LLMD: A Large Language Model for Interpreting Longitudinal Medical Records

[Preprint - June, 2024. Do not distribute.]

Robert Porter, Benjamin Pastel, J. Henry Hinnefeld, Adam Diehl, Lawson Nerenberg,
Pye Maung, Sebastien Kerbrat, Gillian Hanson, Troy Astorino, Stephen J. Tarsa
PicnicHealth
San Francisco, CA, USA

ABSTRACT

This preprint introduces LLMD, a large language model (LLM) that uses medical records to characterize patient health over time, and is deployed today in applications that improve health outcomes and power clinical research. Along with general medical knowledge, LLMD is trained on labeled longitudinal medical records, giving it unique advantages over LLMs trained on knowledge alone, on unlabeled records, or on records from a single health system. We show that LLMD learns to make nuanced connections in information covering years of patient care documented across facilities, and that these are critical to real-world accuracy.

LLMD is trained by *instruction-fine-tuning* a foundational model on millions of records, spanning an average of 10 years and as many as 140 care sites per patient. LLMD's *structuring* tasks jointly identify and normalize metadata, provenance information, clinical named entities, and ontology mappings. *Abstraction* tasks then roll this data into higher-level representations, such a continuous era of time a patient was on a medication. LLMD is deployed within a multi-layered validation system implementing both continual random audits and configurable review by experts, e.g. based on output uncertainty, disease-specific rules, or use-case. This provides both a feedback loop to improve LLMD, as well as fine-grained control over data quality for a spectrum of needs, from lowest cost to regulatory-grade auditability.

LLMD exhibits large gains over both more-powerful generalized models and domain-specific models. On medical knowledge benchmarks like MedMCQA and PubMedQA, LLMD-7B's zero shot accuracy outperforms comparable models. On real-world production tasks, LLMD performs 2x better than GPT-4 when structuring records, and 1.5x better than specialized models like John Snow Labs's "John" when abstracting records. Today, LLMD powers patient-facing tools for care management, as well as research datasets behind 60+ studies, including data submitted to the FDA.

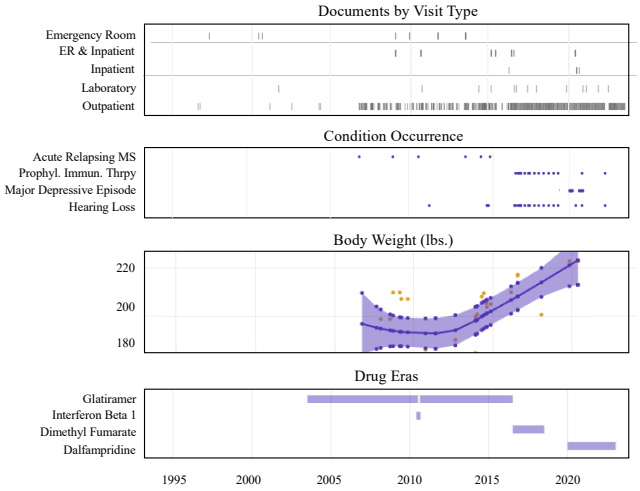

Figure 1: Data from a patient on our platform with Multiple Sclerosis (MS). Their health information spans 27 years of care, 705 visits to 40 specialists providers, and 5 health systems in 2 states. Once longitundinal records are structured, we can abstract measurements and drug eras to intuitively model this patient's care journey.

1 INTRODUCTION

LLMD is the large language model (LLM) behind patient- and research-facing products offered by PicnicHealth. LLMs represent an astonishing breakthrough in Artificial Intelligence (AI) [1–3], and exhibit nuanced pattern matching and broad information recall capabilities. In the medical domain, LLMs fine-tuned on generalized medical knowledge can appropriately respond to licensing exam questions and patient queries [2, 4–6]. Techniques like Retrieval Augmented Generation (RAG) and Chain of Thought (CoT) suggest paths to deployment for applications that demand trustworthy outputs, e.g. by citing evidence to support responses [7, 8]. Building on this promise, this paper presents an LLM that closes the loop for use cases requiring a detailed understanding of a patient's health and care over time at the highest accuracy standards.

Today, narrative text in medical records is the richest source of patient health information available to such applications. However, tapping into it remains difficult due to important patient privacy concerns, technical barriers to access, and the difficulty of modeling contents [9]. Compounding these issues, we show that longitudinality matters – information spanning many records from many facilities is critical to an accurate picture of patient health. For example, we find contradictory diagnoses of disease subtype among 30% of hemophilia patients in our dataset. For them, an LLM answering even simple questions about their primary condition must be able to weigh information from several records based who recorded it, when, and what connections to other evidence can be inferred.

LLMD does just that. Its training combines medical knowledge with labeled longitudinal medical records (Figure 2) from the PicnicHealth platform, which retrieves records for patients and provides tools to help them manage their healthcare journey. Researchers can also leverage the platform to design, monitor, and run observational studies in collaboration with willing patients, producing datasets that support therapy development. By bringing powerful AI advances together with a scalable mechanism for

data access, active patient participation, cross-domain expertise, and need-driven use cases, we demonstrate safe and successful deployment of a medical LLM in the real world.

LLMD is an instruction-fine-tuned [10] version of the `llama` [3] open-weight model trained on three categories of tasks: *(1) general medical knowledge*, *(2) structuring* tasks that produce normalized, validated data from arbitrary record contents, and *(3) abstraction* tasks that mimic clinicians to capture the clinical view of patient health. Only once records are both structured *and* abstracted do we find it possible to draw insights from a patient's data (Figure 1). LLMD is trained on labeled data from millions of records, retrieved from 100,000+ sites, covering decades of care for most patients. Labels for each record include metadata and data provenance information, clinical entities, ontology mappings, and abstracted variables. At its most granular, our training data has 350M labels from nearly 10 years of labeling by human *clinical data abstractors* (CDAs) [11]. For the results in this paper, we curate those labels to 5B tokens, consisting of 4M prompt/completion pairs.

In production, LLMD's outputs are subject to multiple layers of validation to ensure consistency and accuracy. These include secondary models that predict performance relative to CDAs performing the same task, rule-based data conformance and plausibility checks, and manual auditing by CDAs and clinicians. Outputs failing at any layer are corrected or suppressed and folded into future training data. These mechanisms are configurable based on use case, allowing us to process low risk data quickly and efficiently when appropriate, or guarantee that a CDA verifies data using protocols acceptable to regulators when needed.

LLMD's direct outputs outperform comparable models on benchmarks for general medical knowledge and tasks reflecting real-world structuring and abstraction. This includes best-in-class performance on PubMedQA among LLMs with the same parameter count. On production tasks for real-world records, LLMD significantly outperforms both GPT-4 and John Snow Labs's "John", which advertises use on medical records. Importantly, we develop strong evidence that today's LLMs *must be trained on labeled records to accurately model patient health and care* – those trained on general medical knowledge and unlabeled records alone cannot consistently contend with the complexity and nuances of records.

This paper proceeds as follows: Section 2 describes the PicnicHealth platform. Section 3 introduces our tasks, training dataset, and context generation procedures. Section 4 discusses validation mechanisms. Section 5 evaluates LLMD against alternatives on common benchmarks, tasks pulled from our production data, and investigates performance on infrequent but clinically important long-tail concepts.

2 PICNICHEALTH PLATFORM OVERVIEW

PicnicHealth's platform works on behalf of patients to retrieve and manage their medical records, regardless of the format they are in or the facility holding them. This includes electronic records as well as paper-based records, which account for a substantial portion of patient data. Paper records are of particular importance for visits to providers practicing outside of large health systems, for historical records produced before Electronic Health Records (EHRs) were ubiquitous, and for facilities whose systems impede

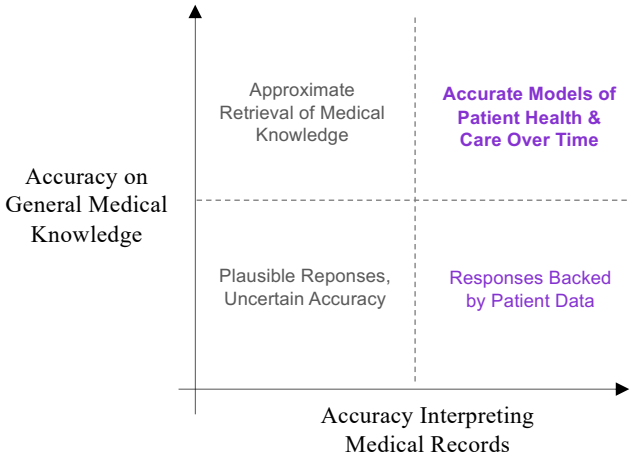

Figure 2: The ability to interpret medical records is critical for LLMs that model patient health using the data available today.

sharing. The platform is format-agnostic under its base case support for processing physical copies of records. It is also facility agnostic based on the legal right patients have to access records on request.

Our platform offers products for patients and researchers. For patients, we retrieve records, organize them, visualize contents, and enable sharing with their doctors. These tools help patients manage day-to-day care and are the basis for concierge care coordination services. For researchers, we offer products that improve the speed, flexibility, and cost of observational studies supporting new therapies. Advantages come from better patient recruiting, more effective data collection, and transparent data management throughout the lifecycle of a study.

3 TASKS & TRAINING DATA

This section introduces the tasks we use to train LLMD and process the records on our platform. Any task performed by our LLM can also be performed by a human CDA using software developed internally. This software implements a heavily optimized user experience to ensure efficiency and consistency among CDAs. CDAs perform training data collection, model auditing, label correction, and inter-rater reliability studies against clinicians.

As detailed in Section 4, our LLM is able of doing a substantial portion of work on incoming records automatically to the same accuracy as CDAs. In cases where our validation system flags tasks that need extra scrutiny, or when data is intended for sensitive use cases like a regulatory filing, CDAs always complete the task, and the LLM outputs act as a tool to make them more efficient.

3.1 Structuring Tasks

Even once available, the challenge of turning medical records into usable, reliable data is daunting [12, 13]. Issues found in their pages include contradictions, errors, omissions, and even notoriously difficult-to-read handwriting for paper records. Pervasive problems like the misdiagnoses in Section 1 can happen for reasons as mundane as a provider choosing the wrong option from a drop down list in EHR software. Or, a Medication List section that is intended

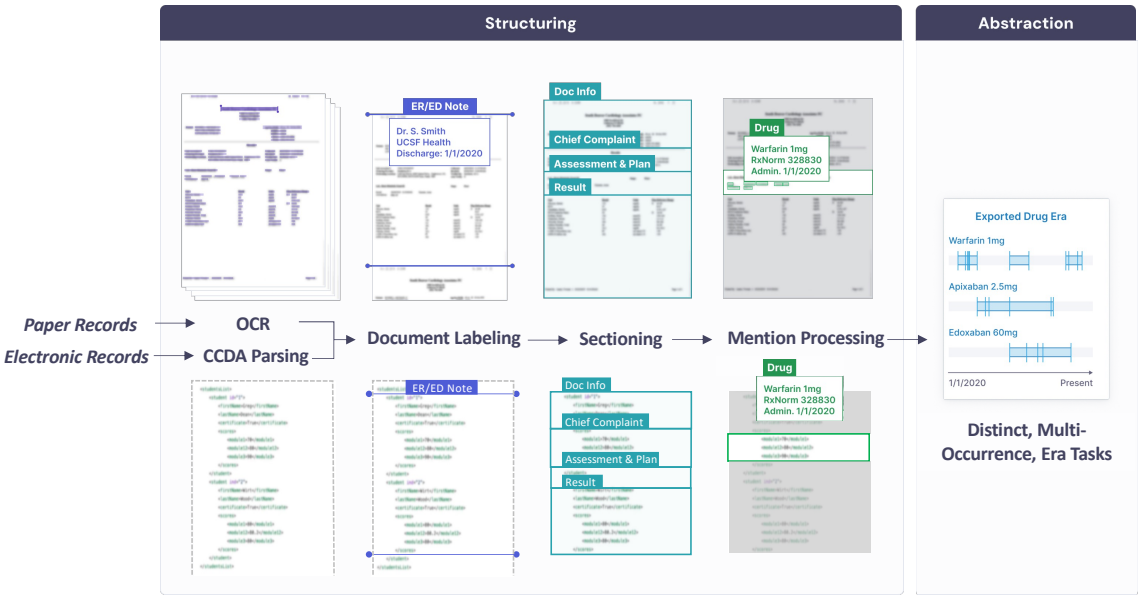

Figure 3: The workflow implemented by our internal software to structure and abstract records sourced electronically and on paper.

to be a universal source of truth may combine patient recollections with provider sourced data. Even aligning data representing the same real-world concept is a challenge, and though ontologies such as SNOMED, RxNorm, and the ICD standard are intended as solutions, coding standards change over time, and across facilities and providers [14–16].

For all of these reasons, getting data out of records and into a structured form suitable for modeling – in our case, following the OMOP Common Data Model – requires more than just digitization for paper records or parsing for electronic records. Our *structuring* tasks implement the following steps (Figure 3):

- **OCR/CCDA Parsing -** Accurate text is critical to all downstream processing. For paper records, we apply an OCR model trained on 6.2B annotated words and bounding boxes, capturing the layouts, styles, artifacts, and language common in our records. We access electronic records in the Consolidated Clinical Document Architecture (CCDA) format and parse them to access text directly [17] [1].
- **Document Labeling -** We tag all incoming records with document-level metadata. This includes document boundaries, e.g. since facilities often consolidate information for many visits, as well as attributes like visit length and type, provider identity and specialty, and normalized facility names.
- **Sectioning -** We then subdivide documents into sections, which offer fine-grained provenance information. This step includes finding section boundaries based on arbitrary contents and, for paper records, physical layouts. We then code section types such as *Progress Note*, *Pathology Report*, etc.
- **Mentions-Processing -** Similar to named entity recognition (NER) [18], we identify clinical concepts, their attributes, and the relationships that link them. We do this for medications, lab tests, vital signs, procedures, and conditions. Example attributes include reference ranges for labs, doses for medications, etc. We then align the resulting *entity mentions* to standard ontologies.

3.2 Abstraction Tasks

While structured data reflects what is *written*, our abstracted data represents the clinical view of a patient's medical history. For example, a structuring task will recognize and normalize every mention of a disease modifying therapy (DMT), but those named-entities alone are not enough to confidently say when a patient started the drug, when they stopped, and why [19].

This is apparent in Figure 4, which shows snippets fed to our LLM as context when abstracting the treatment course for a medication. We see that the most recent note captures a discussion of bladder control difficulties alongside two medications. It definitively notes the date that the DMT was stopped, which is confirmed by the prior note. However, only the oldest note calls out the patient's worsening side effects to give us the stop reason, while leaving unclear what the date of final administration would be. In this example, all three notes must be examined together to model the treatment course – no *one* note nor its structured data tells the full story – and this need motivates abstraction.

Our approach to designing abstraction tasks started from the observation that clinicians intuitively abstract medical information as they read records. Through user study, we discovered that the key to abstracting a nuanced treatment course lay in supplementing structured data with a provider's clinical knowledge and provenance information to contextualize and filter what was written.

[1] Not to be confused with clinical data abstractors (CDA)

New tasks – to be performed by either CDAs or LLMD– are defined by configuring the desired output, the input source material, and the protocol for abstractors to follow. The output is a target concept, e.g. a drug concept code, and one of three data types:

- **Distinct** attributes, such as a primary diagnosis.
- **Multi-occurrence** episodic events, such as pain crises or clinical relapses.
- **Eras** that capture spans of time associated with a clinical status, such as when a patient was on a medication.

The inputs define what context from the patient's full set of records to consider when abstracting, which is critical to consistency, efficiency, and auditability. Inputs are configured based on metadata, such as document type, date, and provider specialty, as well as search hits for concepts related to the output. The protocol consists of definitions, guidelines, and examples to mold clinical expertise into a rigorous, repeatable process. They are designed collaboratively by clinicians and researchers, and can include multiple rounds of training, assessment, feedback, and revision.

In our prior example, drug era abstraction would be instantiated with the DMT's associated RxNorm code and an Era datatype; source material would be configured to look to Progress Note sections from neurology documents, as well as hits of any DMT associated with Multiple Sclerosis (MS); and the protocol would provide guidelines for navigating ambiguities, such as contradictions between primary care and specialist providers, or guidelines on how to identify uncertain drug start dates.

Ultimately, for LLMD, abstraction tasks enable us to train a large language model to mimic clinicians. When a new task is launched to CDAs at scale, their outputs become the labels for training. The configuration of input source material is the basis for LLM context generation. And, the abstraction protocols provided to CDAs become the starting point for task prompts.

3.3 General Medical Knowledge Tasks

Training LLMs for general medical knowledge is well studied, and for completeness, we refer readers to procedures used in OpenBioLLM [2] and Med-Palm2 [20]. In our fine-tune of `llama2`, we focus on public source material related to clinical knowledge and synonym data that helps navigate different coding styles.

3.4 Longitudinal Task Labels

Table 1 categorizes the 86 tasks used to train LLMD today. Figure 5 provides an example prompt for drug NER structuring, while Figure 6 provides an example for drug era abstraction. Each task is paired with labels collected from our corpus, which contains millions of annotated records. These have been both structured and abstracted for research questions associated with 60+ study datasets. Our records are sourced from 100k+ care sites in aggregate. Individual patients have records from 10 facilities on average, but as many as 140 for some patients. On average, patient data spans 10 years, and 20 years at the 90th percentile. Disease specific *completeness* definitions determine when our data fully documents a patient's primary condition over time, e.g. at-least one neurology office visit per 18 months for MS; today our data contains 5 years of complete documentation per patient on average. Data volume is substantial: Paper records are on average 30 pages long, though we have retrieved records as long as 24,000 pages for heavy users of the healthcare system with complex diseases; meanwhile, we electronically retrieve 290 files per patient on average; and for image data, we retrieve on average 5,000 image slices per patient. Labeling was conducted over a period of approximately 10 years by a workforce of several thousand CDAs in sum, and has generated more than 350M labels.

Capability	# Task Types	% Training Samples
General Medical Knowledge	17	29%
Structuring - Metadata & Provenance	14	3%
Structuring - Conditions	18	31%
Structuring - Medications	12	9%
Structuring - Labs	7	4%
Structuring - Vital Signs	6	7%
Abstraction - Conditions	4	4%
Abstraction - Medications	8	13%

Table 1: For each stage of our structuring and abstraction processes, we perform multiple tasks in sequence, e.g. finding vital sign names, and then value for each vital sign name in a second pass.

3.5 Task Decomposition and Context Generation

The capabilities in Table 1 contain multiple task templates. These arise in part when we decompose complex outputs into easier tasks that form dependency chains. For example, with vital signs we use one task to first identify all vital sign names within a section, another to find attributes given a name, and finally a last task to normalize units and align names to an ontology. This decomposition injects a helpful inductive bias between the measurement name and its attributes – given "Body Weight", the model is more likely to latch onto numbers in typical ranges for pounds and kilograms than those for degrees Farenheit. Though simple, such induction greatly improves tolerance to the variations and artifacts we encounter across facilities, e.g. when tables are intermingled with narrative text in unexpected ways.

Beyond exploiting instruction fine-tuning to shape inductive biases, we engineer the context of each task carefully. Similar to REALM [21] and RAG [7] we build context by finding relevant snippets of records and concatenating them into our prompt. As discussed in Section 3.2, context retrieval is configurable based on metadata and search terms, and we implement it today using ElasticSearch. We found that this system design struck the best initial tradeoff between performance and flexibility, allowing us to quickly iterate task design. In the future, we expect to explore end-to-end training [22], though we note that it comes with more complex system interdependencies.

4 SAFETY AND QUALITY CHECKING

Both LLMD's outputs and CDA labels pass through several layers of validation before they are shown to users or included in a research dataset. This section summarizes them.

```
### Snippets
7293 - Progress Note on 2019-12-02 Neurology, found concepts [ocrelizumab]:
   HISTORY OF PRESENT ILLNESS:
   first_name was last seen on 5/31/2019 first_name has decided that since starting the DMT, she has been on a "downhill slide". She has received 5 doses first_name
         reports feeling worse after each dose and is overwhelmed by the side effects. She presents today using her walker to assist with

7296 - Progress Note on 2020-12-09 Neurology, found concepts [ocrelizumab]:
   HISTORY OF PRESENT ILLNESS:
   first_name was last seen on 06/02/2020 first_name_1 her most recent dose of DMT was on 05/31/2019, she has been off DMT since that time. first_name feels that with
         oxybutynin her bladder control is better at night, but she still experiences frequency (probably

7305 - Progress Note on 2021-06-30 Neurology, found concepts [ocrelizumab]:
   first_name was last seen 6/02/2020. She has been off DMT since 5/31/19, when she received her most recent dose of DMT. first_name has some bladder urgency first_name
         wears incontinence pads daily. She takes oxybutynin ER 15mg before bed but still wakes 1-2 times per night to void. She continues to use stool softeners to
         help her move her bowels daily.
```

Figure 4: Neurology notes provided as context to a drug era task that includes *stop reason* for an MS patient. We first turn to notes from specialist physicians for the most definitive account of the disease. Here, the most recent two notes co-mingle a discussion of worsening bladder control with a definitive date the patient stopped the DMT. The oldest note makes a link between consistent worsening feelings and the drug, which provides the stop reason. Only with these three notes together can we piece together the era end date and stop reason.

```
### Section:
 IMPRESSION:
 She continues symptomatically and I'm going to reissue a prescription for
prednisone 1000 mg a day for 5 days. I am hoping this will quiet down her
symptomatology. If she continues with c she will contact me. We had a long
discussion about what to do for her. She has decided she would like to switch
from Copaxone to DMT and I think this is a good choice. She will have blood
today for JC virus and vitamin D level. She will help the paperwork for DMT and
I will see her in followup in January. I have asked her to call as needed.

### Instructions:
In the section above, find all unique mentions of these concepts:
 - Avonex (RxNorm 153326)
 - Copaxone (RxNorm 135779)
 - Ocrevus (RxNorm 1876381)
 - Tecfidera (RxNorm 1373484)
 - DMT (RxNorm DMT_CODE)
 - Zeposia (RxNorm 2288407)
 - dimethyl fumarate (RxNorm 1373478)
 - rituximab-abbs (RxNorm 2105824)
 - teriflunomide (RxNorm 1310520)

Output a list of JSON objects like:
[
    {
        "concept": "Hydrea (RxNorm 151871)",
        "other_fields": {
            ...
        },
    },
    ...
]
With these optional other fields:
 - ANY__SUBJECT (who the mention refers to)
 - DRUG__INSTRUCTION (the frequency a patient is to take a medication; often
                      daily, every other day, etc.)
 - DRUG__STOP_REASON (why the patient stopped taking the drug (a SNOMED item
                      and code))
 - DRUG__STRENGTH (the drug dosage; often in mgs or mg/mL)
```

Figure 5: This task implements clinical named entity recognition, one of our simpler structuring tasks.

4.1 Uncertainty-Driven Manual Review

The goal of LLMD is to automate record processing, while maintaining the same accuracy as clinicians. Even for them, we observe that some records are far more difficult to understand than others and the reasons are complex. For example, a poor quality scan of a decades-old handwritten note likely induces more mistakes due to confusion about the text than the output of a modern, widely used EHR system. At the same time, we see that modern EHRs produce large amounts of redundant information, spreading the most important data sparsely over many pages.

For these reasons, we train secondary *uncertainty models* to classify when the outputs of LLMD are likely to need a second-pass review. Today, these are analytic classifiers, not LLMs, and take into account features such as LLMD's logits, its outputs, information

```
### Snippets
7493 - Progress Note on 2011-06-16 Psychiatry, found concepts [DMT1]:
   use street drugs, doesn't smoke; has caffeine in a.m.
   MEDICAL HISTORY Is on an ex perimental medication (infusions of DMT1)
   for MS under the care of Dr Fox and is doing fairly well with his MS.
   Has had partial

7485 - Progress Note on 2014-02-17 Neurology, found concepts [DMT1]:
   HPI:
   Mr. last_name is here for Camms EXT Visit Month 36. His last
   DMT1 was Month 12 in 1/25/2010. Reports c/o pneumonia
   (start 2/3/2014). Seen by PCP. Chest X-ray abnormal. Started Prednisone 50
   mg po qd, Zithromax Z pak,

7498 - Radiology Report on 2014-02-20, found concepts [DMT2]:
   MRI BRAIN WITH AND WITHOUT CONTRAST: 2/20/2014
   HISTORY: Multiple sclerosis. There is no indication the patient is on DMT2
   COMPARISON: Head CT 03/18/2011 (Seton Northwest Hospital), report, a
      previous outside MRI has been

7483 - Progress Note on 2014-08-19 Neurology, found concepts [DMT1]:
   HPI:
   Mr. last_name is here for CAMMS Extension Visit Month 42. Last
   DMT1 January 2010. No PE to be done today. No new symptoms,
   no relapse. Informed consent given, all questions answered. Signed. Copy
   to patient. No new neurologic



### Instructions:
Predict the expected drug eras after observing these snippets.

Use this format for "start_date" and "end_date":
  {"date": "YYYY-MM-DD",
   "precision": "DAY|MONTH|YEAR",
   "specificity": "KNOWN_DATE|TRUE_DATE"}.

Choose "TRUE_DATE" if the patient started or ended the  drug on that date.
  Otherwise, choose "KNOWN_DATE" for the first or last recorded usage of
  the drug.

Return a JSON list of drug eras. Each drug era should use
the format
{
    "concept_name": "...",
    "start_date": {...},
    "end_date": {...},
    "stop_reason": "...",
}

For this task you should consider the following concepts:
- DMT1 (synonyms BRAND_NAME1)
- DMT2 (synonyms BRAND_NAME2)
```

Figure 6: This is an example of a drug era task for medications associated with MS.

about input text, OCR quality, and document metadata. They are trained to detect when outputs fall short of gold standard labels, while accounting for acceptable variations between CDAs. We audit routing decisions continually in production by randomly selecting extra tasks for review by CDAs – the resulting dataset can then compared to the uncertainty-model's expected performance. Should performance slip, data can be reprocessed and the model retrained

or recalibrated. We remark that these models are highly accurate, though their implementation is not discussed in detail in this paper.

We note that *abstraction* tasks for research studies are *not* processed in a fully automated fashion today. This is because they are often intended for use in research studies that require human verification to meet regulatory standards. LLMD's abstraction outputs are instead hypotheses that can speed CDAs' work and we measure LLMD's impact in CDA task-time for a fixed accuracy bar.

4.2 Rules-Based QC

All outputs of LLMD and CDA work are subjected to rules-based quality control (QC) for data conformance and plausibility. Conformance checks look to ensure basic correctness, e.g. that dates are valid, codes are present in an ontology, and attributes that cannot be null are indeed populated. Plausibility rules incorporate more clinical and disease-specific knowledge, for example that a DMT does not start before a confirmed diagnosis date, or that conditions only possible in females are not associated with male patients. When a rule violation is detected, it is logged with a 'warn' or 'error' priority level. Errors are prevented at point of entry, while warnings are routed to an escalation workflow for manual correction or suppression. Both general and disease-specific QC rules are created and continually expanded by a team of epidimiologists, clinicians, and biostatisticians. An example set our plausibility rules in production today is shown in Table 2.

Cond.	Type	Plausibility Rule Description
PNH	Err	PNH breakthrough hemolysis occurs within a drug era
PNH	Warn	LDH collection date +/- 3 days from breakthrough hemolysis start date
PNH	Warn	Acute kidney injury era has a plausible duration (> 100 days)
PNH	Warn	PNH drug treatment eras should not overlap
PNH	Warn	Eculizumab dose should not be less than 600 mg

Table 2: A subset of plausibility rules applied during abstraction tasks for patients diagnosed with Paroxysmal Nocturnal Hemoglobinuria (PNH). These rules are continually expanded by a team of epidemiologists, clinicians, and biostatisticians.

4.3 Agreement and Accuracy

Labels assigned or verified by CDAs are subject to additional quality-checking (QC) tasks that ensure consistent performance over time and among CDAs [23]. For abstraction, a second blinded task is performed based on a configurable sampling rate. In cases of disagreement, the result is adjudicated by a third CDA. We also perform a smaller number of random audits by clinicians with a higher level of expertise than CDAs to ensure *consistent* results are indeed *correct*. For structuring tasks, which involve smaller units of work and less clinical judgment, QC tasks are not blinded and are performed by team members identified to be high performers. All QC sampling rates are configurable by percent of data volume, by concept, by CDA performance level, and by study in the case of research datasets.

5 LLMD TRAINING & EVALUATION

In this preprint, we present results from a small version of our LLM fine tuned from Meta's Llama2-7B foundational model. These results allow us to characterize our methods relative to others in the literature, though we note that they do not represent top-line system performance nor the most sophisticated foundational model in production today.

Our evaluation model is trained by performing a full fine-tuning of all weights. We use a notably larger dataset than in common practice [24]. This is based on the observation that we are not *just* trying to encode additional knowledge in the model, but that we must build tolerance to an unusual set of artifacts that are absent from the pretraining dataset and evident across the long-tail of our dataset. We train for a single epoch on 5B tokens, regularize using loss smoothing [25], and linearly ramp loss from 0 to $2.0e^{-5}$ over 500 steps before linearly decaying back to 0.

We evaluate LLMD from several angles. We first compare its direct outputs on common benchmarks, including MedMCQA [26] and PubMedQA [27]. We then assess performance on a held-out sample of tasks and data from our production systems, which better characterize real-world applications operating on a wider, more diverse patient population than popular benchmarks [28]. The results of this analysis illuminate performance characteristics on the critical path to deployment. Finally, we evaluate LLMD's performance on long-tail concepts that are infrequent but clinically important.

5.1 Common Benchmarks

Both MedMCQA and PubMedQA's training datasets are included in LLMD's General Medical Knowledge tasks, allowing us to report zero shot results without modifying LLMD's training methodology or dataset. Figure 7 shows leading zero shot accuracies on MedMCQA, a suite of challenges mimicking U.S. medical entrance exams as well as zero shot performance on PubMedQA. Comparison results come from the most recent written analysis by the Open Medical LLM leaderboard for LLMs also trained from 7B parameter foundational models [29]. We also include results for the production version of GPT-4, which we analyze in more detail in Section 5.2.

LLMD-7B outperforms all comparable models on PubMedQA, and performs close to GPT-4's production model, which is much larger and implements a mixture of experts model. On MedMCQA, its performance is similar to the most powerful comparable models, which undergo extensive training and knowledge distillation for medical question answering. This strong performance is notable since our training data mix heavily emphasizes records structuring and abstraction, while following a lighter-weight approach to general medical knowledge. One explanation is that the our real-world records data not only encapsulates similar information (e.g. in the notes of doctors or explanations to patients), but also includes many examples of how medical knowledge manifests in data collected for patients. We find support for this latter case in Section 5.2.

		gemma-7B	BioMistral-7B-Zephyr	Hermes-2-Pro-7B	GPT-4 Prod.	JSL-J	LLMD-7B
General Medical Knowledge *(Open Benchmarks)*	MedMCQA	62.9	61.3	62.3	**74.7**		58.0
	PubMedQA	50.8	51.1	53.1	**67.6**		67.0
	Overall (Avg)	56.9	56.2	**57.7**	71.2		62.5
Interpreting Records *(Production Tasks & Data)*	Structuring				44.0	40.7	**95.6**
	Abstraction				73.0	57.1	**87.0**
	Overall				50.9	45.0	**93.5**

Figure 7: LLMD performs significantly better on the types of tasks needed to structure and abstract medical records in production.

5.2 Production Workload Accuracy

Our production tasks allow us to analyze the strengths and weaknesses of models when working with directly-useful tasks and a broad, representative patient population. In this section, we compare LLMD-7B to OpenAI's GPT-4 production model [6], a best-in-class and heavily used general model, which performed best among the frontier models we evaluated. We also compare against John Snow Labs's "John" LLM (JSL-J), a domain-specific model that claims state-of-the-art performance and is advertised for use interpreting medical records [30]. JSL-J is trained to incorporate structuring approaches considered among the best in the lifesciences industry. We report accuracy against gold standard labels assigned and validated by CDAs.

Overall Performance. First, we compute an overall comparison score based on a sample of the structuring and abstraction tasks in Section 3. Figure 7 shows that LLMD-7B handily beats both comparison models, each of which have much larger parameter count. Given our fine tuning approach, this result suggests that powerful off-the-shelf models do not handle structuring and abstraction tasks well without being explicitly trained on them. We note that both comparison models *are* able to solve some examples satisfactorily, but that aggregate performance leaves room for improvement. Contrasting this result with that of Section 5.1, leads to an important observation: **training on general medical knowledge *may not be enough*** to deploy medical LLMs that must model a patient's health and treatment over time.

Task Type. Figure 7 also shows performance broken out by structuring and abstraction. LLMD-7B beats both GPT-4 and JSL-J in both categories. We observe that GPT-4 performs better than JSL-J on structuring tasks, showing off the power of its strong pattern matching capabilities. We hypothesize that this is partly due to GPT-4's stronger foundational model, since we observe JSL-J struggling to following the instructions of our prompts, which GPT-4 handles well. GPT-4 also performs better than JSL-J on abstraction tasks, supportng the belief that domain knowledge reflected in benchmarks like PubMedQA is not enough to teach an LLM to characterize patient health.

Examining individual responses, we also observe that LLMD-7B's responses are more internally consistent than comparison models. In one example, JSL-J identifies a patient's height value correctly as "6" but improperly assigns the "inches" unit, which also appears in the LLM context window. While that choice is *close* – inches are a valid unit for height – the result is implausible given the patient age. This type of inconsistency is something that LLMD-7B's training dataset directly combats: it includes a massive number examples of vital signs for people in various states of health, as well as enough information to associate them with age, health, etc.

We see strong evidence that LLMD-7B learns these connections. While it may on occasion produce implausible results – motivating the layers of QC in Section 4 – we observe them far less often than with comparison models. The gap is most apparent on tasks that require manipulating lab test codes and medication identifiers. In these cases, we observe a substantial rate of hallucinated codes by GPT-4 and JSL-J. Examples of this type provide evidence of the importance of training on data showing medical knowledge manifest in records, not just on the written ideas themselves.

Figure 8 annotates some tasks based on whether they require nuanced reasoning. These include tasks that are interpretive in nature, for example to determine whether a visit was in an inpatient or outpatient setting. Nuance also comes into play often when records contain contradictory information that requires subtle or disease specific adjudication. For example, should a medication be found in a Medication List, but a narrative Progress Note from the same day says that medication was stopped, an LLM must learn to defer to the Progress Note. Interestingly, we find that GPT-4 does well, but that even with few parameters, LLMD-7B significantly outperforms it.

A few specific examples demonstrate that even the nuances of mundane-seeming metadata can lead to poor application-level behaviors. In many failure cases, we saw comparison models confused by the meaning of dates in notes – these tasks are also categorized in Figure 8. Looking into the records themselves, we see dates and

Measurement Name	Abstracted Variable Type
Type 2 RBC clone size	Occurrence
Monocyte clone size	Occurrence
Granulocyte clone size	Occurrence
Total RBC clone size	Occurrence
Type 3 RBC clone size	Occurrence
Lactate dehydrogenase [Enzymatic activity/volume] in Serum or Plasma	Occurrence

Table 3: A PNH Marker Panel provides an example of infrequent but important measurements. These tests are performed once when a patient is diagnosed; PNH occurs in fewer than 10 in 1M people.

times documenting facility workflows, such as when notes are written, amended, signed, or when test samples are sent off to a lab, returned, etc. Identifying dates using NER is easy, but to interpret records, models must learn the patterns of this workflow information, which do not appear in pre-training data. We observe similar failures navigating the names of personnel listed in records.

For a patient-facing application built on these outputs, this means an LLM will produce plausible results that don't match patient expectations. Patients are quick to recognize their doctors and compare application outputs to their recollection of physical visits; getting these nuances wrong has the potential to undermine trust in these and other outputs of the LLM. We note that LLMD-7B does categorically very well on this type of metadata, providing another example of how training directly on real-world patient data helps us build far more trustworthy applications.

5.3 Long Tail Performance

In this section, we examine LLMD's performance on two specific sets of labs: the top-100 most-common and 100 tests deemed by our clinical team to be both rare and clinically important, which we refer to as long-tail labs. An example of this latter set are measurements associated with the marker panel administered to patients diagnosed with PNH (Table 3). Given the disease's incidence of less than ten per million people, the frequency of these tests is very low in most data samples, but their importance high.

In data audited over the course of April, 2024, we find precision and recall on our top 100 labs strictly above those computed from agreement studies between two CDAs performing manual abstraction. This indicates that LLMD's outputs after validation are as-good or better than a trained human abstractor. Among the set of long tail labs, we find that 60 of the 100 appeared more than 10 times in our audit sample – of these, only 15% of these had an F1 score below 0.80, suggesting that performance in the long tail is good, but not guaranteed. In practice, when we detect this, we are able to flag sections for patients with the associated disease for manual review by CDAs, implement QC rules to ensure we find expected measurements, and ultimately retrain LLMD.

We have experimented both with upsampling and data augmentation to shore up long tail concepts, and for both methods find that LLMD responds smoothly. We find these dynamics supportive of our claim that a large labeled dataset is absolutely critical to good performance: precision and recall on these obscure concepts is *not* a given even among the most powerful models, but we do see that LLMs are well-behaved enough that model blindspots are discoverable and addressable. Our results also highlight how important the input of clinicians is, and suggests that disease-by-disease rollout is likely to produce safe results and incremental generalization for LLMs.

6 CONCLUSION

This paper presented LLMD, an LLM capable of modeling patient health and treatment over time from the data available today. We showed that training on tasks and real-world data from existing patient records is *necessary*: even the most powerful and the most knowledgeable models struggle when working with records. This makes the return-on-investment of building applications on general models unfavorable when they must reflect patient health.

Our results are most exciting because LLMD is operating at levels that are improving patient care *today*. User feedback demonstrates patients discovering new things about their health history, advocating for the highest standards of care for themselves, and making better use of precious time with their doctors. Researchers are working the same underlying data, contributed by willing patients who are highly motivated to improve treatment options for themselves and others. To date, this has produced 60+ datasets covering 50+ rare diseases, and has been the basis for compelling evidence submitted to the FDA.

From a technology standpoint, we demonstrated a feedback loop that allows us to improve LLM performance when modeling important and sometimes obscure aspects of patient health. The systems that make up that feedback loop, as well as the experience designing and collaborating with experts puts LLMD in a leading position to deliver new insights about safe, high accuracy healthcare LLMs.

REFERENCES

[1] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, *et al.*, "A survey on evaluation of large language models," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 3, pp. 1–45, 2024.

[2] M. S. Ankit Pal, "Openbiollms: Advancing open-source large language models for healthcare and life sciences." https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B, 2024.

[3] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[4] D. Brin, V. Sorin, E. Konen, G. Nadkarni, B. S. Glicksberg, and E. Klang, "How large language models perform on the united states medical licensing examination: A systematic review," *medRxiv*, pp. 2023–09, 2023.

[5] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, *et al.*, "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, 2023.

[6] H. Nori, N. King, S. McKinney, D. Carignan, and E. Horvitz, "Capabilities of gpt-4 on medical challenge problems," *ArXiv*, vol. abs/2303.13375, 2023.

[7] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[8] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24824–24837, 2022.

[9] R. E. Sherman, S. Anderson, G. D. D. Pan, G. W. Gray, T. Gross, N. L. Hunter, L. LaVange, D. Marinac-Dabic, P. Marks, M. A. Robb, J. Shuren, R. Temple, J. Woodcock, L. Yue*, and R. Califf, "Real-world evidence - what is it and what can it tell us?," *The New England journal of medicine*, vol. 375 23, pp. 2293–2297, 2016.

[10] J. Wei, M. Bosma, V. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, "Finetuned language models are zero-shot learners," *ArXiv*, vol. abs/2109.01652, 2021.

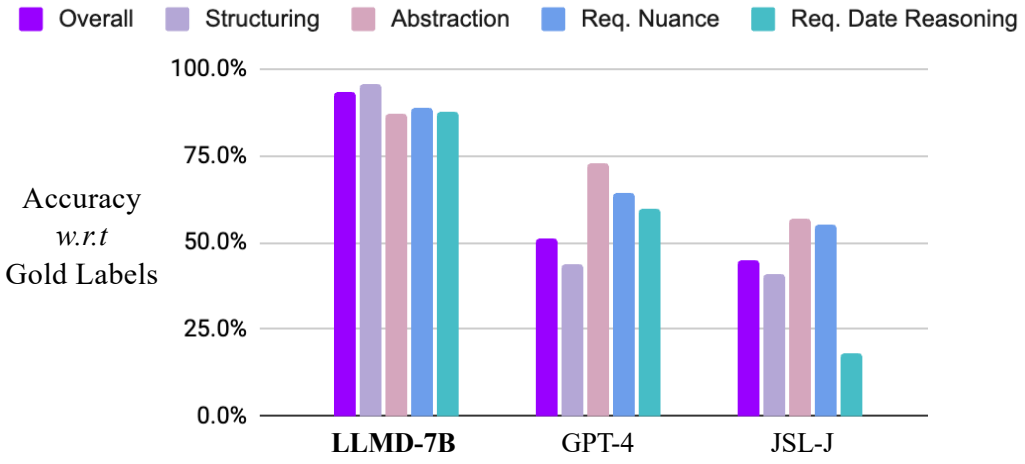

Figure 8: LLMD learns intricate patterns in medical records, including the nuances to resolve conflicting information or to navigate workflow information embedded in records.

[11] "Picnichealth overview and product offerings." http://www.picnichealth.com. Accessed: 2024-06-6.
[12] M. Tayefi, P. D. Ngo, T. Chomutare, H. Dalianis, E. Salvi, A. Budrionis, and F. Godtliebsen, "Challenges and opportunities beyond structured data in analysis of electronic health records," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 13, 2021.
[13] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun, P. Sundberg, H. Yee, K. Zhang, Y. Zhang, G. Flores, G. E. Duggan, J. Irvine, Q. V. Le, K. Litsch, A. Mossin, J. Tansuwan, D. Wang, J. Wexler, J. Wilson, D. Ludwig, S. Volchenboum, K. Chou, M. Pearson, S. Madabushi, N. Shah, A. Butte, M. Howell, C. Cui, G. S. Corrado, and J. Dean, "Scalable and accurate deep learning with electronic health records," *NPJ Digital Medicine*, vol. 1, 2018.
[14] D. Lee, N. D. de Keizer, F. Y. Lau, and R. Cornet, "Literature review of snomed ct use.," *Journal of the American Medical Informatics Association : JAMIA*, vol. 21 e1, pp. e11–9, 2014.
[15] S. J. Nelson, K. Zeng, J. Kilbourne, T. Powell, and R. Moore, "Normalized names for clinical drugs: Rxnorm at 6 years," *Journal of the American Medical Informatics Association : JAMIA*, vol. 18 4, pp. 441–8, 2011.
[16] H. Quan, V. Sundararajan, P. Halfon, A. fong, B. Burnand, J. Luthi, L. D. Saunders, C. Beck, T. Feasby, and W. Ghali, "Coding algorithms for defining comorbidities in icd-9-cm and icd-10 administrative data," *Medical Care*, vol. 43, pp. 1130–1139, 2005.
[17] R. H. Dolin, L. Alschuler, C. Beebe, P. V. Biron, S. L. Boyer, D. Essin, E. Kimber, T. Lincoln, and J. E. Mattison, "The hl7 clinical document architecture," *Journal of the American Medical Informatics Association*, vol. 8, no. 6, pp. 552–569, 2001.
[18] Y. Chen, T. Lasko, Q. Mei, J. Denny, and H. Xu, "A study of active learning methods for named entity recognition in clinical text," *Journal of biomedical informatics*, vol. 58, pp. 11–18, 2015.
[19] Özlem Uzuner, I. Solti, and E. Cadag, "Extracting medication information from clinical text," *Journal of the American Medical Informatics Association : JAMIA*, vol. 17 5, pp. 514–8, 2010.
[20] K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, L. Hou, K. Clark, S. Pfohl, H. Cole-Lewis, D. Neal, M. Schaekermann, A. Wang, M. Amin, S. Lachgar, P. A. Mansfield, S. Prakash, B. Green, E. Dominowska, B. A. Y. Arcas, N. Tomašev, Y. Liu, R. C. Wong, C. Semturs, S. S. Mahdavi, J. Barral, D. Webster, G. Corrado, Y. Matias, S. Azizi, A. Karthikesalingam, and V. Natarajan, "Towards expert-level medical question answering with large language models," *ArXiv*, vol. abs/2305.09617, 2023.
[21] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M.-W. Chang, "Realm: Retrieval-augmented language model pre-training," *ArXiv*, vol. abs/2002.08909, 2020.
[22] Z. Li, R. Guo, and S. Kumar, "Decoupled context processing for context augmented language modeling," *ArXiv*, vol. abs/2210.05758, 2022.
[23] L. Pan, D. Fergusson, I. Schweitzer, and P. Hebert, "Ensuring high accuracy of data abstracted from patient charts: the use of a standardized medical record as a training tool.," *Journal of clinical epidemiology*, vol. 58 9, pp. 918–23, 2005.
[24] S. Zhang, L. Dong, X. Li, S. Zhang, X. Sun, S. Wang, J. Li, R. Hu, T. Zhang, F. Wu, and G. Wang, "Instruction tuning for large language models: A survey," *ArXiv*, vol. abs/2308.10792, 2023.
[25] C.-B. Zhang, P.-T. Jiang, Q. Hou, Y. Wei, Q. Han, Z. Li, and M.-M. Cheng, "Delving deep into label smoothing," *IEEE Transactions on Image Processing*, vol. 30, pp. 5984–5996, 2020.
[26] A. Pal, L. K. Umapathi, and M. Sankarasubbu, "Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering," in *Proceedings of the Conference on Health, Inference, and Learning* (G. Flores, G. H. Chen, T. Pollard, J. C. Ho, and T. Naumann, eds.), vol. 174 of *Proceedings of Machine Learning Research*, pp. 248–260, PMLR, 07–08 Apr 2022.
[27] Q. Jin, B. Dhingra, Z. Liu, W. W. Cohen, and X. Lu, "Pubmedqa: A dataset for biomedical research question answering," pp. 2567–2577, 2019.
[28] T. Panch, T. Pollard, H. Mattie, E. Lindemer, P. Keane, and L. Celi, ""yes, but will it work for my patients?" driving clinically relevant research with benchmark datasets," *NPJ Digital Medicine*, vol. 3, 2020.
[29] A. G. M. A. P. G. Ankit Pal, Pasquale Minervini and B. Alex, "Open life sciences llm leaderboard." https://huggingface.co/spaces//open_medical_llm_leaderboard, 2024.
[30] "New state-of-the-art accuracy for the 3 primary uses of healthcare language models." https://www.johnsnowlabs.com/watch-webinar-new-state-of-the-art-accuracy-for-the-3-primary-uses-of-healthcare-language-models/. Accessed: 2024-06-16.