze how effectively *CPT* allows smaller models to bridge the performance gap with their larger, non-adapted counterparts.

As illustrated in Figures 1b and 2b, domain adaptation consistently narrows the knowledge gap across both model families. We observe a clear capacity-scaling effect: the relative performance gap is significantly easier to close as the base parameter count increases. This trend is especially clear in the *Qwen3* family. While the $0.6B$ and $1.7B$ models achieve approximately 85% of the accuracy of their respective larger counterparts, the $4B$ and $7B$ models reach upwards of 92.5%. The *Qwen2.5* $1.5B$ model serves as a significant exception, achieving over 93% of the $3B$ variant's performance, the highest relative parity observed in the *Qwen2.5* series.

The preference rankings against the respective larger models in Figures 1d and 2d reveal how generative utility evolves differently depending on the model's starting scale.

In the lower parameter regimes, the general-purpose larger models consistently maintain a lead over the adapted smaller variants. However, the performance increase provided by domain adaptation is nonetheless significant. For example, the *Qwen2.5* $3B$ model's ranking against the $7B$ version improved nearly three-fold, rising from 0.055 to 0.142. Despite this trajectory, the wide margin maintained by the larger models indicates that smaller architectures, *CPT* serves primarily to narrow the proficiency gap rather than to achieve actual performance parity. In these instances, the specialized training acts as a powerful optimization of limited capacity, yet it does not fully compensate for the inherent linguistic advantages of the larger generalist model.

As model size increases, the gap closes drastically, mirroring the observations in the *MCQA* benchmarks. In these regimes, domain adaptation transforms the models into highly competitive alternatives to their larger variations, achieving ranking scores upward of 0.3 - 0.4. Interestingly, the the *Qwen3* $4B$ version is significantly more competitive against the $8B$ version than its *Qwen2.5* equivalent, increasing its score from 0.217 to 0.328. The most striking result is found in the *Qwen2.5* $7B$ model, which demonstrates the first instance of scale-surpassing performance. By increasing its preference score from 0.371 to 0.586 through adaptation, the $7B$ variant successfully outperforms the $14B$ generalist model in the medical domain. This indicates that at a certain capacity threshold, domain-specific compute can fully compensate for the advantages of raw parameter scale.

6 Discussion

The central objective of this study was to investigate the scaling behaviors and efficiency trade-offs inherent in domain adaptation. Our results reveal a complex dynamic where objective knowledge acquisition and generative utility are not only decoupled in terms of performance but also in their temporal acquisition rates.

6.1 Scaling Behaviors

A primary finding of this work is that factual recall and generative proficiency do not scale in tandem. This disconnect manifests in two distinct ways:

As seen in the *Qwen2.5* $0.5B$ model, a high generative preference ranking can coexist with stagnant or degrading *MCQA* performance. This suggests that the model successfully internalizes the linguistic patterns and terminology of the German medical domain without a corresponding increase in underlying factual density. This presents a critical risk: a model that adopts an authoritative clinical register but lacks factual grounding is potentially more dangerous than one that openly fails. Investigating the nuances of this risk would provide deeper insights into the mechanics of domain adaptation, yet it necessitates a more granular semantic analysis of model responses and is thus deferred to future work.

Crucially, these metrics also exhibit different evolutionary behaviors across the training trajectory. Generative rankings converge rapidly, often peaking within the first $2B$ to $3B$ tokens with a steep initial trajectory. In contrast, *MCQA* performance converges more slowly, typically requiring approximately $4B$ tokens to stabilize. Furthermore, the knowledge benchmark trajectories are notably shallower. The improvements are overall less pronounced than the ranking gains, indicating that while the model adopts the domain's vocabulary almost immediately, the internalization of clinical facts is a more resistant and gradual process.

This discrepancy in saturation points has profound implications for compute-optimal deployment. Our findings suggest that, if the primary goal of adaptation is stylistic alignment, extended training beyond the initial $2B$ token window yields diminishing returns. Conversely, tasks requiring factual precision demand longer training durations, albeit with shallower improvement curves compared to the rapid initial gains in preference ranking.

Furthermore, the selection of the source model

is as critical as the adaptation volume. The pre-training saturation observed in the *Qwen3* $8B$ model demonstrates that if a model's base training already encompasses a specific distribution, additional *CPT* compute is poorly spent. In such cases, resources are more effectively diverted toward supervised instruction-tuning rather than continued pre-training on raw corpora.

6.2 Scale-Efficiency Crossovers

A pivotal finding of this study is the identification of a capacity threshold where domain-specific compute effectively compensates for the advantages of raw parameter count. Our results indicate that while sub-$3B$ models appear to lack the latent architectural capacity to fully bridge the performance gap to their larger counterparts, mid-to-large-scale architectures exhibit a significant scale-efficiency crossover. The most striking example of this is the adapted *Qwen2.5* $7B$ model, which successfully outperformed the general-purpose $14B$ variant in generative utility. This suggests that the traditional correlation between model scale and performance is not absolute, rather, it is a frontier that can be shifted through targeted continuous pre-training.

This observation implies that for high-stakes domains like medicine, there is a critical degree of architectural depth required to provide the conceptual scaffolding necessary for integrating specialized knowledge into a model's existing reasoning framework. In lower parameter regimes, the models lack the structural complexity to internalize high-density domain information without sacrificing generalist instruction-following capabilities. However, once this threshold is reached, the model gains the ability to leverage domain-specific compute to challenge the practical utility of a larger generalist models. For institutional deployments where hardware constraints or inference costs are a primary concern, these results suggest that investing in the adaptation of a mid-sized model is a more viable strategy than simply scaling up to the next parameter tier.

7 Conclusion

This study provides a systematic characterization of the scaling behaviors and efficiency trade-offs inherent in the domain adaptation of LLMs for the German medical sector. By analyzing the evolution of the *Qwen2.5* and *Qwen3* families, we show that domain-specific refinement follows a non-linear trajectory, with a rapid initial alignment phase followed by convergence plateaus. Notably, we identify a scale-efficiency crossover in mid-to-high parameter regimes, where targeted continuous pre-training enables models to challenge the generative utility of general-purpose architectures with twice the parameter count, suggesting that strategic investment in domain-specific compute can be more resource-efficient than simply increasing model size. Our dual-metric evaluation further reveals a critical divergence between linguistic alignment and factual recall. While stylistic proficiency emerges rapidly, objective knowledge acquisition progresses more slowly and exhibits shallower improvement curves, highlighting a reliability risk for smaller architectures. This disconnect underscores the necessity of multi-dimensional evaluation, as relying solely on generative fluency may mask persistent factual deficits. Together, these findings offer insights into compute-optimal specialization, providing guidance for deploying high-performance, resource-constrained LLMs in clinical environments.

Future research should expand upon these findings in three key directions. First, defining a framework for a more granular semantic analysis of model responses to better characterize the mechanics of the domain adaptation process. Specifically, this inquiry investigate the risk of overconfidence identified in the discussion. Second, evaluating more sophisticated adaptation pipelines designed to mitigate such risks and improve overall performance. This includes the integration of alignment techniques, as well instruction-tuning as an alternative to model merging. Finally, investigating optimal data mixture strategies to address the observed pre-training saturation. Determining the precise ratio of general-purpose to domain-specific tokens is essential to mitigating catastrophic overtraining while ensuring that specialized models preserve the robust reasoning capabilities of their base architectures.

Limitations

This study is subject to several constraints that define the scope of its conclusions. Primarily, the exclusive focus on the German medical domain and the evaluation of the *Qwen* model families up to the $8B$ parameter tier leaves the scaling dynamics of larger architectures and alternative domains unexplored. The potential for data contami-

nation represents a significant confounding variable, specifically, a likely overlap between our specialized medical corpus and the data used in *Qwen3*'s base training may mask the distinction between genuine domain-specific acquisition and the latent reactivation of previously seen tokens. Furthermore, our reliance on *MCQA* and preference rankings provides only a proxy for clinical proficiency. While these metrics track factual density and stylistic alignment, they still do not fully capture the nuanced safety dimensions required for real-world deployment, such as multi-step diagnostic reasoning or the long-term utility of the model in a clinical workflow. Finally, the high frequency of "Ties" observed in low-parameter evaluations points to a resolution limit in the automated judge model. This suggests that the current evaluation framework may lack the sensitivity required to distinguish subtle performance gains in smaller architectures.

References

Peter Belcak, Greg Heinrich, Shizhe Diao, Yonggan Fu, Xin Dong, Saurav Muralidharan, Yingyan Celine Lin, and Pavlo Molchanov. 2025. Small language models are the future of agentic ai. *Preprint*, arXiv:2506.02153.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Jie Chen, Zhipeng Chen, Jiapeng Wang, Kun Zhou, Yutao Zhu, Jinhao Jiang, Yingqian Min, Wayne Xin Zhao, Zhicheng Dou, Jiaxin Mao, and 1 others. 2025. Towards effective and efficient continual pre-training of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5779–5795.

Clément Christophe, Tathagata Raha, Svetlana Maslenkova, Muhammad Umar Salman, Praveen K Kanithi, Marco AF Pimentel, and Shadab Khan. 2024. Beyond fine-tuning: Unleashing the potential of continuous pretraining for clinical llms. *arXiv preprint arXiv:2409.14988*.

Pierre Colombo, Telmo Pessoa Pires, Malik Boudiaf, Dominic Culver, Rui Melo, Caio Corro, Andre F. T. Martins, Fabrizio Esposito, Vera Lúcia Raposo, Sofia Morgado, and Michael Desa. 2024. Saullm-7b: A pioneering large language model for law. *Preprint*, arXiv:2403.03883.

Roger R Davidson. 1970. On extending the bradley-terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association*, 65(329):317–328.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. 2025. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *Preprint*, arXiv:2404.04475.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. A framework for few-shot language model evaluation.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile Van Krieken, and Pasquale Minervini. 2025. Are we done with MMLU? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5069–5096, Albuquerque, New Mexico. Association for Computational Linguistics.

Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vladimir Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. 2024. Arcee's MergeKit: A toolkit for merging large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 477–485, Miami, Florida, US. Association for Computational Linguistics.

Jiawei Gu, Zacc Yang, Chuanghao Ding, Rui Zhao, and Fei Tan. 2024. CMR scaling law: Predicting critical mixture ratios for continual pre-training of language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16143–16162, Miami, Florida, USA. Association for Computational Linguistics.

Yiduo Guo, Jie Fu, Huishuai Zhang, and Dongyan Zhao. 2025. Efficient domain continual pretraining by mitigating the stability gap. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32850–32870.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las

Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14).

Roseval Junior, Ramon Pires, Thales Almeida, Kenzo Sakiyama, Roseli Romero, and Rodrigo Nogueira. 2025. The interplay between domain specialization and model size: a case study in the legal domain.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. *Preprint*, arXiv:2402.10373.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Wei Lu, Rachel K. Luu, and Markus J. Buehler. 2025. Fine-tuning large language models for domain adaptation: exploration of training strategies, scaling, model merging and synergistic capabilities. *npj Computational Materials*, 11(1):84.

Annika Meyer, Janik Riese, and Thomas Streichert. 2024. Comparison of the performance of gpt-3.5 and gpt-4 with that of medical students on the written german medical licensing examination: Observational study. *JMIR Med Educ*, 10:e50965.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2025. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–72.

Mitodru Niyogi and Arnab Bhattacharya. 2024. Paramanu-ayn: Pretrain from scratch or continual pretraining of llms for legal domain adaptation? *Preprint*, arXiv:2403.13681.

Amin Heyrani Nobari, Kaveh Alimohammadi, Ali ArjomandBigdeli, Akash Srivastava, Faez Ahmed, and Navid Azizan. 2025. Activation-informed merging of large language models. *Preprint*, arXiv:2502.02421.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. Fineweb2: One pipeline to scale them all–adapting pre-training data processing to every language. *arXiv preprint arXiv:2506.20920*.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. 2025. Continual learning of large language models: A comprehensive survey. *ACM Computing Surveys*, 58(5):1–42.

Ken Shoemake. 1985. Animating rotation with quaternion curves. In *Proceedings of the 12th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '85, page 245–254, New York, NY, USA. Association for Computing Machinery.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, and 1 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950.

Shamane Siriwardhana, Mark McQuade, Thomas Gauthier, Lucas Atkins, Fernando Fernandes Neto, Luke Meyers, Anneketh Vij, Tyler Odenthal, Charles Goddard, Mary MacCarthy, and Jacob Solawetz. 2024. Domain adaptation of llama3-70b-instruct through continual pre-training and model merging: A comprehensive evaluation. *Preprint*, arXiv:2406.14971.

Jacob Mitchell Springer, Sachin Goyal, Kaiyue Wen, Tanishq Kumar, Xiang Yue, Sadhika Malladi, Graham Neubig, and Aditi Raghunathan. 2025. Overtrained language models are harder to fine-tune. *arXiv preprint arXiv:2503.19206*.

Hao Sun, Yunyi Shen, and Jean-Francois Ton. 2024. Rethinking bradley-terry models in preference-based reward modeling: Foundations, theory, and alternatives. *arXiv preprint arXiv:2411.04991*.

Prateek Yadav, Tu Vu, Jonathan Lai, Alexandra Chronopoulou, Manaal Faruqui, Mohit Bansal, and Tsendsuren Munkhdalai. 2024a. What matters for model merging at scale? *Preprint*, arXiv:2410.03617.

Prateek Yadav, Tu Vu, Jonathan Lai, Alexandra Chronopoulou, Manaal Faruqui, Mohit Bansal, and Tsendsuren Munkhdalai. 2024b. What matters for model merging at scale? *Preprint*, arXiv:2410.03617.

Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. 2024a. Model merging in LLMs, MLLMs, and Beyond: Methods, Theories, Applications and Opportunities. *Preprint*, arXiv:2408.07666.

Guoxing Yang, Xiaohong Liu, Jianyu Shi, Zan Wang, and Guangyu Wang. 2024b. Tcm-gpt: Efficient pre-training of large language models for domain adaptation in traditional chinese medicine. *Computer Methods and Programs in Biomedicine Update*, 6:100158.

Xinlu Zhang, Chenxin Tian, Xianjun Yang, Lichang Chen, Zekun Li, and Linda Ruth Petzold. 2023. Alpacare: Instruction-tuned large language models for medical application. *arXiv preprint arXiv:2310.14558*.

Yifan Zhang, Ge Zhang, Yue Wu, Kangping Xu, and Quanquan Gu. 2024. Beyond bradley-terry models: A general preference model for language model alignment. *arXiv preprint arXiv:2410.02197*.

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, Alban Desmaison, Can Balioglu, Pritam Damania, Bernard Nguyen, Geeta Chauhan, Yuchen Hao, Ajit Mathews, and Shen Li. 2023. Pytorch fsdp: Experiences on scaling fully sharded data parallel. *Preprint*, arXiv:2304.11277.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Furkan Öncel, Matthias Bethge, Betul Ermis, Mirco Ravanelli, Cem Subakan, and Çağlar Yıldız. 2024. Adaptation odyssey in llms: Why does additional pretraining sometimes fail to improve? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 19834–19843.

A Training Configuration

The domain adaptation process is implemented using the *Axolotl* framework [Ref: Axolotl Team, 2023]. For the larger model variants, we utilize *FSDP2* (Fully Sharded Data Parallel) (Zhao et al., 2023) to optimize memory distribution across the compute cluster. All models are trained for exactly one epoch using an *AdamW* optimizer [Ref: Loshchilov & Hutter, 2017] and a cosine learning rate schedule. To analyze the evolution of the adaptation process and map the performance trajectory, we save 18 intermediate checkpoints per experiment, distributed at logarithmic intervals throughout the training epoch.

To maintain consistency in our scaling analysis, we apply a uniform hyperparameter configuration across all experiments, as detailed in Table 1.

Hyperparameter	**Value**
Effective Batch Size	128
Learning Rate	2×10^{-5}
Weight Decay	0.1
Warmup Ratio	0.1

Table 1: Core training hyperparameters for domain adaptation across all model scales.

B Computational Resources

The experiments conducted as part of this study involved a significant allocation of high-performance computing resources. The 18 checkpoints archived for each model scale resulted in a total storage footprint of 2.2 TB of model weights. To quantify the computational investment, we calculate the total Floating Point Operations (FLOPs) and cumulative GPU hours for each training run. Table 2 provides a granular breakdown of these metrics, allowing for a direct comparison of the resources required to achieve specific performance thresholds across different parameter counts.

C Medical Document Filtering

This section details the creation of the medical corpus used for domain adaptation. To address the challenges of domain-specific dataset curation at scale, we combine the high-level reasoning of LLMs with the computational efficiency of classical transformer-based classifiers.

The source data is derived from the German subset of *FineWeb-2* (Penedo et al., 2025), a 640GB

web-crawled corpus encompassing diverse content from forum posts to news articles. To extract medical discourse from this massive collection, we first constructed a supervised dataset by sampling $260k$ documents. These were annotated using *Mixtral-8x7B-Instruct-v0.1* (Jiang et al., 2024) in a zero-shot prompting configuration. Human validation of 100 random samples by three annotators confirmed the high quality of these LLM-generated labels, yielding an F1 score of $91.12.5$. The resulting labeled set, served as the foundational training data for our scalable filtering pipeline.

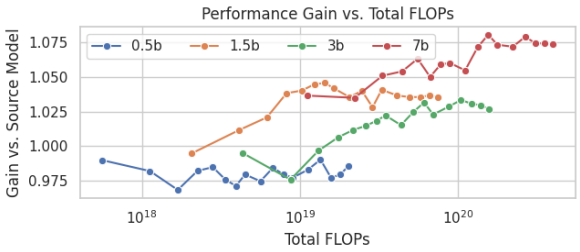
To efficiently process the entire German *FineWeb-2* corpus, we fine-tuned a $279M$ parameter *XLM-RoBERTa-base* model (**?**) as a specialized medical document classifier. This efficient encoder-based model achieved a precision of 0.95 and a recall of 0.80 on our test set. The full-scale classification task was completed in approximately 400 GPU hours using 8 A100 (40GB) GPUs. The final *FineMed-de* corpus comprises approximately $7.3M$ unique documents, totaling $9.5B$ tokens (relative to the *Qwen2.5* tokenizer), providing a high-density factual foundation for continuous medical pre-training.

Model Family	Parameters	Training FLOPs	GPU Hours	GPU Model
Qwen 2.5	$0.5B$	$2.03 \cdot 10^{19}$	$3.63h \times 8$	NVIDIA H100 (80GB)
Qwen 2.5	$1.5B$	$7.42 \cdot 10^{19}$	$8.40h \times 8$	NVIDIA H100 (80GB)
Qwen 2.5	$3B$	$1.57 \cdot 10^{20}$	$15.71h \times 8$	NVIDIA H100 (80GB)
Qwen 2.5	$7B$	$4.00 \cdot 10^{20}$	$31.35h \times 8$	NVIDIA H100 (80GB)
Qwen 3	$0.6B$	$2.49 \cdot 10^{19}$	$27.31h \times 8$	AMD MI250X
Qwen 3	$1.7B$	$7.98 \cdot 10^{19}$	$17.65h \times 32$	AMD MI250X
Qwen 3	$4B$	$2.06 \cdot 10^{20}$	$33.29h \times 32$	AMD MI250X
Qwen 3	$8B$	$4.29 \cdot 10^{20}$	$30.63h \times 64$	AMD MI250X

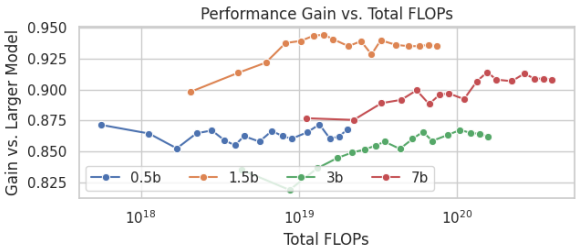
Table 2: Computational expenditure and hardware utilization for domain adaptation. The training load is quantified via total Floating Point Operations (FLOPs) and temporal resource usage. GPU hours are reported as wall-clock time $\times$ number of accelerators to reflect the parallel scale of each training run.

````
messages:
  - role: system
    content: |
      You are an AI assistant that evaluates pairs of responses
      to a given query. Your goal is to determine which response
      is better based on correctness, clarity, completeness,
      and relevance.

  - role: user
    content: |
      Compare the following two responses. Your answer needs
      to follow this JSON schema:

      ```json
      {
        "type": "object",
        "properties": {
          "winner": {
            "type": "string",
            "enum": ["A", "B", "0"],
            "description": "The winning response, i.e. 'A', 'B' or '0' indicating a tie."
          }
        },
      }
      ```

      {% if context %}
      Context:
      {% for msg in context %}
        {{ msg.role | capitalize }}:
          {{ msg.content | indent(width=8) }}
      {% endfor %}
      {% endif %}

      Query:
        {{ query | indent(width=8) }}

      Response A:
        {{ response_a | indent(width=8) }}

      Response B:
        {{ response_b | indent(width=8) }}

      Based on the query and context above,
      which response is better?
````

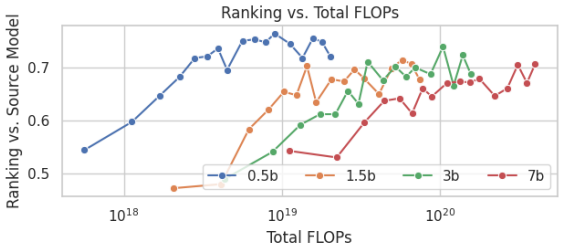
Figure 4: Prompt template for the pairwise Judge model. The structure facilitates head-to-head comparisons by providing the evaluator with query context and two candidate responses, requiring a structured JSON output to determine the winner based on clinical utility.



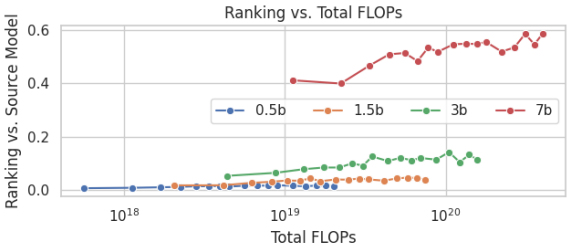
(a) MCQA: against source model



(b) MCQA: against next larger model

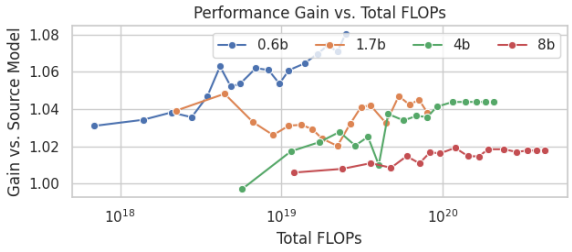


(c) Generative rankings: against source model

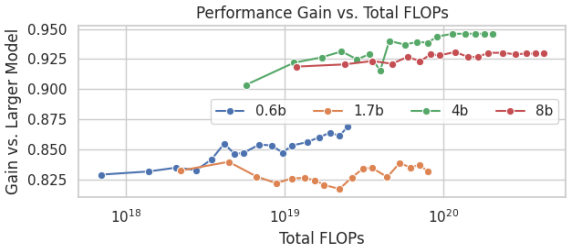


(d) Generative rankings: against next larger model

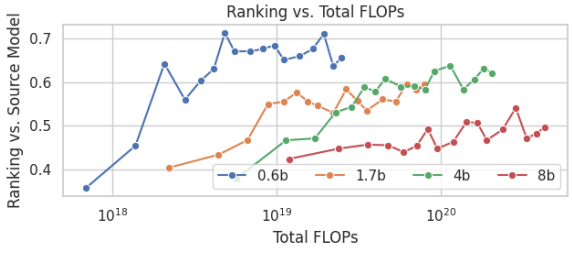
Figure 5: Scaling trajectories for the *Qwen2.5* family as a function of total compute. The top row tracks factual knowledge (*MCQA*) and the bottom row assesses generative utility via Davidson preference rankings. These plots illustrate adaptation dynamics relative to the original source models and relative parity achieved against the next-tier general-purpose variants (e.g., $7B$ vs. $14B$).



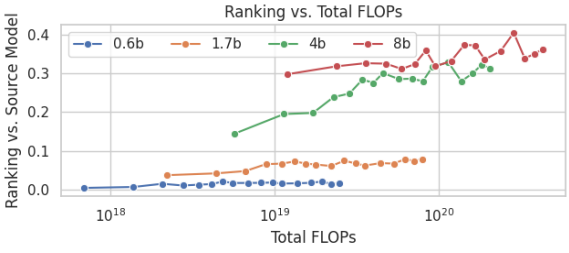
(a) MCQA: against source model



(b) MCQA: against next larger model



(c) Generative rankings: against source model



(d) Generative rankings: against next larger model

Figure 6: Scaling trajectories for the *Qwen3* family across domain-specific token exposure, analogous to Figure 5. The top row tracks factual knowledge while the bottom row assesses generative utility, comparing adapted variants against their original source models and the next-tier general-purpose baselines (e.g., $8B$ vs. $14B$).

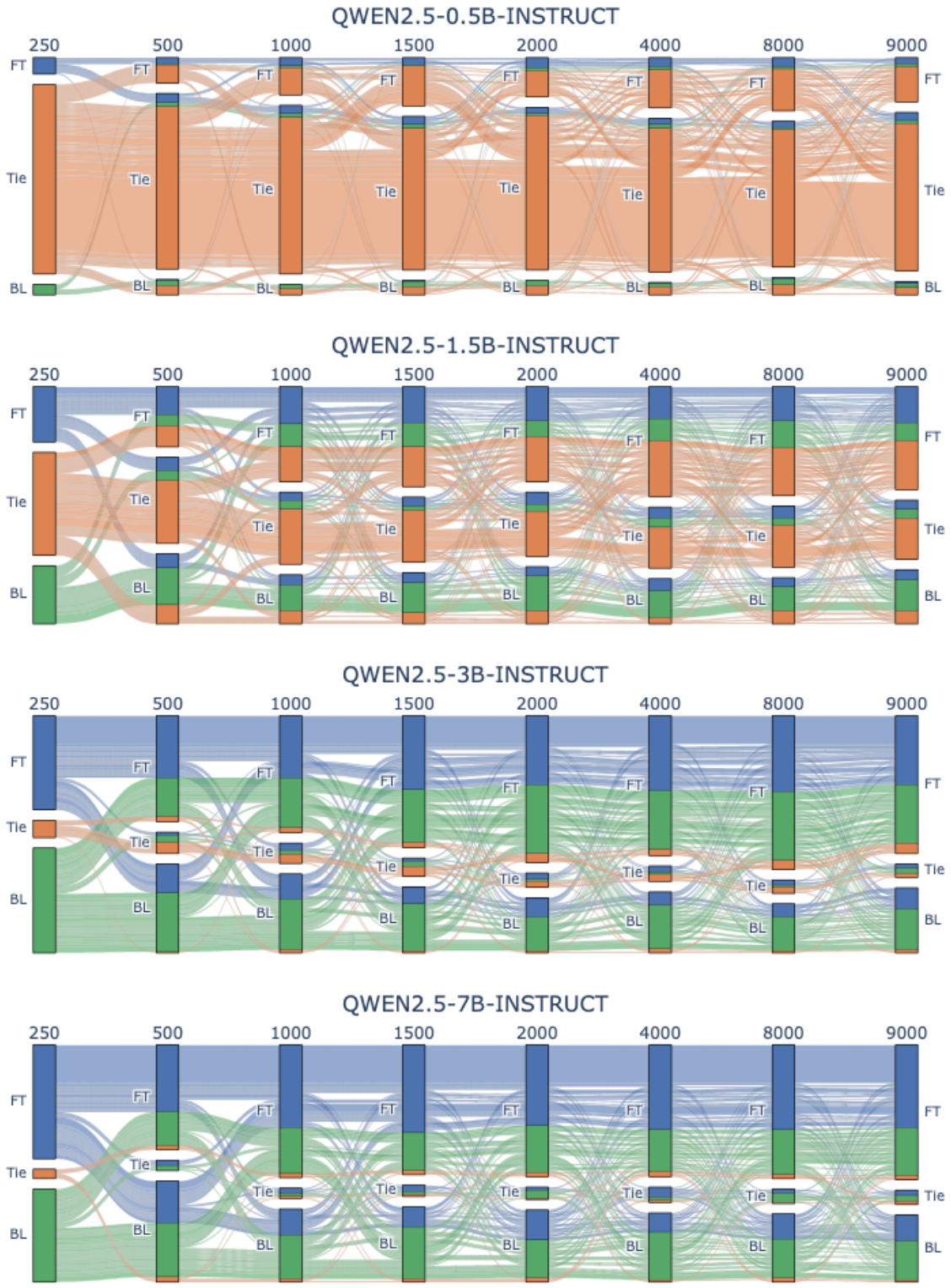


Figure 7: Parallel categories visualization of "Win-Tie-Loss" distributions for the *Qwen2.5* family across different parameter scales. The plots track the flow of judge preferences between the source models (BL) and their domain-adapted counterparts (FT).

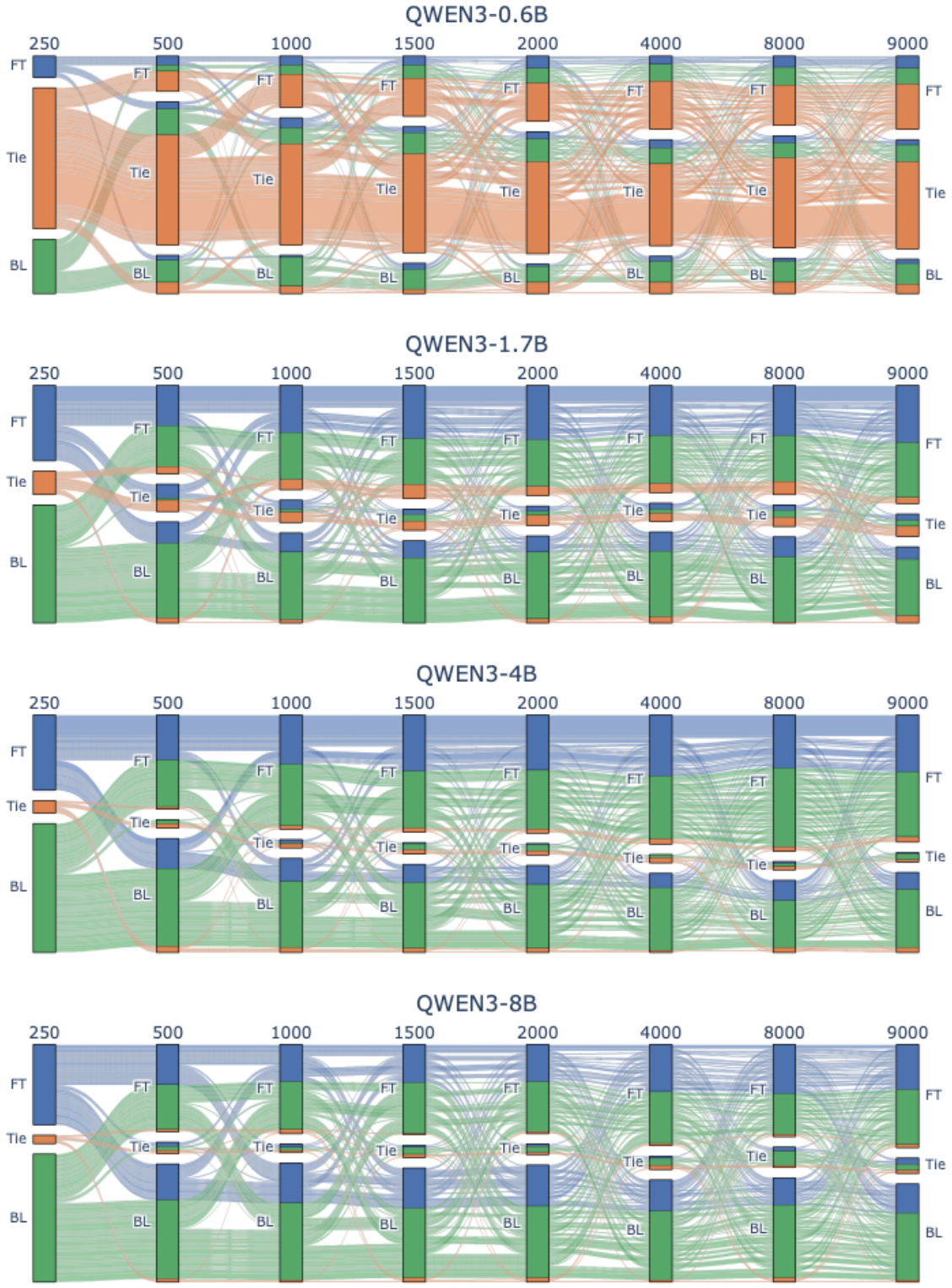


Figure 8: Parallel categories visualization of "Win-Tie-Loss" distributions for the *Qwen3* family across different parameter scales. The plots track the flow of judge preferences between the source models (BL) and their domain-adapted counterparts (FT).