scientific reports

Check for updates

OPEN

Leveraging large language models for knowledge-free weak supervision in clinical natural language processing

Enshuo Hsu[1,2,3] & Kirk Roberts[1]✉

The performance of deep learning-based natural language processing systems is based on large amounts of labeled training data which, in the clinical domain, are not easily available or affordable. Weak supervision and in-context learning offer partial solutions to this issue, particularly using large language models (LLMs), but their performance still trails traditional supervised methods with moderate amounts of gold-standard data. In particular, inferencing with LLMs is computationally heavy. We propose an approach leveraging fine-tuning LLMs and weak supervision with virtually no domain knowledge that still achieves consistently dominant performance. Using a prompt-based approach, the LLM is used to generate weakly-labeled data for training a downstream BERT model. The weakly supervised model is then further fine-tuned on small amounts of gold standard data. We evaluate this approach using Llama2 on three different i2b2/ n2c2 datasets for clinical named entity recognition. With no more than 10 gold standard notes, our final BERT models weakly supervised by fine-tuned Llama2-13B consistently outperformed out-of-the-box PubMedBERT by 4.7–47.9% in F1 scores. With only 50 gold standard notes, our models achieved close performance to fully fine-tuned systems.

Keywords Natural language processing, Large language models, Electronic health records, Weak supervision

Deep learning-based natural language processing (NLP) has achieved remarkable success in the open domain. However, achieving optimal performance in the clinical domain faces many challenges. First, training such complex architectures often requires a large labeled corpus[1]. Second, specific subpopulations (e.g., rare diseases, minority ethnicities) are often under-represented in clinical notes[2], magnifying the consequences of underpowered datasets. Third, even with sufficient notes available in electronic health records (EHRs), the protection of patient privacy makes access to the corpus challenging. Finally, manual annotation of a gold standard is not only a labor-intensive task, it also requires advanced clinical knowledge for interpretation of the text in clinical notes[3,4]. In recent years, approaches including weak supervision and in-context learning have been developed to address this challenge[1,4].

Weak supervision, which utilizes labeling functions (LFs) to generate noisy weak labels for model training, has already been adopted in the clinical domain[3,5–11]. Despite its promise, weak supervision still requires significant resources to construct LFs. The rule-based approach requires domain experts to handcraft decision rules[5–9]. The ontology-based approach requires that the concepts of interest be included in existing ontologies or dictionaries[10,11]. Data programming requires significant efforts from programmers who have a thorough understanding of the clinical data[3].

In-context learning, in which pre-trained large language models (LLMs) gain the ability to generate textual outputs by prompting at inferencing time, is a relatively new method. In theory, it requires few ("few-shot") or even no ("zero-shot") training data[12,13]. However, recent studies raised concerns about underperformance[14–16] and instability[17] in the medical domain. Despite the appealing idea, at this point, there is no strong evidence

[1]McWilliams School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX, USA. [2]Center for Health Data Science and Analytics, Houston Methodist, Houston, TX, USA. [3]Enterprise Development and Integration, University of Texas MD Anderson Cancer Center, Houston, TX, USA. ✉email: kirk.roberts@uth.tmc.edu

to support the use of LLM as the frontline approach in a medical NLP system. Furthermore, due to the model sizes (measured as the number of parameters), LLM inference requires significant computation resources. We estimate that performing few-shot prompting with Llama2-13B, a 13 billion-parameter model[18] for 2018 n2c2 benchmark[19] (a subset of 505 discharge summaries from the MIMIC-III dataset[20]) requires 3.3×10^{12} float point operations (FLOPs) per input sentence. On the other hand, predicting with a Bidirectional Encoder Representations from Transformers (BERT) model with 110 million parameters only requires 4.4×10^{10} FLOPs per input sentence. This computational difference results in a dramatic difference in GPU time such that inferencing the entire collection of MIMIC-III discharge summaries would take an estimated 727 days on an NVIDIA A100 GPU while predicting with BERT would only take around 18 h (Fig. 1).

Recently, a few attempts have been made to combine the benefits of both weak supervision and in-context learning[21,22]. However, to our knowledge, there is no evaluation of an end-to-end approach in the medical domain that prompts an LLM for weak supervision and fine-tunes smaller models on the downstream task gold standard. The benefits and limitations of this method in a practical scenario where a small number of annotated notes are available have not been evaluated. Furthermore, fine-tuning LLM which has shown significant benefits in recent studies[23] has not been considered in such pipelines. Therefore, we propose an LLM-powered weak supervision approach that (1) minimizes domain expertise for rule-crafting and data programming and removes the dependency for ontologies by using the LLM to create weak labels, (2) leverages the latest prompt-based supervised fine-tuning (SFT) techniques to fine-tune LLMs, (3) consistently achieves dominant performances by weakly supervising and fine-tuning BERT[24] models for downstream tasks, and (4) avoids the computational burden of deploying LLMs in the production environment.

In this study, we evaluated four experimental settings as detailed in Table 1. The primary method, Llama-SFT*n*-WS-BERT*n* starts with supervised fine-tuning (SFT) Llama2-13B with a certain number (*n*) of gold standard notes in the training set. The fine-tuned Llama model then performs few-shot prompting on the rest of the training set to generate weak labels. We use the weak labels to perform weak supervision (WS) on BERT, followed by final fine-tuning of BERT with the same *n* gold standards (Fig. 2). Considering the high GPU memory requirement of SFT, we also proposed a compact version, Llama-WS-BERT*n* which the SFT of Llama2 was

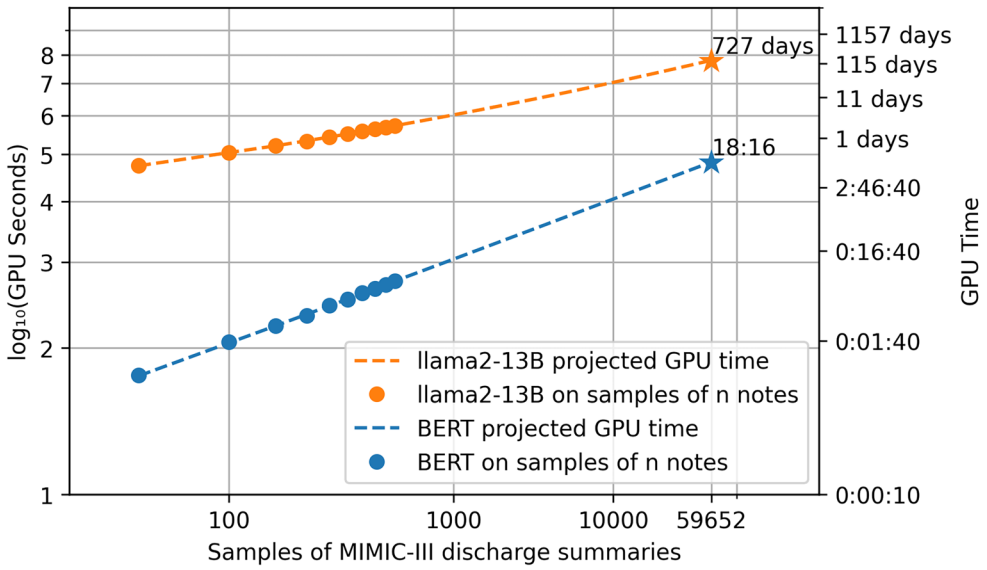

Figure 1. Benchmarking GPU hours with MIMIC-III discharge summaries. The 505 discharge summaries in the 2018 n2c2 challenge were used to project the entire collection of discharge summaries. Running on an NVIDIA A100 GPU, Llama2-13B requires 727 days of GPU time, while PubMedBERT only requires about 18 h.

	Notation	Description	Product
Proposed methods	Llama-SFT*n*-WS-BERT*n*	Llama2-13B is supervised fine-tuned (SFT) with *n* gold standard notes in the training set, then performs few-shot prompting to generate weak labels. Weakly supervise BERT and fine-tune BERT with the same *n* gold standard notes	A weakly supervised and fine-tuned BERT
	Llama-WS-BERT*n*	Llama2-13B out-of-the-box performs few-shot prompting to generate weak labels. Weakly supervise BERT and fine-tune BERT with *n* gold standard notes	A weakly supervised and fine-tuned BERT
Baselines	Llama-SFT*n*	Llama2-13B is supervised fine-tuned with *n* gold standard notes in training set	A fine-tuned Llama
	BERT*n*	Fine-tune BERT with *n* gold standard notes	A fine-tuned BERT

Table 1. Experimental settings.

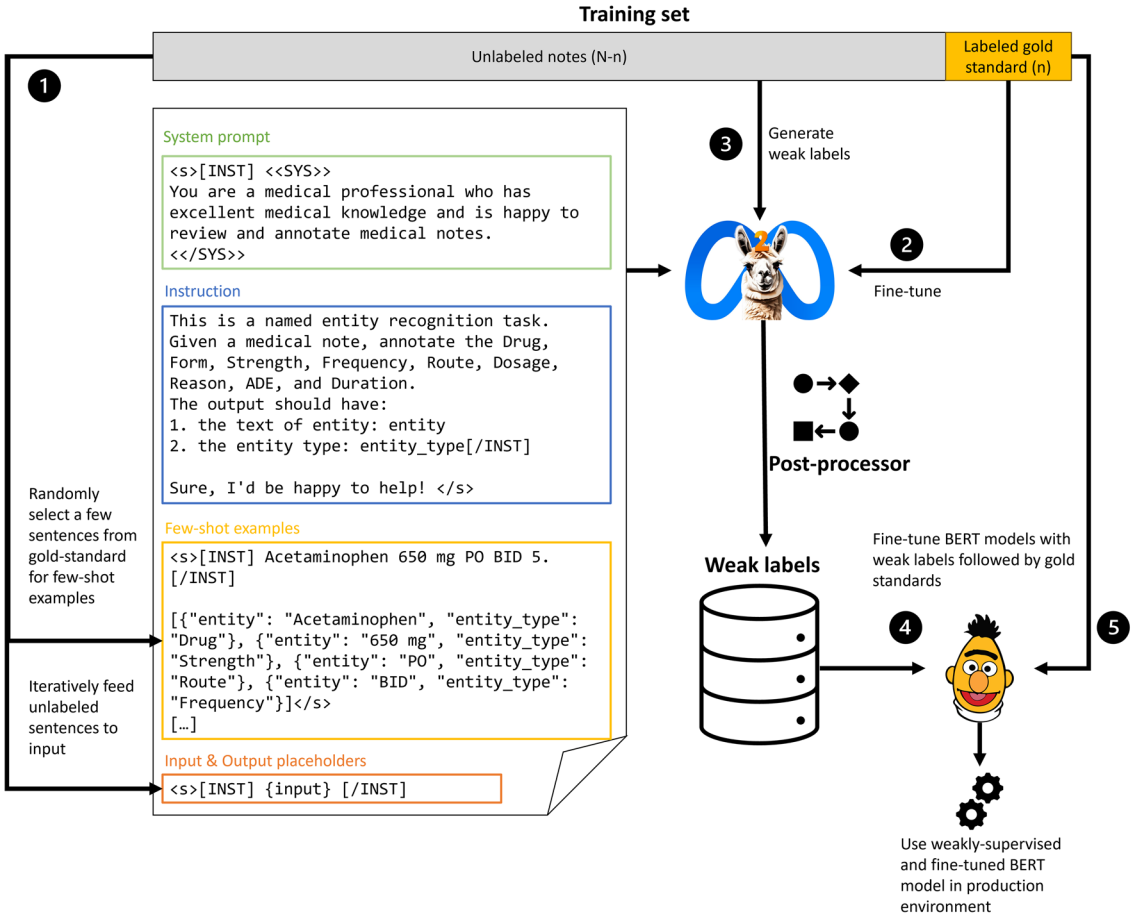

Figure 2. Methodology flowchart. (**A**) prompt template is constructed with a few random sentences from the training set as the few-shot examples. For certain annotated gold standard notes, we first fine-tune Llama2-13B, then use the fine-tuned model to perform few-shot prompting to weakly supervise a PubMedBERT. Finally, we fine-tune the BERT model with gold standard notes and use it in the production environment.

omitted. We use Llama2 out-of-the-box to generate pseudo-labels ("weak labels") and weakly supervise BERT. For comparison, we evaluated two baselines, Llama-SFT*n* and BERT*n* which Llama2-13B and PubMedBERT were fine-tuned with *n* gold standard notes. Details are described in the Methods section.

We evaluated three widely used biomedical benchmarks, 2012[25], 2014[26] Integrating Biology and the Bedside (i2b2), and 2018[19] National NLP Clinical Challenges (n2c2) Natural Language Processing Challenges for temporal relation extraction, protected health information (PHI) de-identification, and adverse drug events (ADEs) and medication properties extraction, respectively. The benchmarks follow the named entity recognition (NER) schema which each named entity is labeled as a frame with entity type, start, and end character positions. Below is an example:

{"entity_type": "Drug", "start_pos": 10227, "end_pos" 10233, "entity": "Ativan"}

This study demonstrates a robust usage of LLMs that requires minimal to zero human input while achieving significant improvement in well-established benchmarks. We hypothesize that this approach is a safe and effective means of augmenting existing supervised clinical NLP approaches by inserting this simple technique between the now-standard pre-training and fine-tuning steps.

Results

LLMs inferencing is computationally expensive

On the 2018 benchmark, which contains a subset of 505 MIMIC-III discharge summaries, Llama2-13B spent 147 GPU hours in total, with a median of 16 min (Q1–Q3 = 11–22 min) to create weak labels on each note. Since the computation for inference is linear to the number of input instances, we fit a linear regression model to project the total GPU time for labeling all the 59,652 discharge summaries in the MIMIC-III dataset. The projected time on a single NVIDIA A100 GPU is 727 days. PubMedBERT took 9 min in total, with a median of 1 s (Q1–Q3 = 0.7–1.4 s) per note. The projected time for labeling all the discharge summaries in MIMIC-III is 18 h and 16 min (Fig. 1).

LLM-generated weak labels

We prompted the out-of-the-box Llama2-13B and the fine-tuned Llama-SFT*n* LLMs to generate a list of entities for each input sentence following the format below:

```
[
 {"entity": "Acetaminophen", "entity_type": "Drug"},
 {"entity": "650 mg", "entity_type": "Strength"},
 {"entity": "PO", "entity_type": "Route"},
 {"entity": "BID", "entity_type": "Frequency"}
]
```

The entities were then filtered and post-processed into frames with standardized entity types and start and end character positions (see Methods, post-processing section). For the 2012 benchmark, the out-of-the-box Llama2-13B and the fine-tuned Llama-SFT*3* generated weak labels with 9804 and 20,402 entities, respectively. The median numbers of entities per sentence were 2 and 3, respectively. For the 2014 benchmark, Llama2-13B and Llama-SFT*3* generated weak labels with 18,062 and 15,190 entities, respectively, with a median of 1 entity per sentence. For the 2018 benchmark, Llama2-13B and Llama-SFT3 generated weak labels with 53,177 and 56,169 entities with 1 and 4 entities per sentence, respectively. Our post-processing algorithm was able to handle the majority of LLM predictions, with less than 1% of sentences failing due to inconsistent output formats (Table 2).

Proposed method: Llama-SFT*n*-WS-BERT*n*

Our primary proposed method, Llama-SFT*n*-WS-BERT*n* consistently achieved dominant performance in most experiments across the three benchmarks. In the extremely low-resourced setting in which only 3 gold standard notes were used, on the 2012 events benchmark, time expression benchmark, 2014 benchmark, and 2018

Benchmarks (Training set)	2012	2014	2018
Features			
Notes	190	790	303
Sentences	5995	34,101	46,228
Total entities	17,933	17,401	50,951
Entities per sentence, median [Q1, Q3]	3 [2, 4]	0 [0, 0]	0 [0, 0]
Entities per sentence, mean (Std Dev)	3.14 (2.34)	0.51 (1.72)	1.1 (3.13)
Entities per note, median [Q1, Q3]	82 [56, 120]	18 [13, 27]	147 [96, 224]
Weak labels generated by Llama2-13B			
Post-processing failed, sentences (%)	2	2	22
Total entities	9804	18,062	53,177
Entities per sentence, median [Q1, Q3]	2 [1, 3]	1 [1]	1 [1, 3]
Entities per note, median [Q1, Q3]	46 [31, 61]	19 [12, 29]	165 [102, 232]
Weak labels generated by Llama-SFT3			
Post-processing failed, sentences (%)	14	27	82
Total entities	20,402	15,190	56,169
Entities per sentence, median [Q1, Q3]	3 [2, 5]	1 [1, 4]	4 [2, 7]
Entities per note, median [Q1, Q3]	88 [59, 132]	15 [10, 21]	181 [111, 246]
Weak labels generated by Llama-SFT5			
Post-processing failed, sentences (%)	76	34	48
Total entities	18,676	15,107	48,143
Entities per sentence, median [Q1, Q3]	3 [2, 5]	1 [1, 3]	4 [2, 7]
Entities per note, median [Q1, Q3]	82 [57, 119]	14 [9, 22]	147 [90, 215]
Weak labels generated by Llama-SFT10			
Post-processing failed, sentences (%)	69	28	25
Total entities	18,386	13,743	46,223
Entities per sentence, median [Q1, Q3]	3 [2, 5]	1 [1, 2]	4 [2, 7]
Entities per note, median [Q1, Q3]	81 [54, 117]	15 [11, 20]	147 [84, 207]
Weak labels generated by Llama-SFT50			
Post-processing failed, sentences (%)	72	54	59
Total entities	18,420	18,415	47,688
Entities per sentence, median [Q1, Q3]	3 [2, 5]	1 [1, 3]	4 [2, 7]
Entities per note, median [Q1, Q3]	77 [57, 118]	17 [12, 25]	150 [87, 212]

Table 2. Summary of LLM-generated weak labels.

benchmark, Llama-SFT*3*-WS-BERT*3* achieved F1 scores of 0.7765, 0.7538, 0.6336, and 0.7747, while the baseline Llama-SFT*3* had 0.7418, 0.6045, 0.5898, and 0.6252; BERT3 had 0.5953, 0.2753, 0.3083, and 0.6555. Llama-SFT*3*-WS-BERT*3* outperformed the Llama-SFT*3* baseline by 3.5% to 15.0% and the BERT*3* baseline by 11.9% to 47.9% in the F1 score. When 10 gold standard notes were used, Llama-SFT*10*-WS-BERT*10* achieved F1 scores of 0.8466, 0.8448, 0.6942, and 0.8005, which is 3.2% to 14.6% higher than the Llama-SFT*10* baseline and 4.7–16.8% higher than the BERT*10* baseline. In the relatively annotation-abundant scenario when 50 gold standard notes were used, the Llama-SFT*50*-WS-BERT*50* achieved close performance to fully supervised BERT models by only 2.8%, 2.5%, 6.1%, and 2.2% lower in F1 score. For the 2012 time expression benchmark, however, the F1 score of Llama-SFT*50*-WS-BERT*50* is slightly lower than BERT*50* by 1.3% (Fig. 3). In almost all the benchmarks and

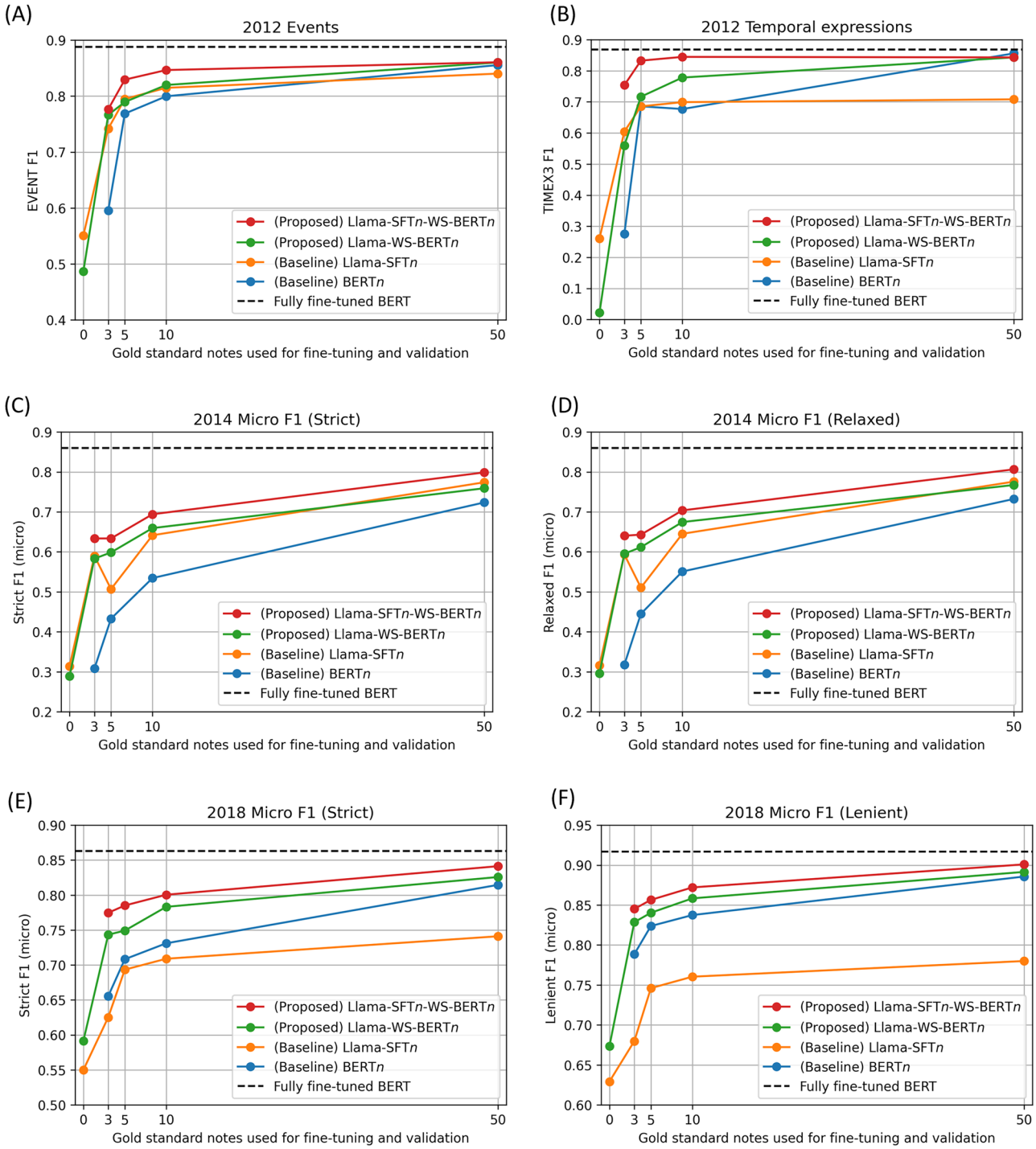

Figure 3. Weakly supervised end models fine-tuned on 3, 5, 10, and 50 gold standard notes from the training set compared to BERT models without weak supervision. (**A**) 2012 i2b2 challenge events extraction F1 score and (**B**) temporal expression extraction F1 score. (**C**) 2014 i2b2 challenge Strict micro F1 score and (**D**) Relaxed micro F1 score. (**E**) 2018 n2c2 challenge Strict micro F1 score and (**F**) Lenient micro F1 score.

subset numbers (n), our proposed Llama-SFT*n*-WS-BERT*n* outperformed the BERT*n* and Llama-SFT*n* baselines in precisions and recalls (Table S3-S5).

Proposed method: Llama-WS-BERT*n*

The compact method Llama-WS-BERT*n* showed improved performance in most benchmarks. On the 2012 event benchmark, Llama-WS-BERT*n* and Llama-SFT*n* had similar performance and the differences are 0.5% to 2.5% for *n* from 3 to 50. While it outperformed the BERT*n* by up to 17.1%. In the 2012 temporal expression benchmark, Llama-WS-BERT*n* and Llama-SFT*n* had similar performances when *n* was less than 10. While Llama-WS-BERT*10* and Llama-WS-BERT*50* outperformed Llama-SFT*10* and Llama-SFT*50* by 7.9% and 13.5%, respectively. On the 2014 benchmark, Llama-WS-BERT*n* and Llama-SFT*n* had similar performances except for *n* of 5. Llama-WS-BERT*n* outperformed BERT*n* by 3.5% to 27.5%. On the 2018 benchmark, Llama-WS-BERT*n* outperformed Llama-SFT*n* and BERT*n* by 5.6% to 11.8% and 1.1% to 8.8%, respectively. Overall, Llama-WS-BERT*n* performs similar to or better than the Llama-SFT*n* baseline while dominating the BERT*n* baseline on most benchmarks (Fig. 3).

Baseline methods

On the 2012 benchmarks, under the low-resource setting ($n < 10$), Llama-SFT*n* performed better than BERT*n*. While when $n = 50$, BERT*50* outperformed Llama-SFT*50*. On the 2014 benchmark, Llama-SFT*n* outperformed BERT*n* across the board by 5–28.2%. On the 2018 benchmark, BERT*n* outperformed Llama-SFT*n* by 1.5% to 7.4% (Fig. 3).

Llama-3 large language models

As a stand-alone sensitivity analysis, we evaluated a recently published large language model, Llama3, with 70 billion parameters[27]. We performed the compact Llama-WS-BERT*n* method, which does not require fine-tuning the LLM. The results of the 2018 benchmark showed a consistent outperformance over the Lama2-13B by 1.1–6.1% under the Llama-WS-BERT*n* setting. While the fine-tuned Llama2 weak supervision (Llama-SFT*n*-WS-BERT*n*) showed higher performance consistently (Figure S1).

Discussion

We proposed an LLM-powered weak supervision system that costs minimal to zero domain knowledge to improve the performance of clinical information extraction by 4.7% to 47.9% from the BERT baseline when no more than 10 gold standard notes were used for training. When 50 gold standard notes were used, our system achieved similar performance as a fully supervised BERT with a 2.2% to 6.1% difference. The method showed an overall benefit of fine-tuning low training sizes across the three benchmarks. Considering the computational burden of fine-tuning LLMs, we also proposed a compact version using Llama2 out-of-the-box and achieved improved performances across the board. The products of our methods are fine-tuned BERT models with 110 million parameters. Compared to modern LLMs which often have billions of parameters, the compact size makes model deployment more computationally efficient. Our framework (i.e., LLM, SFT, prompt templates, and post-processing algorithms) is domain-independent and can be applied to most medical information extraction systems. We expect the performance of this framework will improve further when more medically-focused LLMs become available. We conclude that the proposed method is a generalizable and effortless booster for low-training-size scenarios.

This study is one of the early works exploring the potential use of LLMs in the medical domain. Recent studies have debated the feasibility and performance of in-context learning for information extraction[14,15,17,28]. Following the ideas of LLM-powered labeling functions[29] and clinical knowledge distillation[21], we proposed a robust alternative that combines supervised fine-tuning LLMs, few-shot prompting, and weak supervision to achieve stably dominant performances. As a knowledge-free alternative for labeling functions, our study also points out a direction in which current weak supervision methods could be free from the heavy reliance on domain expert inputs and ontology.

On the 2012 time expression benchmark, when 50 the gold standard notes were used for training, our Llama-SFT*50*-WS-BERT*50* had slightly reduced performances by 1.3% compared to the BERT*50* baseline. This finding is consistent with a recent weak supervision study which showed a negative impact when a large amount of training notes were provided[3]. The most likely explanation is that when gold standards are adequate to provide the model with the correct knowledge, the noise in the weak labels exceeds the benefits. However, the performance drops in such cases with our approach are quite small, suggesting such an approach can have endurance upsides with little chance of catastrophic loss, unlike other LLM use cases.

On the 2014 and the 2018 benchmarks, we observed reversed results between the two baselines Llama-SFT*n* and BERT*n* in which Llama-SFT*n* performed better on the 2014 PHI de-identification task while BERT*n* performed better on the 2018 ADE & medication extraction task. One explanation is that since Llama2 is a general-domain model while PubMedBERT is a biomedical model, the former might have advantages in solving non-medical problems such as PHI identification while the latter has advantages in solving medical problems like medication terms.

In a stand-alone sensitivity analysis, we explored the newer and larger version of Llama, Llama3-70B, and demonstrated that though the choice of LLMs plays a role in the performance, adopting our proposed methods gives consistent benefits.

Despite the promising results, this study does have a few limitations. First, unlike other weak supervision studies in which a large number of unlabeled notes were processed by LFs[2,3], for computational considerations, we chose benchmarks with relatively small sample sizes. We would expect that with larger weakly-labeled datasets

the performance of our approach should increase, though this requires further experimentation. However, even with less than 800 notes, the LLM was able to generate weak labels that dramatically improved performance. Second, as an initial work, we did not evaluate many different LLMs. We selected Llama2-13B as the main LLM and explored Llama3-70B on the side based on the reported performances in the medical domain and their open-source and lightweight features[30]. Other open-source LLMs should be evaluated in future studies. Third, to keep the study focused, we did not evaluate different settings in supervised fine-tuning (e.g., prompt templates, learning rate), few-shot prompting (e.g., prompt templates, the number of few-shot examples), post-processing (e.g., label harmonization), and BERT model fine-tuning. We follow reported best practices for those[12,14,18]. We expect the performance to further improve if those details are carefully tuned.

Conclusion

In conclusion, we proposed a novel method that combines LLMs and weak supervision for high-performance medical information extraction while minimizing domain knowledge dependence. Our method shows a consistent benefit. Further performance improvements are anticipated with more refined few-shot prompting and fine-tuning.

Methods

Figure 2 provides an overview of our approach. We first constructed a prompt template with a system prompt, an instruction, few-shot examples sampled from the training set, and an input/output placeholder. For a given set of *n* gold standard notes, we fine-tuned Llama2-13B via prompt-based supervised fine-tuning (SFT). We then used the fine-tuned Llama2 for few-shot prompting on the unannotated notes to generate weak labels. The weak labels were used to fine-tune ("weakly supervise") a BERT model. The BERT model was then fine-tuned with the gold standard notes to achieve optimal performance.

Benchmarks

We used datasets and tasks from 2012[25] and 2014[26] Integrating Biology and the Bedside (i2b2), and 2018 National NLP Clinical Challenges (n2c2)[19] Natural Language Processing Challenge as benchmarks.

The 2012 i2b2 challenge focused on temporal relation extraction with 310 annotated clinical notes. Entities include (1) clinically significant events ("EVENT"), such as problems, tests, treatments, clinical departments, admissions, and transfers between departments, and (2) temporal expressions ("TIMEX3"), which are dates, times, durations, or frequencies phrases. For this study, the F1 scores for events and time expressions are used as the main metrics, while the temporal relations between events and time expressions are not evaluated.

The 2014 i2b2 challenge de-identification track focused on extracting Health Insurance Portability and Accountability Act (HIPAA) protected health information (PHI) from 1304 annotated clinical notes. We used the i2b2-PHI entities which include 7 types of PHI. We used strict and relaxed micro F1 scores as the main metrics.

The 2018 n2c2 challenge track 2 focused on the extraction of adverse drug events (ADEs) and medication properties from 505 discharge notes. The concept extraction task defined 9 entity types: drug, strength, form, dosage, frequency, route, duration, reason, and ADE. We used the Strict and Lenient micro F1 as the main metrics.

Prompt templates

We prepared a prompt template for each benchmark task as highlighted in Fig. 2 and listed in Table S1. Our design adopted recent studies in prompt engineering[14,21] which includes 4 sections: (1) system prompt, in which a role is assigned to Llama2 to provide the context and to avoid triggering the safety features of the LLM, (2) instruction, which is a narrative description of the background (e.g., medical notes), task (e.g., named entity recognition, entity types), and expected output (i.e., the entity text and the entity type), (3) few-shot examples, where 8 randomly sampled sentences and the corresponding gold standard labels were listed following the JavaScript Object Notation (JSON) format. (4) Input placeholders, where for each sentence the text was placed in the input placeholder while the prompt was fed to LLM. The LLM would output text following the "[/INST]" special token which we collect for post-processing.

Supervised fine-tuning (SFT) Llama2

We used the prompt template described in the previous section to perform SFT. Each sentence in the gold standard notes was placed in the input placeholder and fed to the Llama2. Note that SFT is auto-regressive thus the labels were appended following the "[/INST]" special token after the input sentences. Following the original SFT hyperparameters[18], we use a cosine learning rate schedule with a 2×10^{-5} initial learning rate and a weight decay of 0.1. The sequence length was 4096. We trained for 2 epochs. Due to the limiting GPU memory, we set the batch size to 1.

Few-shot in-context learning

LLMs have limitations in the number of input tokens due to their transformer architecture. Llama2-13B has a limit of 4096 input tokens. Including entire clinical notes in a prompt would often exceed the maximum input length. Therefore, we performed few-shot prompting at the sentence level. We sentence-segmented each note with spaCy 3.5.4 Sentencizer[31]. Sentences were placed in the input placeholder and the output was collected after the "[/INST]" special token for post-processing. To maximize the reproducibility, we set the top-k parameter to 1 which disabled random sampling of generated tokens. To increase the text generation speed, we set the maximum output length to 128 tokens.

Large language models

Clinical notes often include PHI and are restricted from sharing. LLMs that are only available through API (e.g., GPT-3, GPT-4)[17,32] could be limiting in real-world scenarios. An ideal LLM for our system meets the three criteria: (1) is open-source and can be deployed locally, (2) is lightweight enough for making inferences on a local server, and (3) has high performance in the medical domain. Llama2 is a pre-trained open-source large language model that comes with different sizes of architecture from 7 to 65 billion parameters and has demonstrated competitive performances in both open-domain and biomedical NLP benchmarks[30]. The 7 billion parameter version ("Llama2-7B") loaded in 16-bit floating-point can fit in a GPU with 14 GB of vRAM, while the 13 billion parameter version ("Llama2-13B") fits in 26 GB of vRAM. We chose Llama2-13B for a balance of performance and computation cost. As a stand-alone sensitivity analysis, we also evaluated a recently published large language model, Llama3, with 70 billion parameters[27].

Post-processing

To serve the purpose of minimizing human effort, our post-processing was designed to be automatic, robust, and generalizable across tasks. The steps were: (1) generated-text extraction, which extracts all generated text after the "[/INST]" special token. In cases where excessive text was generated after the intended JSON format, for instance, a new "[INST]" was generated by Llama2, we truncated it. (2) JSON formatting, which is a simple regular expression logic that extracts the "\{.*?\}" patterns in a JSON list. (3) Entity recovery, which utilized the extracted entity text to identify the span in the input sentence. (4) Entity type filtering, which filters out irrelevant entity types that Llama2 created and are not one of the entity types for the benchmark tasks. We used exact, case-sensitive string matching to minimize potential bias from human interpretation. By the end of post-processing, we obtained a list of entities with the span, entity text, and entity type for each clinical note.

Weak supervision

We used one of the latest state-of-the-art biomedical BERT models, PubMedBERT[33] (denoted as BERT) in this study. To evaluate the scenario where only a few annotated notes are available for training, the BERT model was first fine-tuned with weak labels from *(N-n_s)* notes followed by fine-tuning with gold labels from n_s notes, where N is the total number of training notes, $n_s \in \{3, 5, 10, 50\}$. To ensure the n_s notes were representative, instead of random sampling, they were selected such as having the closest number of entities to the median number of entities among all notes in the official training set. The formula below defines the selected subset S_{n_s}:

$$S_{n_s} = \{note_i : i \; in \; top \; n_s \; argmin\big(abs\big(\#\; of \; entities \; in \; note_i - median \; \#\; of \; entities\big)\big)\}$$

Fine-tuning BERT

Fine-tuning with weak labels and gold standard data follows similar methods, with a few differences in hyperparameters (Table S2). To segment notes into shorter chunks that the BERT models could process, we sentence-segmented the notes with spaCy. For each sentence, word tokenization was performed using the WordPiece algorithm implemented in the Python *transformers* module (version 4.30.2) and based on a pre-defined dictionary. We set an input length of 256 tokens. Sentences longer than 256 tokens were truncated while sentences shorter than 256 tokens were zero-padded.

For fine-tuning, the development set was divided into a training set (80%) and a validation set (20%), unless specified in Table S2. Model weights were saved as checkpoints after each training period ("epoch"), and optimal checkpoint weights were selected during validation as our final NLP model. For efficiency, an early stop criterion of 8 continuous non-improving epochs was used. The NLP models were implemented using Python 3.9.7, *PyTorch* 2.0.1, and *transformers* 4.30.2. All computations were performed on a server with 8 NVIDIA A100 80GB GPU.

Benchmarking FLOPs and GPU time

The corpus in the 2018 benchmark is a subset of 505 discharge summaries from the MIMIC-III[20] database. We calculated the FLOPs for inferencing one input sentence with Llama2-13B following the formula below[34],

$$\begin{aligned} N_{tokens}\left(2N + 2n_{layer}n_{ctx}d_{attn}\right) &= 128 \times (2 \times (13015864320) + 2 \times 40 \times 400 \times 4096) \\ &\cong 3.348 \times 10^{12} \end{aligned}$$

where N_{tokens} denotes the number of tokens Llama2 outputs; N denotes the total parameters in the model; n_{layer} denotes the number of layers in the model; n_{ctx} denotes the input context token length. We use the length of the prompt template to estimate. d_{attn} denotes the dimension of attention output. We monitored the FLOPs for PubMedBERT with the built-in tool, *profiler* in *PyTorch*. The GPU time for each note was monitored during the inferencing with Llama2-13B and the prediction with PubMedBERT. We randomly sampled 50 to 500 notes and fitted a linear regression line to model the correlation between the number of notes and the total GPU time. A projection was made to estimate the total GPU time required for all the discharge summaries from the MIMIC-III database.

Data availability

The benchmark datasets used in this study are publicly available. Registration is required via the DBMI portal (https://portal.dbmi.hms.harvard.edu/). Once approved, dataset requests can be made through the n2c2 NLP Research Data Sets webpage (https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/).

Received: 10 June 2024; Accepted: 22 July 2024
Published online: 10 March 2025

References

1. Zhang, J., Hsieh, C.-Y., Yu, Y., Zhang, C. & Ratner, A. A Survey on Programmatic Weak Supervision. Preprint at http://arxiv.org/abs/2202.05433 (2022).
2. Dong, H. *et al.* Ontology-driven and weakly supervised rare disease identification from clinical notes. *BMC Med. Inform. Decis. Mak.* **23**, 86 (2023).
3. Datta, S. & Roberts, K. Weakly supervised spatial relation extraction from radiology reports. *JAMIA Open* **6**, ooad027 (2023).
4. Ge, Y., Guo, Y., Das, S., Al-Garadi, M. A. & Sarker, A. Few-shot learning for medical text: A review of advances, trends, and opportunities. *J. Biomed. Inform.* **144**, 104458. https://doi.org/10.1016/j.jbi.2023.104458 (2023).
5. Cusick, M. *et al.* Using weak supervision and deep learning to classify clinical notes for identification of current suicidal ideation. *J. Psychiatr. Res.* **136**, 95–102 (2021).
6. Wang, H. *et al.* A weakly-supervised named entity recognition machine learning approach for emergency medical services clinical audit. *Int. J. Environ. Res. Public Health* **18**, 7776 (2021).
7. Singhal, S., Hegde, B., Karmalkar, P., Muhith, J. & Gurulingappa, H. Weakly supervised learning for categorization of medical inquiries for customer service effectiveness. *Front. Res. Metr. Anal.* **6**, 683400 (2021).
8. Agnikula Kshatriya, B. S. *et al.* Identification of asthma control factor in clinical notes using a hybrid deep learning model. *BMC Med. Inform. Decis. Mak.* **21**, 272 (2021).
9. Shen, Z. *et al.* Classifying the lifestyle status for Alzheimer's disease from clinical notes using deep learning with weak supervision. *BMC Med. Inform. Decis. Mak.* **22**, 88 (2022).
10. Fries, J. A. *et al.* Ontology-driven weak supervision for clinical entity classification in electronic health records. *Nat. Commun.* **12**, 2017 (2021).
11. Dhrangadhariya, A. & Müller, H. Not so weak PICO: Leveraging weak supervision for participants, interventions, and outcomes recognition for systematic review automation. *JAMIA Open* **6**, ooac107 (2023).
12. Brown, T. *et al.* Language models are few-shot learners. *Adv. Neural Inform. Process. Syst.* **33**, 1877–1901 (2020).
13. Dong, Q. *et al.* A Survey on In-context Learning. Preprint at http://arxiv.org/abs/2301.00234 (2023).
14. Agrawal, M., Hegselmann, S., Lang, H., Kim, Y. & Sontag, D. Large Language Models are Few-Shot Clinical Information Extractors. Preprint at https://doi.org/10.48550/arXiv.2205.12689 (2022).
15. Moradi, M., Blagec, K., Haberl, F. & Samwald, M. GPT-3 Models are Poor Few-Shot Learners in the Biomedical Domain. Preprint at https://doi.org/10.48550/arXiv.2109.02555 (2022).
16. Jimenez Gutierrez, B. *et al.* Thinking about GPT-3 In-Context Learning for Biomedical IE? Think Again. in *Findings of the Association for Computational Linguistics: EMNLP 2022* 4497–4512 (Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022). https://doi.org/10.18653/v1/2022.findings-emnlp.329.
17. Liu, Y. *et al.* Summary of ChatGPT/GPT-4 Research and Perspective Towards the Future of Large Language Models. Preprint at http://arxiv.org/abs/2304.01852 (2023).
18. Touvron, H. *et al.* Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv.org* https://arxiv.org/abs/2307.09288v2 (2023).
19. Henry, S., Buchan, K., Filannino, M., Stubbs, A. & Uzuner, O. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *J. Am. Med. Inform. Assoc.* **27**, 3–12 (2019).
20. Johnson, A. E. W. *et al.* MIMIC-III, a freely accessible critical care database. *Sci. Data* **3**, 160035 (2016).
21. Meoni, S., De la Clergerie, E. & Ryffel, T. Large Language Models as Instructors: A Study on Multilingual Clinical Entity Extraction. in *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks* 178–190 (Association for Computational Linguistics, Toronto, Canada, 2023).
22 Frei, J. & Kramer, F. Annotated dataset creation through large language models for non-english medical NLP. *J. Biomed. Inform.* **145**, 104478. https://doi.org/10.1016/j.jbi.2023.104478 (2023).
23. Karkera, N., Acharya, S. & Palaniappan, S. K. Leveraging pre-trained language models for mining microbiome-disease relationships. *BMC Bioinform.* **24**, 290 (2023).
24. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://doi.org/10.48550/arXiv.1810.04805 (2019).
25. Sun, W., Rumshisky, A. & Uzuner, O. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *J. Am. Med. Inform. Assoc.* **20**, 806–813 (2013).
26. Stubbs, A., Kotfila, C. & Uzuner, O. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1. *J. Biomed. Inform.* **58**, S11–S19 (2015).
27. Introducing Meta Llama 3: The most capable openly available LLM to date. *Meta AI* https://ai.meta.com/blog/meta-llama-3/.
28. Gutiérrez, B. J. *et al.* Thinking about GPT-3 In-Context Learning for Biomedical IE? Think Again. Preprint at http://arxiv.org/abs/2203.08410 (2022).
29. Smith, R., Fries, J. A., Hancock, B. & Bach, S. H. Language Models in the Loop: Incorporating Prompting into Weak Supervision. Preprint at http://arxiv.org/abs/2205.02318 (2022).
30. Touvron, H. *et al.* LLaMA: Open and Efficient Foundation Language Models. Preprint at https://doi.org/10.48550/arXiv.2302.13971 (2023).
31. Honnibal, M. & Montani, I. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *To Appear* **7**, 411–420 (2017).
32. OpenAI. GPT-4 Technical Report. Preprint at https://doi.org/10.48550/arXiv.2303.08774 (2023).
33. Gu, Y. *et al.* Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare* **3**, 1–23 (2022).
34. Kaplan, J. *et al.* Scaling Laws for Neural Language Models. Preprint at http://arxiv.org/abs/2001.08361 (2020).

Author contributions

E.H. conceived the research, performed experiments, and wrote the main manuscript text. K.R. helped with study design and manuscript writing. All authors reviewed the manuscript.

Funding

This work was partially supported by awards from the National Institutes of Health, including the National Institute of Biomedical Imaging and Bioengineering (NIBIB: R21EB029575) and the National Institute of Allergy & Infectious Diseases (NIAID: R21AI164100).

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/s41598-024-68168-2.

Correspondence and requests for materials should be addressed to K.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.