Life of Research · Published 30 October 2026 · DOI 10.5281/zenodo.23784965

A Domain Label Is Not a Treatment: Why Specialist and General-Purpose Language Models Trade Places Across Studies, What a 'Medical' Checkpoint Leaves Unspecified, and the Matched Contrasts That Would Assign the Gain

Pranay Mahendrakar
SONYTECH · ORCID 0009-0003-7224-029X
pranaymahendrakar@sonytech.in · mahendrakarpranay@gmail.com · +91 6361723454

Abstract. A September 2026 study compares an 8B general-purpose instruction-tuned model with a medically fine-tuned checkpoint built on the same base and finds that the general model scores higher both at selecting the meaning of medical jargon and at recognising when a word is being used as jargon, significantly so on the second, and that the medical checkpoint's deficit on the second task sits in a few pre-existing components it over-weights. The finding joins a run of 2024-2026 results in which medical models fail to beat their bases, set against an older and still-growing record in which domain-adapted models in code, mathematics, chip design, law and clinical text beat general ones by wide margins. In this analysis the author reviews and synthesises published results to show that these verdicts are not answers to one question, because a specialist-versus-generalist comparison is indexed by five things that the studies set differently: the treatment (what was trained, on how much text the base had not already seen), the comparator (the starting checkpoint, a post-trained sibling, or a model of another lineage), the elicitation (whose prompt format and whose decision threshold), the target skill (recall, terminology, extraction against labels, or multi-step computation), and the statistic (paired tests, intervals, calibration). Read through those indices, the negative medical results hold in one well-populated cell - a modest post-training or continued-pretraining treatment on a web-trained base, compared with a strong sibling under independent prompting on recall or terminology - and the large positive results hold in cells where the domain data are large relative to the base's exposure, absent from the public web, or labelled for the exact task. The paper argues that the anchor's mechanistic result is better read as a shifted decision prior than as lost knowledge, consolidates 27 published results in a table, maps fourteen studies onto four of the five indices, states a classification procedure, and specifies five matched contrasts that would decide what a domain label buys.

Keywords: domain adaptation; domain-specific language models; continued pretraining; medical language models; specialized terminology; evaluation methodology; fine-tuning mechanisms; calibration

SONYTECH ORCID: 0009-0003-7224-029X

1. Introduction

Whether to specialise is one of the oldest practical questions in applied language modelling, and the answer has changed sign at least twice. In the encoder era, a second phase of pretraining on in-domain text improved classification in biomedical, computer-science, news and review domains (Gururangan et al., 2020), and for biomedicine, where unlabelled text is abundant, pretraining from scratch on domain text beat continuing from a general checkpoint (Gu et al., 2020). Clinical text gave the same verdict two years later: small models pretrained on clinical notes from scratch outperformed every in-context approach with general models up to 175B parameters on electronic-health-record tasks (Lehman et al., 2023). Then the generalists grew. Flan-PaLM reached 67.6 percent on MedQA with prompting alone (Singhal et al., 2022); GPT-4 steered by a general prompting recipe topped every benchmark in the MultiMedQA suite and passed 90 percent on MedQA (Nori et al., 2023); and a reasoning model without any steering outperformed that recipe (Nori et al., 2024). By 2026 three frontier general models outperformed two commercial clinical tools on exam questions, on HealthBench items and on real physician queries (Vishwanath et al., 2026).

The open-weight medical models that were built to close this gap have not, on the most careful re-evaluations, done so. Comparing ten medical language models with their own base models under per-model prompt selection and statistical testing, Jeong et al. (2024b) find that on clinical-note question answering in the 3-shot setting the medical model is significantly better in 26.7 percent of cases and significantly worse in 56.7 percent. Dorfner et al. (2024) report an 8B biomedical model scoring 30 percent on NEJM case challenges against 64.3 percent for the general instruction-tuned model of the same family. Dada et al. (2024) find that only one of twelve biomedical models improves on its base across all six of their clinical tasks.

The study that occasions this paper sharpens the question to vocabulary. Keng and Sun (2026) build two benchmarks from electronic-health-record contexts - one asking a model to choose the correct definition of a medical term, one asking whether a word is being used as jargon in context - and compare Llama-3.1-8B-Instruct

© 2026 Pranay Mahendrakar. Licensed under CC BY 4.0.

with Llama-3.1-8B-UltraMedical, a medically fine-tuned checkpoint of the same base. The general model scores higher on both, and the identification gap is significant under a paired test. Using a decomposition of the output logits into per-component contributions (Chang et al., 2024a), the authors locate the specialist's deficit in a small set of attention heads and MLP blocks that it weights more heavily than the general model does, and show that learning per-component weights on a training split recovers part of the gap. They close with the recommendation that domain adaptation should not be assumed to improve performance on specialised terminology.

Terminology is the place where this result should hurt most. Jargon is long-tail knowledge, and language models answer long-tail questions in proportion to how many relevant documents they saw in pretraining (Kandpal et al., 2023); a domain corpus is exactly a corpus in which the long tail is dense. If specialisation does not help with the vocabulary of the specialty, the case for it looks thin.

The question this paper asks is what a specialist-versus-generalist comparison measures, and what it would take to say that a domain treatment helped. In this analysis the author reviews and synthesises published results to show that the answer turns on five choices that the studies above make differently, that the anchor's pair differs on three of them at once, and that the robust negative and robust positive results in the literature sit in different cells of the space those choices define. The paper makes four contributions.

First, it states five indices of a specialist claim - treatment, comparator, elicitation, target and statistic - and shows for each, from published results, that changing it alone has reversed a verdict (Sections 4 and 5).

Second, it reads the medical disagreement through those indices (Section 5), sets it against the domains where specialisation has paid reliably (Sections 6 and 7), and offers a marginal-information account that fits both, labelled as a hypothesis (Section 9).

Third, it argues that the anchor's mechanistic evidence, read alongside the literature on what fine-tuning changes inside a model, points to a shifted decision prior rather than lost or reorganised knowledge, and states the limits of that reading (Section 8).

Fourth, it consolidates the measured values in Table 1 and Figure 1, maps the studies onto the indices in Table 2, states a classification procedure as Algorithm 1, and specifies the matched contrasts that would settle the open part (Sections 10 to 13).

The conclusion is narrower than either camp's. "General models already know medicine" is true of one kind of treatment on one kind of task against one kind of comparator. "Domain data still matter" is true of others. Neither is a property of the domain label, and the anchor study, careful as it is, compares - on the evidence of its makers' training configuration, though not of the specialist's model card - two post-training pipelines rather than a medical model with a model that has had no medical treatment.

2. What the anchor study reports

Keng and Sun (2026) draw their medical items from README (Yao et al., 2024), a collection of more than 50,000 medical terms paired with lay definitions and situated in de-identified clinical-note contexts. From a 113,659-entry subset they filter, using Sentence-BERT and Wiktionary-based checks, down to 5,977 entries judged genuinely specialised and 2,095 judged non-jargon. Two tasks follow. Jargon Understanding (JU) is five-way multiple choice: given a clinical context and a term, pick the correct definition from the true one and four distractors drawn from other training-set definitions. Jargon Identification (JI) is binary: given a context and a word, say whether the word is being used as jargon. The JU test split has 1,495 items; the JI test split has 2,990, balanced between classes. A materials-science JI set built from MatScholar annotations and a BoolQ control reformatted to the same A/B prompt are used for cross-domain analysis.

The pair is Llama-3.1-8B-Instruct, Meta's post-trained release (Grattafiori et al., 2024), and Llama-3.1-8B-UltraMedical from Zhang et al. (2024). On JU the general model scores 76.1 percent (95 percent Wilson interval 73.8-78.2) and the specialist 74.1 (71.8-76.3). On JI the general model scores 58.8 (57.0-60.6) and the specialist 53.1 (51.3-54.9). On the JI items where exactly one model is right, the general model is right on 368 and the specialist on 197, and an exact McNemar test rejects equality. The paper reports the McNemar test for JI; the JU intervals overlap, and no paired test of the JU difference is reported. An appendix reruns both tasks under prompt variants that each change one element of the canonical prompt - system persona, question wording, the definition of jargon and the order of the A/B labels for JI; persona and question wording for JU - and finds that "the relative ordering of the two models is consistent across all tested prompts" (Keng and Sun, 2026, Appendix H). Across the five JI prompts the general model scores 58.0 to 60.3 and the specialist 51.0 to 53.9, the reversed label order included; across the three JU prompts the general model scores 74.7 to 76.3 and the specialist 74.1 to 74.5, so under the persona variant the JU gap narrows to 0.3 points. The error analysis finds the specialist's JI boundary shifted toward calling words in clinical contexts jargon, and the calibration plots show the specialist's accuracy falling below its confidence in the high-confidence bins on both tasks.

The mechanistic part uses the framework of Chang et al. (2024a), in which the final-position logits are written as

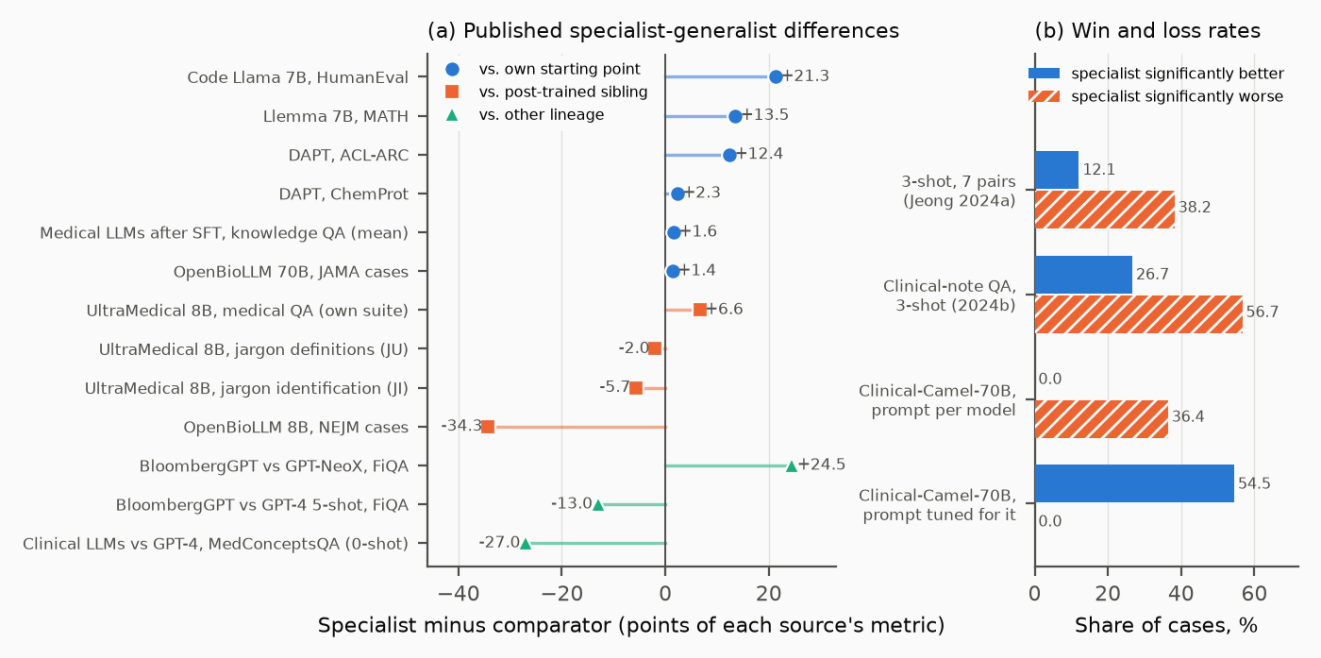

Figure 1. (a) Specialist-minus-comparator differences in points of each source's own metric, computed by subtraction from published values and grouped by comparator type: Code Llama 7B vs Llama 2 7B, HumanEval pass@1 (Roziere et al., 2023); Llemma 7B vs Code Llama 7B, MATH (Azerbayev et al., 2024); domain-adaptive pretraining vs RoBERTa on ACL-ARC and ChemProt (Gururangan et al., 2020); mean gain of medical LLMs over their bases after supervised fine-tuning on knowledge QA (Jeong et al., 2024b); UltraMedical 8B vs Llama-3-8B-Instruct on its makers' medical QA average (Zhang et al., 2024) and Llama-3.1 UltraMedical vs Llama-3.1-8B-Instruct on JU and JI (Keng and Sun, 2026); OpenBioLLM vs Llama-3-Instruct on JAMA cases (70B, own base) and NEJM cases (8B, sibling) (Dorfner et al., 2024); BloombergGPT vs GPT-NeoX and vs 5-shot GPT-4 on FiQA weighted F1 (Li et al., 2023); clinical LLMs vs GPT-4 on MedConceptsQA, zero-shot (Shoham and Rappoport, 2024). Metrics differ across rows, so only signs and rough magnitudes compare. (b) Share of (model pair, dataset) cases in which the medical model is significantly better or worse than its base: seven pairs, 3-shot (Jeong et al., 2024a); clinical-note QA, 3-shot (Jeong et al., 2024b); and Clinical-Camel-70B on knowledge QA under per-model prompt selection versus a prompt tuned only for it (Jeong et al., 2024b).

a sum of direct contributions from each attention head and MLP block, so that each component can be scored on its own and re-weighted. Both checkpoints have similar rankings of jargon-sensitive components; the specialist leans harder on components associated with jargon-favouring predictions. Learning per-component weights with an L1 penalty on the 1,200-item JI training split raises the specialist's JI accuracy by 3.0 points, to 56.1, a change that is significant under the paired test; the same procedure lowers the general model's JI accuracy by 1.0 point, and JU moves by 0.1 points for the specialist and 0.8 for the general model, both inside the intervals. The two components that receive the largest reductions, MLP-29 and head L30H27, are driven to near zero in the specialist and only partly suppressed in the general model; after reweighting, which drives L30H27's weight to zero, two of the specialist's errors on ordinary words ("addiction", "tablet") flip to the non-jargon label. In the cross-domain analysis of the general model, nine components rank in the top fifty on both medical and materials JI but outside the top fifty on BoolQ; two heads, L20H10 and L18H26, score 0.769 and 0.735 on materials JI against a full-model 0.739, while sitting near chance on BoolQ.

Three features of this design matter for what follows. The comparator is probably not the specialist's starting point, although public sources disagree. The UltraMedical recipe, as Zhang et al. (2024) describe it for the Llama-3 generation, applies supervised fine-tuning to the base model using completions from gpt-4-turbo and then preference optimisation, and the supervised fine-tuning configuration in the UltraMedical code repository (github.com/TsinghuaC3I/UltraMedical, file src/finetune/config/sft_config_full.yaml) sets "model_name_or_path: meta-llama/Meta-Llama-3.1-8B" - the pretrained base, not Instruct - with the preference stage starting from that fine-tuned output. Against this, the anchor describes its specialist as the result of "medical fine-tuning of Llama-3.1-8B-Instruct" (Keng and Sun, 2026, Section 5.2), and the checkpoint's Hugging Face card declares "base_model: meta-llama/Meta-Llama-3.1-8B-Instruct" in its metadata. This paper follows the training configuration: on it, the specialist and the general model share a pretrained base and differ in their entire post-training pipeline. If the specialist was instead trained from Instruct, the anchor is an own-base comparison, and its negative result bears on medical post-training directly rather than on a choice between two pipelines. The target is not what the treatment was built for: Zhang et al. (2024) evaluate on a suite of medical question-answering benchmarks, on which their Llama-3-8B SFT model averages 75.20 against 68.56 for Llama-3-8B-Instruct. And the reweighting is fitted per

task on labelled task data, which makes it a supervised probe of how much of the gap is a re-weighting of existing signal, not a deployable fix. The study's own conclusion is appropriately framed as a case study of one checkpoint.

Two further readings are worth stating now, because they shape the rest of the paper. JI at 58.8 percent on a balanced set is not far above chance for either model, so the JI contrast is a contrast between two weak classifiers whose difference lies mostly in where they put the boundary. And the gap after reweighting is 2.7 points (56.1 against 58.8), with intervals that overlap; the abstract's phrase that reweighting "closes the gap" is accurate as a direction and generous as a magnitude.

3. Why adaptation was expected to help

The expectation that a domain checkpoint should know its domain's vocabulary better rests on three published lines.

The first is the encoder-era adaptation record. Gururangan et al. (2020) continued pretraining RoBERTa on domain text and reported consistent gains in the biomedical, computer-science and review domains: ChemProt relation classification rose from 81.9 to 84.2 F1 and ACL-ARC citation-intent from 63.0 to 75.4, while AG News, a high-resource news task, stayed at 93.9; adapting on an irrelevant domain instead lowered ChemProt to 79.4. Gu et al. (2020) went further for biomedicine and showed substantial gains from pretraining from scratch on PubMed over continued pretraining of a general model, evaluated on their BLURB benchmark. MatSciBERT, trained on materials-science papers, outperformed the general-science SciBERT on all three of its downstream tasks, and its authors motivated it by the domain's notation and jargon (Gupta et al., 2022).

The second is the frequency dependence of factual knowledge. Kandpal et al. (2023) linked question-answer pairs to pretraining documents through entity linking and showed, correlationally and causally, that accuracy tracks the count of relevant documents, and estimated that closing the gap on rarely supported questions by scale alone would take orders of magnitude more parameters. Chang et al. (2024b) traced how facts are acquired during pretraining and found that each exposure raises the probability of the fact a little and subsequent training erodes it on a power-law schedule, which they connect to poor long-tail performance. If medical jargon is long-tail in web text, a corpus where it recurs should raise the relevant counts.

The third is the clinical-text record. Agrawal et al. (2022) showed that GPT-3 was already a capable zero- and few-shot extractor of clinical variables, but Lehman et al. (2023), on three electronic-health-record tasks, found the opposite at the frontier of 2023: small clinical models pretrained on MIMIC notes, even when fine-tuned on limited labelled data, beat every in-context approach, and pretraining on clinical tokens bought the parameter efficiency to match much larger general models.

Each line is a real result. What each actually supports is narrower than "a domain model knows the domain better". The encoder results compare a model with its own adapted continuation under task fine-tuning; Kandpal et al. (2023) concern document counts, not domain labels; Lehman et al. (2023) concern text - private clinical notes - that a web-trained model has not seen. None of them predicts the outcome of post-training a web-trained 8B base on synthetic medical instructions and comparing it, by prompting, with a sibling post-trained by its maker on general instructions.

4. Five indices of a specialist claim

A published sentence of the form "the specialist beats (or loses to) the generalist on domain task T" leaves five things unstated. Each has, in the published record, reversed a verdict on its own.

4.1 Treatment: what was trained, and on how much new text

"Medical model" names at least six different treatments. Pretraining from scratch on domain text (BioMedLM, trained only on PubMed: Bolton et al., 2024; the clinical T5 models of Lehman et al., 2023). Continued pretraining of a general base on a domain corpus (MEDITRON, 48.1B tokens on Llama-2: Chen et al., 2023; BioMistral on PubMed Central: Labrak et al., 2024; Me-LLaMA, 129B tokens plus instruction tuning: Xie et al., 2024; PMC-LLaMA, 4.8M papers and 30K textbooks plus instruction tuning: Wu et al., 2023b). Domain instruction tuning and preference optimisation of a base (UltraMedical: Zhang et al., 2024). Parameter-efficient prompt tuning of a general instruction model (Med-PaLM: Singhal et al., 2022). Task fine-tuning on labelled examples (Lawma: Dominguez-Olmedo et al., 2024). And system-level specialisation by retrieval and tooling around a general model (the clinical tools of Vishwanath et al., 2026; the search-augmented Med-Gemini of Saab et al., 2024).

These do not have one expected effect, because they act on different parts of the model's behaviour. Raw-corpus continued pretraining adds exposure to domain text but, on the evidence of Cheng et al. (2024a), damages the model's ability to answer prompted questions about what it read, which they repair by converting the corpus into reading-comprehension exercises. Supervised fine-tuning on question-answer pairs mostly teaches a model how to use what it already knows: Gekhman et al. (2024) find that examples carrying new facts are learned more slowly than examples consistent with the model's knowledge and, once learned, raise its hallucination rate roughly linearly; Ovadia et al. (2024) find unsupervised fine-tuning a weaker way to inject knowledge than

retrieval, for facts seen and unseen in pretraining; Allen-Zhu and Li (2023) show in a controlled setting that a fact stored without paraphrastic variety during pretraining can be memorised yet remain unextractable, whatever instruction tuning follows.

Treatment size matters relative to what the base has already seen. Roziere et al. (2023) add 500B tokens of code to Llama 2's 2T, about a quarter of its original budget, and report HumanEval pass@1 rising from 12.2 to 33.5 percent at 7B. Azerbayev et al. (2024) continue Code Llama 7B for 200B tokens on a mixture dominated by their 55B-token Proof-Pile-2 and report MATH accuracy rising from 4.5 to 18.0 percent; Shao et al. (2024) continue a 7B code model for 500B tokens, more than half drawn from a 120B-token math corpus, and report 36.2 percent on MATH for the resulting base model (51.7 percent after instruction tuning and reinforcement learning). MEDITRON's 48.1B tokens on the same Llama-2 base are about 2.4 percent of that base's budget by the arithmetic of the published counts. The Llama-3 generation was pretrained on about 15T tokens (Grattafiori et al., 2024), which shrinks any fixed domain corpus further as a fraction. Dagan et al. (2024) report that specialising the tokenizer of a pretrained code model pays off, in generation speed and effective context, once fine-tuning runs past about 50B tokens - a dose few medical treatments reach.

The base's own state also enters. Springer et al. (2025) find that an OLMo-1B model pretrained on 3T tokens ends more than 2 percent worse on several benchmarks after instruction tuning than its 2.3T-token checkpoint, and trace this to a rising sensitivity of parameters to modification. Kim et al. (2024) find that the breadth of memory sources a model draws on narrows over pretraining and that the narrowing tracks a reduced capacity to acquire and retain new knowledge. Mishra et al. (2024), adapting LLaMA models to materials science, observe that their adapted Llama-2 outperforms their adapted Llama-3 on domain tasks despite Llama-3's stronger starting point, and suggest an adaptation rigidity in heavily trained bases. Ou et al. (2025) find that how readily continued pretraining installs new knowledge depends on how related it is to what the model already knows.

4.2 Comparator: against whom

Three comparators recur and they answer different questions. The starting checkpoint answers "did the treatment help this model" and is the only comparator that isolates the treatment. A post-trained sibling - the general instruction model built by the base's maker - answers "is this domain post-training better than the maker's general post-training", which compares two pipelines differing in data volume, data quality, preference stages and engineering effort. A model of another lineage or generation answers "is this product better than that product".

The anchor's comparator is the second kind, if the specialist's training configuration is right about its starting point (Section 2). So is Dorfner et al.'s (2024) at 8B: their OpenBioLLM-8B was built on the Llama-3-8B base (as tabulated by Jeong et al., 2024b) and is compared with Llama-3-8B-Instruct. Dada et al. (2024) list that model's base as Meta-Llama-3-8B-Instruct instead, but its model card declares "base_model: meta-llama/Meta-Llama-3-8B", agreeing with Jeong et al. At 70B, Dorfner et al.'s comparison is with the own base: Jeong et al. (2024b) pair "Llama-3-70B-Instruct" with "OpenBioLLM-70B" as general model and medical adaptation, and that model's card declares "base_model: meta-llama/Meta-Llama-3-70B-Instruct". When Jeong et al. (2024a; 2024b) insist on comparing each medical model with its own base, the sibling-instruct confound is removed by design.

The comparator also changes the sign of a single specialist's record. Li et al. (2023), re-scoring financial benchmarks, list BloombergGPT's published 5-shot weighted F1 of 75.07 on FiQA sentiment against 50.59 to 53.12 for GPT-NeoX, OPT-66B and BLOOM-176B, the same-era open models BloombergGPT's authors had compared against (Wu et al., 2023a), and against 88.11 for GPT-4 with five shots and 87.09 for a fine-tuned RoBERTa-large. Same specialist, same benchmark: a 22- to 24-point win against its contemporaries and a 13-point loss against a later generalist. Dada et al. (2024) find the comparator's strength organising their own results: biomedical training on the weaker Mistral-7B-v0.1 base showed small gains, while models built on stronger general bases showed clear losses.

4.3 Elicitation: whose prompt, whose format, whose threshold

A comparison by prompting fixes a prompt format, few-shot exemplars, an answer-extraction rule and a decision rule, and each of these interacts with the model. Sclar et al. (2024) find one open model (LLaMA-2-13B) varying by up to 76 accuracy points across meaning-preserving formats and show that format performance correlates only weakly between models, so a single fixed format is not neutral between two checkpoints. Jeong et al. (2024b) show what this does to a specialist claim: with prompts optimised only for the medical model, Clinical-Camel-70B's win rate over its base on knowledge QA goes from 0 to 54.5 percent and its loss rate from 36.4 to 0, which reverses the conclusion.

Post-training is largely a change in elicitation. Zhou et al. (2023) argue from LIMA that a thousand curated examples teach output format and that nearly all knowledge comes from pretraining; Lin et al. (2024) measure the token-level shift between base and aligned models and find it concentrated in stylistic tokens, and match aligned models with base models plus three in-context examples. Cheng et al. (2024a) report that raw-corpus adaptation hurts prompted QA; Jindal et al.

(2024) report that continuing to pretrain an instruction model catastrophically erodes its instruction following, and recover it by adding the weight difference between the maker's instruction and base checkpoints, as Huang et al. (2024) do with their chat vector and as task arithmetic generalises (Ilharco et al., 2023). A specialist that has lost instruction following will lose a prompted comparison for reasons that have nothing to do with domain knowledge. Zhang et al. (2024) tabulate exactly this for OpenBioLLM-8B: AlpacaEval 2 length-controlled win rate of 0.06 percent and MMLU of 44.2, against 22.9 and 68.4 for Llama-3-8B-Instruct; their own UltraMedical 8B holds MMLU at 68.1.

Calibration and decision thresholds are part of elicitation too. Post-training changes calibration in general, not only in specialists: the GPT-4 technical report shows expected calibration error on an MMLU subset rising from 0.007 for the pretrained model to 0.074 after reinforcement-learning post-training (OpenAI, 2023). A binary task scored by the argmax of two option logits measures, jointly, what the model knows and where its prior puts the boundary. Chang et al. (2024a), whose decomposition the anchor uses, find "label-biased components that always predict the same label" among a model's heads, and their reweighting is in part a way of removing such bias.

4.4 Target: which skill

Domain tasks are not one kind of task. Knowledge recall in exam format (MedQA and its relatives) is the target most medical models were built and reported on. Terminology - knowing what a term means, and noticing when a word is being used as one - is the anchor's target. Extraction and classification against labelled schemes - clinical coding, legal annotation, financial sentiment - reward a model fitted to the label set. Generation and dialogue reward instruction following. Multi-step computation - mathematics, code, chemistry calculations - rewards skills that web text supports thinly.

The same treatment can move these in opposite directions. Dada et al. (2024) list, per biomedical model, the gain on multiple-choice QA reported by its own authors next to the change they measure on clinical tasks: MEDITRON-7B, +6.07 and -7.08; MEDITRON-70B, +3.63 and -4.59. Jeong et al. (2024b) find, after supervised fine-tuning on each task, a small average gain for medical models on knowledge QA (1.6 points) and none on clinical-note QA. The UltraMedical recipe that beats its general sibling by 6.6 points on its makers' medical QA average (Llama-3 generation) loses to Llama-3.1-8B-Instruct by 2.0 and 5.7 points on the anchor's terminology tasks (Llama-3.1 generation).

4.5 Statistic: what counts as a difference

Benchmark differences of one or two points on a few thousand items are within sampling noise unless the comparison is paired. Miller (2024) sets out the paired analysis that LLM evaluations usually omit. Jeong et al. (2024a) show that ignoring both prompt sensitivity and statistical uncertainty manufactures apparent medical gains; under their protocol, across seven medical models in the 3-shot setting, the specialist is significantly better in 12.1 percent of model-dataset cases, tied in 49.8 percent, and significantly worse in 38.2 percent. The tie column is the most informative number in that result: half the time, a careful comparison cannot tell the two apart. The anchor reports intervals for both tasks and a paired test for JI, which is better than most of the literature it joins; its JU difference is inside overlapping intervals.

5. Medicine through the indices

The medical literature is large enough to populate most cells of the index space, and reading it that way removes most of the apparent contradiction.

Specialist-favourable results. MEDITRON-70B reports a 6-point absolute gain over the best public baseline of its size and a 3-point gain over the strongest Llama-2 baseline its authors fine-tuned themselves (Chen et al., 2023); PMC-LLaMA reports its 13B model surpassing ChatGPT on medical QA (Wu et al., 2023b); BioMistral reports outperforming the open medical models before it (Labrak et al., 2024); Me-LLaMA reports beating LLaMA and other open medical models in zero-shot and supervised settings, with task tuning surpassing ChatGPT on seven of eight datasets (Xie et al., 2024); MedGemma reports 2.6 to 10 percent improvements over its Gemma 3 bases on out-of-distribution multimodal medical QA while retaining general capability (Sellergren et al., 2025); Med-PaLM 2, combining a stronger base, medical fine-tuning and ensemble refinement, reaches 86.5 percent on MedQA (Singhal et al., 2023), and Med-Gemini 91.1 percent with uncertainty-guided search (Saab et al., 2024). Read through the indices, these are mostly comparisons against weaker or older comparators (index 2), on exam-format recall (index 4), with prompts and fine-tuning chosen by the specialist's authors (index 3). Several of the strongest - Med-PaLM 2, Med-Gemini, MedGemma - are specialisations built by the makers of the base model, which changes the comparator problem: the specialist's maker also controls the generalist's post-training.

Specialist-unfavourable results. Jeong et al. (2024a; 2024b) fix index 2 at the starting checkpoint, fix index 3 by independent prompt selection, and fix index 5 by paired tests; the medical advantage mostly disappears on knowledge QA and reverses on clinical-note QA. Dorfner et al. (2024) choose tasks they consider unlikely to overlap with medical fine-tuning data and find biomedical models mostly behind their general counterparts, most sharply at 8B. Dada et al. (2024) compare twelve biomedical models with their bases on six practical tasks; they report a performance decline for nine of twelve biomedical models, and only

BioMistral-7B-DARE improves on its base across all six tasks. MedConceptsQA tests knowledge of diagnosis, procedure and drug concepts across coding vocabularies and finds clinical models near random while GPT-4 averages 27 to 37 points higher (Shoham and Rappoport, 2024). Hager et al. (2024), in a simulated clinical workflow on 2,400 MIMIC cases, find current models unable to follow guidelines or interpret laboratory results and sensitive to the quantity and order of information, which is a statement about elicitation as much as about knowledge.

The probing results sit underneath both. BioLAMA's factual probes of biomedical encoders reached at most 18.51 percent Acc@5, and Sung et al. (2021) found predictions largely determined by the prompt template rather than the subject; MedLAMA's probes of UMLS relations reached at most 3 percent acc@10 before a contrastive re-wiring raised it to 28 percent (Meng et al., 2022). Domain models have been hard to read for domain facts for as long as anyone has tried, and the reading has always depended on the probe.

Where the anchor sits. Treatment: domain instruction and preference tuning from the base, with synthetic completions. Comparator: post-trained sibling on the evidence of the training configuration, though the anchor and the model card describe the specialist as a fine-tune of Instruct (Section 2). Elicitation: lettered-option prompts with argmax over option logits; the ordering survives five JI and three JU prompt variants (Appendix H), but every variant is shared by both models rather than selected per model, the decision rule is the same raw argmax over the A/B logits for both with no threshold calibration, and under the persona variant the JU gap shrinks to 0.3 points. Target: terminology, which neither model was tuned for. Statistic: intervals and a paired JI test. It is therefore evidence about one cell: whether the UltraMedical pipeline gives an 8B model better terminology behaviour than Meta's general pipeline, under prompts and a decision rule shared by both models. Its answer, no, is consistent with the post-2024 record in that cell. It does not say what continued pretraining on clinical notes would do for terminology, or what UltraMedical post-training does relative to the bare base.

Figure 1 shows the pattern this section describes. Against its own starting point, a domain treatment is positive in every published row the figure draws, large where the treatment is large and the skill is thin in web text, and small elsewhere. Against a post-trained sibling, the sign depends on the target. Against another lineage, it depends on which generation the comparator comes from. Panel (b) shows one specialist, Clinical-Camel-70B, moving from 0 percent wins and 36.4 percent losses to 54.5 percent wins and no losses when the prompt is tuned for it alone; Jeong et al. (2024b) report the same direction for every pair they examined, writing that "we consistently observe this across all LLM/VLM pairs and medical QA tasks".

6. Where specialisation pays reliably

The negative medical results are often read as a general verdict on specialisation. The record outside medicine does not support that reading, and the cases where specialisation pays share features that the medical treatments mostly lack.

Code and mathematics. Continued pretraining on code and mathematics produces the largest published specialist gains (Figure 1a). Code Llama's 500B code tokens took Llama 2 7B from 12.2 to 33.5 percent pass@1 on HumanEval, and its authors note that the 7B code model roughly matches Llama 2 70B on Python benchmarks (Roziere et al., 2023). Llemma lifted Code Llama 7B on MATH from 4.5 to 18.0 percent and on GSM8k from 10.5 to 36.4 percent (Azerbayev et al., 2024). DeepSeekMath's 7B model reached 51.7 percent on MATH, which its authors attribute to 120B math tokens selected from Common Crawl and a new reinforcement-learning method (Shao et al., 2024). These treatments are tens to hundreds of billions of tokens, a substantial fraction of the base's original budget; the target skills are multi-step procedures under-represented in general web text; and the comparator is the starting checkpoint. Biderman et al. (2024), studying exactly these two domains, find that full fine-tuning on 20B tokens of continued pretraining learns substantially more than low-rank adaptation and that low-rank adaptation forgets less outside the domain, which suggests that the domain gain and the general loss both grow with how far the treatment moves the weights.

Private and proprietary text. ChipNeMo combined a domain-adapted tokenizer, continued pretraining on proprietary hardware code and specifications, alignment on largely public general chat data plus a small amount of domain instruction data, and domain retrieval, and its 70B model outperformed GPT-4 on two of its three chip-design use cases without degrading generic capability, and its authors stress the value of domain-adaptive pretraining for low-volume or proprietary programming languages (Liu et al., 2023). BloombergGPT was trained on a 363B-token financial dataset built from Bloomberg's own sources plus 345B general tokens and outperformed existing models on financial tasks (Wu et al., 2023a). Lehman et al.'s (2023) clinical models were pretrained on MIMIC notes, which are credentialed and not on the open web. In each case the domain corpus carries text that a web-trained generalist is unlikely to have seen.

Labelled tasks. When the target is a fixed labelling scheme with examples, small fine-tuned models win. Dominguez-Olmedo et al. (2024) build 260 legal annotation tasks and find small, lightly fine-tuned models outperforming commercial models, usually with a few hundred to a thousand labelled examples. Bucher and Martini (2024) find fine-tuned smaller models

beating zero-shot GPT-3.5, GPT-4 and Claude Opus on every text-classification task they tried. The specialist here is a task specialist, not a domain specialist, and its advantage comes from supervision rather than from domain text.

Reading-comprehension and synthetic transformations. Where the treatment is designed around the knowledge-extraction problem of Section 4.1, adaptation works better. Cheng et al. (2024a) convert raw domain corpora into reading-comprehension exercises and report gains across biomedicine, finance and law, with a 7B model competitive with BloombergGPT-50B; Cheng et al. (2024b) report that continual pretraining with synthesised instruction-response pairs lets Llama3-8B match or exceed Llama3-70B; Yang et al. (2025) synthesise diverse text around the entities of a small corpus and show that continued pretraining on it lets a model answer questions about documents it no longer has. These are the treatments that the controlled results of Allen-Zhu and Li (2023) say should work, and they are not the treatments most published medical models used.

7. Finance, law, chemistry and materials

The other professional domains show the same structure with fewer studies.

Finance. The BloombergGPT comparison in Section 4.2 is the clearest single case of comparator dependence in the literature. Separately, Tang and Yang (2024) find that general-purpose embedding models drop sharply on a finance benchmark built from domain text, after controlling for dataset complexity, and that their ranking on the general benchmark does not predict their ranking on the financial one; for retrieval representations, as opposed to generation, the domain gap is measurable.

Law. SaulLM-7B, built on Mistral 7B and trained on an English legal corpus of more than 30B tokens, is reported by its authors as state of the art in understanding and processing legal documents (Colombo et al., 2024). LegalBench's 162 expert-built tasks span six types of legal reasoning and were evaluated across 20 models (Guha et al., 2023), giving the domain a shared yardstick that medicine's exam benchmarks only partly provide. The strongest legal specialist result is task-level (Dominguez-Olmedo et al., 2024), not domain-level.

Chemistry. ChemBench's 2,700-plus questions found the best general models outperforming the best human chemists in its study on average, while struggling on some basic tasks and giving overconfident predictions (Mirza et al., 2024). A generalist beating domain experts while being miscalibrated is the anchor's calibration finding seen from the other side: calibration problems are not specific to specialists.

Materials. The anchor's materials analysis finds components in a general model that carry jargon-identification signal across medicine and materials. Mishra et al. (2024) find that their materials-adapted LLaMA models excel at materials information extraction and crystal generation, and that the older base adapted better, which, together with Springer et al. (2025), is evidence that the comparator-generation index interacts with the treatment index: newer bases both know more and move less.

8. What the mechanism result licenses

The anchor's mechanistic claim has two parts: the specialist does not reorganise its jargon knowledge but re-weights a small set of pre-existing components, and suppressing those components recovers part of the gap. A substantial literature predicts the first part for fine-tuning in general.

Jain et al. (2024) study fine-tuning in controlled settings and find that it rarely changes the capabilities a model learned in pretraining; it learns a thin "wrapper" over them, and further fine-tuning can revive a hidden capability in a few steps. Prakash et al. (2024) find that fine-tuning on mathematics improves entity tracking by strengthening the same circuit the base model uses rather than building a new one. Lin et al. (2024) and Zhou et al. (2023) find alignment tuning shifting mostly stylistic tokens. Zheng et al. (2025) argue that much of what continual learning reports as forgetting is a loss of task alignment rather than of knowledge, recoverable with little training. Chang et al. (2024a) find that individual components can outperform the whole model on classification tasks and that re-weighting them from 24 labelled examples improves Llama-2-7B by 6.0 points over 24-shot in-context learning across eight tasks.

Read against this literature, the anchor's finding - similar component rankings, heavier weight on a jargon-favouring subset, a boundary shifted toward "jargon", and a partial recovery from re-weighting - is what one would expect if medical post-training had installed a domain prior: in a clinical context, lean toward treating words as technical. That prior is useful on exam questions, where the technical reading is usually the right one, and harmful on a balanced identification task whose negatives are ordinary words sitting in clinical text. The reading is consistent with the two examples the anchor shows ("addiction" and "tablet" misclassified as jargon, then corrected), with the overconfidence in the high bins, and with the asymmetry of the reweighting (it helps the specialist and hurts the general model, whose components were not over-weighted).

This paper therefore reads the specialist's JI deficit as primarily a shifted decision prior, not lost knowledge, and holds that reading at the level of a hedged inference. Four limits apply. The decomposition measures each component's direct contribution to the output logits and, as the anchor notes, not its indirect influence through later layers, so a component can be important to the

computation without scoring well on its own. The reweighting is fitted on the JI training split, so its gain shows how much of the deficit lies in a linear reweighting of existing directions, not that the specialist knows the same things. The JU difference, which is the closer test of knowing what terms mean, is two points inside overlapping intervals, and reweighting moves JU by 0.1 points for the specialist and 0.8 for the general model, both inside the intervals. And the evidence is one pair at one scale.

The prior-shift reading makes a prediction the anchor's data could test: fitting a single scalar bias on the JI logit margin for each model, rather than hundreds of component weights, should recover most of the 3.0-point reweighting gain if the deficit is mostly a threshold. If it recovers little, the deficit lies in which components carry signal, and the reorganisation reading gains ground.

The cross-domain heads carry a separate implication. If L20H10 and L18H26 in the general model identify jargon in materials text about as well as the whole model does, while sitting near chance on BoolQ, then "jargonness" - the property of being a term of art in the surrounding register - is at least partly represented without domain treatment. That fits Kandpal et al.'s (2023) picture in which what matters is exposure, and the web already contains a great deal of medical text.

9. A marginal-information account

The pattern across Sections 5 to 7 suggests a hypothesis that this paper states as its own and does not claim the literature has tested: a domain treatment helps a model on a target in proportion to the information about that target that the treatment carries and the base lacks, and the information must arrive in a form the model can later extract.

Three quantities enter. The first is novelty: how much of the domain corpus is text the base has not effectively seen. Openly published biomedical literature is plausibly well represented in web-scale pretraining, though for closed corpora nobody outside the model's maker can check; credentialed clinical notes, Bloomberg's own data and ChipNeMo's proprietary hardware code are absent from open crawls by construction. The second is dose: tokens of the treatment relative to the base's budget, and relative to the forgetting the treatment induces (Ibrahim et al., 2024, show that re-warming, re-decaying and replay let continual pretraining match full retraining at 405M and 10B parameters; Que et al., 2024, fit the trade-off between general and domain loss as a function of mixture ratio). The third is form: whether the treatment presents knowledge with the variety that makes it extractable (Allen-Zhu and Li, 2023; Cheng et al., 2024a; Yang et al., 2025), or as instruction-response pairs that mostly teach the model to use what it has (Gekhman et al., 2024).

Under this account, synthetic medical instruction tuning of a 15T-token base scores low on novelty and dose for terminology and high on form only for the question types it was built from, so its expected effect on terminology is small and its main effect is on priors and format, as Section 8 reads the anchor. Code and math continued pretraining scores high on all three. Clinical-note pretraining scores high on novelty. The account predicts that the medical results would turn positive with a treatment that is large, novel and extractable - pretraining on private clinical text with knowledge-augmented formats - and that the anchor's comparison would come out the other way for the rarest terms in README if the specialist had been continued-pretrained on clinical notes.

What this account cannot yet do is quantify novelty. The pretraining corpora of the Llama-3 generation are not published; for open corpora one could count term frequencies, but no study has related specialist gains to term-level pretraining frequency. The account is offered as the most economical reading of the published pattern, at the level of a marked conjecture.

10. Consolidated evidence

Table 1 collects the measured values on which the argument rests, one row per published result, with the comparator type stated so that rows of different kinds are not read as one series. Figure 1 plots the subset that reduces to a single difference.

Three regularities stand out. Every row that scores a specialist against its own starting checkpoint on the kind of task the treatment targeted shows a gain (rows 9, 12 and 13 in their first values, and 18 to 21), although rows 6 and 7 show that under independent prompting and paired tests the medical gains mostly vanish even against the own base. Against a sibling, the sign follows the target: the specialist leads on exam-format medical QA (row 4, on its makers' own suite, and row 27, scored by a third party) and is behind or within two points on other targets (rows 1 to 3, 5, 10); against a later-generation model it is behind or within two points (rows 15 to 17, 23). And the rows with the largest positive effects (20, 21, 22, 24, 25) are treatments with a large dose, novel text, or labels.

Table 2 maps fourteen studies onto four of the five indices; the statistic is discussed in Section 4.5. Cells are coarse where the source reports its protocol coarsely.

11. A procedure for reading a specialist claim

Algorithm 1 states, as a decision procedure, what Sections 4 to 9 argue: which labels a published specialist-versus-generalist result earns, which verdict about domain data it licenses, and what sign the marginal-information hypothesis predicts for it.

Applied to the anchor, the procedure returns "pipeline comparison" (if the training configuration is right about

Table 1. Published specialist-generalist results, by source. "Own base" is the specialist's starting checkpoint; "sibling" is the base maker's general post-trained model (for rows 1 to 3, on the evidence of the specialist's training configuration; its model card names Instruct as the base, which would make them own-base rows, Section 2); "other" is a different lineage or generation. Values are in each source's own metric (percent unless stated).

#	Source	Specialist vs comparator (type)	Task	Specialist / comparator	Note
1	Keng and Sun (2026)	L3.1-8B-UltraMedical vs L3.1-8B-Instruct (sibling)	Jargon definitions (JU)	74.1 / 76.1	Intervals overlap
2	Keng and Sun (2026)	same pair	Jargon identification (JI)	53.1 / 58.8	Paired test significant
3	Keng and Sun (2026)	same, after component reweighting	JI	56.1 / 58.8	Weights fit on JI train split
4	Zhang et al. (2024)	L3-8B UltraMed SFT vs L3-8B-Instruct (sibling)	Medical QA suite, mean	75.20 / 68.56	Makers' benchmarks
5	Zhang et al. (2024)	OpenBioLLM-8B vs L3-8B-Instruct (sibling)	AlpacaEval 2 LC	0.06 / 22.9	MMLU 44.2 / 68.4
6	Jeong et al. (2024a)	7 medical LLMs vs own base	Medical QA, 3-shot	Win 12.1, loss 38.2	Tie 49.8
7	Jeong et al. (2024b)	10 medical LLMs vs own base	Clinical-note QA, 3-shot	Win 26.7, loss 56.7	Tie 16.7
8	Jeong et al. (2024b)	Clinical-Camel-70B vs own base	Knowledge QA win rate	0 → 54.5	Prompt tuned for specialist only
9	Jeong et al. (2024b)	Medical LLMs vs own base, after SFT	Knowledge QA	+1.6 mean	No gain on clinical notes
10	Dorfner et al. (2024)	OpenBioLLM-8B vs L3-8B-Instruct (sibling)	NEJM cases	30 / 64.3	Unseen-data selection
11	Dorfner et al. (2024)	OpenBioLLM-70B vs L3-70B-Instruct (own base)	JAMA cases	66.4 / 65	
12	Dada et al. (2024)	MEDITRON-7B vs Llama-2-7B (own base)	MCQA (reported) ; CLUE L1	+6.07 ; -7.08	Sign flips with task
13	Dada et al. (2024)	MEDITRON-70B vs Llama-2-70B (own base)	MCQA (reported) ; CLUE L1	+3.63 ; -4.59	
14	Chen et al. (2023)	MEDITRON-70B vs best public same size (other)	Medical QA	+6 absolute	Authors' protocol
15	Shoham and Rappoport (2024)	Clinical LLMs vs GPT-4 (other)	MedConceptsQA	GPT-4 +27 to +37	Clinical LLMs near random
16	Nori et al. (2023)	GPT-4 + Medprompt vs specialist SOTA (other)	MedQA	>90; 27% fewer errors	Generalist wins
17	Vishwanath et al. (2026)	2 clinical tools vs 3 frontier LLMs (other)	MedQA, HealthBench, RCQ	Frontier ahead on all 3	Matters Arising filed
18	Gururangan et al. (2020)	RoBERTa+DAPT vs RoBERTa (own base)	ChemProt F1	84.2 / 81.9	Off-domain: 79.4
19	Gururangan et al. (2020)	RoBERTa+DAPT vs RoBERTa (own base)	ACL-ARC F1	75.4 / 63.0	
20	Roziere et al. (2023)	Code Llama 7B vs Llama 2 7B (own base)	HumanEval pass@1	33.5 / 12.2	500B code tokens
21	Azerbayev et al. (2024)	Llemma 7B vs Code Llama 7B (own base)	MATH, 4-shot	18.0 / 4.5	200B tokens
22	Li et al. (2023)	BloombergGPT vs GPT-NeoX (other, same era)	FiQA weighted F1	75.07 / 50.59	5-shot
23	Li et al. (2023)	BloombergGPT vs GPT-4 (other, later)	FiQA weighted F1	75.07 / 88.11	Fine-tuned RoBERTa-L 87.09
24	Lehman et al. (2023)	Clinical T5 vs in-context LLMs to 175B (other)	3 EHR tasks	Specialist ahead	Pretrained on MIMIC
25	Dominguez-Olmedo et al. (2024)	Small fine-tuned vs commercial models (other)	260 legal annotation tasks	Specialist ahead	Hundreds of labels
26	OpenAI (2023)	GPT-4 pretrained vs post-trained	MMLU subset, ECE	0.007 / 0.074	Post-training cost, not domain
27	Zhang et al. (2024)	OpenBioLLM-8B vs L3-8B-Instruct (sibling)	Medical QA suite, mean	72.48 / 68.56	Third-party protocol

the starting point; an own-base comparison if the model card is), "elicitation-confounded" in a qualified sense (the ordering holds across the appendix's prompt variants, but each variant and the argmax decision rule are shared by both models rather than chosen per model), "transfer test" (terminology is not the treatment's target), "unresolved" for JU and resolved for JI, no licensed verdict about medical data, and a predicted sign of near zero or negative, which is what was observed. Applied to Jeong et al. (2024b), it returns a licensed verdict, negative on clinical-note QA, for the treatments those ten models received. Applied to Roziere et al. (2023), whose comparison with the starting checkpoint shares one evaluation setup, it returns "elicitation-confounded"

Table 2. Fourteen studies on four of the five indices (treatment, comparator, elicitation, target).

Study	Treatment	Comparator	Elicitation	Target	Verdict
Keng and Sun (2026)	Domain SFT + pref.	Sibling (start contested)	Shared prompts (5 JI, 3 JU), argmax	Terminology	Specialist worse
Zhang et al. (2024)	Domain SFT + pref.	Sibling	Authors' setup	Exam QA	Specialist better
Jeong et al. (2024a; 2024b)	CPT and SFT, 10 models	Own base	Per-model prompt search	Exam and note QA	Mostly tie or worse
Dorfner et al. (2024)	Mixed	Mixed (8B sibling, 70B own base)	Shared benchmarks	Cases, CLUE	Mostly worse
Dada et al. (2024)	CPT and SFT, 12 models	Own base	Shared tasks	Six clinical tasks	Mostly worse
Chen et al. (2023)	CPT, 48.1B tokens	Own base, others	Authors' setup	Exam QA	Better
Nori et al. (2023)	None (prompting)	Specialists	Medprompt	Exam QA	Generalist better
Vishwanath et al. (2026)	Product systems	Frontier LLMs	Product defaults	Exam, rubric, queries	Generalist better
Lehman et al. (2023)	Scratch, clinical notes	Larger LLMs	ICL vs fine-tuning	EHR tasks	Specialist better
Gururangan et al. (2020)	CPT, encoder	Own base	Task fine-tuning	Classification	Better
Roziere et al. (2023)	CPT, 500B tokens	Own base	Shared	Code synthesis	Much better
Li et al. (2023)	363B fin. + 345B gen. data	Same era; GPT-4	Few-shot	Finance tasks	Sign set by comparator
Liu et al. (2023)	Tokenizer + CPT + align	Own base; GPT-4	Authors' setup	Chip-design tasks	Better on 2 of 3
Dominguez-Olmedo et al. (2024)	Task fine-tuning	Commercial LLMs	Prompting vs FT	Legal annotation	Specialist better

Algorithm 1. Classifying a specialist-versus-generalist claim

```
input: specialist S, comparator G, task T, reported difference d
1  treat <- kind of treatment: scratch | CPT | transformed CPT | domain SFT/pref | task FT | system
2  dose  <- treatment tokens / base pretraining tokens   (UNKNOWN if either is unpublished)
3  novel <- is the treatment text absent from public web text?  yes | no | unknown
4  labels <- {}
5  if G is not the checkpoint S started from:
6      add "pipeline comparison" to labels       # d mixes treatment with G's post-training
7      if G is a later generation or another lineage: add "product comparison"
8  if one prompt/format/threshold is shared, or tuned for only one model:
9      add "elicitation-confounded"
10     if S's instruction following is unmeasured: add "format check missing"
11 tgt <- recall | terminology | labelled extraction | generation | computation
12 if tgt differs from the target S was trained toward: add "transfer test"
13 if no paired test, or intervals overlap: mark d "unresolved"
14 licensed <- G is S's start AND prompts chosen per model AND paired test significant
15 if licensed: verdict <- sign(d) as evidence about the domain treatment
16 else: verdict <- none about domain data; report labels and missing controls
17 predicted <- positive if (dose large OR novel = yes OR treat = task FT) and tgt matches,
18              near zero or negative otherwise                # Section 9 hypothesis
19 return labels, verdict, predicted, missing controls
```

and no formal license, together with a predicted positive sign for 500B tokens of code on code synthesis - and a 21-point effect that no plausible prompt artefact would erase.

12. The strongest case against this reading

Three objections deserve a full statement.

The careful studies already fixed the comparator and the prompt, and medicine still lost. This is correct and the paper does not contest it. Jeong et al. (2024a; 2024b) compare each model with its own base, choose prompts per model and test differences pairwise, and Dada et al. (2024) compare with bases on practical tasks; both find the typical open medical model no better and often worse. Those are the best-controlled verdicts in the medical literature. What they establish is a statement about the treatments those models received - mostly tens of billions of tokens of public biomedical literature, or instruction data, on bases trained on trillions of web tokens - not about domain adaptation as such. The paper's claim is that the negative result has a location, and that the location is the low-dose, low-novelty corner of Section 9.

For anyone choosing a model, the product comparison is the right one. Also correct. A clinician or developer chooses between downloadable or purchasable systems, not between a treatment and its counterfactual, and on the product comparison frontier generalists currently win: GPT-4 with general prompting beat the specialist systems of its time on every MultiMedQA dataset (Nori et al., 2023), a reasoning model beat that (Nori et al., 2024), and three frontier

models outperformed two commercial clinical tools on exam questions, HealthBench and real physician queries (Vishwanath et al., 2026). That last study's article page lists a Matters Arising, "Limited benchmarks constrain the conclusions of a general-purpose versus clinical AI comparison", with the authors' reply dated September 2026, so even the product verdict is disputed at the level of benchmark coverage. The anchor's sibling comparison is, likewise, the right comparison for someone deciding whether to download UltraMedical rather than Llama-3.1-Instruct. The paper's point is only that such comparisons are not evidence about what domain data do, and that they are often read as if they were.

The marginal-information account is unfalsifiable as stated. Novelty cannot be measured for closed corpora, and dose and form are loosely defined. The objection is fair for the closed models. For open corpora the account makes a sharp, checkable prediction (Section 13, Study 3): specialist gains should rise with the rarity of a term in the base's pretraining data. If they do not, the account is wrong.

13. Studies that would settle the open part

Five studies, each runnable with open weights and published data, would assign the effects this paper has separated.

Study 1: a treatment ladder on one base. Take Llama-3.1-8B base and build, with matched compute: (a) the base; (b) Meta's Instruct; (c) the base with UltraMedical's SFT and preference stages; (d) the base with a general instruction set of the same size, format and teacher as UltraMedical's, but without medical content; (e) the base continued-pretrained on medical text, then given the instruction residual of Jindal et al. (2024). Score all five on JU, JI, exam QA and clinical-note QA under per-model prompt selection and paired tests (Jeong et al., 2024b). The difference (c) minus (d) is the effect of medical content holding the post-training pipeline fixed, which neither the anchor nor most of the literature measures; (b) minus (d) is the effect of Meta's pipeline over a comparable open one; (e) minus (a) is the effect of domain text with instruction following restored.

Study 2: threshold versus components. On the anchor's JI data, fit a single scalar bias on each model's jargon-minus-non-jargon logit margin using the same 1,200 training items, and compare the recovered accuracy with the 3.0 points recovered by component reweighting. If a scalar recovers most of it, the specialist's deficit is a decision prior; if not, it lies in the components. Report expected calibration error before and after each.

Study 3: frequency-stratified terminology. Bin README terms by their frequency in an open pretraining corpus, score general and medical models of a fully open family on each bin, and test whether specialist minus generalist rises toward the rare end. This is the direct test of the marginal-information hypothesis and of Kandpal et al.'s (2023) mechanism applied to vocabulary.

Study 4: a generation ladder. Apply one fixed medical recipe to two generations of one model family at the same size and compare gains over each base, testing whether more heavily trained bases gain less from the same treatment, as Mishra et al. (2024) observe for materials and Springer et al. (2025) lead one to expect from parameter sensitivity.

Study 5: form at fixed content. Continue-pretrain one base on one medical corpus in three forms - raw, converted to reading-comprehension exercises (Cheng et al., 2024a), and expanded by entity-centred synthesis (Yang et al., 2025) - and score terminology and recall. If form matters at fixed content, the medical literature's null results are partly a statement about raw-text continued pretraining.

Each study should report, at minimum, win, tie and loss rates under paired tests (Miller, 2024), the spread across plausible prompt formats (Sclar et al., 2024), an instruction-following check, and calibration on the domain task.

14. What is not known

No published study known to the author measures the effect of medical content on terminology while holding the post-training pipeline fixed. No published study known to the author relates a specialist's gain on a term to the term's frequency in the base's pretraining data. The fraction of public biomedical literature present in the pretraining corpora of current open-weight bases is not published. Whether the anchor's specialist deficit is a scalar threshold shift or a component-level reorganisation has not been tested; the evidence in hand is compatible with both, and favours the first. Whether the jargon-sensitive heads found in the general model are causally necessary for jargon identification, as opposed to scoring well on their direct effects, is not known. How the verdicts of Table 1 change at frontier scale under the controls of Section 13 is not known, because no maker of a frontier model has published a matched medical-versus-general contrast on its own base.

15. Limitations

This analysis consolidates published numbers that were measured under different protocols, and Figure 1 places differences in different metrics on one axis; only signs and rough magnitudes compare across rows, as the caption states. The selection of studies was organised around the anchor's question and over-represents medicine; law, finance and chemistry are covered by a handful of studies each. Several cited results come from preprints. The paper's two interpretive claims - that the anchor's JI deficit is mainly a shifted prior, and the

marginal-information account - are a hedged inference and a conjecture respectively, and Sections 8 and 13 state how each could fail. The author's coding of Table 2's cells relies on what each source reports about its protocol, which is sometimes incomplete; coarse cells reflect that.

16. Conclusion

A medical checkpoint that loses to its general sibling on medical vocabulary is a real and useful result. It shows that one widely used domain post-training pipeline does not improve, and may harm, an 8B model's handling of terms it was not trained to define, and the anchor's component analysis gives a plausible reason: the pipeline sharpens a prior toward reading clinical words as technical, which is right on exams and wrong on ordinary words in clinical text. It does not show that medical data fail to teach medical vocabulary, because, if the specialist's training configuration is right about its starting point, the comparison changes the whole post-training pipeline at once, and the treatment carried little text the base is likely not to have seen.

Read across the literature, specialisation behaves the way a treatment should and not the way a label would. It pays when the dose is large, when the text is new to the model, when the knowledge arrives in an extractable form, or when the task comes with labels; it does little or harm when a modest instruction or continued-pretraining treatment is laid on a base that has probably read much of the public domain literature already, and is then judged by prompt against a stronger sibling. The disagreement between "general models already know medicine" and "domain data still matter" is mostly a disagreement about which of those cells one is standing in. The controlled ladder of Study 1 and the frequency test of Study 3 would show which.

References

Agrawal, M., Hegselmann, S., Lang, H., Kim, Y., & Sontag, D. (2022). Large Language Models are Few-Shot Clinical Information Extractors. EMNLP 2022. arXiv:2205.12689. doi:10.48550/arXiv.2205.12689.

Allen-Zhu, Z., & Li, Y. (2023). Physics of Language Models: Part 3.1, Knowledge Storage and Extraction. arXiv:2309.14316. doi:10.48550/arXiv.2309.14316.

Azerbayev, Z., Schoelkopf, H., Paster, K., Dos Santos, M., McAleer, S., Jiang, A. Q., Deng, J., Biderman, S., & Welleck, S. (2024). Llemma: An Open Language Model For Mathematics. ICLR 2024. arXiv:2310.10631. doi:10.48550/arXiv.2310.10631.

Biderman, D., Portes, J., Ortiz, J. J. G., Paul, M., Greengard, P., Jennings, C., King, D., Havens, S., Chiley, V., Frankle, J., Blakeney, C., & Cunningham, J. P. (2024). LoRA Learns Less and Forgets Less. arXiv:2405.09673. doi:10.48550/arXiv.2405.09673.

Bolton, E., Venigalla, A., Yasunaga, M., Hall, D., Xiong, B., Lee, T., Daneshjou, R., Frankle, J., Liang, P., Carbin, M., & Manning, C. D. (2024). BioMedLM: A 2.7B Parameter Language Model Trained On Biomedical Text. arXiv:2403.18421. doi:10.48550/arXiv.2403.18421.

Bucher, M. J. J., & Martini, M. (2024). Fine-Tuned 'Small' LLMs (Still) Significantly Outperform Zero-Shot Generative AI Models in Text Classification. arXiv:2406.08660. doi:10.48550/arXiv.2406.08660.

Chang, T.-Y., Thomason, J., & Jia, R. (2024a). When Parts Are Greater Than Sums: Individual LLM Components Can Outperform Full Models. EMNLP 2024. arXiv:2406.13131. doi:10.48550/arXiv.2406.13131.

Chang, H., Park, J., Ye, S., Yang, S., Seo, Y., Chang, D.-S., & Seo, M. (2024b). How Do Large Language Models Acquire Factual Knowledge During Pretraining? arXiv:2406.11813. doi:10.48550/arXiv.2406.11813.

Chen, Z., Hernandez Cano, A., Romanou, A., Bonnet, A., Matoba, K., Salvi, F., et al. (2023). MEDITRON-70B: Scaling Medical Pretraining for Large Language Models. arXiv:2311.16079. doi:10.48550/arXiv.2311.16079.

Cheng, D., Huang, S., & Wei, F. (2024a). Adapting Large Language Models to Domains via Reading Comprehension. arXiv:2309.09530. doi:10.48550/arXiv.2309.09530.

Cheng, D., Gu, Y., Huang, S., Bi, J., Huang, M., & Wei, F. (2024b). Instruction Pre-Training: Language Models are Supervised Multitask Learners. EMNLP 2024. arXiv:2406.14491. doi:10.48550/arXiv.2406.14491.

Colombo, P., Pires, T. P., Boudiaf, M., Culver, D., Melo, R., Corro, C., Martins, A. F. T., Esposito, F., Raposo, V. L., Morgado, S., & Desa, M. (2024). SaulLM-7B: A pioneering Large Language Model for Law. arXiv:2403.03883. doi:10.48550/arXiv.2403.03883.

Dada, A., Bauer, M., Contreras, A. B., Koras, O. A., Seibold, C. M., Smith, K. E., & Kleesiek, J. (2024). Does Biomedical Training Lead to Better Medical Performance? arXiv:2404.04067. doi:10.48550/arXiv.2404.04067.

Dagan, G., Synnaeve, G., & Roziere, B. (2024). Getting the most out of your tokenizer for pre-training and domain adaptation. arXiv:2402.01035. doi:10.48550/arXiv.2402.01035.

Dominguez-Olmedo, R., Nanda, V., Abebe, R., Bechtold, S., Engel, C., Frankenreiter, J., Gummadi, K., Hardt, M., & Livermore, M. (2024). Lawma: The Power of Specialization for Legal Annotation. arXiv:2407.16615. doi:10.48550/arXiv.2407.16615.

Dorfner, F. J., Dada, A., Busch, F., Makowski, M. R., Han, T., Truhn, D., Kleesiek, J., Sushil, M., Lammert, J., Adams, L. C., & Bressem, K. K. (2024). Biomedical Large Languages Models Seem not to be Superior to Generalist Models on Unseen Medical Data. arXiv:2408.13833. doi:10.48550/arXiv.2408.13833.

Gekhman, Z., Yona, G., Aharoni, R., Eyal, M., Feder, A., Reichart, R., & Herzig, J. (2024). Does Fine-Tuning LLMs on New Knowledge Encourage Hallucinations? EMNLP 2024. arXiv:2405.05904. doi:10.48550/arXiv.2405.05904.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., et al. (2024). The Llama 3 Herd of Models. arXiv:2407.21783. doi:10.48550/arXiv.2407.21783.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2020). Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. arXiv:2007.15779. doi:10.48550/arXiv.2007.15779.

Guha, N., Nyarko, J., Ho, D. E., Re, C., Chilton, A., Narayana, A., et al. (2023). LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models. arXiv:2308.11462. doi:10.48550/arXiv.2308.11462.

Gupta, T., Zaki, M., Krishnan, N. M. A., & Mausam. (2022). MatSciBERT: A Materials Domain Language Model for Text Mining and Information Extraction. npj Computational Materials. arXiv:2109.15290. doi:10.48550/arXiv.2109.15290.

Gururangan, S., Marasovic, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. ACL 2020. arXiv:2004.10964. doi:10.48550/arXiv.2004.10964.

Hager, P., Jungmann, F., Holland, R., Bhagat, K., Hubrecht, I., Knauer, M., Vielhauer, J., Makowski, M., Braren, R., Kaissis, G., & Rueckert, D. (2024). Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nature Medicine. doi:10.1038/s41591-024-03097-1.

Huang, S.-C., Li, P.-Z., Hsu, Y.-C., Chen, K.-M., Lin, Y. T., Hsiao, S.-K., Tsai, R. T.-H., & Lee, H.-y. (2024). Chat Vector: A Simple Approach to Equip LLMs with Instruction Following and Model Alignment in New Languages. ACL 2024. arXiv:2310.04799. doi:10.48550/arXiv.2310.04799.

Ibrahim, A., Therien, B., Gupta, K., Richter, M. L., Anthony, Q., Lesort, T., Belilovsky, E., & Rish, I. (2024). Simple and Scalable Strategies to Continually Pre-train Large Language Models. arXiv:2403.08763. doi:10.48550/arXiv.2403.08763.

Ilharco, G., Ribeiro, M. T., Wortsman, M., Gururangan, S., Schmidt, L., Hajishirzi, H., & Farhadi, A. (2023). Editing Models with Task Arithmetic. arXiv:2212.04089. doi:10.48550/arXiv.2212.04089.

Jain, S., Kirk, R., Lubana, E. S., Dick, R. P., Tanaka, H., Grefenstette, E., Rocktaschel, T., & Krueger, D. S. (2024). Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks. arXiv:2311.12786. doi:10.48550/arXiv.2311.12786.

Jeong, D. P., Garg, S., Lipton, Z. C., & Oberst, M. (2024a). Medical Adaptation of Large Language and Vision-Language Models: Are We Making Progress? EMNLP 2024. arXiv:2411.04118. doi:10.48550/arXiv.2411.04118.

Jeong, D. P., Mani, P., Garg, S., Lipton, Z. C., & Oberst, M. (2024b). The Limited Impact of Medical Adaptation of Large Language and Vision-Language Models. arXiv:2411.08870. doi:10.48550/arXiv.2411.08870.

Jindal, I., Badrinath, C., Bharti, P., Vinay, L., & Sharma, S. D. (2024). Balancing Continuous Pre-Training and Instruction Fine-Tuning: Optimizing Instruction-Following in LLMs. arXiv:2410.10739. doi:10.48550/arXiv.2410.10739.

Kandpal, N., Deng, H., Roberts, A., Wallace, E., & Raffel, C. (2023). Large Language Models Struggle to Learn Long-Tail Knowledge. arXiv:2211.08411. doi:10.48550/arXiv.2211.08411.

Keng, D., & Sun, Z. (2026). Domain-Specific Jargon in Large Language Models: A Comparative Analysis between General-Purpose and Specialist Models. arXiv:2609.13556. doi:10.48550/arXiv.2609.13556.

Kim, J., Lee, H., Cho, H., Jang, J., Hwang, H., Won, S., Ahn, Y., Lee, D., & Seo, M. (2024). Knowledge Entropy Decay during Language Model Pretraining Hinders New Knowledge Acquisition. arXiv:2410.01380. doi:10.48550/arXiv.2410.01380.

Labrak, Y., Bazoge, A., Morin, E., Gourraud, P.-A., Rouvier, M., & Dufour, R. (2024). BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains. Findings of ACL 2024. arXiv:2402.10373. doi:10.48550/arXiv.2402.10373.

Lehman, E., Hernandez, E., Mahajan, D., Wulff, J., Smith, M. J., Ziegler, Z., Nadler, D., Szolovits, P., Johnson, A., & Alsentzer, E. (2023). Do We Still Need Clinical Language Models? arXiv:2302.08091. doi:10.48550/arXiv.2302.08091.

Li, X., Chan, S., Zhu, X., Pei, Y., Ma, Z., Liu, X., & Shah, S. (2023). Are ChatGPT and GPT-4 General-Purpose Solvers for Financial Text Analytics? A Study on Several Typical Tasks. EMNLP 2023 Industry Track. arXiv:2305.05862. doi:10.48550/arXiv.2305.05862.

Lin, B. Y., Ravichander, A., Lu, X., Dziri, N., Sclar, M., Chandu, K., Bhagavatula, C., & Choi, Y. (2024). The Unlocking Spell on Base LLMs: Rethinking Alignment via In-Context Learning. arXiv:2312.01552. doi:10.48550/arXiv.2312.01552.

Liu, M., Ene, T.-D., Kirby, R., Cheng, C., Pinckney, N., Liang, R., et al. (2023). ChipNeMo: Domain-Adapted LLMs for Chip Design. arXiv:2311.00176. doi:10.48550/arXiv.2311.00176.

Meng, Z., Liu, F., Shareghi, E., Su, Y., Collins, C., & Collier, N. (2022). Rewire-then-Probe: A Contrastive Recipe for Probing Biomedical Knowledge of Pre-trained Language Models. ACL 2022. arXiv:2110.08173. doi:10.48550/arXiv.2110.08173.

Miller, E. (2024). Adding Error Bars to Evals: A Statistical Approach to Language Model Evaluations. arXiv:2411.00640. doi:10.48550/arXiv.2411.00640.

Mirza, A., Alampara, N., Kunchapu, S., Rios-Garcia, M., Emoekabu, B., Krishnan, A., et al. (2024). Are large language models superhuman chemists? arXiv:2404.01475. doi:10.48550/arXiv.2404.01475.

Mishra, V., Singh, S., Ahlawat, D., Zaki, M., Bihani, V., Grover, H. S., Mishra, B., Miret, S., Mausam, & Krishnan, N. M. A. (2024). Foundational Large Language Models for Materials Research. arXiv:2412.09560. doi:10.48550/arXiv.2412.09560.

Nori, H., Lee, Y. T., Zhang, S., Carignan, D., Edgar, R., Fusi, N., et al. (2023). Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine. arXiv:2311.16452. doi:10.48550/arXiv.2311.16452.

Nori, H., Usuyama, N., King, N., McKinney, S. M., Fernandes, X., Zhang, S., & Horvitz, E. (2024). From Medprompt to o1: Exploration of Run-Time Strategies for Medical Challenge Problems and Beyond. arXiv:2411.03590. doi:10.48550/arXiv.2411.03590.

OpenAI. (2023). GPT-4 Technical Report. arXiv:2303.08774. doi:10.48550/arXiv.2303.08774.

Ou, Y., Yao, Y., Zhang, N., Jin, H., Sun, J., Deng, S., Li, Z., & Chen, H. (2025). How Do LLMs Acquire New Knowledge? A Knowledge Circuits Perspective on Continual Pre-Training. arXiv:2502.11196. doi:10.48550/arXiv.2502.11196.

Ovadia, O., Brief, M., Mishaeli, M., & Elisha, O. (2024). Fine-Tuning or Retrieval? Comparing Knowledge Injection in LLMs. EMNLP 2024. arXiv:2312.05934. doi:10.48550/arXiv.2312.05934.

Prakash, N., Shaham, T. R., Haklay, T., Belinkov, Y., & Bau, D. (2024). Fine-Tuning Enhances Existing Mechanisms: A Case Study on Entity Tracking. arXiv:2402.14811. doi:10.48550/arXiv.2402.14811.

Que, H., Liu, J., Zhang, G., Zhang, C., Qu, X., Ma, Y., et al. (2024). D-CPT Law: Domain-specific Continual Pre-Training Scaling Law for Large Language Models. arXiv:2406.01375. doi:10.48550/arXiv.2406.01375.

Roziere, B., Gehring, J., Gloeckle, F., Sootla, S., Gat, I., Tan, X. E., et al. (2023). Code Llama: Open Foundation Models for Code. arXiv:2308.12950. doi:10.48550/arXiv.2308.12950.

Saab, K., Tu, T., Weng, W.-H., Tanno, R., Stutz, D., Wulczyn, E., et al. (2024). Capabilities of Gemini Models in Medicine. arXiv:2404.18416. doi:10.48550/arXiv.2404.18416.

Sclar, M., Choi, Y., Tsvetkov, Y., & Suhr, A. (2024). Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. arXiv:2310.11324. doi:10.48550/arXiv.2310.11324.

Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., et al. (2025). MedGemma Technical Report. arXiv:2507.05201. doi:10.48550/arXiv.2507.05201.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y. K., Wu, Y., & Guo, D. (2024). DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv:2402.03300. doi:10.48550/arXiv.2402.03300.

Shoham, O. B., & Rappoport, N. (2024). MedConceptsQA: Open Source Medical Concepts QA Benchmark. arXiv:2405.07348. doi:10.48550/arXiv.2405.07348.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., et al. (2022). Large Language Models Encode Clinical Knowledge. arXiv:2212.13138. doi:10.48550/arXiv.2212.13138.

Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., et al. (2023). Towards Expert-Level Medical Question Answering with Large Language Models. arXiv:2305.09617. doi:10.48550/arXiv.2305.09617.

Springer, J. M., Goyal, S., Wen, K., Kumar, T., Yue, X., Malladi, S., Neubig, G., & Raghunathan, A. (2025). Overtrained Language Models Are Harder to Fine-Tune. arXiv:2503.19206. doi:10.48550/arXiv.2503.19206.

Sung, M., Lee, J., Yi, S., Jeon, M., Kim, S., & Kang, J. (2021). Can Language Models be Biomedical Knowledge Bases? EMNLP 2021. arXiv:2109.07154. doi:10.48550/arXiv.2109.07154.

Tang, Y., & Yang, Y. (2024). Do We Need Domain-Specific Embedding Models? An Empirical Investigation. arXiv:2409.18511. doi:10.48550/arXiv.2409.18511.

Vishwanath, K., Alyakin, A., Ghosh, M., Hage, A., Neifert, S. N., Orillac, C., et al. (2026). General-purpose large language models outperform specialized clinical AI tools on medical benchmarks. Nature Medicine, 32, 2405-2409. doi:10.1038/s41591-026-04431-5.

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., & Mann, G. (2023a). BloombergGPT: A Large Language Model for Finance. arXiv:2303.17564. doi:10.48550/arXiv.2303.17564.

Wu, C., Lin, W., Zhang, X., Zhang, Y., Wang, Y., & Xie, W. (2023b). PMC-LLaMA: Towards Building Open-source Language Models for Medicine. arXiv:2304.14454. doi:10.48550/arXiv.2304.14454.

Xie, Q., Chen, Q., Chen, A., Peng, C., Hu, Y., Lin, F., et al. (2024). Me LLaMA: Foundation Large Language Models for Medical Applications. arXiv:2402.12749. doi:10.48550/arXiv.2402.12749.

Yang, Z., Band, N., Li, S., Candes, E., & Hashimoto, T. (2025). Synthetic continued pretraining. arXiv:2409.07431. doi:10.48550/arXiv.2409.07431.

Yao, Z., Kantu, N. S., Wei, G., Tran, H., Duan, Z., Kwon, S., Yang, Z., README annotation team, & Yu, H. (2024). README: Bridging Medical Jargon and Lay Understanding for Patient Education through Data-Centric NLP. Findings of EMNLP 2024. arXiv:2312.15561. doi:10.48550/arXiv.2312.15561.

Zhang, K., Zeng, S., Hua, E., Ding, N., Chen, Z.-R., Ma, Z., et al. (2024). UltraMedical: Building Specialized Generalists in Biomedicine. arXiv:2406.03949. doi:10.48550/arXiv.2406.03949.

Zheng, J., Cai, X., Qiu, S., & Ma, Q. (2025). Spurious Forgetting in Continual Learning of Language Models. arXiv:2501.13453. doi:10.48550/arXiv.2501.13453.

Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., et al. (2023). LIMA: Less Is More for Alignment. NeurIPS 2023. arXiv:2305.11206. doi:10.48550/arXiv.2305.11206.

© 2026 Pranay Mahendrakar. Licensed under CC BY 4.0.

DOI: 10.5281/zenodo.23784965