Best of Both Worlds: Combining General and Clinical Language Models for Classification and Text Generation

Sasha Ronaghi SRONAGHI@STANFORD.EDU
Asad Aali ASADAALI@STANFORD.EDU
Chloe Stanwyck CHLOEO@STANFORD.EDU
Miguel Fuentes MIGUFUEN@STANFORD.EDU
Tina Hernandez-Boussard* BOUSSARD@STANFORD.EDU
Emily Alsentzer* EALSENTZER@STANFORD.EDU
300 Pasteur Dr, Stanford, CA 94305
Department of Biomedical Data Science, Stanford School of Medicine
**Jointly supervised this work.*

Abstract

We study proxy tuning, a training-free, decoding-time method that combines the strengths of general and clinical language models. Across three classification and four text generation tasks, zero-shot proxy tuning consistently improves performance over baselines, yielding an average 6.5% Macro-F1 gain over a large general model on classification tasks and surpassing a 70B clinical model on all generative tasks. Our analysis reveals that proxy tuning isolating clinical continued pretraining produces the largest gains on medical knowledge-intensive tasks. We additionally introduce Cross-Architecture Proxy Tuning (CAPT), which enables proxy tuning across models with different architectures and limited logit distribution access. CAPT with a new-generation base model (Qwen3-30B) achieves performance comparable to supervised fine-tuning with 2,600 samples on classification tasks and produces 90% clinically safe outputs on generation tasks. Our findings demonstrate that proxy tuning offers a practical, efficient path to clinical domain adaptation without model retraining.

Keywords: Efficient Domain Adaptation, Clinical Adaptation, Language Models

Data and Code Availability Code is available at: https://github.com/sronaghi/bestofbothworlds

Institutional Review Board (IRB) This work does not require IRB approval.

1. Introduction

General-domain language models (LMs) show promise for clinical applications but often hallucinate, omit critical clinical details, or struggle with domain-specific reasoning (Lehman et al., 2023; Hager et al., 2024; Asgari et al., 2025; Kim et al., 2025; Dorfner et al., 2024). These shortcomings stem from the limited representation of clinical data in pretraining corpora and reliance on large-scale Internet text, which can encode biases, outdated information, or incorrect medical knowledge (Alber et al., 2025; Wu et al., 2025).

Existing adaptation techniques such as fine-tuning and continued pretraining require large labeled datasets, full access to model parameters, and substantial computational resources (Xie et al., 2024; Sellergren et al., 2025). However, as data requirements for adaptation grow (Kaplan et al., 2020; Zhang et al., 2024), limited availability of labeled clinical data (Xiao et al., 2018) creates a bottleneck. Furthermore, LLMs are increasingly deployed in zero-shot clinical settings (Small et al., 2024; Mandal et al., 2025; Armitage, 2025). These challenges highlight the need for training-free adaptation strategies.

Moreover, clinical models continually pre-trained from general base models lose the advantages of broad, general-domain instruction tuning, and risk catastrophic forgetting (Kirkpatrick et al., 2017). General and clinical models thus exhibit complementary strengths: reasoning breadth versus domain expertise (Gururangan et al., 2020). Figure 1 illustrates this complementarity on the MedNLI task (Romanov and Shivade, 2018): the general LM correctly predicts

12% of examples that the clinical model misses, the clinical LM correctly predicts 22% of examples that the general model misses, and their union of correct predictions cover 82% of the dataset.

Proxy tuning (Liu et al., 2024) combines the benefits of general and clinical models by projecting the token probability difference between a clinical expert and its untuned counterpart onto a general LM at decoding-time. Prior work on proxy tuning and similar decoding-time approaches has focused on domains closer to general-model training data, such as math, programming, and law (Liu et al., 2024; Ormazabal et al., 2023). Whether these approaches transfer to the more out-of-distribution clinical domain remains unknown. We examine whether proxy tuning can effectively adapt LMs to clinical settings, where there is a need for training-free adaptation methods.

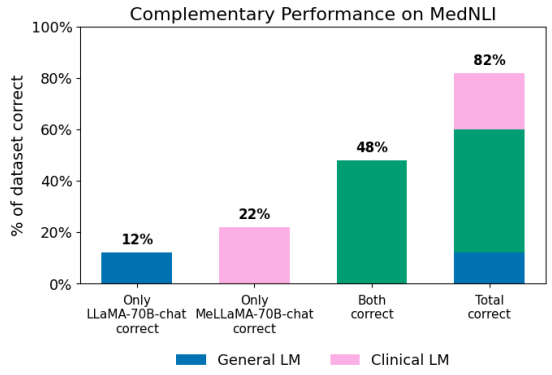

Figure 1: **Complementary Performance of General-Domain and Clinical Models on MedNLI Classification Task.** The bars from left to right represent: (1) general model's unique correct predictions (12%), (2) clinical model's correct predictions (22%), (3) intersection of correct predictions from both models (48%), and (4) the union of correct predictions from both models (82%).

Our contributions are threefold. First, we comprehensively evaluate proxy tuning across three classification and four text generation tasks. In classification, proxy tuning yields an average 6.5% Macro-F1 improvement over the large general model and recovers 64% of the performance gap between the large general and large clinical models. In text generation, proxy tuning achieves an average 36% gain in LLM-jury scores over the Large Clinical model while increasing clinical safety by an average of 10% relative to the Large General model. Second, we conduct token-level analysis, revealing that clinical proxy tuning produces more retrospective, evidence-oriented clinical note tokens. Third, we introduce Cross-Architecture Proxy Tuning (CAPT), which enables proxy tuning across models with different architectures without requiring full logit distribution access. Combining a new-generation general LM (Qwen3-30B) with a previous-generation clinical LM (Me-LLaMA-13B-chat) outperforms supervised fine-tuning with 2,800 samples on a medical natural language inference classification task (MedNLI) and increases risk-free outputs by 30% of the general LM in a clinical note generation task (ACI-Bench).

1.1. Related Works

Models such as Meditron, PMC-LLaMA, and Me-LLaMA have demonstrated the efficacy of continued pretraining on large volumes of unlabeled medical text (Chen et al., 2023; Wu et al., 2024; Xie et al., 2024). While powerful, continued pretraining is computationally expensive: Me-LLaMA performed continued pretraining on LLaMA-2 base models using 129B tokens and 160 × 80 GB A100 GPUs.

Fine-tuning offers a more efficient approach to adapt to specific clinical tasks, improve instruction-following, or align more closely with human preferences, as demonstrated by MedAlpaca (Han et al., 2023), AlpaCare (Zhang et al., 2023b), MedPaLM (Singhal et al., 2023), LLaMA-Clinic (Wang et al., 2024), and HuatuoGPT (Zhang et al., 2023a). However, previous work suggests that fine-tuning primarily affects style and preference rather than introducing domain-specific knowledge (Gekhman et al., 2024; Ouyang et al., 2022). Additionally, Wu et al. (2025) find that fine-tuning frontier models with recent medical evidence shows limited generalizability.

Continued pretraining and supervised fine-tuning both require modifying model weights, which is costly and often infeasible for frontier models. Decoding-time methods offer an alternative by transferring knowledge through output token probabilities without updating model parameters. Mitchell et al. (2023), Ormazabal et al. (2023), and Liu et al. (2024) explore proxy tuning, which combines the logit distribution of a large general model with the delta distribution of a small expert and its untuned counterpart for each token. Mitchell et al. (2023) showed

that combining a small instruction-tuned expert with a large base model improves output helpfulness on a helpfulness-harmfulness dataset; Ormazabal et al. (2023) applied the method with experts trained on small law and energy datasets; Liu et al. (2024) demonstrated proxy tuning's broad effectiveness for instruction following, mathematical reasoning, QA tasks, and code domain adaptation.

2. Methods

Proxy Tuning. The proxy tuning approach combines a large base model M with a small model M^- fine-tuned into M^+ (Liu et al., 2024). At decoding time for each new token, the logit difference $(M^+ - M^-)$ is added to M, projecting the smaller model's learned behavior onto the large model. To perform direct logit-level addition, this approach requires a shared tokenizer and vocabulary across the three models and access to the full logit distribution.

Cross-Architecture Proxy Tuning (CAPT). CAPT allows the base model to use different architectures, tokenizers, and vocabularies from the expert and anti-expert models. Additionally, rather than relying on full logits for all tokens, CAPT uses only the top-20 log probabilities returned by most black-box APIs, such as OpenAI GPT-5 and Google Gemini 2.5 (OpenAI, 2025; Dong, 2025). For each of the base model's top-20 tokens, we retrieve the corresponding log-probability differences from the expert and anti-expert models. If a token does not appear in their vocabularies, we tokenize the base token using the expert/anti-expert tokenizer and use the log-probability of the first resulting subtoken.

Experimental Setup. We use Me-LLaMA models as our clinical LLMs, state-of-the-art open-source models trained on MIMIC clinical notes (Xie et al., 2024). Me-LLaMA base models were obtained through continued pretraining of LLaMA-2 on 129B tokens (15:1:4 ratio of biomedical, clinical, and general text), while chat models were produced by instruction tuning these base models with 214k medical instructions. Both variants are available in 13B and 70B parameter sizes.

Table 1 shows expert–anti-expert configurations that isolate the effects of Me-LLaMA's continued pretraining (CPT), instruction tuning (IT), and their combination (CPT+IT). For base models, we use LLaMA-2-70B-chat (Touvron et al., 2023) in proxy tuning experiments and Qwen3-30B-A3B-Instruct-2507 (Yang et al., 2025) in CAPT experiments.

Configuration	Expert	Anti-expert
CPT	Me-LLaMA-13B-base	LLaMA-13B-base
IT	Me-LLaMA-13B-chat	Me-LLaMA-13B-base
CPT + IT	Me-LLaMA-13B-chat	LLaMA-13B-base

Table 1: **Proxy Tuning Configurations.**

Evaluation Tasks and Metrics.

We evaluate on three classification tasks and four text generation tasks. For classification, we report Macro-F1 on 200 samples from: **MedNLI** (Romanov and Shivade, 2018), classifying hypothesis–premise pairs; **MTSample-Specialty** (MTSamples), identifying medical specialties from transcriptions; and **Fall Event Extraction** (Pillai et al., 2024), detecting fall events in clinical notes from Stanford Health Care surgical patients. To compare supervised fine-tuning in limited data settings with zero-shot proxy tuning, we fine-tuned Me-LLaMA-13B-base on increasing subsets of training data for each classification task until performance surpasses proxy tuning.

For text generation, we report MedHELM LLM-jury scores, which assess accuracy, clarity, and completeness against gold-standard responses and have been validated against clinician evaluations (Bedi et al., 2025). We evaluate clinical safety using MedVAL, a fine-tuned model that assigns risk ratings to generated outputs and has shown strong correlation with physician assessments (r=0.833) (Aali et al., 2025). We report the percentage of risk-free outputs across models. We evaluate on 120 samples from four text generation tasks: **ACI-Bench** (Yim et al., 2023), generating clinical notes from patient-doctor conversations; **MTSample-Replicate** (MTSamples), generating treatment plans from clinical notes; **MedDialog** (Zeng et al., 2020), summarizing patient–doctor conversations; and **MediQA** (Abacha et al., 2019), answering consumer health questions.

Token-level Analysis. The top quartile of tokens generated in the **ACI-Bench** task were grouped into semantic categories by a physician. We report the mean probability difference between the proxy-tuned and base models for each token category. Positive differences indicate greater influence from the expert model, while negative differences indicate greater influence from the base model.

Table 2: **Model Performance.** Performance of baseline and proxy-tuned models on classification and text generation tasks (200 and 120 samples per task, respectively). Models: Large General-chat = LLaMA-2-70B-chat, Large Clinical-chat = Me-LLaMA-70B-chat, Small Clinical-base = Me-LLaMA-13B-base, Small Clinical-chat = Me-LLaMA-13B-chat. Proxy-tuning configurations are in Table 1. We report Macro-F1 for classification tasks and MedHELM LLM jury scores and % of risk-free outputs assigned by MedVAL. Best performance in each model category is in bold.

	Classification			**Generative**							
	MedNLI	MTSample-Specialty	Fall Event Extraction	ACI-Bench		MTSample-Replicate		Med-Dialog		MediQA	
	Macro-F1	Macro-F1	Macro-F1	LLM-jury	% Risk-Free	LLM-jury	% Risk-Free	LLM-jury	% Risk-Free	LLM-jury	% Risk-Free
Baseline											
Large General-chat	0.587	0.101	0.778	**3.88**	25.83	**3.83**	24.17	**4.12**	85.0	**4.27**	**60.83**
Large Clinical-chat	**0.687**	**0.115**	**0.791**	3.49	**50.83**	1.93	**63.33**	4.07	70.0	3.37	47.5
Small Clinical-base	0.335	0.0376	0.389	2.02	19.17	1.43	20.0	2.61	**91.6**	2.33	34.17
Small Clinical-chat	0.488	0.0694	0.613	2.85	30.0	1.49	22.5	2.74	77.5	2.78	40.83
Proxy Tuning											
CPT	**0.624**	**0.113**	0.742	**3.94**	20.83	**3.88**	**28.33**	4.08	**80.83**	4.30	66.67
CPT+IT	0.596	0.106	0.777	3.93	**33.33**	3.87	22.5	**4.09**	79.17	4.30	70.83
IT	0.584	0.105	**0.787**	3.91	40.0	3.86	**28.33**	4.07	77.5	**4.32**	**74.17**

	MedNLI	MTSample-Specialty	Fall Event Extraction
	Macro-F1	Macro-F1	Macro-F1
0.5%	0.549	0.089	0.551
1%	0.524	0.093	0.531
5%	**0.791**	0.095	0.531
10%	-	0.109	0.355
25%	-	**0.136**	0.402
50%	-	-	**0.823**

Table 3: Supervised fine-tuning results across classification tasks with increasing subsets of the training data.

3. Results

Proxy Tuning Improves Performance on Classification and Text Generation Tasks. Proxy tuning consistently improved over the large general baseline. As shown in Table 2, In classification, proxy-IT models recovered 32% of the average performance gap between large general and clinical models, while proxy-CPT recovered 86% on MTSample-Specialty. On generative tasks, proxy-IT models exceeded the large clinical baseline by 25.5% on average LLM-jury scores and increased risk-free outputs by 13% on average over the large general baseline. As shown in Table 3, we find that supervised fine-tuning requires 5% of MedNLI training data (∼560 samples), 25% of MTSample-Specialty training data (∼1250 samples), and 50% of Fall Event Extraction training data (∼100 samples) to surpass zero-shot proxy-tuning performance.

Proxy-CPT excelled on tasks requiring medical terminology (MTSample-Specialty), reasoning (ACI-Bench, MTSample-Replicate), and inference (MedNLI). Proxy-IT performed best on tasks similar to those in Me-LLaMA's instruction-tuning dataset: Fall Event Extraction and MediQA resemble mortality classification and patient question-answering tasks from that dataset. Proxy-CPT outperformed both Proxy-CPT+IT and Proxy-IT on MedNLI despite its similarity to relation-extraction tasks in instruction tuning, suggesting alignment benefits may be greater for tasks requiring less medical knowledge.

Performance trends also reveal when clinical proxy tuning is less effective: Proxy-CPT and Proxy-CPT+IT reduced performance on Fall Event Extraction despite Me-LLaMA's clinical note pretraining, possibly due to distributional differences between Stanford Hospital task data and Me-LLaMA's Beth Israel training data. MedDialog showed uniformly small negative effects from proxy tuning, likely because the task requires concise one-sentence responses that are uncommon in clinical notes.

Proxy Tuning Shifts Token Generation Toward Clinical Documentation Styles. Figure 2 shows that clinical proxy tuning shifted the base model toward retrospective, evidence-oriented clinical note styles: it decreased present-tense observational verbs, increased past-tense observational verbs, encounter/workflow terms, and hedges. Proxy-CPT+IT resembled Proxy-IT more closely than Proxy-CPT (46% smaller mean token shift), indicating that inference-time adaptation is more sensitive

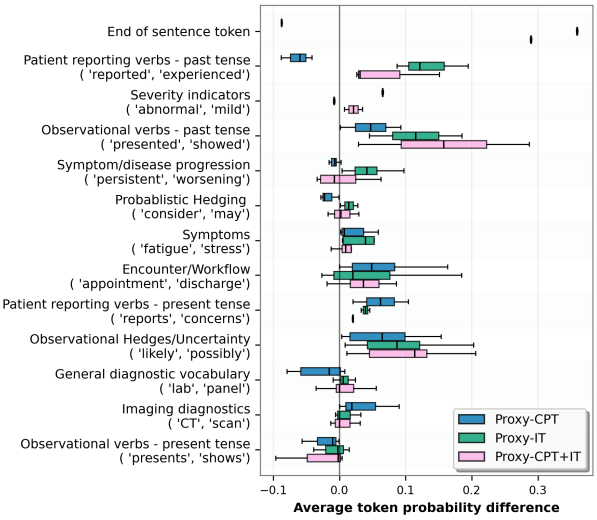

Figure 2: **Average token probability difference across categories between proxy-tuned and base models.** Positive difference indicates stronger expert influence while negative difference indicates stronger base influence.

Table 4: **Model Performance** Comparison of CAPT and original proxy tuning on MedNLI and ACI-Bench. Proxy-tuning configurations are in Table 1.

	MedNLI	ACI-Bench	
	Macro-F1	LLM-jury	% Risk-Free
Cross-Architecture Proxy Tuning (CAPT)			
Qwen3-30B	0.716	4.36	60.83
CAPT - CPT	0.824	4.331	85.83
CAPT - CPT + IT	**0.899**	**4.379**	**89.17**
CAPT - IT	0.881	4.3805	87.5
Original Proxy Tuning (PT)			
LLaMA-70B-chat	0.587	3.88	30.83
MeLLaMA-70B-chat	0.687	3.49	54.17
PT - CPT	0.623	3.943	37.5
PT - CPT + IT	0.596	3.929	35
PT - IT	0.584	3.912	40.83
Supervised Fine-Tuning (% of Dataset)			
10% (∼1,120 samples)	0.813	–	–
25% (∼2,800 samples)	0.860	–	–

to instruction-tuned alignment compared to continued pretraining.

Proxy-IT and Proxy-CPT showed distinct emphases. The end-of-sentence token shifted strongly positive under Proxy-IT but negative under Proxy-CPT. This likely reflects the differences learned from the short responses of question-answering tasks in instruction-tuning, versus the longer, information-dense lines of clinical documentation in the continued pretraining data. Proxy-IT favored tokens which support clinical reasoning, but may not be necessary for factual completeness: severity terms (e.g., "abnormal"), symptom-progression markers (e.g., "persistent"), and probabilistic hedging (e.g., "may"). Proxy-CPT strengthened tokens tied to factual grounding and procedural documentation: encounter/workflow terms (e.g., "admission," "discharge"), present-tense patient reporting (e.g., "reports," "concerns"), and modality-specific diagnostics (e.g., "CT," "x-ray").

Cross-Architecture Proxy Tuning Improves Performance and Increases Applicability. As shown in Table 4, on MedNLI and ACI-Bench, CAPT with Qwen3-30B outperformed Qwen3-30B alone, the original proxy-tuned models using LLaMA-2-70B, and the large clinical model (Me-LLaMA-70B-chat). CAPT also outperformed Me-LLaMA-13B fine-tuned on 25% of the MedNLI dataset (∼2,800 samples) and achieved ∼90% risk-free outputs on ACI-Bench compared to ∼60% for Qwen3-30B alone. These results suggest that proxy tuning can be an effective approach as general models continue to improve.

4. Discussion

Proxy tuning is a promising training-free method to combine the benefits of general-domain and clinical language models, showing performance gains across classification and generation tasks. Our token-level analyses reveal mechanisms of how proxy tuning which isolate continued pretraining and instruction tuning affects style and tone. We introduce Cross-Architecture Proxy Tuning (CAPT) for broader applicability across model architectures and proprietary/open model setups. Supervised fine-tuning comparisons show that CAPT matches performance requiring ∼2,600 training samples, establishing its potential for efficient clinical domain adaptation in limited data settings.

Future Work. Our token-level analysis reveals disproportionate proxy tuning effects across tokens, motivating adaptive scaling mechanisms that adjust expert contributions token-by-token to improve performance and reduce inference costs. Ablation studies across expert model sizes would clarify sensitivity to scale.

References

Asad Aali, Vasiliki Bikia, Maya Varma, Nicole Chiou, Sophie Ostmeier, Arnav Singhvi, Magdalini Paschali, Ashwin Kumar, Andrew Johnston, Karimar Amador-Martinez, Eduardo Juan Perez Guerrero, Paola Naovi Cruz Rivera, Sergios Gatidis, Christian Bluethgen, Eduardo Pontes Reis, Eddy D. Zandee van Rilland, Poonam Laxmappa Hosamani, Kevin R Keet, Minjoung Go, Evelyn Ling, David B. Larson, Curtis Langlotz, Roxana Daneshjou, Jason Hom, Sanmi Koyejo, Emily Alsentzer, and Akshay S. Chaudhari. Medval: Toward expert-level medical text validation with language models, 2025. URL https://arxiv.org/abs/2507.03152.

Asma Abacha et al. Overview of the mediqa 2019 shared task on consumer health question answering. In *ACL-BioNLP Workshop*, 2019.

Daniel A. Alber, Zeyi Yang, Andrey Alyakin, et al. Medical large language models are vulnerable to data-poisoning attacks. *Nature Medicine*, 31:618–626, 2025. doi: 10.1038/s41591-024-03445-1.

Hanae Armitage. Clinicians can 'chat' with medical records through new AI software, ChatEHR, June 2025. URL https://med.stanford.edu/news/all-news/2025/06/chatehr.html.

Ehsaneddin Asgari, N. Montaña-Brown, M. Dubois, et al. A framework to assess clinical safety and hallucination rates of llms for medical text summarisation. *npj Digital Medicine*, 8:274, 2025. doi: 10.1038/s41746-025-01670-7.

R. Bedi et al. Medhelm: Evaluation framework for medical llms. *arXiv preprint*, 2025.

Zvika Chen, Andoni H. Cano, Angelos Romanou, et al. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint*, 2023.

Eric Dong. Unlock gemini's reasoning: A step-by-step guide to logprobs on vertex ai, July 2025. URL https://developers.googleblog.com/en/unlock-gemini-reasoning-with-logprobs-on-vertex-ai/. Google Developers Blog, accessed 2025-11-03.

F. J. Dorfner, L. Jürgensen, L. Donle, F. A. Mohamad, T. R. Bodenmann, M. C. Cleveland, C. P. Bridge, et al. Is open-source there yet? a comparative study on commercial and open-source llms in their ability to label chest x-ray reports. *arXiv preprint arXiv:2402.12298*, 2024.

Ziv Gekhman, Gal Yona, Roee Aharoni, et al. Does fine-tuning llms on new knowledge encourage hallucinations? *arXiv preprint*, 2024.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.740.

P. Hager, F. Jungmann, R. Holland, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine*, 30:2613–2622, 2024. doi: 10.1038/s41591-024-03097-1.

Ting Han et al. Medalpaca: An open-source collection of medical conversational ai models and training data. *arXiv preprint*, 2023.

Jared Kaplan et al. Scaling laws for neural language models. *arXiv preprint*, 2020.

T. S. Kim, Y. Lee, Y. Park, J. Kim, Y. H. Kim, and J. Kim. Cupid: Evaluating personalized and contextualized alignment of llms from interactions. *arXiv preprint arXiv:2508.01674*, 2025.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, March 2017. ISSN 1091-6490. doi: 10.1073/pnas.1611835114. URL http://dx.doi.org/10.1073/pnas.1611835114.

Eric Lehman et al. Do we still need clinical language models? In *Conference on Health, Inference, and Learning (CHIL)*. PMLR, 2023. URL https://arxiv.org/abs/2302.08091.

A. Liu, X. Han, Y. Wang, Y. Tsvetkov, Y. Choi, and N. A. Smith. Tuning language models by proxy. *arXiv preprint arXiv:2401.08565*, 2024.

S. Mandal, B. M. Wiesenfeld, A. C. Szerencsy, et al. Utilization of generative ai-drafted responses for managing patient-provider communication. *npj Digital Medicine*, 8:591, 2025. doi: 10.1038/s41746-025-01972-w.

Eric Mitchell, Rafael Rafailov, Archit Sharma, Chelsea Finn, and Christopher D. Manning. An emulator for fine-tuning large language models using small language models. *arXiv preprint arXiv:2310.12962*, 2023.

MTSamples. Mtsamples: Medical transcription samples. https://www.mtsamples.com/. Accessed 2025-09-08.

OpenAI. GPT-5 System Card. Technical report, OpenAI, August 2025. URL https://cdn.openai.com/gpt-5-system-card.pdf. Accessed: 2025-11-03.

Aitor Ormazabal, Mikel Artetxe, and Eneko Agirre. CombLM: Adapting black-box language models through small fine-tuned models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2961–2974, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.180. URL https://aclanthology.org/2023.emnlp-main.180/.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

M. Pillai, T. L. Blumke, J. Studnia, Y. Wang, Z. P. Veigulis, A. D. Ware, P. J. Hoover, I. R. Carroll, K. Humphreys, T. F. Osborne, S. M. Asch, T. Hernandez-Boussard, and C. M. Curtin. Improving postsurgical fall detection for older americans using llm-driven analysis of clinical narratives. *medRxiv*, June 2024. doi: 10.1101/2024.06.25.24309480. URL https://doi.org/10.1101/2024.06.25.24309480. Preprint.

Alexey Romanov and Chaitanya Shivade. Lessons from natural language inference in the clinical domain, 2018. URL https://arxiv.org/abs/1808.06752.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, Lu Yang, Kejia Chen, Per Bjornsson, Shashir Reddy, Ryan Brush, Kenneth Philbrick, Mercy Asiedu, Ines Mezerreg, Howard Hu, Howard Yang, Richa Tiwari, Sunny Jansen, Preeti Singh, Yun Liu, Shekoofeh Azizi, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Riviere, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Elena Buchatskaya, Jean-Baptiste Alayrac, Dmitry Lepikhin, Vlad Feinberg, Sebastian Borgeaud, Alek Andreev, Cassidy Hardin, Robert Dadashi, Léonard Hussenot, Armand Joulin, Olivier Bachem, Yossi Matias, Katherine Chou, Avinatan Hassidim, Kavi Goel, Clement Farabet, Joelle Barral, Tris Warkentin, Jonathon Shlens, David Fleet, Victor Cotruta, Omar Sanseviero, Gus Martins, Phoebe Kirk, Anand Rao, Shravya Shetty, David F. Steiner, Can Kirmizibayrak, Rory Pilgrim, Daniel Golden, and Lin Yang. Medgemma technical report, 2025. URL https://arxiv.org/abs/2507.05201.

Karan Singhal, Shekoofeh Azizi, Tu Tu, et al. Large language models encode clinical knowledge. *Nature*, 620:172–180, 2023. doi: 10.1038/s41586-023-06291-2.

Warren R. Small, Benjamin Wiesenfeld, Ben Brandfield-Harvey, et al. Large language model–based responses to patients' in-basket messages. *JAMA Network Open*, 7(7):e2422399, 2024. doi: 10.1001/jamanetworkopen.2024.22399.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin

Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

Han Wang et al. Towards adapting open-source large language models for expert-level clinical note generation. *arXiv preprint*, 2024.

Chen Wu et al. Pmc-llama: Toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, 31: 1833–1843, 2024.

Eric Wu, Kevin Wu, and James Zou. Limitations of learning new and updated medical knowledge with commercial fine-tuning large language models. *NEJM AI*, 2, 2025. doi: 10.1056/AIcs2401155.

Cao Xiao, Edward Choi, and Jimeng Sun. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *J. Am. Medical Informatics Assoc.*, 25(10):1419–1428, 2018. URL https://doi.org/10.1093/jamia/ocy068.

Qiang Xie, Qian Chen, Ailin Chen, Cheng Peng, Yutong Hu, Fang Lin, and Jiang Bian. Me-llama: Foundation large language models for medical applications. Research Square preprint, 2024. rs-3.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Wen Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation, 2023. URL https://arxiv.org/abs/2306.02022.

Ying Zeng et al. Meddialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (Findings)*, 2020.

Biao Zhang et al. When scaling meets llm finetuning: The effect of data, model and finetuning method. *arXiv preprint*, 2024.

Hongyi Zhang et al. Huatuogpt: Towards taming language model to be a doctor. *arXiv preprint*, 2023a.

Xiang Zhang, Chuan Tian, Xia Yang, et al. Alpacare: Instruction-tuned large language models for medical application. *arXiv preprint*, 2023b.

5. Appendix

5.1. Additional Experimental Details

Experiments were conducted on NVIDIA A100 40GB GPUs and V100 32GB GPUs. For classification tasks, we implemented constrained decoding. For generative tasks, we implement a custom proxy tuning model client within the MedHELM evaluation framework. (Bedi et al., 2025). Table 5 contains a description of each task and the prompt used. Table 6 contains the token categories for token-level analysis.

5.2. Significance Testing

In Table 7, we report the coefficient of variation (CV Δ) for each model configuration to quantify the relative variability of gains and losses across tasks. Specifically, CV Δ (Clinical Tasks) reflects consistency across the tasks that involved clinical text

Table 5: **Evaluation Tasks.** Each task includes its description and the prompt template used.

Task	Description	Prompt
MedNLI	A natural language inference task in which the goal is to determine whether a hypothesis written by a doctor can be inferred from a premise taken directly from a clinical note (multi-class classification with labels entailment, neutral, or contradiction).	"TASK: Please classify the relationship between the given premise and hypothesis into one of the following labels: entailment, contradiction, or neutral. Return only the label. INPUT:{text} OUTPUT:"
MTSample	A multi-class classification task in which the goal is to determine the medical specialty or domain that a medical transcription belongs to from 40 medical specialties and domains.	"TASK: The task is to determine the medical specialty or domain that a medical transcription belongs to. The input is a medical transcription. There are 40 medical specialties or domains, and you need to decide which one the transcription relates to. The medical specialties or domains are: 'Surgery', 'Allergy / Immunology', ..., 'Obstetrics / Gynecology'. The output should be only one medical specialty or domain. INPUT:{text} OUTPUT:"
Fall Extraction	A binary classification task in which the goal is to determine whether or not the patient had a fall event based on a postoperative clinical note.	"TASK: Classify whether a patient fell or not after surgery into one of the following labels: fall, no fall. Historical falls, fall risk/precautions, or other miscellaneous mentions of falls like blood pressure falling are not fall events and the output should be 'no fall' unless a fall event is also indicated in the note."
ACI-Bench	Generating a structured clinical note from patient-doctor conversations.	"Summarize the conversation to generate a clinical note with four sections: 1. HISTORY OF PRESENT ILLNESS 2. PHYSICAL EXAM 3. RESULTS 4. ASSESSMENT AND PLAN. Conversation: Doctor-patient dialogue: [doctor] hi , andrew . how are you ? ..."
MTSample-Replicate	Generating an appropriate treatment plan from a clinical note.	"Given various information about a patient, return a reasonable treatment plan for the patient. Medical Specialty: Orthopedic ... PREOPERATIVE DIAGNOSES: 1. Cellulitis with associated abscess, right foot. ..."
MedDialog	Generating a one sentence summary of a patient-doctor conversation.	"Generate a one sentence summary of this patient-doctor conversation. Patient-Doctor: Patient: Can rabies be transferred through blood? ..."
MediQA	Answering a consumer health question.	"Answer the following consumer health question. Question: what are known causes of bipolar disorder"

(MedNLI, MTSample-Specialty, Fall Event, ACI-Bench, MTSample-Replicate), and CV Δ (All Tasks) extends to include Med-Dialog and MediQA. The relatively low CV values indicate stable performance of our proxy tuning approach on clinical tasks.

For the classification tasks, we computed macro-F1, 95% confidence intervals using bootstrap resampling, bootstrap standard error, and paired-bootstrap p-values relative to the general-domain model baseline. Improvements with Qwen were statistically significant for both MedNLI and MTSample. Results are shown in Table 8.

Table 6: **Token Categories and Tokens.**

Category	**Tokens**
Encounter/Workflow	appointment, manage, managed, follows, follow, routine, review, appoint, Reg, return, informed, admitted, discharge, inpatient, outpatient, follow-up, Sch, Enc, Ext, Copy, copy, order, pending, referred, management
Observational Hedges/Uncertainty	likely, appears, particularly, possibly, sometimes, indicating, consistent, associated, related, possible
Symptom/disease progression	improved, worsening, improve, improvement, persistent, progress, controlled
End of sentence token	$< eos >$
Negations	no, not, without, denies, none, absent, negative, No
Severity indicators	abnormal, severe, moderate, mild, mildly, moderately, severely, good
Symptoms	feeling, feels, tired, fatigue, stress, dizy, dizziness, nausea, pain, stress, breath, sympt
General diagnostic vocabulary	lab, labs, panel, test, tests, level, levels, results, testing, evaluation
Imaging diagnostics	CT, MRI, X-ray, ray, scan, ultrasound
Clinical directive verbs	follow, evaluate, Order, prevent, control, take, use
Patient reporting verbs – present tense	reports, concerns
Patient reporting verbs – past tense	reported, experienced, described
Observational verbs – present tense	presents, shows, has
Observational verbs – past tense	presented, showed
Probabilistic Hedging	consider, confirm, may

Table 7: **Coefficient of variation (CV Δ) and mean percentage improvement across tasks.**

Model	**Mean % Improvement (Clinical Tasks)**	**CV Δ (Clinical Tasks)**	**Mean % Improvement (All Tasks)**	**CV Δ (All Tasks)**
Large Clinical	−5.30	3.4	−7.99	2.9
Small Clinical-CPT	−53.24	0.15	−49.74	0.18
Small Clinical-CPT+IT	−31.46	0.53	−32.24	0.43
Proxy-CPT	3.26	1.6	2.56	2.0
Proxy-CPT+IT	1.54	1.1	1.29	1.5
Proxy-IT	1.13	1.3	0.62	2.7

Table 8: **Macro-F1, 95% confidence intervals using bootstrap resampling, bootstrap standard error, and paired-bootstrap p-values relative to the general-domain model baseline for classification tasks with both Qwen3-30B and LLaMA70B-chat as base models. Bold** indicates statistically significant results (P-value < 0.05).

	MedNLI				**MTSample-Specialty**				**Fall Event Extraction**			
	Macro-F1	**95% CI**	**STD**	**P-value**	**Macro-F1**	**95% CI**	**STD**	**P-value**	**Macro-F1**	**95% CI**	**STD**	**P-value**
Cross Architecture Proxy Tuning												
Qwen3-30B Base	0.716	(0.652–0.771)	0.0304	1	0.103	(0.059–0.126)	0.0175	1	0.803	(0.744–0.856)	0.0283	1
CAPT-CPT	**0.824**	**(0.772–0.869)**	0.0248	**0.0248**	0.117	(0.071–0.137)	0.0169	0.264	0.807	(0.748–0.860)	0.0282	0.733
CAPT-CPT + IT	**0.899**	**(0.856–0.938)**	0.0212	**0.0212**	**0.134**	**(0.093–0.150)**	**0.0148**	**0.004**	0.817	(0.758–0.870)	0.0278	0.091
CAPT-IT	**0.881**	**(0.837–0.922)**	0.0221	**0.0221**	0.112	(0.068–0.134)	0.0172	0.300	0.812	(0.753–0.865)	0.0279	0.266
Proxy Tuning												
LLaMA-70b-chat Base	0.587	(0.523–0.650)	0.0329	1	0.101	(0.056–0.124)	0.0178	1	0.778	(0.715–0.833)	0.0299	1
PT-CPT	0.624	(0.556–0.689)	0.0335	0.128	0.113	(0.067–0.135)	0.0178	0.190	0.742	(0.676–0.799)	0.0316	0.101
PT-CPT + IT	0.596	(0.533–0.658)	0.0317	0.740	0.106	(0.058–0.130)	0.0187	0.730	0.777	(0.717–0.833)	0.0297	0.958
PT-IT	0.584	(0.521–0.648)	0.0320	0.932	0.105	(0.057–0.128)	0.0186	0.855	0.787	(0.728–0.841)	0.0293	0.605