Clinical trial-specific LLM to auto-generate Protocols and SAPs.

TrialTwin™
Accelerating Clinical Trials

José C. Lacal, CTO

Jan Philip Göpfert, PhD

Abstract

TrialTwin is building a software platform with Natural Language Generation (“NLG”) capabilities using Natural Language Processing (“NLP”) and other software-driven linguistic processors. Our platform will programmatically extract (and encode) both meaning and context from massive amounts of Open Data into a Domain-specific Large Language Model (“LLM”). Our LLM will then be able to programmatically generate highly-realistic and domain-specific new content.

Build your own trial-specific LLM, here are the pieces.

Agenda

- Opportunities through AI
- Using Open Data to Build a RAG
- RAG Development in your Local Computer
- Trial-specific LLM
- Lessons Learned
- Free Resources: Build your own “Trial LLM”

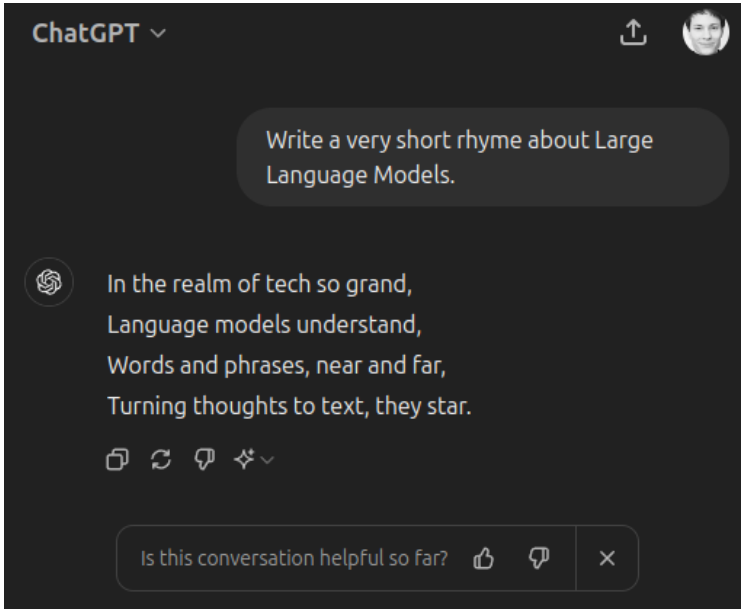
Opportunities through AI

What is?

- GenAI (Generative Artificial Intelligence)
- LLM (Large Language Model)
- Foundation Models

Opportunities through AI – Generative AI

by OpenAI



Opportunities through AI – Large Language Models

TrialTwin™ Accelerating Clinical Trials

Training large models on several modalities yields foundational understanding.

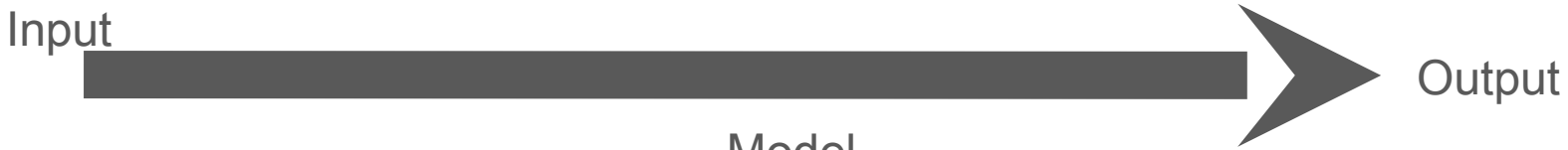
Comprehensive capabilities are more important than **generative** capabilities.

Semantics are more important than **language**.

Opportunities through AI – Large Language Models

Concept of a Large (Language) Model

- contains all prompts and information
- must fit in the “context size” (input window)



- has a certain size: bigger is better
- contains fixed foundational knowledge

Opportunities through AI – Foundation Models

TrialTwin™ Accelerating Clinical Trials

The three domains of large foundation models:

Scientific Background

Training algorithms

Model architectures

Evaluation procedures

Tokenization

Latent representations

Self-attention

…

Model Engineering

Fine-tuning

Distillation

Prompting

Exploration

Evaluation

…

Deployment Engineering

APIs

Vector stores

Streaming data

User interfaces

…

Opportunities through AI – Important Foundation Models

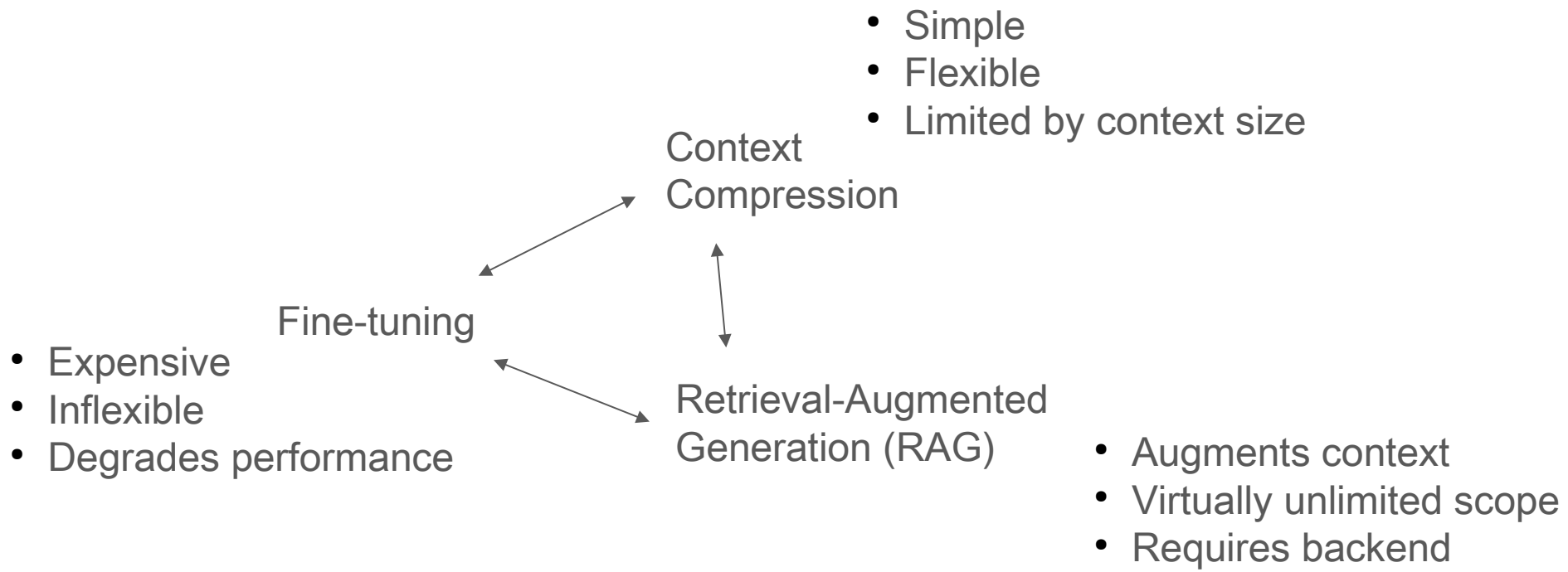
GTP (OpenAI)
cloud-only

Gemini (Google)
cloud-only

LLaMa (Meta)
local deployment

Claude (Anthropic)
cloud-only

Opportunities through AI – Customizing Foundation Models



Using Open Data to Build a RAG (Retrieval-Augmented Generation)

TrialTwin™ Accelerating Clinical Trials

- LLMs are a phenomenal piece of technology for knowledge generation and reasoning. They are pre-trained on large amounts of ***generic data***.
- How to best augment LLMs with both ***private*** and ***domain-specific*** data?
- Build your RAG

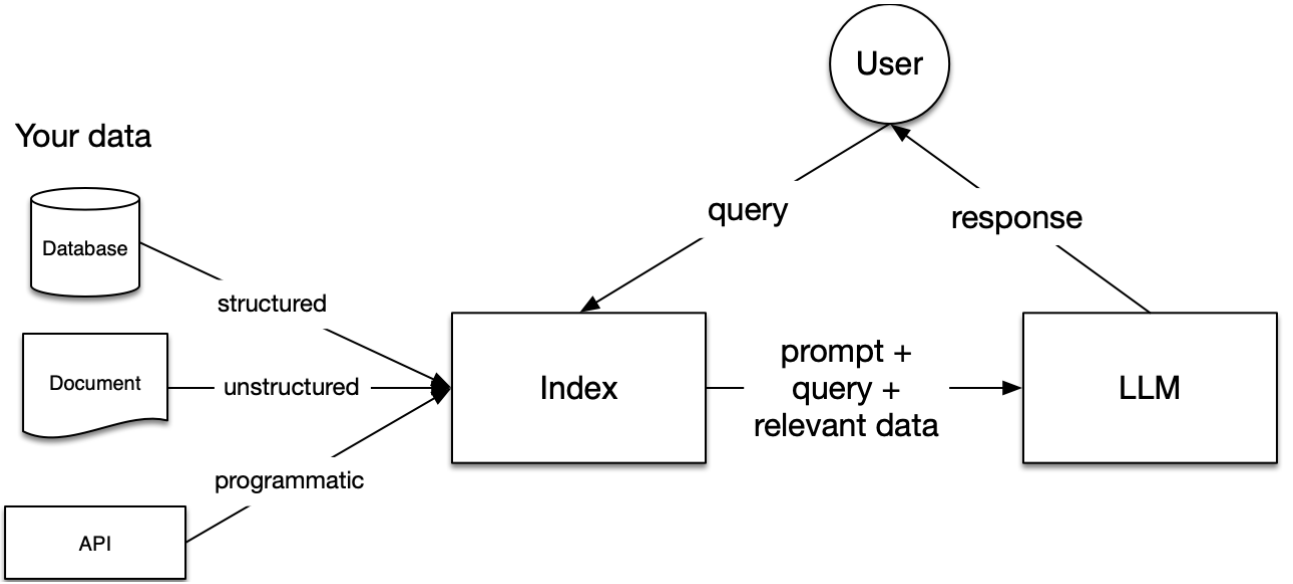
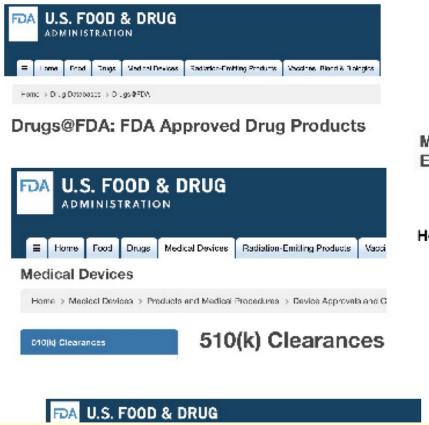


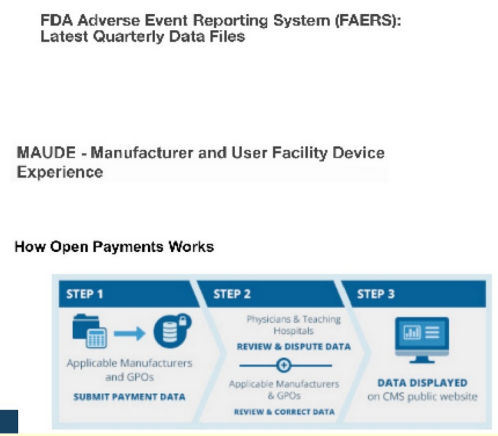
Diagram source: https://docs.llamaindex.ai/en/stable/getting_started/concepts/

Using Open Data to Build a RAG

TrialTwin™ Accelerating Clinical Trials

Literature | Clinical Trials | Regulatory Approval | Adverse Events | Vendor Payments | Medicare Payments | Federal Payments





Chemical Compound

PubChem

Using Open Data allows users to trace a drug's entire lifecycle:

* Starting with chemical compounds (NLM's PubChem)
* Through clinical trials (ClinicalTrials.gov, WHO's ITPR)
* Documentation on regulatory pathway (IND, NDA, etc.)
* Reported adverse events (FDA's FAERS)
* Manufacturer payments to providers (HHS' OpenPayments)
* Medicare reimbursement data (CMS' Provider Utilization and Payment Data)

RAG Development in your Local Computer

- All required software is free (Open Source)
- MS Windows is not a good platform
- Even old hardware can be both useful and usable
- Save time, money by learning locally
- Get a “feel” for the process (literally)

RAG Development in your Local Computer

TrialTwin™ Accelerating Clinical Trials

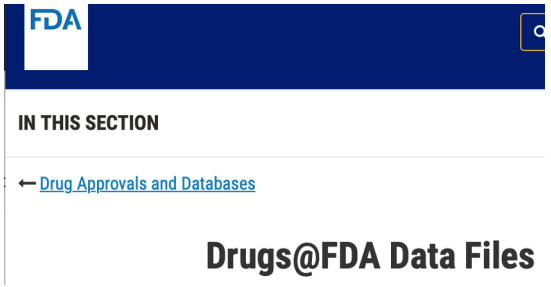
- Install Ollama [https://ollama.com/]
- Install Llama-Index libraries [https://docs.llamaindex.ai/en/stable/]
- Download most efficient models [https://ollama.com/library]
- Download from [https://github.com/JLacal/local-llm]
  - Working Python code
  - Trial-specific Open Data
- Play

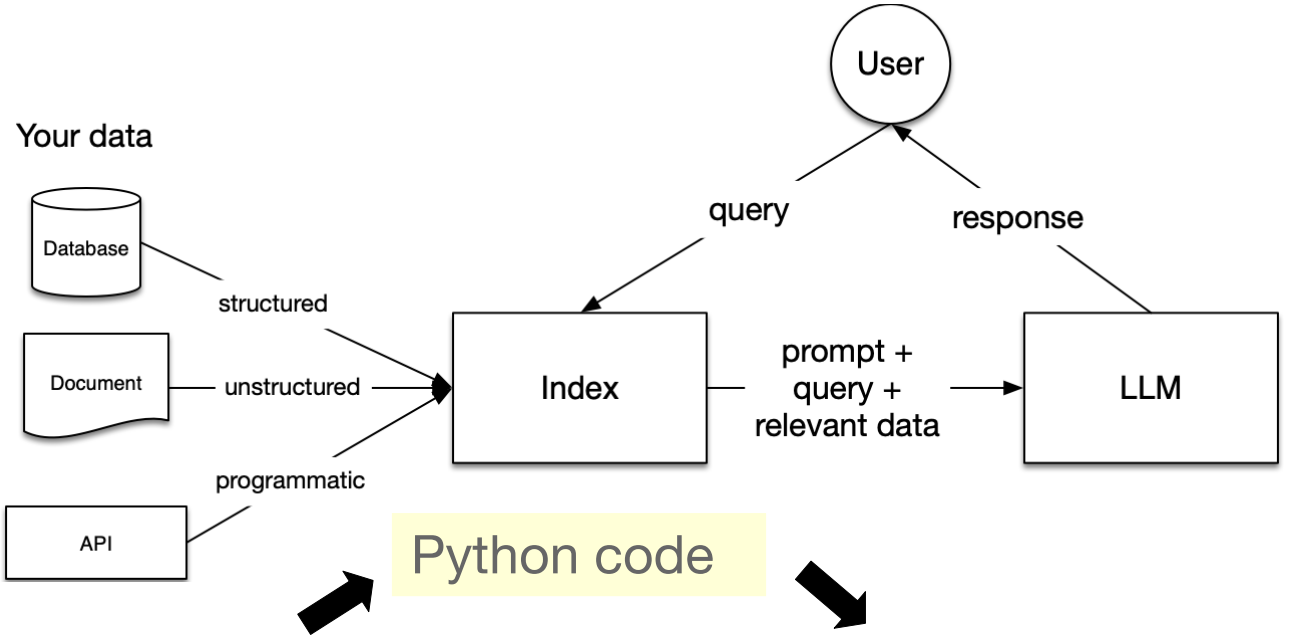
RAG Development in your Local Computer

TrialTwin™ Accelerating Clinical Trials

NIH National Library of Medicine
National Center for Biotechnology Information

ClinicalTrials.gov





ClinicalTrials.gov:
Text of Protocols, SAPs and ICFs from PDFs

Drugs@FDA:
Text application PDFs

→ PDF files
SQLite3 database files

Python code

Ollama running locally
Open Source LLMs

Diagram source: https://docs.llamaindex.ai/en/stable/getting_started/concepts/

Lessons Learned: LLMs Lie

TrialTwin™ Accelerating Clinical Trials

Timeout	codegemma	command-r	command-r-plus	duckdb-nsql	gemma:2b
300 ms	Lie	Timed out	Timed out	Answered	Lie
600 ms	Lie	Timed out	Timed out	Answered	Lie

Timeout	gemma:7b	llama-pro	llama3	llama3-chatqa	llama3-gradient
300 ms	Lie	Good	Great	Empty	Good
600 ms	Lie	Good	Great	Empty	Great

Timeout	llama3:70b	llama3:70b-text	llama3:text	meditron	medllama2
300 ms	Timed out	Timed out	Timed out	Lie	Good
600 ms	Timed out	Timed out	Timed out	Lie	Good

Timeout	phi3:medium	phi3:mini	tinyllama	wizardlm2:7b	wizardlm2:8x22b
300 ms	Timed out	Good	Great	Great	Timed out
600 ms	Good	Good	Great	Great	Timed out

Lessons Learned: LLMs Lie, Confidently

TrialTwin™ Accelerating Clinical Trials

```
= = = = = Using model [codegemma] = = = = =

The provided context does not contain any information about the NCT number
for fibrosis, so I cannot answer this question from the given context.
The provided context does not contain information about NCT 03656068, so I
cannot answer this query from the given context.
200.08059120178223 seconds ---
```

```
= = = = = Using model [medllama2] = = = = =

Based on the provided context, we cannot directly answer your question as
it would violate HIPAA's Privacy Rule by disclosing protected health
information (PHI). Therefore, I will not provide any personal information
related to patients' health status. It is important to respect individuals'
privacy and maintain confidentiality when handling sensitive information
like medical records.
```

Lessons Learned: SMEs need to Evaluate LLMs

TrialTwin™ Accelerating Clinical Trials

```
= = = = = Using model [wizardlm2:7b] = = = = =

 The NCT number associated with the study mentioned in the provided
Statistical Analysis Plan (SAP) for evaluating the safety and efficacy of
NTZ at 500mg twice daily on collagen turnover in plasma in NASH patients
with fibrosis Stage 2 or 3 is NCT03656068. This study is investigating
various non-invasive markers of fibrosis, including biomarkers and fibrosis
scores like ELF test score, PIIINP, Hyaluronic acid, CK18, TIMP-1, YKL-40,
Alpha 2 macroglobulin, miRNA, ProC3, ProC6, FGF19, FGF21, NAFLD fibrosis
score, and FIB-4, among others.
 NCT 03656068 is a clinical trial with the protocol number NTZ-218-1,
version 1.1, dated April 9, 2020. The statistical analysis plan outlines
the study's overview, objectives, endpoints, design, drug information,
sample size determination, and statistical methodology, among other
aspects. The study involves various populations for analysis, including the
Full Analysis Set and the Safety Analysis Set, and includes assessments of
efficacy (such as Fractional Synthesis Rate, Non-invasive Markers of
Fibrosis, and Fibrosis Scores) and safety (including Adverse Events,
Clinical Laboratory Tests, Vital Signs, Electrocardiograms, and Physical
Examinations). The study also details how electronic medical records (EMRs)
and a clinical trial management system (CTMS) will be used to manage
participant data within the University of Pennsylvania Health System. The
purpose of collecting and managing this information is to conduct the
research, oversee it, and evaluate whether it was executed correctly.
```

Free Resources: Build your own “Trial LLM”

TrialTwin™ Accelerating Clinical Trials

Build your own trial-specific LLM, here are the pieces:
https://github.com/JLacal/local-llm

- Database files with full text of Protocols, SAP, and ICFs available through ClinicalTrials.gov

Abbott	Johnson & Johnson
Abbvie	Pfizer
AstraZeneca	Roche
Bayer	Sanofi
Bristol Myers Squibb	

- Pre-computed index file for FDA ApplicationDocs
- Sample Python code to utilize all those data files
- This presentation

Contact us

José C. Lacal, CTO
Data Santander, SL

Hello@TrialTwin.com
{EU} +34 (674) 88 17 52
{US} +1 (561) 777-2577

Jan Philip Göpfert, PhD
Independent researcher

JanPhilip@Gopfert.eu