



WHITEPAPER

Truveta Language Model

SPRING 2023



Introduction

Truveta connects health systems and life sciences to advance the mission of saving lives with data. Offering the most timely, complete, and cleanest data on US health combined with AI-driven analytics, Truveta accelerates insights in patient care for every drug, disease, or device. Truveta is used to monitor safety, study comparative effectiveness, find clinical trial participants, study care quality and health equity, and many more use cases consistent with [our ethics policy](#).

This whitepaper introduces the Truveta Language Model (TLM), a large-language, multi-modal AI model used to clean billions of daily Electronic Health Record (EHR) data points for health research. TLM's healthcare expertise is trained on the largest collection of complete medical records representing the full diversity of the United States. It is the first large-language model specifically designed to empower researchers to study all patient care and outcomes.

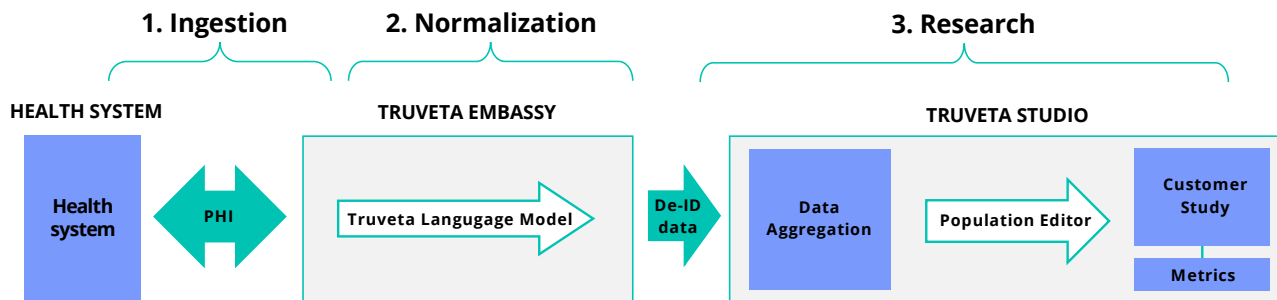
*Truveta Data is updated daily from 28 health system members, who provide **over 16% of patient care in the United States** in more than 20,000 clinics and 700 hospitals from all 50 states.*

Truveta Data provides unparalleled breadth and depth, including EHR with full diagnoses, vital signs, lab tests, clinical notes, and images. Truveta Data is updated daily from 28 health system members, who provide over 16% of patient care in the United States in more than 20,000 clinics and 700 hospitals from all 50 states. This data is linked across health systems and augmented with social drivers of health (SDOH), mortality, and claims data for a complete view of patient journeys. This data is normalized, de-identified, and made available for research daily, enabling researchers to discover insights on yesterday's care, today.

TLM is an intrinsic part of our overall Truveta Data quality process, which you can read more about [here](#).

The Truveta Data Pipeline

First, a reminder of how the Truveta Data pipeline works.



- **Ingestion**

Data from health system members is sent to their own secure, cloud-based environment, called a Truveta Embassy. This data includes raw medical records recorded by busy clinicians and many different machines. This data is messy, hard to analyze, and generally not useful for research.

- **Normalization**

Once medical records are secure in an Embassy, these records are normalized for consistency and measured for data quality. This stage is where TLM transforms the data. Records are then de-identified to maintain patient privacy.

- **Research**

Records from each Embassy are sent to Truveta Studio, where they are aggregated with data from other health system members. Researchers use Studio to precisely define populations of interest and to conduct fast AI-driven analytics.

Truveta Language Model

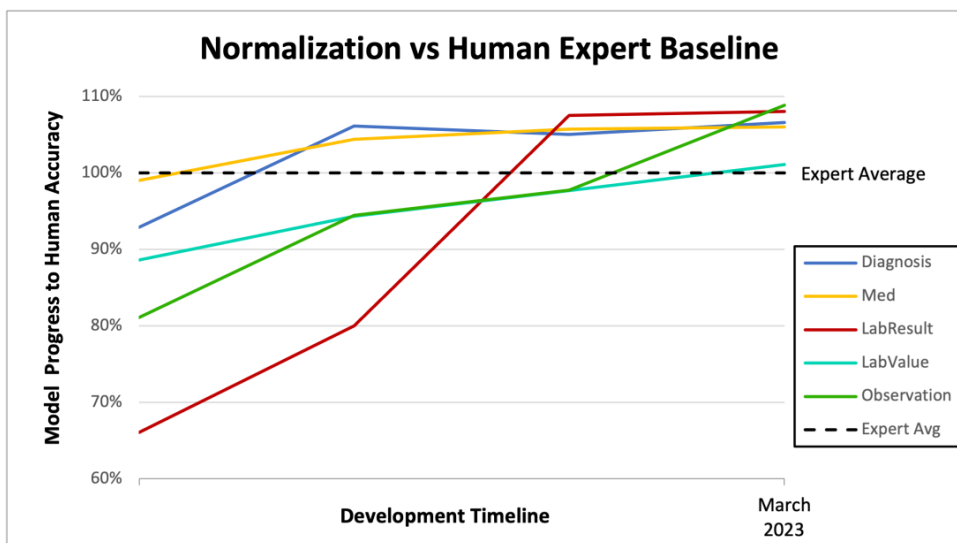
As healthcare considers the potential of AI and real-world data, the opportunities and potential consequences are real. General large language models understand language but are inaccurate within the medical domain due to being trained on the public Internet, which contains no real medical records. In contrast, TLM combines pre-trained open large language models with additional training on the most complete and representative clinical data set to achieve above 90% accuracy on diagnoses, medications, lab results, lab values, clinical observations, and more. TLM's healthcare expertise is trained on Truveta Data, the largest collection of complete medical records, representing the full diversity of the United States. TLM also builds on top of general large language models to understand clinician's notes.

While claims data are the standard of data used in health research today, they are created by normalizing EHR data to maximize revenue reimbursement for encounters, medications, and labs, resulting in commercial bias in all claims data-based health research. Instead, TLM normalizes EHR data to maximize clinical accuracy and is trained without commercial bias, helping ensure research is conducted with data focused on clinical outcomes, not billing.

*TLM's healthcare expertise is trained on Truveta Data, **the largest collection of complete medical records, representing the full diversity of the United States.***

Training Truveta Language Model

The goal of TLM is to exceed the accuracy of clinical experts reviewing medical records. When greater accuracy than clinical experts in a particular healthcare domain is achieved, we deploy the model into Truveta Embassies to start normalizing data. TLM is currently achieving more than 92% accuracy on diagnoses, medications, lab results, lab values, clinical observations, and more.

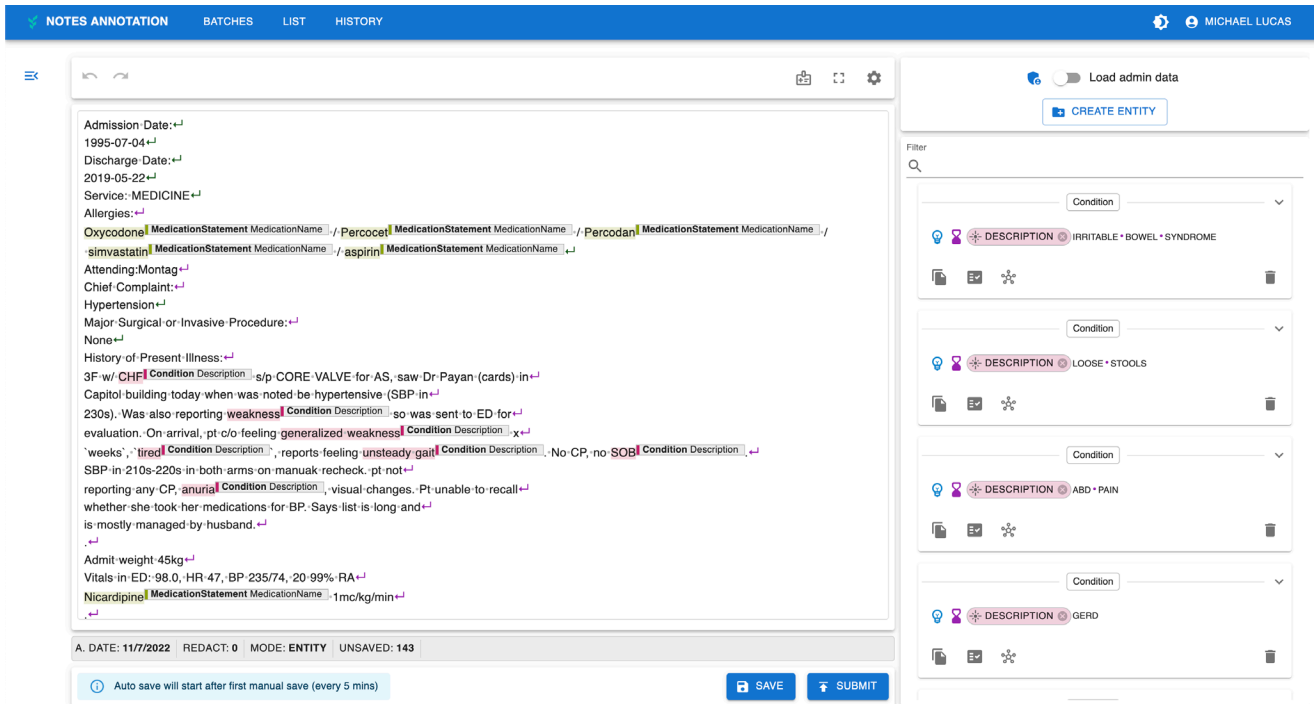


This unprecedented accuracy is due to the ongoing training TLM receives. TLM is trained upon data from Truveta's health system members, currently representing more than 80 million patient journeys, including 5.5 billion diagnoses, 3.1 billion encounters, and 2.4 billion medication orders. Truveta Data combines this EHR data with insurance claims, mortality, and social drivers of health data, for unmatched breadth and depth of data for research. Truveta Data is updated daily and growing every day.

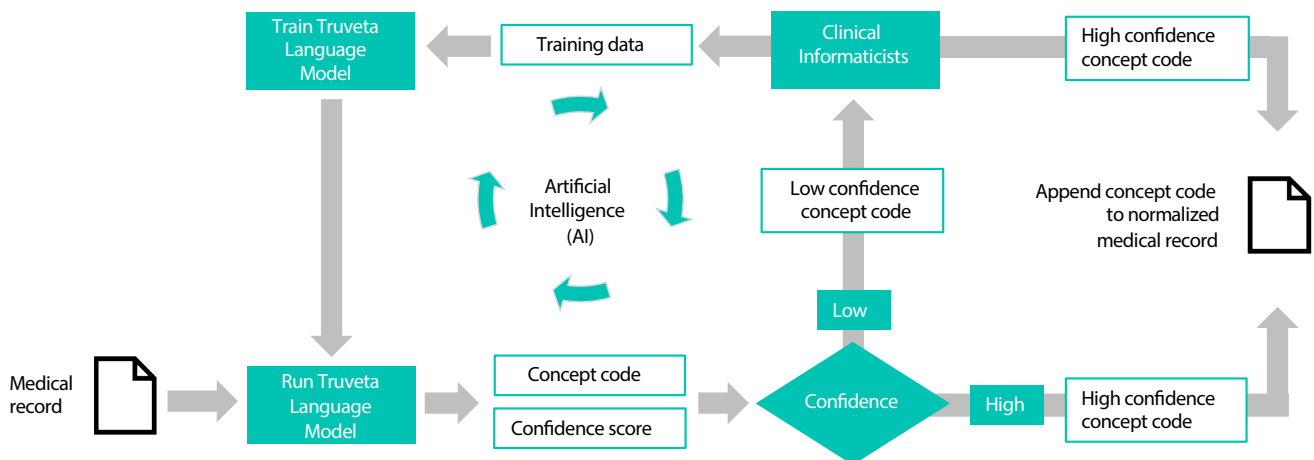
Using this unprecedented data, Truveta's clinical expert annotation team labels thousands of raw clinical terms, including misspellings and abbreviations, to train TLM to normalize healthcare data for clinical research. Note that this labeling is 100% focused on clinical accuracy with no commercial bias. With different types of data, TLM learns how to normalize raw medical text to the most appropriate medical information ontology:

Concept Type	Ontology
Diagnosis and Observations	SNOMED
Lab Tests	LOINC
Lab Test Results Units of Measure	UCUM
Drugs	RxNorm
Devices	GUDID
Immunizations	CVX
Genomics	HGNC

TLM's training is done with a tool custom-designed to train the AI for clinical accuracy.



The results of the model are continuously checked as it runs through a workflow that involves dozens of clinicians continuously refining the accuracy of concepts being actively researched.



Cleaning the data

Raw medical records are recorded in heterogeneous systems with billions of different ways clinicians, hospitals, and health systems express observations, diagnoses, medication plans, and more. Clinicians use different terms based on their location, training, and expertise. “Acute COVID-19,” “COVID,” “COVID-19,” “COVID infection,” and “COVID19 _ acute infection” (and hundreds of other variations) all refer to COVID-19, and “600mg Ibuprofen” and “Ibuprofen 600mg” are the same thing. Before TLM, this unstructured data presented a very expensive data cleaning challenge for analytics.

TLM accurately, without commercial bias, cleans the complete EHR medical record for analytics. For example, consider two blood lab tests which TLM structures into four rows of the LabResults table within our Truveta Data Model. Each test is mapped to a standard medical ontology with standard units for the measurements.

TLM accurately, without commercial bias, cleans the complete EHR medical record for analytics.

Raw medical record text	Structured TDM Label Results table		
	Lab Name (LOINC)	Unit (UCUM)	Value
RBC COUNT,RBC CBC WITH AUTOMATED DIFF 3.80 M/uL 2.70 4.90	789-8	10*6/uL	3.80
CBC: 3/9 07:45PM WBC-8.1 RBC-3.89 Hgb-11.7	6690-2	10*3/uL	8.1
	789-8	10*6/uL	3.89
	718-7	g/dL	11.7

Unlocking the depth of clinician notes

Clinician notes hold critical information about the patient journey, such as disease stages, adverse events, medication change rationales, and disease symptoms. These concepts are not found in claims data sets, and not found in most structured EHR analytics data. For example, a claims dataset would include a medication and later a diagnosis of a rash, but the clinician note is the only place where those two concepts are connected, showing the rash as an adverse reaction to the medication.

To unlock these concepts for analytics, TLM combines general large language models that understand English with rich medical expertise to structure concepts from clinician notes. Truveta Data today includes more than 2.5 billion notes and is growing daily.

Respiratory failure, acute (not ARDS)

Assessment:

Resp status stable on PSV 18/5, no changes. Sats 93-99, RR 22-35. Continues to cough freq. asks to be sx freq, but less than previously. Exp wheezing t/o lung fields consistently all noc, not improving after MDIs. Pt c/o pain in throat and sore face from ETT.

Action:

Sx pt q1-3hrs for minimal thin secretions. Lidocaine 2mls down ETT q4hrs. Asked pt to take Ativan and/or Morphine to help symptoms but pt took only 1mg ativan this 12 hr shift. ETT tube retaped but not rotated. pt does not like tube over on L side. Mouth examined and intact.

Progress Notes

IN-111 WHITE BLOOD CELL STUDY

...

RADIOPHARMACEUTICAL DATA:

480.0 uCi In-111 WBCs ({date}):

HISTORY: Patient with coronary artery disease post STEMI with mental status change. Assess for occult infection.

INTERPRETATION: Following the injection of autologous white blood cells labeled with In-111, images of the whole body were obtained at 24 hours.

These images show physiologic distribution of labelled white cells in the liver, spleen, and bone marrow. There are no abnormal foci of tracer to suggest occult infection. Note is made of a right below the knee amputation.

IMPRESSION: No evidence of occult infection. Normal WBC study.

...

Lab/Study Results

Discharge Plan:

1. Follow up with Dr. {practitioner_name} in a couple of weeks.
2. Follow up with pcp in one week.

Medications on Admission:

diovan 160mg po daily
HCTZ 25mg po daily
terazosin 5mg po daily
metoprolol XL 50mg po daily
ibuprofen PRN

Discharge Medications:

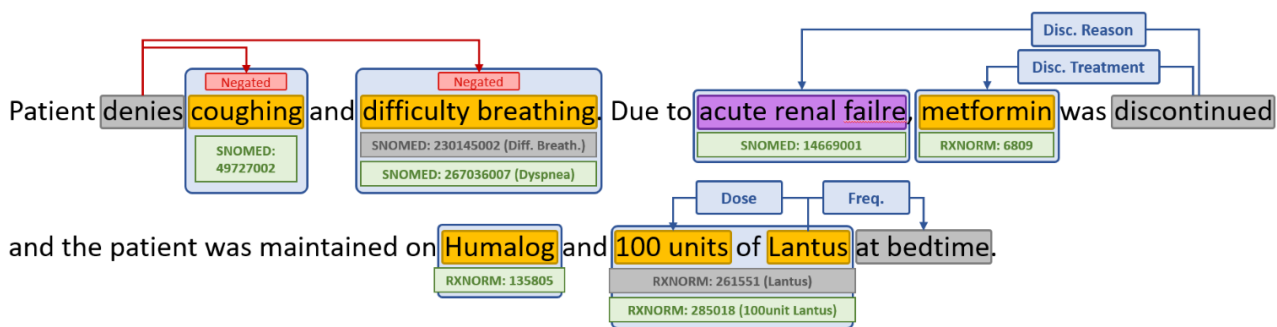
1. Metoprolol Succinate 50 mg Tablet Sustained Release 24 hr Sig: One (1) Tablet Sustained Release 24 hr PO DAILY (Daily).
2. Terazosin 5 mg Capsule Sig: One (1) Capsule PO HS (at bedtime).
3. Oxycodone-Acetaminophen 5-325 mg Tablet Sig: One (1) Tablet PO Q4H (every 4 hours) as needed for PAIN.
Disp:*30 Tablet(s)* Refills:*0*
4. Docusate Sodium 100 mg Capsule Sig: One (1) Capsule PO BID (2 times a day).
Disp:*60 Capsule(s)* Refills:*0*
5. Cipro 500 mg Tablet Sig: One (1) Tablet PO twice a day: for one week.
Disp:*14 Tablet(s)* Refills:*0*
6. Flagyl 500 mg Tablet Sig: One (1) Tablet PO three times a day: for one week.
Disp:*21 Tablet(s)* Refills:*0*
7. Hydrochlorothiazide 25 mg Tablet Sig: One (1) Tablet PO once a day.
8. Diovan 160 mg Tablet Sig: 1/8 Tablet PO once a day.

Discharge Notes

I called the patient's husband, Dr. {practitioner_name}, to let him know the preliminary findings on the CT Scan, which were concerning for pneumatisis and possible mesenteric ischemia. He asked that he be called if a decision for surgery were to be made. He can be reached at {phone_number}.

Telephone Encounters

TLM identifies and normalizes clinical concepts within clinician notes, while accounting for typos/misspellings (e.g., “glipizide” vs “glipizide”) and clinical nuances such as negation (e.g., “patient denies feeling fatigued”), hypotheticals/conditionals (e.g., “Will consider starting low-dose glypizide if A1C still grossly elevated”), and family history (e.g., “Family Hx: Mother: Diabetes, Father/son: bipolar disorder”). TLM reasons over the entire medical record, accounting for changes over time to ensure the most accurate and complete information is structured.



State of the art AI

This massive transformation of healthcare data is only possible with state-of-the-art AI. Truveta's team of technologists and clinical experts have decades of experience and are at the cutting edge of using AI to accurately make health data useful for research.

TLM treats cleaning the data as a translation task. Ontologies are represented as graphs, and the translation is performed from a node in the source ontology graph to a path in the target ontology graph. TLM leverages a multi-task sequence-to-sequence transformer model to perform alignment across multiple ontologies in a zero-shot, unified, and end-to-end manner. Multi-tasking enables the model to implicitly learn the relationship between different ontologies via transfer-learning without requiring any explicit cross-ontology manually labeled data. This also enables the formulated framework to outperform existing solutions for both runtime latency and alignment quality.

TLM outperforms state-of-the-art approaches, including GPT-4, EditSimilarity, LogMap, AML, BERTMap, and the recently presented new OM frameworks in Ontology Alignment Evaluation Initiative, offering log-linear complexity, in contrast to quadratic, in the existing end-to-end methods. This makes the ontology matching task efficient and more straightforward without much post-processing involving mapping extension or mapping repair. Read more about the underlying AI [here](#).

Conclusion

TLM is a profound innovation for making healthcare data trustworthy and useful for analytics. With TLM, Truveta's community of healthcare and life sciences are studying complete, timely, and clean data to achieve our mission of Saving Lives with Data. We look forward to researchers finding cures faster, empowering every clinician to build expertise, and families being able to make more informed decisions about their care.

We also look forward to more industry models being built that compose well with foundation large language models to achieve the full potential of AI to improve human productivity. We have exciting years ahead.



Truveta connects healthcare and life sciences to advance the mission of saving lives with data. Truveta offers the world's first health data and analytics solution to study patient care and outcomes. To learn more, please follow us on LinkedIn and visit truveta.com.

[Truveta.com](https://truveta.com) | info@truveta.com