Thesis Proposal: LLMs post-training for multilingual medical tasks. Instruction-Tuning, Continual-Pretraining or Reasoning?

Pietro Ferrazzi
Fondazione Bruno Kessler
University of Padova
pferrazz@fbk.eu

Alberto Lavelli
Fondazione Bruno Kessler
Trento, Italy

Bernardo Magnini
Fondazione Bruno Kessler
Trento, Italy

Abstract

Adapting Large Language Models to the medical domain remains an active area of research, with multiple strategies proposed to leverage annotated and unannotated data effectively. In this work, we propose a thesis outline to compare three common adaptation approaches—Instruction Tuning, Continual Pretraining, and Reasoning-oriented Training. We identify 5 dimensions to analyse: i) the interaction between the adaptation technique and the tasks; ii) the impact of the data size on the downstream performance; iii) the differences between datasets required by the three techniques; iv) the impact of the techniques given the model size; v) the impact of the techniques given the language. We construct an evaluation framework composed by 5 multilingual medical NLP tasks (named entity recognition, relation extraction, question answering, case report form filling, argument mining), spanning on 21 datasets in English, Italian, and Spanish, for a total of 61 combinations of language and sub-task.

1 Introduction

Pretrained Large Language Models (LLMs) have demonstrated strong performance on a wide range of medical NLP tasks; however, their effectiveness remains uneven across tasks, languages, and data conditions, and they often require substantial adaptation to reliably handle domain-specific terminology, reasoning requirements, and low-resource clinical settings. Over recent years, several distinct trends have emerged following the rise of decoder-only architectures, focused on developing domain-specific language models tailored to biomedical and clinical data. Through different adaptation procedures, several models have been presented, including BioGPT (Luo et al., 2022a), ClinicalGPT (Wang et al., 2023), MedPALM-1 (Singhal et al., 2023), Meditron (Chen et al., 2023), PMC-LLaMA (Wu et al., 2024), MedPalm-2 (Singhal et al., 2025), MedGemma (Sellergren et al., 2025), Huatuo-o1 (Chen et al., 2025). The proposed models are all built on top of foundational LLMs, relying on different post-training methodologies for domain and task adaptation. In the context of large language models, we refer to **adaptation** as the set of techniques used to adjust a pretrained model so that it performs more effectively on certain data, tasks, or usage conditions. Adaptation can target different aspects of model behaviour, such as aligning outputs to task formulations, improving robustness to domain-specific terminology, or inducing particular reasoning patterns. We identify three methodologies commonly exploited for adapting LLMs to the medical domain. **Instruction tuning** consists in further training a model using instruction–response supervision, with the goal of aligning model outputs to task descriptions and expected formats, without explicitly altering domain-level language representations. **Continual pretraining** extends next token prediction pretraining on large collections of domain-specific raw text, incrementally refining internal representations to better capture linguistic and semantic characteristics of the target domain. Finally, **reasoning-oriented training** seeks to induce explicit multi-step inference by exposing models to examples that include intermediate reasoning or structured problem-solving processes, aiming to influence not only task performance but also the manner in which predictions are produced, particularly for complex tasks.

In this work, we are interested in determining what is the impact of different methods when applied in the context of the medical domain. In other words, we want to study how to "adapt" a model to perform medical tasks. Direct comparison is non-trivial, as IT is trained on target tasks while CP and RoT are not. To ensure fairness, we apply IT after CP to isolate its contribution, whereas for RoT we reformulate tasks as question answering to preserve learned reasoning behaviors.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 110–122

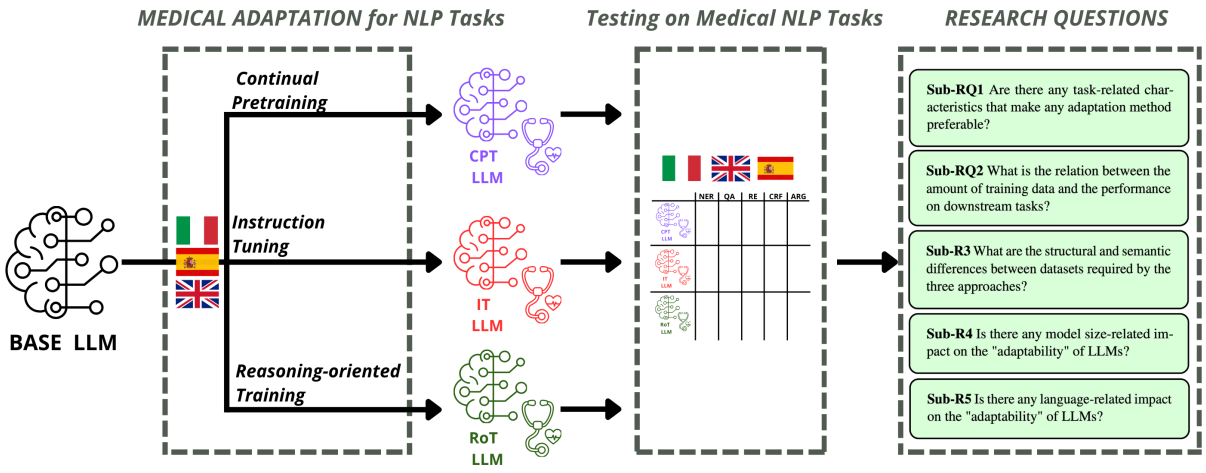

Figure 1: Overview of the proposed comparison, including Continual Pretraing, Instruction Tuning, and Reasoning-oriented Training. We report a summary of the dimensions we aim to analyze experimentally as research questions in the right box.

Our main research question can then be framed as follows:

RQ1: *How do recent advancements in LLM post-training methodologies address medical NLP tasks?*

This overarching research question guides and motivates the comparison of the three target methodologies. However, such a comparison remains inherently broad. To make the analysis more focused and informative, we formulate a set of sub–research questions that target specific dimensions of interest and provide more granular insights into the behavior of each approach.

Sub-RQ1: *Are there any task-related characteristics that make any adaptation method preferable?*

Sub-RQ2: *What is the relation between the amount of training data and the performance on downstream tasks?*

Sub-RQ3: *What are the structural and semantic differences between datasets models learn from in the three approaches?*

Sub-RQ4: *Is there any model size-related impact on the "adaptability" of LLMs?*

Sub-RQ5: *Is there any language-related impact on the "adaptability" of LLMs?*

To answer these questions, we construct an evaluation setting composed of 5 tasks among 21 datasets in Italian, Spanish, and English, for a total of 61 combinations of language and sub-task (e.g, entity extraction on bodyparts in Italian). We apply Instruction-Tuning, Continual Pretraining, and Reasoning-oriented training to ~1B and ~8B open-weight LLMs and analyse their performance.

2 Related work

Existing work underscores the role of domain alignment in medical NLP. Luo et al. (2022b); Wu et al. (2024) indicate that domain adaptation of large language models can significantly enhance performance on medical NLP tasks and, in controlled settings, may reach or surpass expert-level results on certain benchmarks (Singhal et al., 2025).

Instruction Tuning Instruction tuning consists of training models to align with complex instructions (Zhang et al., 2023). Examples of alignment of models to the biomedical domain have been proposed for encoder-only models (BioMedROBERTA Gururangan et al. (2020)), and decoder-only models (Galactica Taylor et al. (2022), ClinicalGPT Wang et al. (2023), Med-PALM Singhal et al. (2023), BioMedLM Bolton et al. (2024)).

Continual Pretraining One approach to tailor LLMs to domain-specific knowledge is continual pretraining (Cossu et al., 2024; Shi et al., 2025). Training happens on next token prediction on corpora of raw text, enhancing the model's

knowledge about the language structure that characterises the field, acquiring semantical knowledge, use of syntax, words distribution, meaning, co-occurrences, etc. Through these techniques, models are enabled to improve performance on specialised tasks, without incurring the costs and downsides of training from scratch on large corpora (Jin et al., 2022; Çağatay Yıldız et al., 2025). It has been observed that continual pretraining can mitigate the risk of catastrophic forgetting with respect to instruction tuning on downstream tasks (Rongali et al., 2021), which occurs when the model loses its general language-related capabilities in favour of high specialization. Examples of models adapted to the biomedical domain via continual pretraining are BioBERT (Lee et al., 2019), BioLM (Lewis et al., 2020), BioMedROBERTA (Gururangan et al., 2020), ClinicalBERT (Huang et al., 2020) for encoder-only models; decoder-only models as BioGPT (Luo et al., 2022a), Meditron (Chen et al., 2023); encoder-decoder models as Med-mT5 (García-Ferrero et al., 2024).

Reasoning in the Medical Domain Reasoning is commonly understood as the ability to solve problems through multi-step inference (Zhang et al., 2025). Examples of models capable of reasoning are OpenAI-o1 (OpenAI et al., 2024), marco-o1 (Zhao et al., 2024), skywork-o1 (He et al., 2024), QwQ (Team, 2025), DeepSeek-R1 (Guo et al., 2025) . A line of research investigates training models to generate structured reasoning traces in the medical domain. Efforts exploring this direction via distillation and medical knowledge ingestion are presented by Huatuo (Chen et al., 2025), MedReason (Wu et al., 2025), m1 (Huang et al., 2025), ReasonMed (Sun et al., 2025), FineMedLM-o1 (Yu et al., 2025). Alternative directions seek to elicit or refine reasoning abilities without extensive distillation pipelines. Some methods aim to optimize the use of existing reasoning traces (Thapa et al., 2025), while others investigate lightweight strategies to induce reasoning behaviours with reduced computational cost (Liu et al., 2025). A comprehensive overview of these trends is provided by Wang et al. (2025).

Multilingual Medical NLP Névéol et al. (2018) showed that specific adaptation is required when handling NLP tasks in languages other than English. This is due to barriers posed by the differences in medical terminology among languages, as efforts towards multilingual interoperability have highlighted (Noll et al., 2025; Kandala et al., 2025).

Effectiveness medical-oriented LLMs Given the growing capabilities of general-purpose LLMs, a question raises about the effectiveness of medical-oriented training. Jeong et al. (2024) tested 7 LLMs adapted to the medical domain against their base models on medical question answering tasks in English, and found that often cases there is no significant performance increase. Dada et al. (2025) extend the analysis to seven tasks, obtaining similar results. Hence, the role of multi-task in determining whether the effectivness of post-training methods.

3 Methods

In this section, we present the three methods we compare. In addition, we highlight our design and implementation choices. We acknowledge that the three methods differ in data requirements, computational cost, complexity, and objectives. by order to present a comprehensive list of the pros and cons of all approaches, we restrict our analysis to the dimensions that help answer our research question *RQ1*. To do so, we build on the consolidated training data structures and procedures that we describe below.

3.1 Instruction Tuning

Instruction Tuning (IT) consists of aligning a model towards a desired instruction-following behaviour. In the context of our work, this means specializing a model to perform one or more tasks, making sure to produce structurally and semantically correct outputs.

Data & Training IT output depends on the training data structure. Previous works described in Section 2 suggest formatting the data according to a user-assistant template, where training sequences are composed of inputs/questions answered via textual outputs. Gao et al. (2025) shows that performance on downstream tasks increases when input-output pairs are shaped in a JSON-like format. We maintain a hybrid approach by keeping textual input and producing JSON-structured outputs (Figure 2, left). We convert the training splits of all datasets into instruction-completion pairs, and shuffle the resulting joint dataset of sequences. The training routine consists in learning the desired answers via next token prediction. Following the findings of Shi et al. (2024), we consider both the instruction and the completion in the training loss calculation. Prior work has shown that task-oriented

<bos><start_of_turn>user You are an AI assistant trained for named entity recognition. You are given a text and you need to extract body parts mentioned in the text.
Analyze the given text and extract all named entities that are bodyparts.

Rules:
- Extract clinical entities exactly as they appear in the text.
- Do not include entities not mentioned in the text.
- If no entities are mentioned, return an empty list.

Output Format:
Provide your answer as a JSON object. ONLY the JSON object, without any additional text:
{
"entities": ["entity1", "entity2", ...]
}

The case concerns a 12-year-old boy admitted to the Paediatric Surgery Unit with acute abdomen and chest pain. <start_of_turn>
assistant
{
"entities": ["chest"]
}<eos>

<bos>Insomnia is common, but undertreated, among primary care patients. Within the Veterans Health Administration (VA), increasing attention has been given to the treatment of insomnia within primary care settings, but little research has examined Veterans' treatment preferences. We examined preferences for sleep treatment among VA primary care patients. Participants (N = 126: 98% male, 89% white; M age = 60 years) completed a brief survey. On the basis of Insomnia Severity Index scores, 22% reported subthreshold and 13% moderate insomnia. Fifty percent reported having issues with sleep (falling asleep, staying asleep, or sleeping too much) in the past 12 months; among these, only 44% reported any discussion of medication (34%) or other strategies (32%) to improve sleep with medical providers. The most preferred treatment approach was to work it out on one's own, followed by consulting the primary care provider (PCP). The most preferred modality was a one-on-one meeting with the PCP, followed by a one-on-one meeting with the behavioral health provider. In conclusion, VA primary care patients preferred handling sleep problems on their own, but if seeking help, they preferred working with PCPs over behavioral health providers. The majority of Veterans preferred individual treatment and strategies other than medication. <eos>

<bos><start_of_turn>user
<question>A 45-year-old woman has [...]. Which of the following best explains these findings?</question> <possible answers>
1) Ascites
2) Ileus
3) Necrotizing pancreatitis
4) Pancreatic pseudocyst</possible answers>
<start_of_turn>
assistant

Search for potential reasoning traces:
- Painless abdominal distension 2 days after admission for acute pancreatitis
- Absent *[...]*

Reasoning process:
1. The patient has a known diagnosis of acute pancreatitis, which is a recognized cause of paralytic ileus due to inflammation and neurohumoral effects on gut motility.
[...]
9. Ascites would not typically explain the absence of bowel sounds or the clinical picture of ileus.

Conclusion:
2<eos>

Figure 2: Training sequences examples for **Instruction Tuning (left)**, **Continual Pretraining (center)**, and **Reasoning-oriented training (right)**. Special tokens are highlighted in green and are model-dependent.

instruction tuning using LoRA (Hu et al., 2022) is reliable in both general (Gema et al., 2024) and medical domains (Gao et al., 2024). The method consists in training a set of adapters with significant less paramenters than the model itself, which can then be added to the base model enhancing it's capabilities and knowledge. Multiple adapters can be trained on multiple tasks, allowing for a task-based choice of the best one at inference time.

3.2 Continual Pretraining

In the context of our work, Continual Pretraining (CP) is employed relying on the intuition that domain shift through exposure to large medical data can improve performances on downstream tasks.

Data & Training The amount of data needed for CP is significantly higher than what is required by task-oriented instruction tuning. For instance, Xie et al. (2024) showed that the performance in downstream tasks of continual pretrained models only plateaus after +2B tokens. To explore this direction, we rely on large medical datasets from García-Ferrero et al. (2024). This resource collects data from different medical fields from various sources (PubMed, Wikipedia, drug descriptions, etc.). For English, it contains 1.1B words, and 950M for Spanish, while for Italian the size is notably smaller (140M). Given this, we expanded the Italian medical dataset with additional sources from both the scientific literature (140M words) and clinical notes from the Emergency Department of an Italian hospital (125M words), resulting in a final size of 405M words (Ferrazzi et al., 2026). We train *i)* separately and *ii)* jointly on the three languages. An example of the training sequence is presented in Figure 2 (center). To assess the relative impact of CP on downstream task performance, we test models that underwent both simple CP, and CP+IT. This comparison isolates the contribution of CP by evaluating whether acquired domain-level knowledge yields additional benefits beyond those provided by task-specific abilities introduced via IT.

3.3 Reasoning-oriented Training

We refer to Reasoning-oriented Training (RoT) as the approach that involves exposing an LLM to reasoning-like outputs, where prompts are completed through a long sequence of steps that logically lead to a conclusion. Recent works suggest that reasoning capabilities can be acquired through simple next token prediction, and further enhanced via Reinforcement Learning (RL) (Chen et al., 2025). Nevertheless, the inclusion of RL would misalign the training procedures of the methods we compare and, therefore, bias the analysis. Therefore, we limit our analysis of RoT to next token prediction. Moreover, we are interested in utilizing reasoning enhancement as a mean to obtain the correct output for a given task, and do not include any analysis of intermediate steps faithfulness, robustness, clinical correctness[1].

Data & Training The data required for RoT are similar to those described for IT. Each train-

[1]Initial analysis of these dimensions can be found at Ferrazzi et al. (2025b)

ing sequence consists of a user request and a desired model answer. Building on prior work, we analyse this direction using a large, multilingual corpus of reasoning traces that answer medical queries. We utilize the resources provided by Ferrazzi et al. (2025b), resulting in a dataset of 530k training pairs, evenly distributed between languages. Each pair is formatted as a multiple-choice question with a reasoning-intensive answer (Figure 2, right). Models trained on this resource learn to answer medical questions. Therefore, we re-frame each downstream task to resemble question answering, following a structure such as *<question> {medical text} What are the [entities | relations | arguments | etc] in the text?</question> <possible_answers>{task description}</possible_answers>*.

4 Multilingual Medical NLP Tasks

To analyze performance on multilingual medical NLP tasks, we carefully select datasets that enable controlled cross-lingual comparison on a diversity of medical scenarios. We focus on three languages—English, Italian, and Spanish—that share substantial overlap in available medical resources, enabling comparable evaluations across languages. Dataset selection is guided by the goal of covering a broad range of tasks and medical domains, while acknowledging that the study's scope is inherently constrained by existing annotated resources. Overall, we evaluate 5 medical NLP tasks, spanning 21 datasets and 35 task variants (Table 1). In what follows, we present the selected tasks and datasets.

Named Entity Recognition Named Entity Recognition (NER) aims to identify and classify medically relevant entities in text, and can be evaluated using the F1 metric on the entities themselves. The datasets we select cover clinical narratives, pharmaceutical documents, psychiatric records, cardiology case reports, and research documents. Each dataset defines a distinct set of entity types, including clinical entities, body parts, drugs, diseases, anatomical parts, symptoms, medications, disabilities, cognitive symptoms, ages, and tumor morphologies. We collect 3 datasets in the three languages (E3C by Magnini et al. (2023), DisteMIST by Miranda-Escalada et al. (2022), and CardioCCC by Lima-López et al. (2024)); one dataset in English and Spanish (DIANN-2018 by Fabregat et al. (2018)); 3 datasets in Italian (the projected version of E3C by Ghosh et al. (2025), PsyNIT by Crema et al. (2023), PharmaER by Zugarini and Rigutini (2025)); 3 datasets in English (NCBI by Doğan et al. (2014), BC5-disease by Wei et al. (2016), n2b2-2010 dataset by Uzuner et al. (2010)); 3 datasets in Spanish (MEDDOCAN by Marimon et al. (2019); CANTEMIST by Miranda-Escalada et al. (2020), PharmaconNER by Gonzalez-Agirre et al. (2019)).

Case Report Form Filling Case Report Form (CRF) filling focuses on extracting structured patient information from unstructured clinical narratives. Given a patient clinical record, the objective is to fill a predefined list of medical items such as "temperature", "history of diabetes", etc. There is no available dataset for such a task in Spanish, while for English and Italian we use a derivative of E3C (Ferrazzi et al., 2025a), which maps clinical notes to predefined CRF fields. Our experiments target the extraction of diagnoses, clinical history, and examination-related information. Furthermore, we used the eCream dataset, composed of private data from the S. Giovanni Bosco Hospital (Turin, Italy). The evaluation of this task is performed using the F1 macro metric.

Medical Question Answering Medical Question Answering (QA) requires selecting the correct answer to clinically relevant questions, typically drawn from medical examinations. For the three languages, the evaluation is performed on MedExpQA (Alonso et al., 2024), using both standard multiple-choice questions and a variant augmented with retrieved contextual evidence. We also employ the automatically translated versions of MedMCQA (Pal et al., 2022) and MedQA (Jin et al., 2021) proposed by Ferrazzi et al. (2025b). For Italian, we include AT, medical admission tests collected by Casola et al. (2023). For English, we use PubMedQA (Jin et al., 2019). As these are all multiple-choice datasets, the performances can be calculated using accuracy.

Relation Extraction Relation Extraction (RE) aims to identify semantic relations between medical entities mentioned in text. For the three languages, we focus on the E3C dataset, targeting relations that link laboratory exams and tests to their corresponding results. For English, we utilize the n2b2-2010 dataset (Uzuner et al., 2010) (rela-

Dataset	Language	n	Example type	Annotaion type	Examples		
					train	val	test
Named Entity Recognition							
e3c	IT-EN-ES	2	clinical note	clinical; body part	451	67	628
distemist	IT-EN-ES	1	clinical case	disease	500	100	150
cardioccc	IT-EN-ES	1	cardiology report	medications	250	100	150
diann	EN-ES	1	abstract	disability	300	100	100
bc5	EN	1	pubmed	disease	500	500	500
n2b2-2010	EN	1	clinical note	problem & treatment & test	394	177	300
ncbi	EN	1	pubmed	disease	500	93	200
pharmaer	IT	4	drug label	drug; disease; symptom; anatomical part	1316	404	64
psynit	IT	1	psychiatric report	cognitive symptom	200	80	120
e3c-proj	IT	1	clinical case	clinical entity	632	101	738
meddocan	ES	1	synthetic clinical case	age	700	100	200
cantemist	ES	1	clinical case	tumor morphology	501	500	300
pharmaconer	ES	1	clinical case	drug	700	100	200
total	**all**	**16**	**all**	**all**	**6144**	**2322**	**3550**
Question Answering							
medexpqa	IT-EN-ES	2	exam question	multiple choice; rag-like	434	63	125
medmcqa	IT-EN-ES	1	exam question	multiple choice	178k	4183	4183
medqa	IT-EN-ES	1	exam question	multiple choice	10k	1272	1273
at	IT	1	exam question	multiple choice	300	71	150
pubmedqa	EN	1	pubmed	multiple choice	700	100	200
total	**all**	**6**	**all**	**all**	**191k**	**5589**	**5931**
Relation Extraction							
e3c	IT-EN-ES	1	clinical note	exam-result	451	67	628
n2b2-2010	EN	2	clinical note	problem-treatment; problem-test	253	50	202
n2b2-2022	EN	3	clinical note	dosage-drug; form-drug; frequency-drug	394	177	300
total	**all**	**6**	**all**	**all**	**1098**	**294**	**1130**
Case Report Form Filling							
e3c	IT-EN	3	clinical note	history; diagnosis; exams	451	67	628
ecream	IT-EN	3	clinical note	history; diagnosis; tests clinical examin; treatment	80	10	200
total	**all**	**6**	**all**	**all**	**531**	**77**	**828**
Argument Mining							
casimed	IT-EN-ES	1	clinical note	claim & premise	434	63	125

Table 1: Datasets selected for five medical NLP tasks. For each dataset, we report the languages, the number of sub-tasks defined per dataset (*n*), the data type each example represents (*Example type*), the type of annotations we considered in our evaluation (*Annotation type*, each of which defines a sub-task), and the number of *Examples*. Examples across different datasets can vary widely in length (for instance, questions in *medqa* tend to be much shorter than clinical cases in *e3c*). We keep the original train-val-test split when available.

method	example type	train words (M)	train TFLOPs
CPT	raw text	2400	57936
IT	input + output	4	97
RoT	question +reasoning +answer	46	1110

Table 2: Differences in training requirements by method. Train FLOPs are calculated using the `calflops` package (xiaoju ye, 2023) for Llama-3.2-1B.

tions between medical problems and treatments, tests), the n2b2-2022 ADE dataset (Henry et al., 2019) (relations between drugs and dosage, form, frequency). The evaluation of relation extraction can be done via F1 calculation on the extracted items.

Argument Mining Argument mining seeks to identify and structure reasoning components in medical and clinical text. Experiments are conducted on the tri-lingual Casimedicos-Arg dataset (Sviridova et al., 2024), which provides annotations for argumentative elements such as claims and premises on clinical cases collected from medical exams. The evaluation can be performed via F1.

5 Experimental comparison

To compare Instruction Tuning (IT), Continual Pretraining (CP), and Reasoning-oriented Training (RoT), we evaluate all models on the test sets of the datasets presented in Section 4. A direct comparison based solely on downstream performance would be potentially misleading, as IT involves explicit supervision on the target tasks, whereas CP and RoT do not. To mitigate this discrepancy, we apply IT on top of CP, allowing us to isolate and quantify the contribution of continual pretraining beyond task-oriented instruction tuning. In contrast, applying standard IT after RoT would risk overriding the reasoning behaviors acquired during reasoning-oriented training. For this reason, we reframe each task as question-answering as described in Section 3.3.

We present a short summary of computational requirements for each method under our setups in Table 2, to help readers contextualize the differences among the three. We select two models from three families of models, Llama-3 -1B, -8B (Dubey et al., 2024), Gemma-3 -1B, -8B (Team et al., 2025), and Qwen-3 -1.7B, -8B (Yang et al., 2025). All models are used in both their base and instructed versions, for a total of 12 models.

Task-datasets size imbalance The datasets presented in Section 4 (Table 1) are characterized by high size imbalances, with more than 95% of the examples from QA tasks. To prevent model collapsing to single task handling at training time, we limit MedMCQA and MedQA at 1k training examples. At test time, this issue is overcome by the evaluation metric, which we calculate for each dataset and then average among sub-tasks.

Learning Objective All methods involve training via next token prediction and cross-entropy-based back-propagation. What differs are the training data, as described in Section 3

5.1 Method impact by task

> **Sub-RQ1** Are there any task-related characteristics that make any adaptation method preferable?

Directions of analysis We plan to experiment with all three approaches on the 35 sub-tasks, and group the results by the 5 tasks. By examining aggregated results, we can understand the impact of IT, CP, and RoT on each task. These comparisons will help shed light on the following hypotheses. **Hypothesis 1)** RoT is the best approach to handle *question answering*, as such models have acquired both medical knowledge and multi-step answering capabilities; **Hypothesis 2)** CP improves the general medical understanding, and should enhance performance on *all tasks* when followed by IT with respect to mere IT; **Hypothesis 3)** *Argument mining* is more related to logical capabilities than medical language understanding, and therefore should not gain much from domain adaptation; **Hypothesis 4)** *Case report form filling* is inheritently similar to question answering; **Hypothesis 5)** *Named entity recognition* and *relation extraction* do not require deep medical knowledge, and IT could be enough.

5.2 Data size impact

> **Sub-RQ2** What is the relation between the amount of training data and the performance on downstream tasks?

Directions of analysis Previous work has investigated the role of data quantity in model adaptation

and proposed heuristics to estimate performance gains as a function of training data for CP (Xie et al., 2024), IT (Li et al., 2024), and RoT (Thapa et al., 2025; Huang et al., 2025). These three approaches exhibit substantially different data requirements, which complicates direct comparisons based solely on data scale. For instance, the amount of data typically required for effective CP is orders of magnitude larger than that used for IT, making comparisons based on matched token counts neither practical nor informative. Therefore, we plan to conduct a two-fold analysis. First, we analyze the pros and cons of all methods with respect to data requirements from a theoretical perspective. Then, we perform ablation on the number of tokens utilized to train models via each method, and determine what are the practical, overall tradeoffs on the specific tasks we tackle. From both analysis, we aim to verify these hypotheses: **Hypothesis 1)** IT is the most data-efficient approach, while CP is the most data-expensive; **Hypothesis 2)** RoT requires more training tokens than IT, as training examples that include reasoning are significantly longer than instruction completions; **Hypothesis 3)** IT is effective even with a low number of examples; **Hypothesis 4)** CP plateau performance on the downstream task plateaus after a certain amount of medical training data.

5.3 Data differences among methods

Sub-R3 What are the structural and semantic differences between datasets required by the three approaches?

Directions of analysis Across all three methodologies, learning is ultimately driven by the same underlying objective, namely next-token prediction. What differentiates IT, CP, and RoT in our implementations is not the learning mechanism itself, but the nature of the data they are exposed to. IT data primarily aims to align the model with task descriptions and output formats, providing explicit signals about how to respond to structured instructions. CP data focuses instead on adapting the model to the linguistic patterns, terminology, and discourse characteristics of a target domain. Data used for RoT emphasizes the generation of explicit reasoning processes, teaching the model how to articulate multi-step inference within the domain. Despite these differences in intent, the optimization process remains almost identical across methods. As a result, understanding how the training data differ in form and content is key to understanding how these adaptation strategies diverge in practice. To this end, we analyze the training examples used by each methodology, with the goal of characterizing the source of what is effectively being taught to the model. In particular, we examine **1)** structural differences in input–output format, and **2)** semantic differences in the type of information and reasoning conveyed.

5.4 Model size impact

Sub-R4 Is there any model size-related impact on the "adaptability" of LLMs?

Directions of analysis For each of the three model families, we compare the performance of two pairs of models. For instance, for the Llama family we compare `Llama-3.2-`**`1B`** with `Llama-3.1-`**`8B`**, and `Llama-3.2-`**`1B`**`-Instruct` with `Llama-3.1-`**`8B`**`-Instruct`. This provides us with six pairs of -1B and -8B models to compare, offering a general understanding of trends in scaling in these two sizes. We are interested in verifying two hypotheses. **Hypothesis 1)** The parametric knowledge of smaller models has a higher degree of "saturation" than bigger ones, resulting in lower adaptability to the new domain; **Hypothesis 2)** Bigger models are more capable and knowledgeable of smaller ones; therefore, the relative improvement due to domain adaptation is smaller.

5.5 Language impact

Sub-R5 Is there any language-related impact on the "adaptability" of LLMs?

Directions of analysis Given the multilingual structure of our evaluation, some hypotheses can be verified by training models on one or more languages at a time. We can determine whether **1)** Multilingual transfer enhances performance on downstream tasks, and **2)** Models exhibit different trends and behaviors in languages other than English (the one models are most exposed to).

6 Conclusion

This PhD research investigated how different post-training methodologies adapt large language models to multilingual medical NLP tasks. We focused on three widely adopted approaches—Instruction

Tuning, Continual Pretraining, and Reasoning-oriented Training—and compared them across five medical task families, 21 multilingual datasets, multiple model sizes, and three languages.

Our comparison layout is framed among five main evaluation dimensions: *i)* verify if there exists a significant interaction between the adaptation technique and the performances on specific tasks; *ii)* determine the impact of the data size on the downstream performance for each approach; *iii)* analyse both theoretically and experimentally the differences between the datasets required by the three techniques; *iv)* determine the impact of the model size on the downstream performance for each approach; *v)* determine the impact of the techniques given the language .

Limitations

This work presents several limitations that should be considered when interpreting the results. First, our implementation of Reasoning-oriented Training (RoT) relies exclusively on supervised learning over reasoning traces and does not incorporate reinforcement learning techniques. As a result, the performance of RoT models may underestimate what is achievable with more advanced training paradigms.

Second, our analysis is restricted to three languages that belong to the same broad linguistic family. This limits the extent to which our findings can be generalized to typologically distant languages or to low-resource scenarios.

Third, we limit our analysis to downstream performances. We do not consider models' answers as a whole, but simply parse models' answers to obtain the expected tasks' output. More detailed analysis could take into account reasoning traces and overall answers.

Finally, our evaluation focuses on a fixed set of medical tasks and datasets, all reformulated within a unified instruction-following framework. Although this design choice enables controlled comparisons between adaptation strategies, it may not capture all aspects of real-world clinical applications, where task formulations, output requirements, and interaction patterns can vary substantially.

References

Iñigo Alonso, Maite Oronoz, and Rodrigo Agerri. 2024. MedExpQA: Multilingual benchmarking of large language models for medical question answering. *Artificial Intelligence in Medicine*, 155:102938.

Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, Jonathan Frankle, Percy Liang, Michael Carbin, and Christopher D. Manning. 2024. Biomedlm: A 2.7b parameter language model trained on biomedical text. *Preprint*, arXiv:2403.18421.

Silvia Casola, Tiziano Labruna, Alberto Lavelli, Bernardo Magnini, and 1 others. 2023. Testing ChatGPT for stability and reasoning: A case study using Italian medical specialty tests. In *Proceedings of the Nineth Italian Conference on Computational Linguistics*.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, and Benyou Wang. 2025. Towards medical complex reasoning with LLMs through medical verifiable problems. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14552–14573, Vienna, Austria.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023. Meditron-70b: Scaling medical pretraining for large language models. *Preprint*, arXiv:2311.16079.

Andrea Cossu, Antonio Carta, Lucia Passaro, Vincenzo Lomonaco, Tinne Tuytelaars, and Davide Bacciu. 2024. Continual pre-training mitigates forgetting in language and vision. *Neural Networks*, 179:106492.

Claudio Crema, Tommaso Mario Buonocore, Silvia Fostinelli, Enea Parimbelli, Federico Verde, Cira Fundarò, Marina Manera, Matteo Cotta Ramusino, Marco Capelli, Alfredo Costa, Giuliano Binetti, Riccardo Bellazzi, and Alberto Redolfi. 2023. Advancing italian biomedical information extraction with transformers-based models: Methodological insights and multicenter practical application. *Journal of Biomedical Informatics*, 148:104557.

Amin Dada, Osman Alperen Koraş, Marie Bauer, Jean-Philippe Corbeil, Amanda Butler Contreras, Constantin Marc Seibold, Kaleb E Smith, Julian Friedrich, and Jens Kleesiek. 2025. Does biomedical training lead to better medical performance? In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 46–59, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 82 others. 2024. The Llama 3 herd of models. *CoRR*, abs/2407.21783.

Hermenegildo Fabregat, Juan Martínez-Romo, and Lourdes Araujo. 2018. Overview of the diann task: Disability annotation task. In *IberEval@SEPLN*.

Pietro Ferrazzi, Mattia Franzin, Alberto Lavelli, and Bernardo Magnini. 2026. Small llms for medical nlp: a systematic analysis of few-shot, constraint decoding, fine-tuning and continual pre-training in italian. *Preprint*, arXiv:2602.17475.

Pietro Ferrazzi, Alberto Lavelli, and Bernardo Magnini. 2025a. Converting annotated clinical cases into structured case report forms. In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 307–318, Vienna, Austria. Association for Computational Linguistics.

Pietro Ferrazzi, Aitor Soroa, and Rodrigo Agerri. 2025b. Grounded multilingual medical reasoning for question answering with large language models. *Preprint*, arXiv:2512.05658.

Chang Gao, Wenxuan Zhang, Guizhen Chen, and Wai Lam. 2025. JsonTuning: Towards generalizable, robust, and controllable instruction tuning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24029–24055, Vienna, Austria. Association for Computational Linguistics.

Dehong Gao, Yufei Ma, Sen Liu, Mengfei Song, Linbo Jin, Wen Jiang, Xin Wang, Wei Ning, Shanqing Yu, Qi Xuan, Xiaoyan Cai, and Libin Yang. 2024. Fashiongpt: Llm instruction fine-tuning with multiple lora-adapter fusion. *Knowledge-Based Systems*, 299:112043.

Iker García-Ferrero, Rodrigo Agerri, Aitziber Atutxa Salazar, Elena Cabrio, Iker de la Iglesia, Alberto Lavelli, Bernardo Magnini, Benjamin Molinet, Johana Ramirez-Romero, German Rigau, Jose Maria Villa-Gonzalez, Serena Villata, and Andrea Zaninello. 2024. MedMT5: An open-source multilingual text-to-text LLM for the medical domain. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11165–11177, Torino, Italia. ELRA and ICCL.

Aryo Gema, Pasquale Minervini, Luke Daines, Tom Hope, and Beatrice Alex. 2024. Parameter-efficient fine-tuning of LLaMA for the clinical domain. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 91–104, Mexico City, Mexico. Association for Computational Linguistics.

Soumitra Ghosh, Begoña Altuna, Saeed Farzi, Pietro Ferrazzi, Alberto Lavelli, Giulia Mezzanotte, Manuela Speranza, and Bernardo Magnini. 2025. Low-resource information extraction with the European Clinical Case Corpus. *CoRR*, abs/2503.20568.

Aitor Gonzalez-Agirre, Montserrat Marimon, Ander Intxaurrondo, Obdulia Rabal, Marta Villegas, and Martin Krallinger. 2019. PharmaCoNER: Pharmacological substances, compounds and proteins named entity recognition track. In *Proceedings of the 5th Workshop on BioNLP Open Shared Tasks*, pages 1–10, Hong Kong, China. Association for Computational Linguistics.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of ACL*.

Jujie He, Tianwen Wei, Rui Yan, Jiacai Liu, Chaojie Wang, Yimeng Gan, Shiwen Tu, Chris Yuhao Liu, Liang Zeng, Xiaokun Wang, Boyang Wang, Yongcong Li, Fuxiang Zhang, Jiacheng Xu, Bo An, Yang Liu, and Yahui Zhou. 2024. Skywork-o1 open series.

Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner. 2019. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1):3–12.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *ICLR 2022*.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2020. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *Preprint*, arXiv:1904.05342.

Xiaoke Huang, Juncheng Wu, Hui Liu, Xianfeng Tang, and Yuyin Zhou. 2025. m1: Unleash the potential of test-time scaling for medical reasoning in large language models. In *The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance*.

Daniel P Jeong, Saurabh Garg, Zachary Chase Lipton, and Michael Oberst. 2024. Medical adaptation of large language and vision-language models: Are we making progress? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12143–12170, Miami, Florida, USA. Association for Computational Linguistics.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain

question answering dataset from medical exams. *Applied Sciences*, 11:6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China.

Xisen Jin, Dejiao Zhang, Henghui Zhu, Wei Xiao, Shang-Wen Li, Xiaokai Wei, Andrew Arnold, and Xiang Ren. 2022. Lifelong pretraining: Continually adapting language models to emerging corpora. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4764–4780, Seattle, United States. Association for Computational Linguistics.

Ananth Kandala, Ratna Kandala, Akshata Kishore Moharir, Niva Manchanda, and Sunaina Singh Rathod. 2025. Cross-lingual mental health ontologies for Indian languages: Bridging patient expression and clinical understanding through explainable AI and human-in-the-loop validation. In *NLP-AI4Health*, pages 16–24, Mumbai, India. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. 2020. Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 146–157, Online. Association for Computational Linguistics.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. 2024. From quantity to quality: Boosting LLM performance with self-guided data selection for instruction tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7602–7635, Mexico City, Mexico. Association for Computational Linguistics.

Salvador Lima-López, Eulàlia Farré-Maduell, Jan Rodríguez-Miret, Miguel Rodríguez-Ortega, Livia Lilli, Jacopo Lenkowicz, Giovanna Ceroni, Jonathan Kossoff, Anoop Shah, Anastasios Nentidis, Anastasia Krithara, Georgios Katsimpras, Georgios Paliouras, and Martin Krallinger. 2024. Overview of MultiCardioNER task at BioASQ 2024 on medical specialty and language adaptation of clinical ner systems for Spanish, English and Italian. In *Conference and Labs of the Evaluation Forum*.

Che Liu, Haozhe Wang, Jiazhen Pan, Zhongwei Wan, Yong Dai, Fangzhen Lin, Wenjia Bai, Daniel Rueckert, and Rossella Arcucci. 2025. Beyond distillation: Pushing the limits of medical llm reasoning with minimalist rule-based rl. *Preprint*, arXiv:2505.17952.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022a. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6).

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022b. BioGPT: Generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6).

Bernardo Magnini, Begoña Altuna, Alberto Lavelli, Anne-Lyse Minard, Manuela Speranza, and Roberto Zanoli. 2023. *European Clinical Case Corpus*, pages 283–288. Springer International Publishing, Cham.

Montserrat Marimon, Aitor Gonzalez-Agirre, Ander Intxaurrondo, Heidy Rodriguez, Jose Lopez Martin, Marta Villegas, and Martin Krallinger. 2019. Automatic de-identification of medical texts in Spanish: the MEDDOCAN track, corpus, guidelines, methods and evaluation of results. In *IberLEF@ SEPLN*, pages 618–638.

Antonio Miranda-Escalada, Eulalia Farré-Maduell, and Martin Krallinger. 2020. Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in Spanish, corpus, guidelines, methods and results.

Antonio Miranda-Escalada, Luis Gasco, Salvador Lima-López, Eulàlia Farré-Maduell, Darryl Estrada, Anastasios Nentidis, Anastasia Krithara, Georgios Katsimpras, Georgios Paliouras, and Martin Krallinger. 2022. Overview of DisTEMIST at BioASQ: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources. In *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings*.

A. Névéol, H. Dalianis, S. Velupillai, and 1 others. 2018. Clinical natural language processing in languages other than english: opportunities and challenges. *Journal of Biomedical Semantics*, 9:12.

R. Noll, H. Storf, and J. Schaaf. 2025. Healthtermfinder: Enhancing multilingual interoperability. *Studies in Health Technology and Informatics*, 327:2–6.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, and 244 others. 2024. Openai o1 system card. *arXiv:2412.16720*.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of Machine Learning Research*, volume 174, pages 248–260. PMLR.

Subendhu Rongali, Abhyuday Jagannatha, Bhanu Pratap Singh Rawat, and Hong Yu. 2021. Continual domain-tuning for pretrained language models. *Preprint*, arXiv:2004.02288.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cian Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, and 60 others. 2025. Medgemma technical report. *Preprint*, arXiv:2507.05201.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. 2025. Continual learning of large language models: A comprehensive survey. *ACM Comput. Surv.* Just Accepted.

Zhengyan Shi, Adam X. Yang, Bin Wu, Laurence Aitchison, Emine Yilmaz, and Aldo Lipani. 2024. Instruction tuning with loss over instructions. In *Advances in Neural Information Processing Systems*, volume 37, pages 69176–69205. Curran Associates, Inc.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, and 13 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R. Pfohl, Heather Cole-Lewis, Darlene Neal, Qazi Mamunur Rashid, Mike Schaekermann, Amy Wang, Dev Dash, Jonathan H. Chen, Nigam H. Shah, Sami Lachgar, Philip Andrew Mansfield, and 16 others. 2025. Toward expert-level medical question answering with large language models.

Yu Sun, Xingyu Qian, Weiwen Xu, Hao Zhang, Chenghao Xiao, Long Li, Deli Zhao, Wenbing Huang, Tingyang Xu, Qifeng Bai, and Yu Rong. 2025. ReasonMed: A 370K multi-agent generated dataset for advancing medical reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26457–26478, Suzhou, China.

Ekaterina Sviridova, Anar Yeginbergen, Ainara Estarrona, Elena Cabrio, Serena Villata, and Rodrigo Agerri. 2024. CasiMedicos-Arg: A Medical Question Answering Dataset Annotated with Explanatory Argumentative Structures. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18463–18475.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *Preprint*, arXiv:2211.09085.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Qwen Team. 2025. Qwq-32b: Embracing the power of reinforcement learning.

Rahul Thapa, Qingyang Wu, Kevin Wu, Harrison Zhang, Angela Zhang, Eric Wu, Haotian Ye, Suhana Bedi, Nevin Aresh, Joseph Boen, Shriya Reddy, Ben Athiwaratkun, Shuaiwen Leon Song, and James Zou. 2025. Disentangling reasoning and knowledge in medical large language models. *Preprint*, arXiv:2505.11462.

Özlem Uzuner, Imre Solti, Fei Xia, and Eithon Cadag. 2010. Community annotation experiment for ground truth generation for the i2b2 medication challenge. *Journal of the American Medical Informatics Association*, 17(5):519–523.

Guangyu Wang, Guoxing Yang, Zongxin Du, Longjun Fan, and Xiaohu Li. 2023. Clinicalgpt: Large language models finetuned with diverse medical data and comprehensive evaluation. *Preprint*, arXiv:2306.09968.

Wenxuan Wang, Zizhan Ma, Meidan Ding, Shiyi Zheng, Shengyuan Liu, Jie Liu, Jiaming Ji, Wenting Chen, Xiang Li, Linlin Shen, and Yixuan Yuan. 2025. Medical reasoning in the era of llms: A systematic review of enhancement techniques and applications. *Preprint*, arXiv:2508.00669.

Chih-Hsuan Wei, Yifan Peng, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Jiao Li, Thomas C. Wiegers, and Zhiyong Lu. 2016. Assessing the state of the art in biomedical relation extraction: overview of the biocreative v chemical-disease relation (cdr) task. *Database*, 2016:baw032.

C. Wu, W. Lin, X. Zhang, Y. Zhang, W. Xie, and Y. Wang. 2024. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, 31(9):1833–1843.

Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu, Hyunjin Cho, Chang-In Choi, Yihan Cao, Hui Ren, Xiang Li, Xiaoxiao Li, and Yuyin Zhou. 2025. Medreason:

Eliciting factual medical reasoning steps in llms via knowledge graphs. *Preprint*, arXiv:2504.00993.

xiaoju ye. 2023. calflops: a flops and params calculate tool for neural networks in pytorch framework.

Yong Xie, Karan Aggarwal, and Aitzaz Ahmad. 2024. Efficient continual pre-training for building domain specific large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10184–10201, Bangkok, Thailand. Association for Computational Linguistics.

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, and 40 others. 2025. Qwen3 technical report. *CoRR*, abs/2505.09388.

Hongzhou Yu, Tianhao Cheng, Yingwen Wang, Wen He, Qing Wang, Ying Cheng, Yuejie Zhang, Rui Feng, and Xiaobo Zhang. 2025. FinemedLM-o1: Enhancing medical knowledge reasoning ability of LLM from supervised fine-tuning to test-time training. In *Second Conference on Language Modeling*.

Duzhen Zhang, Zhong-Zhi Li, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Xiuyi Chen, Yingying Zhang, Fei Yin, Jiahua Dong, Zhijiang Guo, Le Song, and Cheng-Lin Liu. 2025. From system 1 to system 2: A survey of reasoning large language models. *IEEE Trans. Pattern Anal. Mach. Intell.*, PP:1–20.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2023. Instruction tuning for large language models: A survey. *CoRR*, abs/2308.10792.

Yu Zhao, Huifeng Yin, Bo Zeng, Hao Wang, Tianqi Shi, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2024. Marco-o1: Towards open reasoning models for open-ended solutions. *Preprint*, arXiv:2411.14405.

Andrea Zugarini and Leonardo Rigutini. 2025. PharmaER.IT: an Italian dataset for entity recognition in the pharmaceutical domain. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics*.

Çağatay Yıldız, Nishaanth Kanna Ravichandran, Nitin Sharma, Matthias Bethge, and Beyza Ermis. 2025. Investigating continual pretraining in large language models: Insights and implications. *Preprint*, arXiv:2402.17400.