npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01681-4

Synthetic data distillation enables the extraction of clinical information at scale

Elizabeth Geena Woo[1,2,3,5], Michael C. Burkhart[1,3,5], Emily Alsentzer[4] & Brett K. Beaulieu-Jones[1,3] ✉

Large-language models (LLMs) show promise for clinical note information extraction, but deployment challenges include high computational costs and privacy concerns. We used synthetic data distillation to fine-tune smaller, open-source LLMs to achieve performance comparable to larger models while enabling local hardware deployment or reduced cloud costs. Using Llama-3.1-70B-Instruct, we generated synthetic question-answer training pairs to fine-tune smaller Llama models. We evaluated performance across three tasks: synthetic clinical trial criteria, the i2b2 2018 Clinical Trial Eligibility Challenge, and apixaban trial criteria questions. The 8B-parameter model achieved high accuracy across all tasks and sometimes outperformed the 70B-Instruct teacher model. Fine-tuning with only the most challenging questions still improved performance, demonstrating the value of targeted training. Results from 3B- and 1B-parameter models showed a clear size-performance tradeoff. This work demonstrates synthetic data distillation's potential for enabling scalable clinical information extraction.

Research with real-world data typically relies on human-labeled data for training and validation. Though effective, human annotation can be costly, time-consuming, and prone to errors. Recent research suggests that the few-shot capabilities of generative large language models (LLMs) can be used to annotate text data with reduced time and cost burden[1–4]. These capabilities of generative LLMs can be applied to information extraction from patient clinical notes. Traditional methods for information extraction include rule-based approaches, which can be limited by low recall due to user-defined rules and variability of medical texts, and supervised machine learning models, which can be limited by a lack of labeled training data[5–7]. The zero- and few-shot capabilities of LLMs can enable more flexible and scalable information extraction from clinical notes without the need for extensive manual annotation.

While promising, state-of-the-art LLMs (such as GPT-4[8]) are challenging to deploy in a scalable way in healthcare systems. Many of these models (including those from OpenAI, Anthropic, and Google) are proprietary and come with limited license terms. Concerns about patient privacy and lack of transparency in these proprietary models also lead to some hesitancy in their adoption for healthcare institutions[9]. Additionally, these models can be extremely large and require substantial computational resources (e.g., Llama 405B), limiting their deployment within typical health system IT settings[10]. So far, many of the successful deployments have been through partnerships where industry partners subsidize costs or provide in-kind contributions in terms of computing and engineering. This limits the number and type of institutions that are able to participate and the use cases to which generative AI can be applied. Additionally, setting up these partnerships can require additional administrative lift (e.g., legal negotiation and information security evaluation) compared to performing analyses in existing environments, whether institution-hosted or existing private cloud deployments[11]. Even where solutions have been widely available, such as partnerships for draft inbox responses[12], the ability to achieve similar performance with smaller models will make customizing models to a specific institution, as well as serving inference requests at scale, substantially cheaper and less cumbersome.

Challenges in generative AI around scalability necessitate cost-effective and privacy-conscious solutions, which could be addressed through the development of open-source LLMs that can be integrated into existing healthcare system infrastructure. Open-source LLMs historically did not perform as well as their proprietary counterparts[13], but recent progress has led to very competitive models across most evaluation metrics[14]. Recent efforts have been made to evaluate the capacity of locally deployable LLMs to extract clinical information with low hardware requirements[15].

Synthetic data generation, distillation, and instruction tuning offer an opportunity to close the gap between open-source and proprietary models. Larger models can generate synthetic data that can be used to fine-tune a smaller model for a given task, with the idea that the smaller model could

[1]Department of Medicine, Biological Sciences Division, University of Chicago, Chicago, IL, USA. [2]Committee on Genetics, Genomics and Systems Biology, University of Chicago, Chicago, IL, USA. [3]Center for Computational Medicine and Clinical AI, University of Chicago, Chicago, IL, USA. [4]Department of Biomedical Data Science, Stanford University, Palo Alto, CA, USA. [5]These authors contributed equally: Elizabeth Geena Woo, Michael C. Burkhart.
✉e-mail: beaulieujones@uchicago.edu

mirror the performance of the larger model for that task. This process, called *distillation*, has been shown to improve the performance of these models[16,17]. It allows researchers to develop models with the potential for wider adoption by reducing computational cost without sacrificing performance. Knowledge and data distillation approaches have been used for medical applications, including health event prediction[18], medical dataset sharing[19], and medical image analysis[20,21]. It can be challenging to access annotated data for medical applications due to privacy concerns, regulatory constraints, and the time- and resource-intensive process of manual annotation. Synthetic data generated by LLMs can serve as a potential alternative and have successfully been used as training data for knowledge distillation approaches[22–25]. Data augmentation approaches with synthetic data can be used to enhance model performance by producing high-quality, diverse training examples from which the LLM can learn[26]. For example, fine-tuning on synthetic data generated by GPT-4 improved zero-shot performance of Llama models[27].

While synthetic data distillation approaches are actively evolving, there have been few approaches specifically focused on clinical information extraction from unstructured clinical notes. The ability to extract clinical information at scale from unstructured clinical notes could enhance patient phenotyping, which is important for research and clinical applications. Current phenotyping approaches often rely on structured data such as ICD codes, which are used for billing purposes and may not reflect the nuances of the patient's condition. This can limit analytical precision and potentially introduce biases when studying research outcomes. Unstructured clinical notes, which contain information including medical, social, and family history that may not be captured by structured data, could offer more granular and reliable insight into patient history, particularly in heterogeneous populations where there can be large differences in disease manifestation and progression[28,29]. LLMs can perform zero-shot information extraction from notes to improve phenotyping accuracy over the use of ICD codes, without the need for extensive manual annotation[30].

Another application for these methods is in clinical trial recruitment, which requires a comprehensive evaluation of both clinical trial eligibility criteria and patient medical histories in order to appropriately match patients who meet trial requirements[31–34]. Synthetic data distillation is particularly useful in this case where less labeled data (i.e., paired patient-criterion matching annotations) is available. A recent study developed an LLM framework that used GPT-4 to predict patient eligibility on a criterion-level basis with explanations and achieved near expert-level performance[35]. Recent work comparing proprietary and open-source models suggested that distillation, along with fine-tuning, can improve the performance of open-source LLMs for patient trial matching, approaching that of GPT-4[36]. As opposed to Nievas et al.[36], we used an open-source model to generate synthetic data, generated our data with MIMIC-III notes, and fine-tuned with QLoRA[37]. The fine-tuned models were evaluated against both the data used to create the synthetic question-answer pairs (MIMIC-III) as well as external data. Additionally, it is critical to use open-source models, even as a teacher. Deploying a model fine-tuned on GPT-4 outputs is likely against OpenAI's terms of service[38] as this would be deemed competing with OpenAI. As a whole, these developments show promise for the capacity of LLMs to aid in clinical information extraction for patient-trial matching.

Related work has also explored context distillation[39] and the inclusion of intermediate reasoning steps and rationales[40,41]. For example, Huang et al.[40] conducted ablation studies to show the effect of fine-tuning on reasoning for self-improvement. Hsieh et al.[41] extracted chain-of-thought rationales and labels, which they used to fine-tune smaller T5 models. It can be informative to consider the impact of including model-generated rationales as well as other subsets of synthetic data. Examining model performance in answering single-order questions (e.g., what was the patient's highest creatinine value) compared to questions requiring multiple steps (e.g., does this patient fit this trial's eligibility criteria?) could provide additional insights.

In this work, we demonstrate the ability to perform synthetic data distillation for scalable clinical note annotation, using a large open-source model to generate realistic questions based on patient clinical records, which can be used to train a smaller model that can perform inference. Additionally, we perform an ablation study to understand which types of synthetic data yield optimal performance and the tradeoff between model size and performance. We conduct comprehensive evaluations against multiple datasets. This is critical because we observe it is substantially easier to achieve strong performance against synthetic data with manual review as opposed to fully human-generated evaluations. Alongside the work, we release source code which provides a framework for meaningful, clinical information extraction, synthetic data generation (https://github.com/bbj-lab/clinical-synthetic-data-distil), and an annotation tool built around making the annotation process faster, particularly when LLM predicted annotations are already available (https://github.com/bbj-lab/annotation-ui). We are also releasing two newly manually annotated datasets to PhysioNet, which will be available via the same data use agreement as MIMIC-III/IV : (1) **Annotated Synthetic Trial Criteria Questions**: 1000 questions generated by the large 70B model as Synthetic Data, which have been human-reviewed, and (2) **Apixaban Trial Criteria Questions:** 2300 questions based on trial criteria from the ARISTOTLE apixaban clinical trial[42,43].

Results

The process of knowledge distillation by generating synthetic question and answer pairs using a large model (Llama 3.1 70B-Instruct) to teach a smaller model (e.g., Llama 3.1 8B-Instruct, Llama 3.2 3B-Instruct, Llama 3.2 1B-Instruct) is described in Fig. 1. We then evaluated the distillation process on three distinct tasks, **1.)** a set of synthetic trial criteria questions (1000) which we manually reviewed (Table 2), 2.) real world data from the i2b2 n2c2 clinical trial cohort challenge[44] (Fig. 2, Table 3), and 3.) a set of 2,300 questions derived from the MIMIC real world dataset to emulate the eligibility criteria of the ARISTOTLE apixaban clinical trial[42,43] (Table 4).

The knowledge distillation process worked by passing in a discharge summary to Llama 3.1 70B-Instruct along with prompt instructions (Supplementary Table 1) to create questions meeting specific criteria (e.g., yes/no, numeric, or questions that can not be answered based on the content of the note). In addition to questions, the model was tasked with providing the section of the discharge summary an answer could be found (e.g., Pertinent Results), the source or exact text that allowed the model to answer the question, and an explanation of why the answer was correct based on the source and rest of the note. The model was also tasked with estimating the difficulty of the question it created (Supplementary Table 2).

Next, these questions were filtered depending on which model was being fine-tuned (Table 1). For example, 8B-All includes all of the generated synthetic question and answer pairs (as do 3B-All and 1B-All), 8B-H-25K includes only the 25,000 questions the 70B-Instruct model ranked hardest within each category, 8B-NB-Only includes the 25,000 hardest numeric and boolean (yes/no) questions, and 8B-No-S includes the 25,000 hardest questions of each type but does not finetune on any of the supporting information (namely, the explanation, the section the model believed the answer was in when generating the question, or the source, which is the exact text which allowed for the model to answer the question). Next, QLoRA fine-tuning (detailed in Methods) was performed for each of the question categories to result in six fine-tuned models (8B-All, 8B-H-25k, 8B-NB-Only, 8-No-S, 3B-All, and 1B-All) in addition to the four instruct models open-sourced by Meta (70B-Instruct, 8B-Instruct, 3B-Instruct, and 1B-Instruct) (Table 1).

Each model was evaluated on three tasks: (i) annotated synthetic trial criteria questions, (ii) i2b2 Clinical Trial Eligibility Criteria Cohort Selection shared task from the 2018 National NLP Clinical Challenges, and (iii) apixaban trial criteria. We report performance metrics including Balanced Accuracy, which measures the average between sensitivity and specificity and can be used on imbalanced datasets, and Micro-F1 score. Micro-F1 was the primary metric used to judge the i2b2 challenge, which permits direct comparison between our results and challenge entries (for the test set).

Synthetic Data Evaluation

We evaluated model performance on a manually annotated subset of 1000 generated examples from the hold-out test set described in the methods

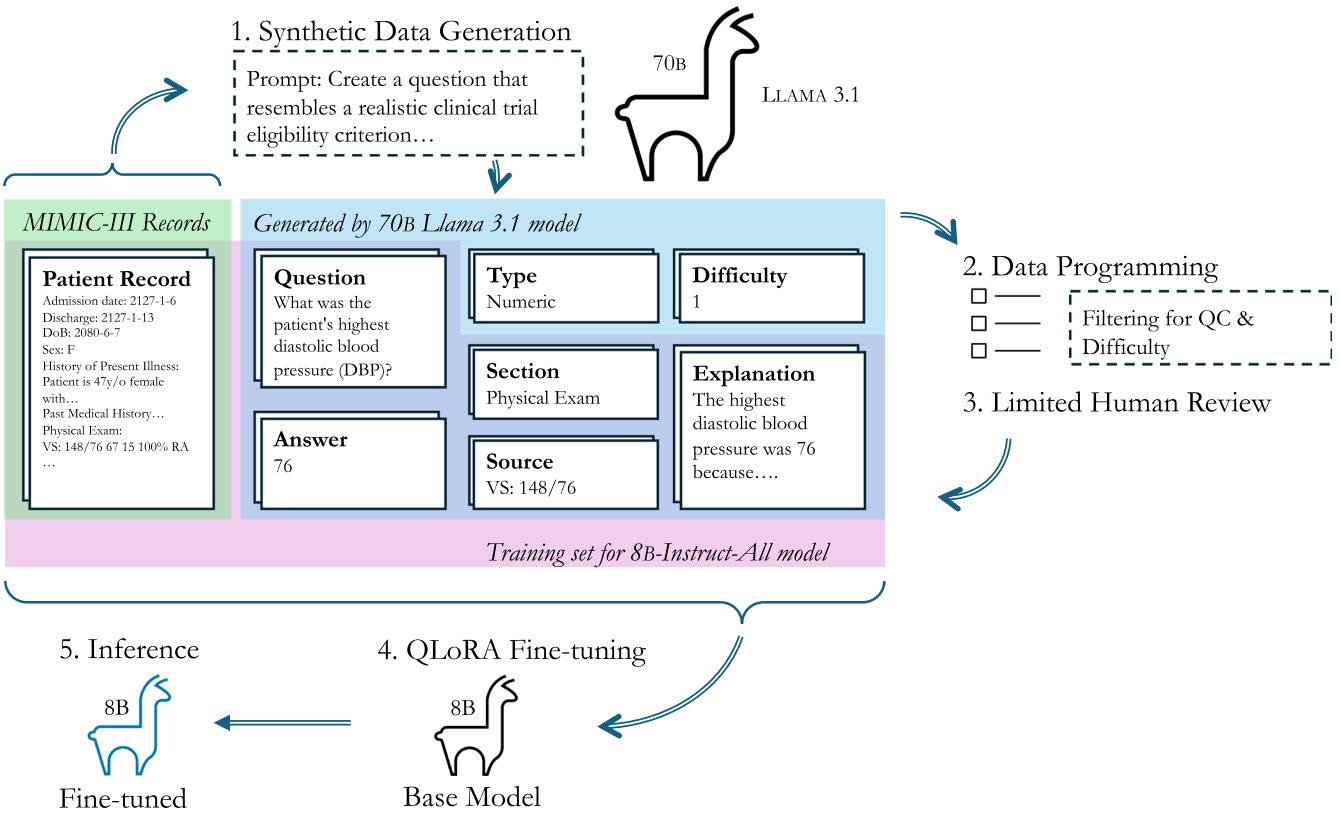

Fig. 1 | Synthetic distillation training workflow. MIMIC-III records, outlined in green, are provided to the 70B-parameter Llama-3.1 model, which in turn generates the elements outlined in blue. After post-processing, the elements outlined in purple are provided to the 8B-parameter Llama-3.1 model (or 3B- or 1B-parameter models) for fine-tuning the "All" version of the model.

datasets subsection (Table 2). The 8B-All model achieves the best overall accuracy (89.30%), outperforming even the 70B-Instruct model used for creating the synthetic data (76.20%). This was especially visible in the "NA" categories, where there appears to be a strong impact of training models explicitly on questions that cannot be answered based on the context (note) provided. Within each category, 8B-All and 8B-H-25k improved over 8B-Instruct, reflecting the impact of fine-tuning. 8B-H-25k also outperformed 70B-Instruct overall, suggesting that while the model benefits from further fine-tuning, a relatively small dataset of 25k examples can still provide an appreciable benefit. Unsurprisingly, the 8B-NB-Only model which was not fine-tuned on any "NA" data struggles in both of the NA columns, but it does perform very well on questions of numeric and boolean type and is actually the top performer for numeric questions. When comparing between the 8B-All, 3B-All, and 1B-All models, we find a general tradeoff between model size and performance, with a notable exception of the ability of the 3B-All and 1B-All models to identify questions that it could not answer (the NA-type questions).

i2b2 clinical trial eligibility challenge evaluation

We next evaluated the performance of all base and fine-tuned models on the i2b2 2018 Clinical Trial Eligibility Challenge (Fig. 2). Because we did not train on or otherwise use these data in our fine-tuning process we were able to assess the performance of models across both the train and test sets for the original i2b2 challenge.

We evaluated two different values of two parameters, temperature and top_p (see Methods). We had a hypothesis that sampling strategies (i.e., higher temperature) might work well to force the model to provide an answer that aligned well with the explanation. However, we observed that the temperature did not have a large impact, and a temperature of 0 seems to slightly outperform higher temperatures (Supplementary Table 3). The 70B-Instruct model performed the best on both train and test data. The two fine-tuned models, which included all types and supporting information (8B-All and 8B-H-25K) outperformed the base 8B-Instruct model. The fine-tuned models that either did not include all types (8B-NB-Only) or did not include supporting information (8B-No-S) had worse performance than the base 8B-Instruct model. When comparing the 8B-All, 3B-All, and 1B-All models, we find that performance decreases as model size decreases. This held on both the training and test folds, for both balanced accuracy and micro-F1 score.

An interesting trend we observed throughout this work was the need to isolate criteria and thus the prompts provided to the models into questions that required only single order answers. This was illustrated when comparing the performance of both the base models and fine-tuned models for their ability to either a.) directly answer a prompt question for a given criterion (i.e. direct boolean "yes" or "no") vs. b.) extracting the numeric value relevant to the criterion and then performing post-processing to arrive at a boolean "yes" or "no" answer (Table 3). Within the i2b2 n2c2 challenge, two questions asked whether labs were abnormal (serum creatinine and hemoglobin levels). Across all models, numeric extraction followed by post-processing achieved higher performance compared to asking the model to directly answer the question.

Trial Criteria Evaluation

As the third evaluation task, we compared the performance of the base and fine-tuned models using manual annotations based on 23 questions resembling eligibility criteria from the apixaban clinical trial for a random sample of 100 patient notes from MIMIC-IV (Table 4). The fine-tuned 8B-All model achieved high performance, exceeding Balanced Accuracy and Micro-F1 of 0.8 across all criteria assessed, with an overall average Balanced Accuracy of 0.93 and Micro-F1 of 0.94. This fine-tuned model outperformed the 8B-Instruct (Balanced Accuracy = 0.84, Micro-F1 = 0.86) and even the 70B-Instruct model (Balanced Accuracy = 0.89, Micro-F1 = 0.92). The model fine-tuned on the most difficult 25,000 questions, 8B-Instruct-H-25K, achieved a similarly high performance across criteria

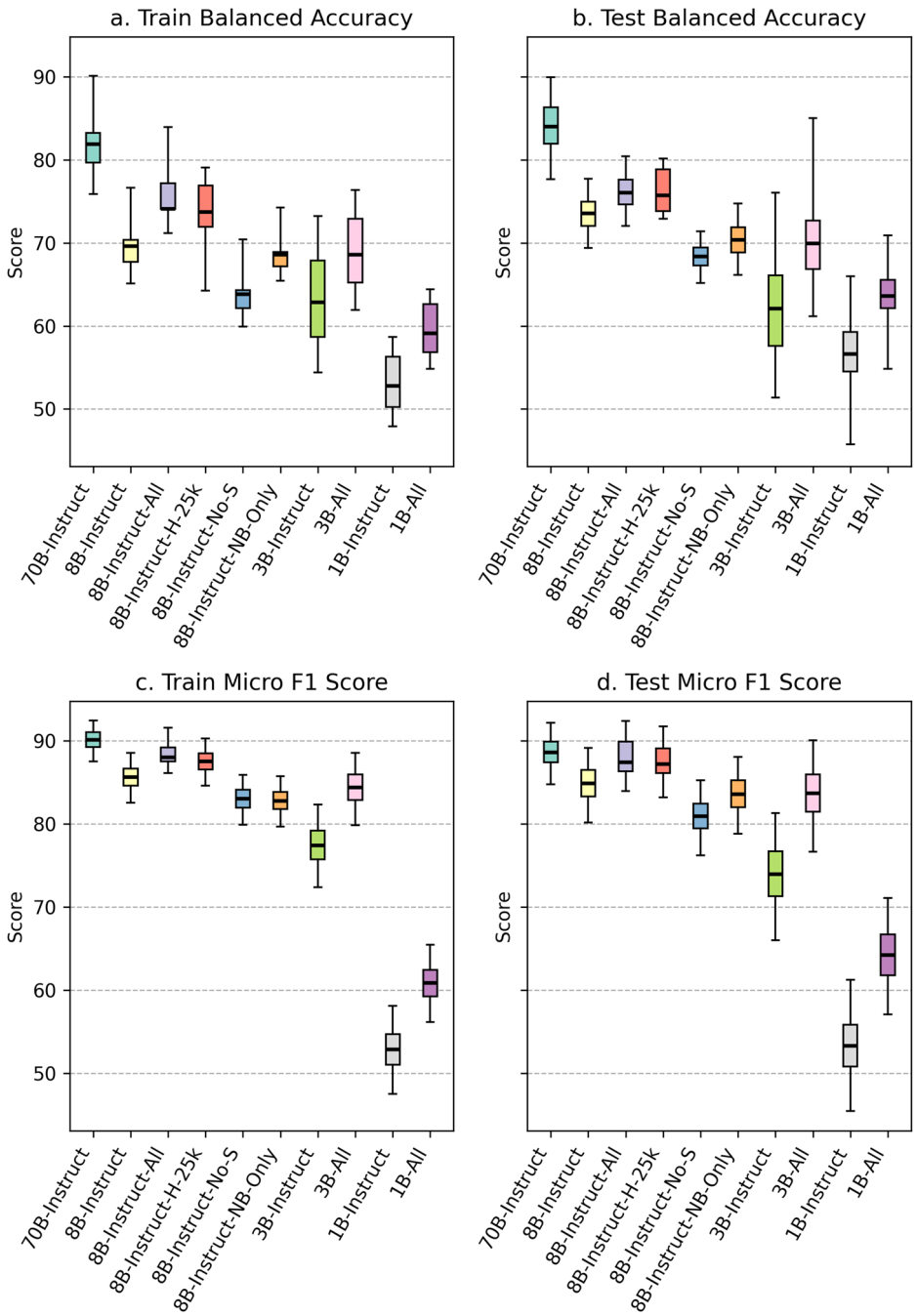

Fig. 2 | Comparison of model performance for the i2b2 (n2c2) Clinical Trial Eligibility Challenge. Evaluation includes the Training Set (**a**, **c**) because these data were not included during any of the pre-processing, hyperparameter selection or fine-tuning process of the models. All evaluations are zero-shot, but performance on Training (**a**, **c**) is separated from Test set (**b**, **d**) for clarity. (70B and 8B are Llama-3.1, 3B and 1B are Llama-3.2).

(average Balanced Accuracy = 0.95, Micro-F1 = 0.94), suggesting that either fewer total questions may be needed for fine-tuning, or that more difficult questions offer greater value in fine-tuning. Average performance (both balanced accuracy and balanced micro-F1) decreased monotonically with model size over the 8B-All, 3B-All, and 1B-All models. For each model size, there was a sizeable performance improvement due to finetuning (comparing the Instruct and All versions for each model size).

There were some criteria where base 8B-Instruct model had relatively lower performance, including extraction of aspartate aminotransferase (AST) (Balanced Accuracy = 0.54, Micro-F1 = 0.54), blood glucose (Balanced Accuracy = 0.25, Micro-F1 = 0.25), and left ventricular ejection fraction (Balanced Accuracy = 0.72, Micro-F1 = 0.72). The use of the larger 70B-Instruct model dramatically improved performance for these criteria, exceeding Balanced Accuracy and Micro-F1 of 0.84. The fine-tuned models 8B-All and 8B-H-25k performed comparably to the 70B model, and in some cases outperformed it. All three models for the AST criteria led to Balanced Accuracy and Micro-F1 scores of 0.94 and above. For blood glucose, the fine-tuned models 8B-All (Balanced Accuracy = 0.98, Micro-F1 = 0.98) and 8B-H-25k (Balanced Accuracy = 0.94, Micro-F1 = 0.94) achieved higher performance than the 70B-Instruct model (Balanced Accuracy = 0.84, Micro-F1 = 0.84). For identification of hemorrhagic tendencies, the model fine-tuned on the 25k most difficult questions led to the biggest performance improvement (Balanced Accuracy = 0.96, Micro-F1 = 0.92) compared to both the 8B-All (Balanced Accuracy = 0.96, Micro-F1 = 0.92) and 70B-Instruct models (Balanced Accuracy = 0.96, Micro-F1 = 0.92).

For some criteria, the 70B-Instruct model did not perform as well as any of the 8B-Instruct models, including the base model. This was the case when detecting the presence of atrial fibrillation (**8B-Instruct:** Balanced Accuracy = 0.98, Micro-F1 = 0.97; **70B-Instruct:** Balanced Accuracy = 0.65, Micro-F1 = 0.84) and whether there was planned/past ablation for atrial

Table 1 | Comparison of the different models that were compared throughout the clinical information extraction tasks

Model Name	Base Model	Fine-Tuned	Question difficulty	Question Type				Supporting information (Section, Source, Explanation)
				Boolean	Numeric	Boolean-NA	Numeric-NA	
70B-Instruct	Llama-3.1 70B-Instruct (Meta)		-	✓	✓	✓	✓	✓
8B-Instruct	Llama-3.1 8B-Instruct (Meta)		-	✓	✓	✓	✓	✓
8B-All	Llama-3.1 8B-Instruct (Meta)	✓	All	✓ *N* = 212,132	✓ *N* = 209,637	✓ *N* = 106,288	✓ *N* = 106,245	✓
8B-H-25K	Llama-3.1 8B-Instruct (Meta)	✓	25 K highest difficulty	✓ *N* = 25,000	✓ *N* = 25,000	✓ *N* = 25,000	✓ *N* = 25,000	✓
8B-No-S	Llama-3.1 8B-Instruct (Meta)	✓	25 K highest difficulty	✓ *N* = 25,000	✓ *N* = 25,000	✓ *N* = 25,000	✓ *N* = 25,000	
8B-NB-Only	Llama-3.1 8B-Instruct (Meta)	✓	25 K highest difficulty	✓ *N* = 25,000	✓ *N* = 25,000			✓
3B-Instruct	Llama-3.2 3B-Instruct (Meta)			✓	✓	✓	✓	✓
3B-All	Llama-3.2 3B-Instruct (Meta)	✓	All	✓ *N* = 212,132	✓ *N* = 209,637	✓ *N* = 106,288	✓ *N* = 106,245	✓
1B-Instruct	Llama-3.2 1B-Instruct (Meta)			✓	✓	✓	✓	✓
1B-All	Llama-3.2 1B-Instruct (Meta)	✓	All	✓ *N* = 212,132	✓ *N* = 209,637	✓ *N* = 106,288	✓ *N* = 106,245	✓

fibrillation (**8B-Instruct:** Balanced Accuracy = 0.89, Micro-F1 = 0.98; **70B-Instruct:** Balanced Accuracy = 0.65, Micro-F1 = 0.94). There were also some criteria, including creatinine and platelets, where the models did not perform as well as other criteria as no model exceeded 0.85 for either balanced accuracy or micro-F1. Of the manually annotated notes, 60% did not have a numeric value for platelet count available in the note, while only 3% did not have a serum creatinine value available (Supplementary Table 4). This rate may be at least in part due to the fact that the de-identification process for MIMIC-III seemed to accidentally redact some platelet values. During the manual annotation process we did not observe this occurring with other laboratory values.

Resource requirements

Data distillation allowed the models to be run with vastly reduced resource requirements compared to the 70B-Instruct model. All model evaluation was done on the Center for Research Informatics' "Randi" cluster at the University of Chicago. The cluster's GPU nodes each contain 8 Nvidia A100 GPU's with two 16-core 3.0-GHz AMD Milan processors. We monitored seconds/example, tokens in/second, and tokens out/second for both the 8B-parameter and 70-B parameter architectures and reported these in Fig. 3. These differences could translate into meaningful cost savings. For example, performing a study of the Apixaban criteria (23 questions) for 10,000 patients to identify a cohort on the least expensive cloud provider would be $3132 less expensive for the 8B vs. 70B parameter models (see Supplementary Table 5 for a comparison of current rates among the main providers). In this example, running the 8B-parameter model would cost less than $1000 (0.535 sec./ex. * 230k ex. * 1/3600 hr./sec. * $27.2/hr. = $929), while the 70B-parameter model would cost over $4000 (2.34 sec./ex. * 230k ex. * 1/3600 hr./sec. * $27.2/hr. = $4066).

Discussion

In this study, we present an approach to improve the scalability of open-source LLMs for clinical information extraction using synthetic data distillation. We used the larger Llama-3.1-70B-Instruct to generate synthetic data, consisting of question-answer pairs with supporting information and difficulty scores. These were used to fine-tune smaller models: Llama-3.1-8B-Instruct, Llama-3.2-3B-Instruct, and Llama-3.2-1B-Instruct. We found a general tradeoff between the size and performance of the finetuned models. We also explored the impact of fine-tuning on different amounts and subsets of synthetic data (including one fine-tuned with all data, one fine-tuned with only the hardest 25 K questions, one fine-tuned without questions where the note does not contain the answer - NA, and one fine-tuned without any supporting information). We observe that the inclusion of NA and supporting information was critical to the high performance of fine-tuned models, especially when applied to fully human-generated evaluations as opposed to synthetic data with human review. When evaluating the accuracy of these models based on manually annotated synthetic data, we found that the model fine-tuned on all synthetic data (8B-All) achieved a high overall accuracy that exceeded that of a larger base model (70B-Instruct). We found that these fine-tuned models also performed well across different clinical tasks, including the i2b2 Clinical Trial Eligibility Challenge and a dataset designed to resemble real eligibility criteria from the apixaban clinical trial. The fine-tuned models can achieve performance comparable to, and in some cases exceeding, that of even the larger model that served as the teacher. Even when fine-tuning is performed using only a subset of the hardest questions in the synthetic dataset, the performance still improves over base models, suggesting that targeted fine-tuning with less data can still be beneficial. Finally, we release several artifacts we believe will be beneficial to researchers further developing approaches for clinical information extraction: (a) source code - both the framework for synthetic data

Table 2 | Model Accuracy on a subset of manually annotated Synthetic Labels (70B)

	Accuracy Reported by Question Type				
	NA – Boolean (*N* = 241)	NA – Numeric (*N* = 232)	Numeric (*N* = 236)	Boolean (*N* = 291)	All Questions (*N* = 1000)
70B-Instruct	69.5% (63.5%, 75.1%)	81.8% (76.7%, 86.6%)	61.7% (55.5%, 67.8%)	88.6% (84.9%, 92.1%)	76.1% (65.6%, 85.6%)
8B-Instruct	27.7% (21.6%, 33.6%)	78.4% (73.3%, 83.2%)	79.2% (74.2%, 84.3%)	87.3% (83.5%, 91.1%)	68.4% (42.9%, 85.0%)
8B-All	88.0% (83.8%, 91.7%)	98.3% (96.6%, 99.6%)	83.9% (78.8%, 88.1%)	84.7% (80.8%, 88.7%)	89.1% (84.3%, 95.4%)
8B-H-25k	80.4% (74.9%, 86.1%)	85.5% (80.6%, 90.1%)	84.2% (79.2%, 88.6%)	88.0% (84.2%, 91.1%)	84.60% (79.9%, 90.3%)
8B-No-S	78.9% (73.8%, 83.8%)	89.3% (85.3%, 93.1%)	80.6% (75.4%, 85.2%)	83.5% (79.4%, 88.0%)	83.0% (79.7%, 87.1%)
8B-NB-Only	0.0% (0.0%, 0.0%)	40.0% (33.6%, 46.6%)	84.4% (79.2%, 88.6%)	87.6% (83.5%, 91.1%)	54.0% (22.2%, 87.0%)
3B-Instruct	87.7% (83.8%, 91.7%)	74.1% (68.5%, 79.7%)	66.0% (60.2%, 72.0%)	57.0% (50.9%, 62.5%)	71.1% (61.2%, 80.9%)
3B-All	85.1% (80.5%, 89.2%)	99.2% (97.8%,100.0%)	77.9% (72.5%, 83.1%)	85.2% (81.4%, 89.3%)	86.7% (79.7%, 95.6%)
1B-Instruct	21.0% (16.2%, 26.6%)	29.1% (23.3%, 35.3%)	17.1% (12.7%, 22.0%)	55.7% (49.8%, 60.8%)	31.0% (19.1%, 47.0%)
1B-All	93.0% (89.6%, 95.9%)	99.2% (97.8%, 100.0%)	40.2% (34.3%, 46.6%)	53.6% (47.8%, 59.5%)	71.2% (46.9%, 96.1%)

Reported values include the mean accuracy and 95% CI.

Table 3 | Comparison between directly answering clinical trial criteria about laboratory value ranges vs. extracting a number and applying rules-based post processing to determine whether to answer "yes" or "no" (i.e., ask the model to return a number, if that number is above a range answer yes, otherwise answer no)

Criterion Title	PromptType	Prompt Question	Extracted Value Processing	Performance					
				Balanced Accuracy			Micro-F1		
				70B	8B	8B-All	70B	8B	8B-All
Creatinine	Numeric	What was the patient's highest recorded creatinine level? Answer NA if there are no values.	<= 1.3: No > 1.3: Yes (Does not account for Sex)	0.893	0.870	0.894	0.878	0.844	0.899
	Boolean	Has the patient ever had a serum creatinine level above the upper normal limit? (Typically > 1.3 mg/dL for men and 1.1 mg/dL for women).	None	0.825	0.763	0.819	0.788	0.715	0.791
HbA1c	Numeric	What was the patient's highest recorded hemoglobin A1c (HbA1c) value? Answer NA if there are no values.	>= 6.5: Yes Else: No	0.949	0.783	0.896	0.937	0.729	0.875
	Boolean	Has the patient ever had a hemoglobin A1c (HbA1c) level between 6.5 and 9.5 inclusive?	None	0.774	0.583	0.743	0.774	0.462	0.760

generation for clinical information extraction model fine-tuning as well as the annotation tool which allowed for faster manual review of LLM pre-annotated notes, and (b) datasets - two manually annotated datasets (Annotated Synthetic Trial Criteria Questions and Apixaban Trial Criteria Questions) which will allow for researchers to evaluate future methods for clinical information extraction.

The use of LLMs to extract information from clinical notes has already demonstrated the potential to improve upon traditional methods relying on rule-based methods or extensive manual annotation. While proprietary models, such as GPT-3 and GPT-4, have shown strong performance for this purpose, their deployment in healthcare settings can be limited by computational costs and licensing barriers[35]. Our findings align with recent research suggesting that fine-tuning open-source models with synthetic data can improve their performance across clinical information extraction tasks, bringing it closer to that of proprietary models[36]. By generating synthetic data, this approach also reduces reliance on manually labeled data. Our findings also align with those of concurrent work exploring the utility of synthetic data distillation using Llama-3.1-405B-Instruct as a teacher model[24]. The use of larger teacher models such as the 405B-Instruct poses a challenge in terms of necessary computational requirements, particularly in academic or resource-constrained settings. We find that the smaller 70B-Instruct can successfully be used as a teacher model for distillation. The use of smaller, open-source models that can perform well opens the door for broader adoption of LLMs in healthcare settings in a cost-effective and privacy-compliant manner. The reduced computational requirements of these smaller models can make them more accessible to hospitals with limited healthcare IT infrastructure. Another advantage of this approach is the adaptability, as smaller models can be better tailored to the specific needs of individual hospitals. As resource-efficient LLMs continue to evolve, this approach enables rapid (e.g., <12 hours on 8 gpus) finetuning[45].

Our work focuses specifically on clinical trial eligibility criteria as an evaluation task because these criteria are well-defined, but our proposed framework is more broadly adaptable and could be applied to other areas including cohort identification, patient phenotyping, or feature extraction. Many observational and retrospective studies have specific inclusion criteria that currently require some form of manual chart review. A solution that works for clinical trials would also apply in these settings. The goal of this work is to bring us closer to minimizing the need for manual chart review and annotation. The finetuned models we developed in this study would not be able to finalize candidate selection on their own but could be used to screen a large number of candidates, narrowing an initial set down to a smaller pool of patients much more likely to qualify. Manual review would only need to be performed on the smaller pool, allowing medical professionals to avoid having to look at a majority of the records. Recent work has explored information extraction tasks with minimal human oversight or annotations[46], but a significant amount of work remains to be done before models could be accurate enough to perform the entire screening process without human oversight.

By enabling scalable information extraction from unstructured notes, this approach presents a promising opportunity for retrospective research

Table 4 | Performance on clinical trial eligibility criteria for apixaban

	Balanced Accuracy							
Criterion	**70B-Instruct**	**8B-Instruct**	**8B-All**	**8B-H-25k**	**3B-Instruct**	**3B-All**	**1B-Instruct**	**1B-All**
AST	97.0%	54.0%	94.3%	99.4%	48.0%	98.5%	30.0%	91.7%
Bilirubin	99.0%	100.0%	100.0%	99.3%	50.0%	100.0%	32.0%	91.9%
Creatinine	80.0%	85.0%	84.0%	85.0%	52.0%	86.0%	3.0%	83.3%
Hemoglobin	90.0%	96.0%	98.0%	96.0%	49.0%	98.5%	1.0%	91.4%
Platelets	87.0%	78.0%	79.6%	79.1%	75.0%	82.0%	36.4%	37.8%
AFib	81.7%	98.0%	98.0%	98.0%	81.2%	80.5%	29.0%	15.6%
Ablation for Afib	64.6%	89.5%	81.8%	98.0%	73.9%	85.2%	95.0%	100.0%
Arterial Hypertension	95.0%	99.4%	97.7%	97.1%	73.5%	90.9%	63.7%	58.5%
Bipolar Disorder	100.0%	100.0%	100.0%	100.0%	98.0%	99.7%	94.9%	100.0%
Bleeding	92.1%	91.6%	89.4%	80.0%	91.2%	61.8%	80.2%	37.8%
Blood Glucose	84.0%	25.0%	98.5%	94.0%	58.2%	93.0%	18.0%	81.8%
Chads2	94.0%	89.0%	94.9%	97.6%	53.7%	97.7%	80.0%	94.8%
Heart Failure	98.0%	96.5%	98.2%	99.0%	95.0%	98.0%	75.9%	94.0%
Hemorrhagic Tendencies	62.7%	60.3%	85.6%	96.1%	91.4%	82.0%	79.5%	80.7%
Left ventricular ejection fraction	90.0%	72.0%	94.9%	96.0%	67.0%	91.8%	49.5%	57.0%
Depression	95.3%	100.0%	98.7%	97.5%	91.8%	93.1%	77.1%	71.4%
Makes Medical decisions	81.4%	79.5%	95.3%	93.4%	79.4%	94.9%	43.8%	87.8%
Peptic Ulcer Disease	99.5%	75.0%	92.9%	99.5%	81.3%	92.5%	46.9%	98.9%
Prior Stroke	86.2%	81.7%	88.5%	89.2%	97.1%	95.6%	79.6%	88.3%
Recent Stroke	94.2%	75.3%	94.2%	94.2%	91.1%	96.2%	82.8%	78.7%
Schizophrenia	100.0%	100.0%	100.0%	100.0%	99.5%	100.0%	48.9%	100.0%
Valvular Disease requiring Surgery	83.9%	87.2%	86.8%	86.8%	70.4%	88.7%	45.4%	90.7%
Diabetes	82.2%	98.3%	99.1%	98.3%	91.8%	100.0%	77.7%	77.4%
Average	88.7%	84.0%	93.5%	94.2%	73.9%	91.9%	55.0%	78.4%
95% C.I.	(84.3%, 92.6%)	(75.9%, 90.5%)	(90.9%, 96.1%)	(90.9%, 96.9%)	(66.7%, 81.1%)	(87.5%, 95.4%)	(44.1%, 66.1%)	(69.7%, 86.0%)
Micro-F1								
AST	0.97	0.54	0.94	0.99	0.48	0.99	0.30	0.92
Bilirubin	0.99	1.00	1.00	0.99	0.50	1.00	0.32	0.92
Creatinine	0.80	0.85	0.84	0.85	0.52	0.86	0.03	0.83
Hemoglobin	0.90	0.96	0.98	0.96	0.49	0.98	0.01	0.91
Platelets	0.87	0.75	0.80	0.81	0.75	0.82	0.36	0.38
AFib	0.84	0.97	0.97	0.97	0.84	0.80	0.29	0.16
Ablation for Afib	0.94	0.98	0.96	0.96	0.95	0.93	0.95	1.00
Arterial Hypertension	0.98	0.99	0.96	0.95	0.81	0.96	0.82	0.31
Bipolar Disorder	1.00	1.00	1.00	1.00	0.96	0.99	0.95	1.00
Bleeding	0.85	0.91	0.89	0.80	0.83	0.62	0.80	0.38
Blood Glucose	0.84	0.25	0.98	0.94	0.58	0.93	0.18	0.82
Chads2	0.94	0.89	0.95	0.98	0.54	0.98	0.80	0.95
Heart Failure	0.98	0.96	0.98	0.99	0.95	0.98	0.56	0.95
Hemorrhagic Tendencies	0.78	0.52	0.89	0.92	0.83	0.82	0.84	0.81
Left ventricular ejection fraction	0.90	0.72	0.95	0.96	0.67	0.92	0.49	0.57
Depression	0.92	1.00	0.98	0.96	0.85	0.97	0.77	0.88
Makes Medical decisions	0.90	0.91	0.91	0.93	0.91	0.90	0.87	0.88
Peptic Ulcer Disease	0.99	0.94	0.99	0.99	0.95	0.98	0.93	0.98
Prior Stroke	0.92	0.89	0.94	0.94	0.95	0.97	0.80	0.79
Recent Stroke	0.89	0.86	0.89	0.89	0.92	0.93	0.83	0.79
Schizophrenia	1.00	1.00	1.00	1.00	0.99	1.00	0.97	1.00

Table 4 (continued) | Performance on clinical trial eligibility criteria for apixaban

	Balanced Accuracy							
Criterion	**70B-Instruct**	**8B-Instruct**	**8B-All**	**8B-H-25k**	**3B-Instruct**	**3B-All**	**1B-Instruct**	**1B-All**
Valvular Disease requiring Surgery	0.92	0.95	0.93	0.93	0.90	0.93	0.90	0.85
Diabetes	1.00	0.97	0.99	0.98	0.89	1.00	0.56	0.59
Average	0.90	0.86	0.95	0.94	0.76	0.93	0.62	0.78
95% C.I.	(0.873, 0.934)	(0.783, 0.934)	(0.920, 0.966)	(0.912, 0.971)	(0.677, 0.828)	(0.885, 0.963)	(0.484, 0.741)	(0.678, 0.857)

Columns are grouped by model size. Blue color indicates models that received fine-tuning

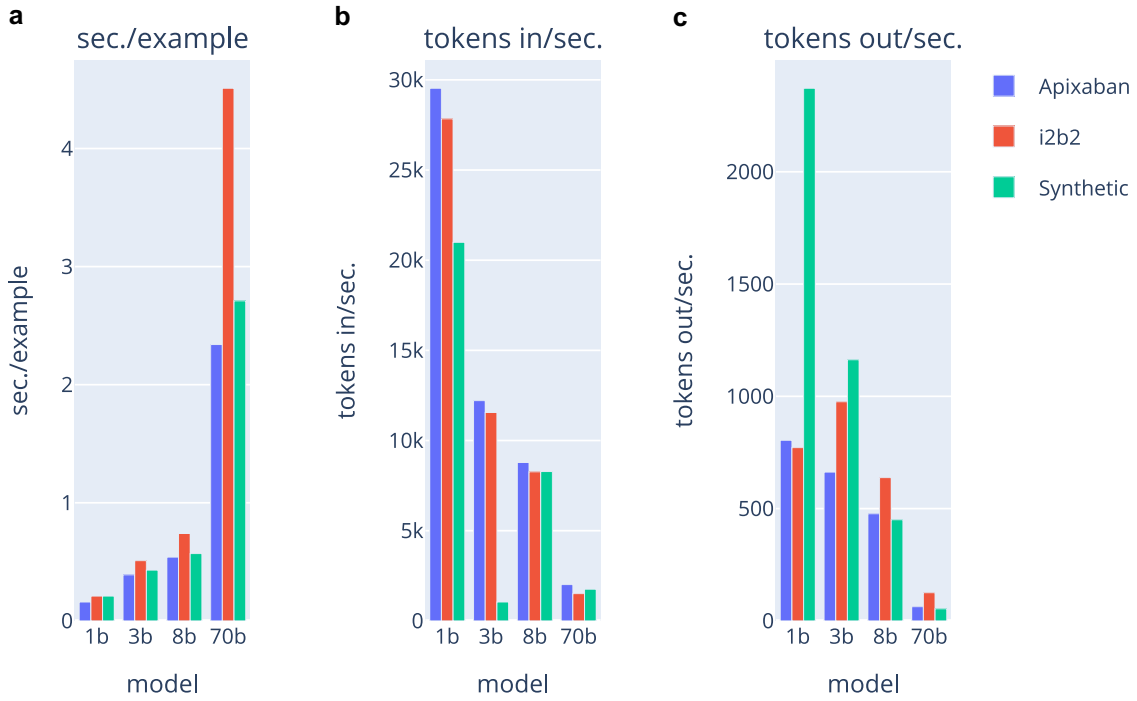

Fig. 3 | Comparison of inference speed across model sizes and evaluation tasks. **a** illustrates the average number of seconds needed to process an example for each dataset and model, **b** shows the average number of tokens read or ingested per second, and (**c**) depicts the average number of tokens generated per second. When comparing the center and right panels, note that token generation tends to be more time-consuming than token ingestion.

through its potential impact on enhancing patient phenotyping. This is particularly important when studying complex and heterogeneous patient populations, where phenotyping approaches relying solely on structured data, such as ICD codes, fall short. Better phenotyping can result in improved quality and relevance of retrospective studies.

While this has exciting potential, we also note some of the limitations and challenges identified through manual review of the synthetic data generated by Llama-3.1-70B-Instruct that may begin to elucidate failure modes for these models. In general, the model struggled with ranges when forming numeric questions. In multiple instances, a range (e.g., 60-70%) would be collapsed to one of its limits (60% or 70%) in a numerical answer. In at least one instance, the model had difficulty comparing a range and a given value outside that range (e.g., concluding that >70% precludes 50%). This contrasts with the model's generally consistent ability to locate the highest or lowest value in a sequence of measurements (e.g., finding the highest blood pressure recorded in a note containing multiple readings). Numeric ranges of values are a known area of difficulty for model reasoning[47]. Hager et al. explicitly provided example lab results along with reference ranges for those labs and asked multiple LLM's to determine if the result fell below, within, or above the range; they concluded that "all LLMs performed very poorly"[48]. To avoid having the model reason about ranges of values, questions could be reworded to ask the model to return the patient's measurement for a given metric, and then evaluate if that measurement falls within a certain reference range as a separate step (as in Table 3). For ranges of values that appear within notes, separate questions could be used to determine the maximum and minimum estimated values.

The model also sometimes struggled with redacted data and contextual understanding. In one instance, the model identified numbers in a redaction tag as the answer to a question. This tag would have contained the correct answer prior to redaction. The tag itself, "[**3-22**]", contained numbers, and this may have contributed to the model's confusion. Non-numerical redaction tokens may help to alleviate this sort of issue. Additionally, token representation should be taken into account when deciding how to include redactions, as the existing format generates several extra tokens per redaction requiring greater context size. In another example, the model successfully identified the inappropriately partially-redacted "[churgg [**Doctor Last Name **] disease" as Churg-Strauss disease. In another case, the model correctly identified a patient's hemoglobin value but then incorrectly concluded that it fell below the normal range. This conclusion would have been correct had the patient been male; however, the patient was female, and the reference range is lower for females. In another case, the model asked if a female over 70 was "a candidate for future pregnancy?" Interestingly, the model

was also able to identify and parse a fishbone diagram within a note, correctly answering questions about lab values contained within the diagram.

The model also sometimes lacked creativity when generating questions with unspecified answers. To generate questions that could not be answered using the contents of a note, the model seemed to commonly inquire about BMI (height and weight measurements are recorded separately from these notes and so are often not contained in the text) and the results from a 6-minute walking test (6MWT). In the full test set containing 42,498 instances, we found 997 questions related to BMI (98.7% of which resolved n/a) and 666 questions related to a 6-minute walk test (all of which resolved n/a). The model would also ask about measurements from a patient prior to them seeking medical attention, which are typically unavailable in these notes. Additionally, the models would sometimes struggle with repetitive generation. In the test set, we found 1,676 (3.94%) questions containing "creatinine". Admittedly, our prompt for numeric type questions included an example "What was the patient's highest creatinine measurement recorded in the note?" However, a majority (1151) of these questions were of na-numeric, boolean, or na-boolean type, and none of those prompts mention creatinine. Future work could potentially introduce a mechanism to deprioritize questions that are highly repetitive or likely to yield non-informative answers. With the development of models that support longer context windows, it is possible to keep track of previously generated question-answer pairs and use this to avoid redundancy during question generation. Retrieval-augmented generation (RAG), which combines LLMs with knowledge from external sources[49,50], could also be incorporated into the question generation process. For example, it could be used to retrieve data (including lab values, medications, and diagnoses) and use this information to generate question-answer pairs more relevant to the recorded information. This could reduce the risk of generating unanswerable questions like those about BMI when height and weight are missing. RAG could also use information from other external sources including notes and question-answer pairs from other similar patients or clinical knowledge bases to help guide toward more diverse and contextually appropriate question generation.

We found that carefully worded prompts could help to avoid some of the incorrect model outputs described in the previous section. By rewording questions, we could deter the model from drawing inferences and obtain less ambiguous question-answer pairs. For questions that asked if a patient had a history of X, where X was not mentioned in the note, the model would sometimes conclude that a patient did not have a history of X because X was not mentioned in the note, and other times conclude that the question could not be definitively answered from the contents of the note. This ambiguity could be resolved by modifying the question to ask if a patient's history of X could be found in the note. This is especially critical because it allows us to use a combination of clinical expertise and post-processing to knowingly make assumptions where appropriate about whether X would have been in the note if they had it, as opposed to the model making this assumption for us without our knowledge. We observed in multiple evaluations that performance is substantially higher when asking the model to answer single-order questions (e.g., what was the patient's highest creatinine value?) as opposed to questions which require multiple steps (e.g., does this patient fit this trial's eligibility criteria?). In future work, we will determine if chain-of-thought prompting[51] or multi-step inference[52] can circumvent the need for manual postprocessing.

Developing resource-efficient LLMs to extract relevant information from clinical notes is a rapidly advancing discipline with many open questions. For example, there may be better ways to make the distillation process more data-efficient. In this work, we showed how fine-tuning on only a fraction of the synthetic dataset (e.g., 8B-H-25k) still appreciably enhances the base 8B-Instruct model. Different criteria for selecting a subset of the fine-tuning data may better maintain performance while decreasing data requirements[53,54]. Ordering the fine-tuning set by increasing difficulty and interleaving question types may also help[55].

Future work could consider whether a metric besides micro-F1 could better characterize good performance. We used micro-F1 in part because it benchmarked the original i2b2 challenge. However, some researchers view patient-clinical trial matching as a ranking problem and consequently report metrics like normalized discounted cumulative gain at k and precision at k[35,36]. We could also consider the optimal way to handle ambiguity in notes. Unlike tabular or structured data that typically complies with a strict format, notes often include estimates and conjectures, especially when discussing medical history. There are often question marks next to past diagnoses and values reported. Another potentially interesting extension of this work could look into how data from multiple notes could be combined. Many people have a medical history spanning decades. For selection criteria involving disease progression or patient history, multiple notes may be required to obtain a complete answer. Combining records in a time-aware manner remains an open problem.

In this study, we use synthetic data for its capacity to enhance datasets for distillation. We also note additional considerations that come with synthetic data use. While synthetic data is often used as a more privacy-protecting alternative to real data, it is important to consider how synthetic datasets are generated and their regulatory compliance[56]. For example, the European Union's General Data Protection Regulation (GDPR) requires that generated data cannot be used to re-identify any individuals[57]. An additional consideration for synthetic data use is the potential for IP contamination[58]. By using open-source models in this study, we minimize the risk of IP contamination compared to alternatives using proprietary models.

It is also important to note the potential issues of representativeness and biases when using synthetic data. Representational biases introduced through the synthetic generation process have the potential to be exacerbated if the synthetic data does not accurately represent the patient population[59]. We use real-world data from the i2b2 n2c2 2018 challenge[44] which was derived from Partners Health (now Mass General Brigham) and MIMIC, derived from Beth Israel. Evaluation in real-world data may avoid some of the potential bias exacerbation concern from synthetic data, but these datasets are both from health systems in Boston and may lack a population with sufficient representation to ensure generalization. Unfortunately, the barriers to releasing clinical notes in public or gated systems limit the datasets researchers have access to. In future studies, we aim to perform evaluation in additional real-world datasets from diverse health systems to work toward better generalizability and portability. We also aim to explore additional bias mitigation strategies that can intervene at various stages of the LLM workflow[60]. For example, prior knowledge distillation work has used data filtering and reweighting to produce more equitable teacher outputs[61–63]. The teacher's predicted token probabilities can be reweighted before being passed to the student model to reduce the inheritance or amplification of bias in the student model. Data augmentation techniques, such as data balancing[64–66], selective replacement[67,68], or interpolation[69,70] can also mitigate bias through the addition of training examples that may otherwise be underrepresented. RAG has also been explored for its potential to address biases in generative AI for health care through retrieving more inclusive or population-specific information (for example, gender-based reference ranges) to help generate more representative outputs[71].

Synthetic data distillation and fine-tuning of smaller, open-source LLMs that can be locally deployed within existing healthcare IT infrastructures can serve as a scalable alternative to more resource-intensive, proprietary models for clinical information extraction. The ability for scalable extraction of information from unstructured clinical notes allows for broader adoption in diverse healthcare system settings, with the potential to strengthen retrospective research by enabling more precise and accurate phenotyping. This work contributes to efforts to support the effective and practical integration of LLMs in healthcare settings, with the ultimate goal of supporting medical research to improve patient outcomes.

Methods

In this section, we describe our knowledge distillation process which uses a large model, Llama-3.1-70B-Instruct, to generate training examples for the smaller model, Llama-3.1-8B-Instruct (or Llama-3.2-3B-Instruct or Llama-3.2-1B-Instruct; Fig. 1). We chose the Llama family of models over other open-source alternatives due to both their benchmarked performance metrics[72] and the extent to which they have been integrated into software frameworks for finetuning and inference. Additionally, the QLoRA finetuning method[37] that we describe in subsection 4 was originally tested with the Llama family of models.

Synthetic data generation

For each patient record, we used Llama-3.1-70B-Instruct[72] to generate different, patient note-specific questions similar to clinical trial eligibility criteria of a given type (Supplementary Table 1). We prompted the model to supply its answers in json format. Each JSON includes the following: (1) the *question*; (2) the *question type*; (3) the *answer*, (4) the *section* of the note containing the answer (*e.g.*, Past Medical History, Plan, etc.); (5) the verbatim *source* of the answer from the clinical note; (6) a *difficulty level* for the question on a scale of 1-10; and (7) an *explanation* justifying the answer choice, including how the source helped to answer the question.

We included the following question types: "boolean" (answer "Yes" or "No"), "numeric", "na-boolean", and "na-numeric", where the "na" types corresponded to questions that could not be answered relying on the information in the note but seemed like they would be applicable to this patient and are similar to clinical trial eligibility criteria. For "na" type questions, we stipulated the section to be "Not Found" and the source was "Not in Note." The purpose of the "na" types as well as the supporting data, was to try to teach the model not to provide seemingly confident answers (i.e., hallucinations) when there doesn't exist sufficient evidence in the note to draw a conclusion. We provide example questions of each type (Supplementary Table 6) as well as a specific example supplied in the prompt to demonstrate the specific language used to generate each question type (Supplementary Table 1).

We generated 212,132 boolean question and answer (Q&A) pairings, 209,637 numeric Q&A pairings, 106,288 "na-boolean" Q&A pairings, and 106,245 "na-numeric" Q&A pairings. The number of questions arose from running the synthetic data generation process on 10,000 discharge summaries, where the model was asked to generate 20 boolean questions (10 with yes as the answer and 10 with no), 20 numeric questions, and 10 of each "na" category. The model tended to provide slightly more than the requested number of questions per note. The number of questions of each type per difficulty score assigned by Llama-3.1-70B are described in the supplement (Supplementary Table 2).

Data programming

For each question type, we select the 25,000 most difficult questions according to the LLM-estimated difficulty rating and randomly split them into a training and test set at a 90–10% ratio. We perform post-processing to extract our requests from the JSON response and handle malformed JSON outputs. The datasets are randomly shuffled prior to fine-tuning.

Limited human review

To ensure data quality for the fine-tuning process, we manually reviewed a random sample containing 1000 questions generated by Llama-3.1-70B-Instruct. For this purpose, we developed an open-source tool that facilitates record review from within a web browser (https://github.com/bbj-lab/annotation-ui). Users with minimal technical experience can check patient records against the generated question-answer pairs and refine answers if needed. Statistics about the number of questions that required refinement are available in the supplement (Supplementary Table 1). Manual review allowed us to both profile the accuracy of the synthetic data generation process and to better understand common failure modes.

QLoRA fine-tuning. After data programming and limited human review, we used the refined synthetic dataset to perform supervised fine-tuning on an instance of Llama-3.1-8B[72]. Specifically, we fine-tuned with QLoRA[37], a quantized version of Low-Rank Adaptation (LoRA: [73]). LoRA fine-tunes the attention weights in a pre-trained transformer with a low-rank update (a $d \times k$ matrix BA, where B is $d \times r$ and A is $r \times k$ where $r \ll min\{d,k\}$) that significantly reduces the number of required parameters and does not add to inference latency. QLoRA operates on a quantized transformer, i.e., one that uses 4-bit as opposed to 16-bit parameters, to further reduce memory requirements and uses paged optimizers that manage the exchange of memory between GPU and CPU components.

Inference - Sampling hyperparameter selection

During generation, we tested different values of *temperature* and *top_p* (specifically temperatures of 0 and 1 and top_p of 0.5 and 0.95). Temperature controls the randomness of sampling, with higher temperatures corresponding to more novelty in generated output. However, increasing temperature may also make text less coherent and hallucinations more likely. Consequently, higher values of temperature are often used for creative tasks, while lower values are used for dialoguing about matters of fact. Chang et al.[74] hypothesized that lower values of temperature may be better suited to question-answering with attribution. However, Renze and Guven's recent work[75] indicates that LLM problem-solving performance does not significantly vary for temperature values between 0 and 1. The top_p parameter controls nuclear sampling[76], with higher values corresponding to a more permissive threshold for filtering.

Setting temperature = 0 and top_p = 1 results in a nearly deterministic, greedy sampling strategy that aims to select the most likely token given the current context. Setting temperature = 1 and top_p = 0.5 restricts tokens to come from a likely subset of the token set, but otherwise samples according to the predicted odds. We limit this parameter evaluation to the *i2b2 n2c2* challenge and report the full results for all parameters (Supplementary Table 3). Because we did not see a benefit when increasing the temperature we fixed temperature = 0 and top_p = 1 for all other evaluations.

Versions of Fine-tuned Models

The following models resulted from finetuning Llama-3.1-8B-Instruct released by Meta as the base model (*Ablation study on fine-tuning data selection* - Table 1):

a. All (Labeled 8B-All). Fine-tuning was performed with the complete dataset, using question, answer, question type, section, source, and explanation as described in the methods.
b. Hardest (Labeled 8B-H-25k). To determine the performance impact of reducing training set size, we selected the 25,000 questions the model determined had the highest difficulty in Step #1 (*Synthetic Data Generation)* for each question type. This subset of the original training data was then used to fine-tune the model.
c. Hardest Boolean and Numeric (Labeled 8B-NB-Only). To determine the impact that n/a questions have on model fine tuning, we selected the most difficult 25,000 questions for only the boolean and numeric types (dropping "na-boolean" and "na-numeric") from the original dataset for fine-tuning.
d. No Support (Labeled 8B-No-S). To determine the usefulness of including textual references and an explanation of the correct answer, we dropped the section, source, and explanation from the original training set and fine-tuned the model with this data.

We also fine-tuned smaller versions of Llama-3.2 released by Meta:

a. All (Labeled 3B-All). Finetuning was performed using the complete dataset (as with 8B-All), except with Llama-3.2-3B-Instruct as the base model.
b. All (Labeled 1B-All). Finetuning was performed using the complete dataset (as with 8B-All), except with Llama-3.2-1B-Instruct as the base model.

Model evaluation

We took the models finetuned on synthetic training data and evaluated them on three separate datasets, including one synthetic dataset and two real-world datasets, as follows:

First, we evaluate methods on a held-out set of 42,498 synthetic examples generated in an identical manner to the dataset used for fine-tuning. The breakdown of examples by type was as follows: 10,722 (25.2%) boolean, 10,666 (25.1%) numeric, 10,664 (25.1%) na-boolean, and 10,446 (24.6%) na-numeric. From this set, we drew a random sample containing 1000 examples and manually annotated it as described in the "Limited Human Review" subsection of our methods, correcting questions, answers, and explanations when necessary. We calculated the accuracy for these questions, as given in the results section. We provided a summary of this dataset (Supplementary Table 1) and have released a copy of it on PhysioNet. Results for this dataset allow us to evaluate the extent to which the fine-tuning objective was successfully optimized. The next two subsections describe tests on real-world data.

Next, we evaluate methods on the clinical trial eligibility criteria cohort selection shared task from the i2b2 2018 National NLP Clinical Challenges (n2c2)[44]. Track 1 contains 288 de-identified longitudinal medical records for patients with diabetes, many of whom are at risk for heart disease. The records are manually annotated according to 13 selection criteria adapted from real clinical trials and split into a 202-patient training set and an 86-patient test set. We calculated balanced accuracy and micro-F1 score on both the training and test datasets corresponding to the original challenge. At the time of the challenge, the top-performing team adopted a rule-based method to obtain a micro-F_1 score of 0.91 on the test set. Other teams achieved similar results ($F_1 > 0.9$) with hybrid approaches; for example, cTakes[77] was used by 3 of the top 5 teams to extract knowledge from the text. Because we only use this dataset to test zero-shot extraction and do not train on it, we are able to evaluate the model performance on both the training and test sets to have a larger sample size.

We also evaluate methods on clinical trial eligibility criteria resembling those of the 2011 ARISTOTLE clinical trial comparing apixaban to warfarin[42]. We developed 23 human-generated boolean and numeric questions assessing these criteria (Supplementary Table 4). Using these questions, we manually annotated notes for 2300 total question-answer pairs within MIMIC-IV[78,79]. Notes from MIMIC-IV were taken from after 2012 to ensure no overlap with any of the notes from MIMIC-III, which were used to generate synthetic data. We evaluated the models on these question-answer pairs and calculated both balanced accuracy and micro-F1 score. We are releasing the dataset and manual annotations to PhysioNet and will make them available under the same data use terms as MIMIC-III/IV.

Data availability

The MIMIC-III [Johnson, et al., 2016] and MIMIC-IV [Johnson, et al., 2023] datasets are available from PhysioNet. The datasets of the Annotated Synthetic Questions and the Apixaban Trial Criteria Questionsare are available from physionet: https://physionet.org/content/mimic-iv-ext-apixaban-trial/1.0.0/; https://physionet.org/content/mimic-ext-synth-trial-question/1.0.0/.

Code availability

Source code for clinical information extraction and synthetic data generation can be accessed at https://github.com/bbj-lab/clinical-synthetic-data-distil. Source code for the annotation tool, when LLM predicted annotations are already available, is available at https://github.com/bbj-lab/annotation-ui.

Received: 27 September 2024; Accepted: 25 April 2025;
Published online: 10 May 2025

References

1. Goel, A. et al. LLMs Accelerate Annotation for Medical Information Extraction. *arXiv [cs.CL]* (2023).
2. Pangakis, N., Wolken, S. & Fasching, N. Automated annotation with generative AI requires validation. *arXiv [cs.CL]* (2023).
3. Agrawal, M., Hegselmann, S., Lang, H., Kim, Y. & Sontag, D. Large language models are few-shot clinical information extractors. *arXiv [cs.CL]* (2022).
4. McInerney, D. J., Young, G., van de Meent, J.-W. & Wallace, B. C. CHiLL: Zero-shot custom interpretable feature extraction from clinical notes with large language models. *arXiv [cs.CL]* (2023).
5. He, K. et al. A survey of large language models for Healthcare: From data, technology, and applications to accountability and ethics. *arXiv [cs.CL]* (2023).
6. Chapman, W. W. et al. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *J. Am. Med. Inform. Assoc.: JAMIA* **18**, 540–543 (2011).
7. Wang, Y. et al. Clinical information extraction applications: A literature review. *J. Biomed. Inform.* **77**, 34–49 (2018).
8. OpenAI, et al. GPT-4 Technical Report. *arXiv [cs.CL]* (2023).
9. Toma, A., Senkaiahliyan, S., Lawler, P. R., Rubin, B. & Wang, B. Generative AI could revolutionize health care - but not if control is ceded to big tech. *Nature* **624**, 36–38 (2023).
10. Reddy, S. Generative AI in healthcare: an implementation science informed translational path on application, integration and governance. *Implement. Sci.* **19**, 27 (2024).
11. Minssen, T., Vayena, E. & Cohen, I. G. The challenges for regulating medical use of ChatGPT and other large language models. *JAMA* **330**, 315–316 (2023).
12. Blogs, M. C. Microsoft and Epic expand AI collaboration to accelerate generative AI's impact in healthcare, addressing the industry's most pressing needs. *The Official Microsoft Blog* https://blogs.microsoft.com/blog/2023/08/22/microsoft-and-epic-expand-ai-collaboration-to-accelerate-generative-ais-impact-in-healthcare-addressing-the-industrys-most-pressing-needs/ (2023).
13. Zhang, G. et al. Closing the gap between open source and commercial large language models for medical evidence summarization. *NPJ Digit. Med.* **7**, 239 (2024).
14. Chatbot Arena (formerly LMSYS): Free AI Chat to Compare & Test Best AI Chatbots. https://gradio.app/.
15. Wiest, I. C. et al. Privacy-preserving large language models for structured medical information retrieval. *NPJ Digit. Med.* **7**, 1–9 (2024).
16. Hinton, G., Vinyals, O. & Dean, J. Distilling the knowledge in a neural network. *arXiv [stat.ML]* (2015).
17. Papamakarios, G. Distilling model knowledge. *arXiv [stat.ML]* (2015).
18. Ding, S., Ye, J., Hu, X. & Zou, N. Distilling the knowledge from large-language model for health event prediction. *Health Inform.* **14**, 30675 (2024).
19. Li, R., Wang, X. & Yu, H. LlamaCare: An instruction fine-tuned large language model for clinical NLP. In *Proc. of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (eds Calzolari, N. et al.) 10632–10641 (ELRA and ICCL, Torino, Italia, 2024).
20. Qin, D. et al. Efficient medical image segmentation based on knowledge distillation. *arXiv [eess.IV]* https://doi.org/10.1109/TMI.2021.3098703 (2021).
21. Qi, X. et al. Exploring generalizable distillation for efficient medical image segmentation. *IEEE J. Biomed. Health Inform.* **28**, 4170–4183 (2024).
22. Wang, T., Zhu, J.-Y., Torralba, A. & Efros, A. A. Dataset Distillation. *arXiv [cs.LG]* (2018).
23. Wang, Z., Yu, A. W., Firat, O. & Cao, Y. Towards zero-label language learning. *arXiv [cs.CL]* (2021).

24. Shirgaonkar, A., Pandey, N., Abay, N. C., Aktas, T. & Aski, V. Knowledge distillation using frontier open-source LLMs: Generalizability and the role of synthetic data. *arXiv [cs.LG]* (2024).
25. Yu, P., Xu, J., Weston, J. & Kulikov, I. Distilling System 2 into System 1. *arXiv [cs.CL]* (2024).
26. Ding, B. et al. Data augmentation using large language models: Data perspectives, learning paradigms and challenges. *arXiv [cs.CL]* (2024).
27. Peng, B., Li, C., He, P., Galley, M. & Gao, J. Instruction Tuning with GPT-4. *arXiv [cs.CL]* (2023).
28. Fitzsimmons, L., Frau, F., Bozzi, S., Chandross, K. J. & Beaulieu-Jones, B. K. Characterizing the connection between Parkinson's disease progression and healthcare utilization. *medRxiv* 2024.09.15.24313708 https://doi.org/10.1101/2024.09.15.24313708 (2024).
29. Beaulieu-Jones, B. K. et al. Disease progression strikingly differs in research and real-world Parkinson's populations. *NPJ Parkinsons Dis.* **10**, 58 (2024).
30. Alsentzer, E. et al. Zero-shot interpretable phenotyping of postpartum hemorrhage using large language models. *NPJ Digit. Med.* **6**, 212 (2023).
31. Peikos, G., Symeonidis, S., Kasela, P. & Pasi, G. Utilizing ChatGPT to enhance clinical trial enrollment. *arXiv [cs.IR]* (2023).
32. Yuan, J., Tang, R., Jiang, X. & Hu, X. Large language models for healthcare data augmentation: An example on patient-trial matching. *AMIA Annu. Symp. Proc.* **2023**, 1324–1333 (2023).
33. Wong, C. *et al*. Scaling Clinical Trial Matching Using Large Language Models: A Case Study in Oncology. in *Machine Learning for Healthcare Conference* 846–862 (PMLR, 2023).
34. Gupta, S. et al. PRISM: Patient Records Interpretation for Semantic clinical trial Matching system using large language models. *NPJ Digit. Med.* **7**, 305 (2024).
35. Jin, Q. et al. Matching patients to clinical trials with large language models. *arXiv [cs.CL]* (2023).
36. Nievas, M., Basu, A., Wang, Y. & Singh, H. Distilling large language models for matching patients to clinical trials. *J. Am. Med. Inform. Assoc.* **31**, 1953–1963 (2024).
37. Dettmers, T., et al. vol. 36 10088–10115 (Curran Associates, Inc., 2023).
38. Terms of use. https://openai.com/policies/row-terms-of-use/.
39. Snell, C., Klein, D. & Zhong, R. Learning by distilling context. *arXiv [cs.CL]* (2022).
40. Huang, J. et al. Large Language Models can self-improve. *arXiv [cs.CL]* (2022).
41. Hsieh, C.-Y. et al. Distilling step-by-step! Outperforming larger language models with less training data and smaller model sizes. *arXiv [cs.CL]* (2023).
42. Granger, C. B. et al. Apixaban versus warfarin in patients with atrial fibrillation. *N. Engl. J. Med.* **365**, 981–992 (2011).
43. Study Details. https://www.clinicaltrials.gov/study/NCT00496769#participation-criteria.
44. Stubbs, A., Filannino, M., Soysal, E., Henry, S. & Uzuner, Ö Cohort selection for clinical trials: n2c2 2018 shared task track 1. *J. Am. Med. Inform. Assoc.* **26**, 1163–1171 (2019).
45. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models. *Meta AI* https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/.
46. Wang, T. et al. Self-Taught Evaluators. *arXiv [cs.CL]* (2024).
47. Goodell, A. J., Chu, S. N., Rouholiman, D. & Chu, L. F. Large language model agents can use tools to perform clinical calculations. *NPJ Digit. Med.* **8**, 163 (2025).
48. Hager, P. et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat. Med.* **30**, 2613–2622 (2024).
49. Lewis, P. et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *arXiv [cs.CL]* (2020).
50. Gao, Y. et al. Retrieval-Augmented Generation for large Language Models: A survey. *arXiv [cs.CL]* (2023).
51. Wei, J. et al. Chain-of-thought prompting elicits reasoning in large language models. *arXiv [cs.CL]* (2022).
52. Kojima, T., Gu, S. S., Reid, M., Matsuo, Y. & Iwasawa, Y. Large Language Models are Zero-Shot Reasoners. *arXiv [cs.CL]* (2022).
53. Paul, M., Ganguli, S. & Dziugaite, G. K. Deep Learning on a Data Diet: Finding Important Examples Early in Training. in *Advances in Neural Information Processing Systems* (eds. Beygelzimer, A., Dauphin, Y., Liang, P. & Vaughan, J. W.) (2021).
54. Sorscher, B., Geirhos, R., Shekhar, S., Ganguli, S. & Morcos, A. S. Beyond neural scaling laws: beating power law scaling via data pruning. in *Advances in Neural Information Processing Systems* (eds. Oh, A. H., Agarwal, A., Belgrave, D. & Cho, K.) (2022).
55. Yang, Y., Bean, A. M., McCraith, R. & Mahdi, A. Fine-tuning Large Language Models with human-inspired learning strategies in medical question answering. *arXiv [cs.CL]* (2024).
56. Arora, A. & Arora, A. Synthetic patient data in health care: a widening legal loophole. *Lancet* **399**, 1601–1602 (2022).
57. Beduschi, A. Synthetic data protection: Towards a paradigm change in data regulation? *Big Data Soc*. **11**, https://doi.org/10.1177/20539517241231277 (2024).
58. Marwala, T. *The Use of Synthetic Data to Train AI Models: Opportunities and Risks for Sustainable Development*. https://unu.edu/publication/use-synthetic-data-train-ai-models-opportunities-and-risks-sustainable-development (2023).
59. Draghi, B., Wang, Z., Myles, P. & Tucker, A. Identifying and handling data bias within primary healthcare data using synthetic data generators. *Heliyon* **10**, e24164 (2024).
60. Gallegos, I. O. et al. Bias and fairness in large language models: A survey. *arXiv [cs.CL]* (2023).
61. Gupta, U. et al. Mitigating gender bias in distilled language models via counterfactual role reversal. in *Findings of the Association for Computational Linguistics: ACL 2022* 658–678 (Association for Computational Linguistics, Stroudsburg, PA, USA, 2022).
62. Delobelle, P. & Berendt, B. FairDistillation: Mitigating stereotyping in language models. *arXiv [cs.CL]* (2022).
63. Ahn, J., Lee, H., Kim, J. & Oh, A. Why knowledge distillation amplifies gender bias and how to mitigate from the perspective of DistilBERT. in *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)* 266–272 (Association for Computational Linguistics, Stroudsburg, PA, USA, 2022).
64. Webster, K. et al. Measuring and reducing gendered correlations in pre-trained models. *arXiv [cs.CL]* (2020).
65. Ghanbarzadeh, S., Huang, Y., Palangi, H., Cruz Moreno, R. & Khanpour, H. Gender-tuning: Empowering fine-tuning for debiasing pre-trained language models. in *Findings of the Association for Computational Linguistics: ACL 2023* 5448–5458 (Association for Computational Linguistics, Stroudsburg, PA, USA, 2023).
66. Dixon, L., Li, J., Sorensen, J., Thain, N. & Vasserman, L. Measuring and mitigating unintended bias in text classification. in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society.* https://doi.org/10.1145/3278721.3278729 (ACM, New York, NY, USA, 2018).
67. Hall Maudslay, R., Gonen, H., Cotterell, R. & Teufel, S. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. 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)* 5267–5275 (Association for Computational Linguistics, Stroudsburg, PA, USA, 2019).
68. Zayed, A. et al. Deep learning on a healthy data diet: Finding important examples for fairness. *arXiv [cs.CL]* (2022).
69. Zhang, H., Cisse, M., Dauphin, Y. N. & Lopez-Paz, D. mixup: Beyond Empirical Risk Minimization. *arXiv [cs.LG]* (2017).

70. Yu, L., Mao, Y., Wu, J. & Zhou, F. Mixup-based unified framework to overcome gender bias resurgence. in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* 1755–1759 (ACM, New York, NY, USA, 2023).
71. Yang, R. et al. Retrieval-augmented generation for generative artificial intelligence in health care. *npj Health Syst.* **2**, 1–5 (2025).
72. Dubey, A. et al. The Llama 3 herd of models. *arXiv [cs.AI]* https://doi.org/10.48550/arXiv.2309.03882 (2024).
73. Hu, E. J. et al. LoRA: Low-Rank Adaptation of Large Language Models. in *International Conference on Learning Representations* (2022).
74. Chang, C.-C., Reitter, D., Aksitov, R. & Sung, Y.-H. KL-divergence guided temperature sampling. *arXiv [cs.CL]* (2023).
75. Renze, M. & Guven, E. The effect of sampling temperature on problem solving in Large Language Models. *arXiv [cs.CL]* (2024).
76. Holtzman, A., Buys, J., Du, L., Forbes, M. & Choi, Y. The Curious Case of Neural Text Degeneration. in *International Conference on Learning Representations* (2020).
77. Savova, G. K. et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J. Am. Med. Inform. Assoc.* **17**, 507–513 (2010).
78. Johnson, A. et al. MIMIC-IV. PhysioNet https://doi.org/10.13026/HXP0-HG59 (2024).
79. Johnson, A. E. W. et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci. Data* **10**, 1 (2023).

Acknowledgements

This work was funded in part by the National Institutes of Health, specifically the National Institute of Neurological Disorders and Stroke grant number R00NS114850 to BKB. This project would not have been possible without the support of the Center for Research Informatics at the University of Chicago and, particularly the High-Performance Computing team. The authors are grateful for the resources and support this team provided throughout the duration of the project. The Center for Research Informatics is funded by the Biological Sciences Division at the University of Chicago with additional funding provided by the Institute for Translational Medicine, CTSA grant number UL1 TR000430 from the National Institutes of Health.

Author contributions

B.K.B. and E.A. conceived and designed the study. B.K.B., E.G.W., and M.B. performed data annotation, analysis, interpretation, and drafted the initial manuscript. All authors substantially revised the manuscript and approved the final version.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-025-01681-4.

Correspondence and requests for materials should be addressed to Brett K. Beaulieu-Jones.