Large Language Models for Parsing Clinical Text

Monica Agrawal
MIT + Stealth Startup

Based on *Large Language Models are Few-shot Clinical Information Extractors*

Oral presentation at EMNLP 2022

Stefan Hegselmann

Hunter Lang

Yoon Kim

David Sontag

Clinical information extraction

Electronic health record data could help answer questions in personalized medicine

- What treatment is **best for you?**

Unfortunately, many important variables are **not structured**

- Treatments
- Outcomes
- Side effects

Trapped in messy clinical text: **"Pt will dc carbo ia for TNBC"**

Extraction requires multiple hops of logic

discontinue
~~D/C current~~
~~discharge~~

Intra-arterial
~~Intra-articular~~

"Pt will dc carbo ia for TNBC"

Patient
~~Physical therapy~~
~~Prothrombin time~~

Carboplatin
~~Carbodome~~

Triple-negative breast cancer

Subject: patient

Medication: carboplatin

Reason: triple-negative breast cancer

Status: discontinued

Difficulty of Clinical IE

Labeled data can be prohibitively expensive:

- Not a natural byproduct of clinical care
- Requisite domain expertise
- Difficulty of sharing across institutions

This impedes progress in clinical information extraction

Status Quo in Clinical NLP

1. Define your task
   - *E.g. patient phenotyping: does this patient have condition A?*
2. Create labeled training data
   - *For each input x (e.g. a note), label output y*
3. Train a model to output y, given x
   - *E.g. logistic regression*
4. Use model on new inputs

Huge advances in language modeling

MOTHERBOARD
TECH BY VICE

Students Are Using AI to Write Their Papers, Because Of Course They Are

Essays written by AI language tools like OpenAI's Playground are often hard to tell apart from text written by humans.

By Claire Woodcock

Opinion
Artificial intelligence (AI)

This article is more than **2 years old**

A robot wrote this entire article. Are you scared yet, human?
GPT-3

Tue 8 Sep 2020 04.45 EDT

1188

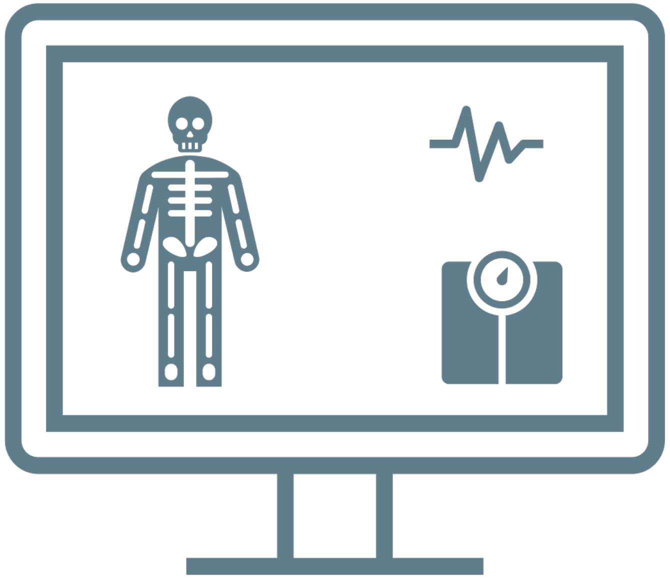
We asked GPT-3, OpenAI's powerful new language generator, to write an essay for us from scratch. The assignment? To convince us robots come in peace

For more about GPT-3 and how this essay was written and edited, please read our editor's note below

The Drum

Ryan Reynolds enlists AI-powered ChatGPT in 'mildly terrifying' new Mint Mobile ad

1 day ago



Background: In-context learning

Traditional Paradigm

The acting

-1

Can this improve clinical data mining?

New Paradig

```
Review: this movie was great.
Positive or Negative? Positive

Review: the acting was subpar.
Positive or Negative?
```


Negative

-1

Brown et al, NeurIPS 2020

Challenge #1: Clinical Text Availability

Due to **patient privacy**, there are likely not significant corpora of clinical notes in the training data

Most existing labeled data sets are under **data use agreements** and can't be sent over APIs

Creation of Benchmark Datasets

We re-annotate the existing publicly available CASI dataset to release **three new** few-shot extraction **datasets**:

- Clinical coreference resolution
- Medication extraction + status classification
- Medication + attribute relation extraction

Each contains 5 examples for development (e.g. prompt design) and 100 examples for test

**Moon et al*, " A sense inventory for clinical abbreviations… "

Challenge #2: Obtaining structured, evidence-backed output

Goal: List medications, and their reason, dosage, and frequency, as available.

Input: "[…] 500mg of metformin b.i.d. […]"

Expected completion: *"Medication: metformin*
Dosage: 500mg
Frequency: b.i.d."

Reality: *"The medication taken is metformin for the reason of diabetes at a dosage of 500mg…"*

Issue #1: Narrative format

Issue #2: Hallucinations

Encouraging structured output

Zero-shot prompt:

Naïve:

```
Input: Pt will dc carbo for TNBC.
Prompt: Label medications. Include dosage, route, …
The medication taken was carbo…
```

Complex post-processing (resolver) of LM output → **“Carbo”: {reason:“TNBC”}**

Encouraging structured output

Our approach:

One-shot example + guidance:

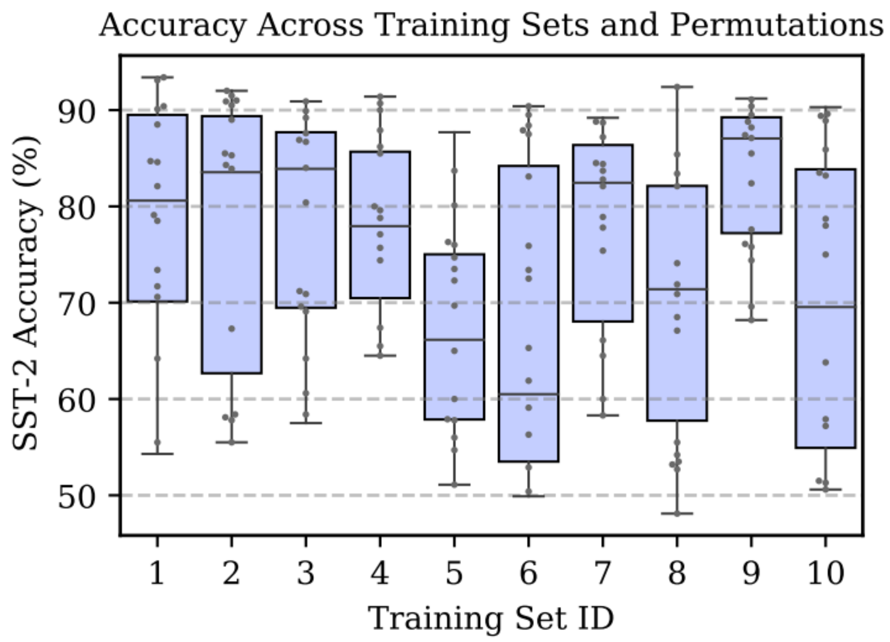
```
Input: He is on 500mg of ibuprofen daily […].
Prompt: Label medications. Include dosage, route, …
-medication: “statin”, dosage: “500mg”, frequency: “daily”
Input: Pt will dc carbo for TNBC […].
Prompt: Label medications. Include dosage, route, …
-medication:“carbo”, reason: “TNBC”
```

Minimal post-processing (resolver) of LM output → **“Carbo”: {reason: “TNBC”}**

Challenge #3: Deployability Concerns

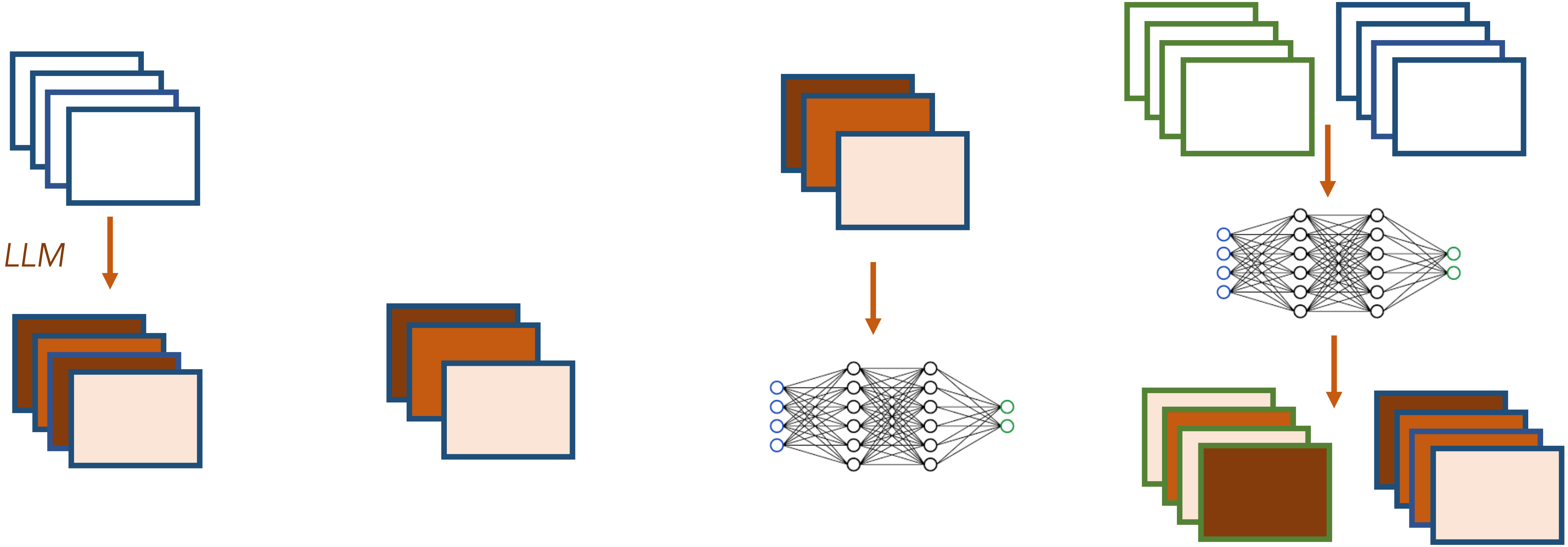
Concerns:

- HIPAA compliance
- Unwieldy size of models
- Sensitivity to wording
- Model miscalibration, when available



Zhao et al, ICML 2021

Weak Supervision + Distillation



Step 1: Get LLM outputs on publicly available data

Step 2: Identify confident outputs* (for classification only, so far)

Step 3: Train smaller model on confident outputs

Step 4: Run smaller model on unlabeled data set

*Lang et al., "Training subset selection for weak supervision" NeurIPS 2022.

Results

OpenAI Engines: *`text-davinci-edit-001`*, *`text-davinci-002`*

Task 1: Zero-shot Clinical Acronym Disambiguation

Input: Clinical Text Snippet + Overloaded Acronym
Output: Multiple-choice Expansion of Acronym

Example: …CARDIAC: Regular rate and rhythm. No murmurs. LUNGS: *CTA*, intubated. ABDOMEN: Obese, nontender, positive bowel sounds…

↓

Clear to auscultation

Task 1: Zero-shot Clinical Acronym Disambiguation

Input: Clinical Text Snippet + Overloaded Acronym
Output: Multiple-choice Expansion of Acronym

Algorithm	CASI Acc.	CASI Macro F1
Random	0.31	0.23
Most Common	0.79	0.28
BERT (from Adams et al. (2020))	0.42	0.23
ELMo (from Adams et al. (2020))	0.55	0.38
LMC (from Adams et al. (2020))	0.71	0.51
GPT-3 edit + R: 0-shot	0.86	0.69
GPT-3 edit + R + weak sup	**0.90**	**0.76**

Zero-shot LM baseline trained on MIMIC data

Task 1: Zero-shot Clinical Acronym Disambiguation

Input: Clinical Text Snippet + Overloaded Acronym
Output: Multiple-choice Expansion of Acronym

Algorithm	CASI Acc.	CASI Macro F1	MIMIC Accuracy	MIMIC Macro F1
Random	0.31	0.23	0.32	0.28
Most Common	0.79	0.28	0.51	0.23
BERT (from Adams et al. (2020))	0.42	0.23	0.40	0.33
ELMo (from Adams et al. (2020))	0.55	0.38	0.58	0.53
LMC (from Adams et al. (2020))	0.71	0.51	0.74	**0.69**
GPT-3 edit + R: 0-shot	0.86	0.69	*	*
GPT-3 edit + R + weak sup	**0.90**	**0.76**	**0.78**	**0.69**

Task 2: Biomedical Evidence Extraction

Input: Clinical trial summary from PICO dataset
Output: List of interventions/arms of trial

Example: …were blood-sampled immediately without anesthesia (control) or subjected to following anesthesia procedure: 40, 120, and 240 s exposure to 3,000, 700, and 500 mg l^{-1} clove solution, respectively. Blood samples were collected…

↓

- Without anesthesia (control)
- 3,000 mg l^{-1} clove solution
- 700 mg l^{-1} clove solution
- 500 mg l^{-1} clove solution

Task 2: Biomedical Evidence Extraction

Input: Clinical trial summary from PICO dataset
Output: List of interventions/arms of trial

On a set of 20 manually scored clinical trials:
GPT-3 scored perfectly: 85%
Supervised PubMedBERT + oracle coreference: 35%

Task 3: Clinical Coreference Resolution

Input: Clinical Text Snippet + Pronoun
Output: Quoted Antecedent of Pronoun

Example: …Her current regimen for her MS is Rebif Monday, Wednesday, and Friday and 1 gram of methylprednisolone p.o. every month. This had been working

↓

"This" refers to "her current regimen for her MS"

Task 3: Clinical Coreference Resolution

Input: Clinical Text Snippet + Pronoun
Output: Quoted Antecedent of Pronoun

Algorithm	**Recall**	**Precision**
Toshniwal et al. (2020, 2021)	0.73	0.60
GPT-3 + R (50 LOC): 0-shot	**0.78**	0.58
GPT-3 + R (1 LOC): 1-shot (incorrect)	$0.76_{.02}$	$\mathbf{0.78}_{.04}$
GPT-3 + R (1 LOC): 1-shot (correct)	$0.75_{.04}$	$0.77_{.04}$

Baseline supervised on non-clinical datasets

Task 4: Medication + status extraction

Input: Clinical text snippet
Output: List of medications + status (active, discontinued, neither)

Example: Assessment and Plan: Therefore, we have recommend Citrucel one tablespoon p.o. q. day and we decided to dc the Colace.

"Citrucel": active
"Colace": discontinued

Task 4: Medication + status extraction

Input: Clinical text snippet
Output: List of medications + status (active, discontinued, neither)

Algorithm	Recall	Precision
ScispaCy (Neumann et al., 2019)	0.73	0.67
GPT-3 + R (32 LOC) (0-Shot)	0.87	0.83
GPT-3 + R (8 LOC) (1-Shot)	**0.90**$_{.01}$	**0.92**$_{.01}$

Task 4: Medication + status extraction

Input: Clinical text snippet
Output: List of medications + status (active, discontinued, neither)

Algorithm	Conditional Accuracy	Conditional Macro F1
T-Few (20-shot)	0.86	0.57
GPT-3 + R (32 LOC) (0-Shot)	0.85	0.69
GPT-3 + R (8 LOC) (1-shot)	**0.89** $_{.01}$	0.62 $_{.04}$
GPT-3 + R (8 LOC) (1-shot) + added classes	0.88 $_{.02}$	**0.71** $_{.03}$
GPT-3 + R (8 LOC) (1-shot) with shuffled classes	0.88 $_{.01}$	0.66 $_{.03}$

Task 5: Medication + attribute relations

Input: Clinical text snippet
Output: Medications, dosage, route, frequency, reason, duration
- Token-level labels
- Phrase-level labels (with chunking)
- Relation extraction setup

Example: "…she was taking 325 mg of aspirin per day for three years for a TIA…"

aspirin: {dose: 325 mg, freq: per day, duration: three years, reason: TIA}

Task 5: Medication + attribute relations

Input: Clinical text snippet
Output: Medications, dosage, route, frequency, reason, duration
- Token-level labels
- Phrase-level labels (with chunking)
- Relation extraction setup

Baseline supervised on different clinical dataset

Subtask	Algorithm	Medication	Dosage	Route	Frequency	Reason	Duration
Token-level	PubMedBERT + CRF (Sup.)	0.82	0.92	0.77	0.76	0.35	**0.57**
	GPT-3 + R: 1-shot	**0.85**	0.92	**0.87**	**0.91**	**0.38**	0.52

What data are LLMs learning from?

We classified sources of colloquial clinical jargon (“fx”, “fracture”) in a subset of Common Crawl data

Source	Median %
Research Articles	16%
Patient Health Resources	15%
Commercial Health	14%
Clinician Forums	13%
Patient Blogs + Forums	6%

43% of mentions for qhs + bedtime

41% of mentions for carbo + carboplatin

Conclusion

Despite not being trained specifically for the clinical domain, LLMs can do quite well at a variety of clinical informatics tasks

However, naïve application of these methods is insufficient:

- *Guiding* the structure of generation is key for structured data output, but correct examples aren't always necessary

- For classification tasks, weak supervision and distillation can improve performance and enable transfer to private data