”Medical Trends Through Entity Recognition & Synthetic Datasets”

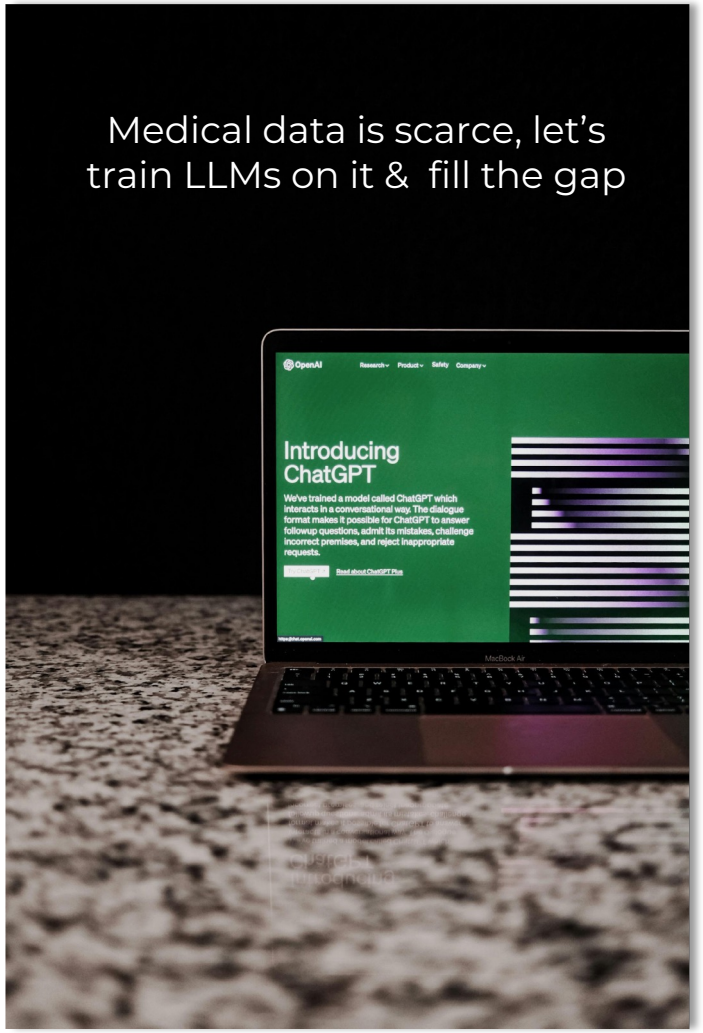
By: Ignacio Alonso, Morten Aas-Lyngby, Harel Ben David, Kangjie Yu, Andreu Artigues and Guillermo Brun

Made at ie

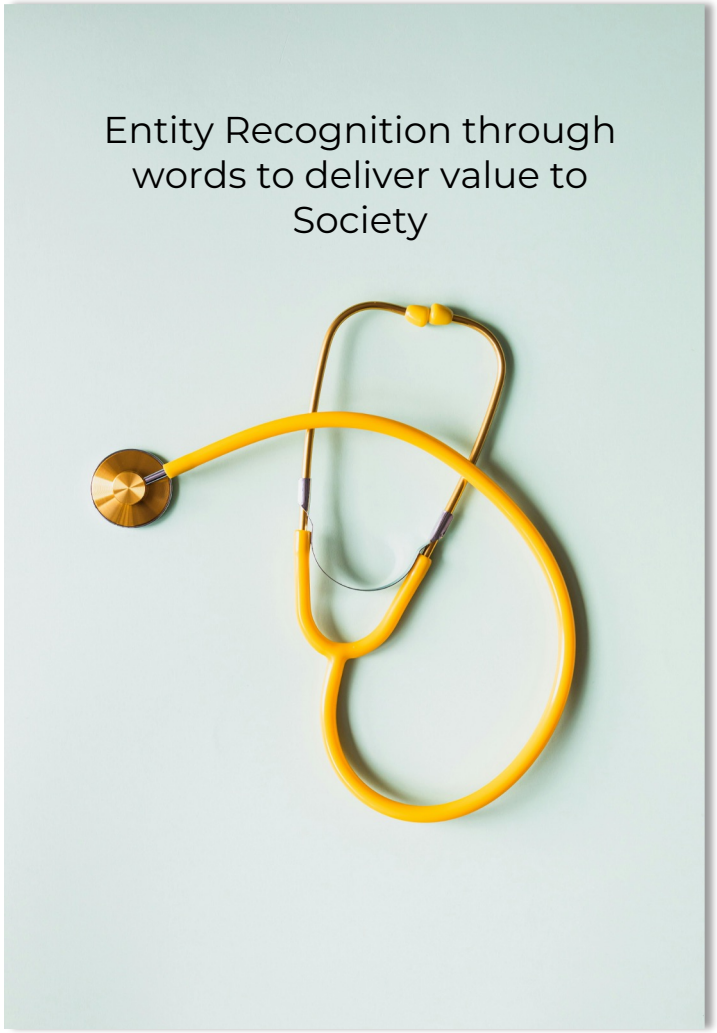
AGENDA

1. **Creating a Synthetic Medical Dataset**
   - A Transformers: creating our own Chat GPT-2 with data
   - B API: Using Chat GPT-3.5T
2. **Medical Entity Recognition**
   - A Disease counts and trends
   - B Geo-patterns
3. **Application Showcase**
4. **Annex & Additional Materials**

The Use Case



Making Synthetic Datasets



Analyzing Trends through NLP diagnostics

The Final Dataset

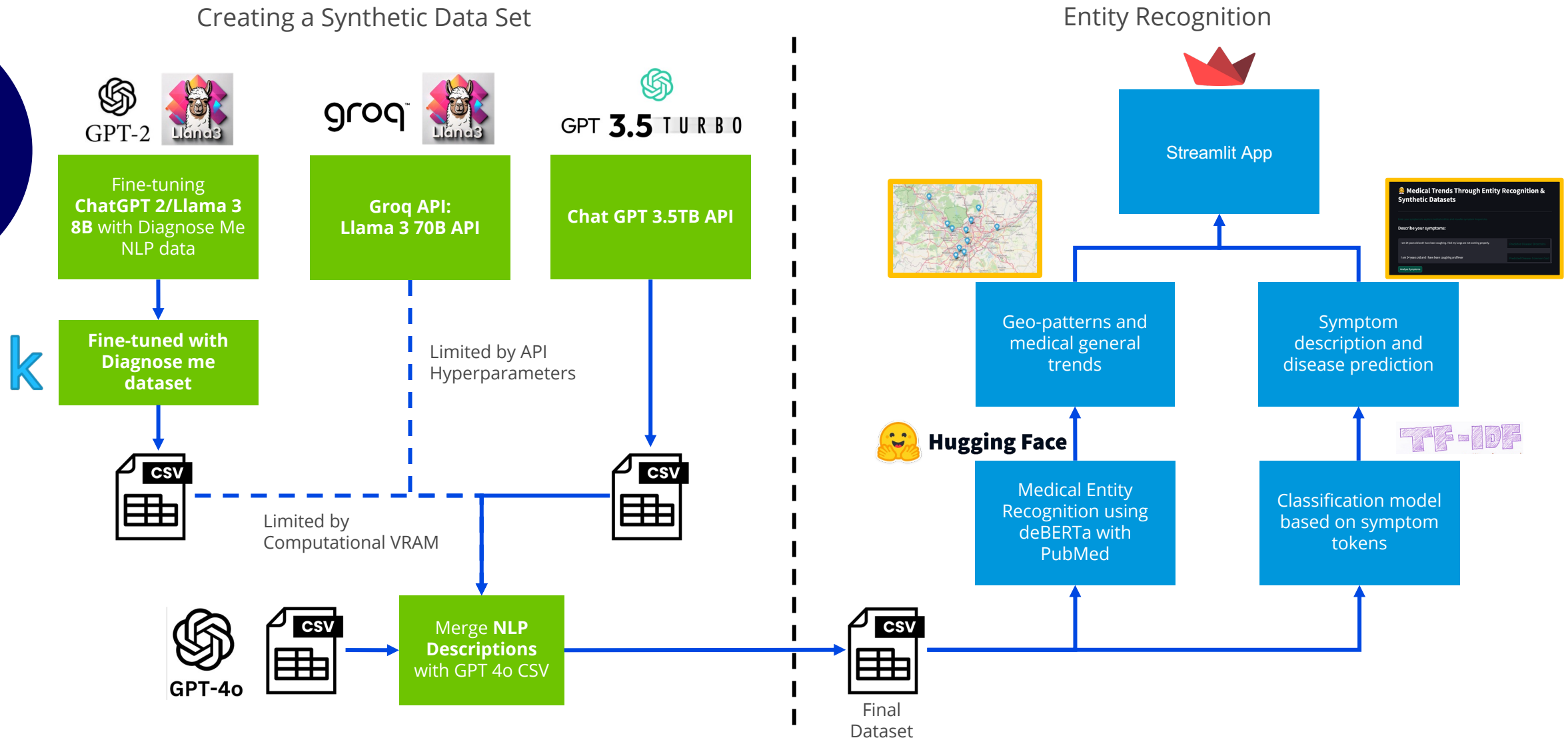
hospital_name	hospital_id	longitude	latitude	identified_disease	patient_id	patient_comment	symptoms	reported_timestamp
Hospital Universitario Fundacion Jimenez Diaz	6	40.4395	-3.7178	Allergic Rhinitis	497718	Patient, a 30-year-old female office worker, presented with complaints of	Sneezing, runny or stuffy nose, itchy eye	57:16.9
Hospital Universitario de Collado Villalba	19	40.6362	-4.0064	Allergic Rhinitis	124835	Patient is a 25-year-old female college student who complains of frequent	Sneezing, runny or stuffy nose, itchy eye	33:55.3
Hospital Universitario de Getafe	12	40.3083	-3.7324	Allergic Rhinitis	317339	A 30-year-old female patient presented with complaints of sneezing, nasal	Sneezing, runny or stuffy nose, itchy eye	19:14.9
Hospital Universitario Vithas Madrid Arturo Soria	26	40.4457	-3.6484	Allergic Rhinitis	28425	Patient is a 30-year-old female who presents with complaints of sneezing,	Sneezing, runny or stuffy nose, itchy eye	27:38.9
Hospital Universitario Ramon y Cajal	4	40.4956	-3.6781	Allergic Rhinitis	481955	Patient is a 25-year-old female who complains of frequent sneezing, runny	Sneezing, runny or stuffy nose, itchy eye	36:41.8
Hospital Universitario Quironsalud Madrid	20	40.4507	-3.8265	Allergic Rhinitis	306013	Patient is a 25-year-old female who presents with complaints of sneezing,	Sneezing, runny or stuffy nose, itchy eye	54:56.4
Hospital Universitario La Zarzuela	23	40.4481	-3.8265	Allergic Rhinitis	42044	Patient is a 25-year-old female who presents with complaints of frequent	Sneezing, runny or stuffy nose, itchy eye	57:13.8
Hospital Universitario Principe de Asturias	11	40.4847	-3.3661	Allergic Rhinitis	146662	Patient is a 35-year-old female who complains of frequent sneezing, nasal c	Sneezing, runny or stuffy nose, itchy eye	38:51.9
Hospital Universitario Fundacion Jimenez Diaz	6	40.4395	-3.7178	Allergic Rhinitis	272795	Patient, 25-year-old female, presents with complaints of nasal	Sneezing, runny or stuffy nose, itchy eye	32:23.7
Hospital Universitario de Torrejon	15	40.4562	-3.4753	Allergic Rhinitis	248421	Patient's Symptoms:	Sneezing, runny or stuffy nose, itchy eye	05:22.5
Hospital Universitario Severo Ochoa	10	40.3317	-3.7694	Allergic Rhinitis	470789	Patient is a 30-year-old male who complains of frequent sneezing, runny	Sneezing, runny or stuffy nose, itchy eye	23:16.2
Hospital Universitario Infanta Cristina	17	40.2808	-3.8291	Allergic Rhinitis	337294	Patient, a 35-year-old female, presented with complaints of sneezing,	Sneezing, runny or stuffy nose, itchy eye	12:22.9
Hospital Universitario Vithas Madrid Arturo Soria	26	40.4457	-3.6484	Allergic Rhinitis	357841	Patient is a 25-year-old female who complains of sneezing, runny nose,	Sneezing, runny or stuffy nose, itchy eye	18:25.5
Hospital Universitario Severo Ochoa	10	40.3317	-3.7694	Allergic Rhinitis	461095	Patient is a 25-year-old female who presents with complaints of sneezing,	Sneezing, runny or stuffy nose, itchy eye	20:10.8
Hospital Universitario de Torrejon	15	40.4562	-3.4753	Allergic Rhinitis	115002	Patient is a 25-year-old female who complains of frequent sneezing, runny	Sneezing, runny or stuffy nose, itchy eye	56:02.5
Hospital Universitario 12 de Octubre	2	40.3733	-3.7074	Allergic Rhinitis	436579	Patient is a 25-year-old male who complains of frequent sneezing, a runny	Sneezing, runny or stuffy nose, itchy eye	23:47.4
Hospital Universitario de Fuenlabrada	16	40.2875	-3.7978	Allergic Rhinitis	268326	Patient is a 35-year-old male who presents with complaints of frequent	Sneezing, runny or stuffy nose, itchy eye	39:15.5
Hospital Universitario Infanta Sofia	7	40.5705	-3.6201	Allergic Rhinitis	428586	Patient is a 25-year-old female who presents with complaints of sneezing,	Sneezing, runny or stuffy nose, itchy eye	30:39.0
Hospital Universitario Puerta de Hierro Majadahonda	8	40.476	-3.8626	Allergic Rhinitis	305639	Patient is a 25-year-old female who presents with complaints of sneezing,	Sneezing, runny or stuffy nose, itchy eye	56:06.3
Hospital Universitario de Collado Villalba	19	40.6362	-4.0064	Allergic Rhinitis	128344	Patient, a 30-year-old female, presented with complaints of frequent	Sneezing, runny or stuffy nose, itchy eye	28:08.3
Hospital Universitario del Henares	14	40.4184	-3.5297	Allergic Rhinitis	16814	Patient is a 35-year-old female who complains of frequent sneezing, itchy	Sneezing, runny or stuffy nose, itchy eye	21:28.7

We built a dataset with ChatGPT 4o. The idea is to have a **patient describing symtoms to a doctor** that annotates them in "**patient_comment**".

The patient comments are **generated through Open AI's API**, with ChatGPT 3.5Turbo. The API looks at the identified disease column and describes the symptom based on this.

We also have included the "**identified_disease**" for classification purposes. The dataset includes 37 diseases of 32 Madrid Hospitals.

Process Outline



Creating our own GPT-2 Model


```
[
    {
        "id":0,
        "Description":"Q. What does abutment of the nerve root mean?",
        "Doctor":"Hi. I have gone through your query with diligence and would like you to know that I am here to help you. For further information consult a neurologist online --> https:\/\/www.icliniq.com\/ask-a-doctor-online\/neurologist  ",
        "Patient":"Hi doctor,I am just wondering what is abutting and abutment of the nerve root means in a back issue. Please explain. What treatment is required for\u00a0annular bulging and tear?"
    },
    {
        "id":1,
        "Description":"Q. Every time I eat spicy food, I poop blood. Why?",
        "Doctor":"Hello. I have gone through your information and test reports (attachment removed to protect patient identity). So, in view of that, there are a couple of things that I can opine upon: Hope that helps. For more information consult a general surgeon online --> https:\/\/icliniq.com.\/ask-a-doctor-online\/general-surgeon  ",
        "Patient":"Hi doctor, I am a 26 year old male. I am 5 feet and 9 inches tall and weigh 255 pounds. When I eat spicy\u00a0food, I poop blood. Sometimes when I have constipation as well, I poop a little bit of blood. I am really scared that I have colon cancer. I do have diarrhea often. I\u00a0do not have a\u00a0family history of colon\u00a0cancer. I got blood tests done last night. Please find my reports attached."
    },
    {
        "id":2,
        "Description":"Q. Will Nano-Leo give permanent solution for erection problem?",
        "Doctor":"Hi. For further doubts consult a sexologist online --> https:\/\/www.icliniq.com\/ask-a-doctor-online\/sexologist  ",
        "Patient":"Hello doctor, I am 48 years old. I am experiencing weak erection and difficulty in sustaining the same. This condition was observed 10 years back. Also, there is premature ejaculation. Other physical ailments that I have are, I am suffering from hypertension and taking Amlopres-L (Amlodipine and Lisinopril) for the last 10 years, high cholesterol and triglycerides. My cholesterol level is 225 and triglyceride is 200 for the last 12 years. I used to do frequent masturbation in early age.\u00a0I do have erection during morning hours many times, particularly after sound sleep or if I had long walk previous day. I am having Sildenafil 25 mg or 5 mg Cialis, which is effective enough. But, I wish to get rid of tablet support and live natural way.\u00a0I consulted urologist today and he prescribed me Nano-Leo capsules and Modula 5 mg for 10 days.\u00a0I wish to have your second opinion on this. Please guide."
    },
    {
        "id":3,
        "Description":"Q. Will Kalarchikai cure multiple ovarian cysts in PCOD?",
        "Doctor":"Hello. I just read your query. See Kalarachi Kai choornam is helpful in amenorrhea. As far as small cysts are concerned they are unmatured eggs which failed to induce menstrual cycle previously, as a result, they got collected in the ovary and they will remain in the ovary. Now, you have got your periods you can start trying for conception. But I advise you to do it under the supervision of a nearby gynecologist because egg size is important while conception and that you can know by ovulation study. Ovulation study is performed under the supervision of a gynecologist. For gall stones, surgical intervention is required generally. Medicine is not of much help.  ",
        "Patient":"Hello doctor, I have multiple small cysts in both ovaries (PCOS). Our family friend suggested me to consume Kalarchikai medicine, powdered mixed with pepper and honey to have it for one mandalam (48 days). I have completed nearly 44 days by consuming one small ball of the mixture every morning and before two weeks I got my periods. Then I had checked whether the cysts were dissolved by having an abdomen scan. But still, small multiple follicules have been seen in it with both ovaries enlarged. Kindly suggest me what I can do further. I got married before one and a half years and I am much worried about my pregnancy. I have gallstones as well."
    },
    {
        "id":4,
```


```python
import json
from datasets import Dataset
from transformers import GPT2Tokenizer, GPT2LMHeadModel, Trainer, TrainingArguments, DataCollatorForLanguageModeling

# Load the dataset
data_file_path = './fine-tuning data/en_medical_dialog.json'
with open(data_file_path, 'r') as f:
    data = json.load(f)

# Prepare the data for fine-tuning
# Extracting Patient and Description fields
dialogs = [{"text": entry['Description'] + " " + entry['Patient']} for entry in data]

# Create a Dataset object
dataset = Dataset.from_list(dialogs)

# Load the pre-trained GPT-2 tokenizer
model_name = 'gpt2'
tokenizer = GPT2Tokenizer.from_pretrained(model_name)

# Set padding token to be the end-of-sequence token
tokenizer.pad_token = tokenizer.eos_token

# Function to tokenize the data
def tokenize_function(examples):
    return tokenizer(examples['text'], padding='max_length', truncation=True, max_length=512)

# Tokenize the dataset
tokenized_datasets = dataset.map(tokenize_function, batched=True)

# Set the format for PyTorch
tokenized_datasets.set_format(type='torch', columns=['input_ids', 'attention_mask'])

# Data collator that will dynamically pad the inputs received
data_collator = DataCollatorForLanguageModeling(
    tokenizer=tokenizer,
    mlm=False,
)

# Split the data into training and evaluation sets
split_dataset = tokenized_datasets.train_test_split(test_size=0.1)
train_dataset = split_dataset['train']
eval_dataset = split_dataset['test']

# Load pre-trained GPT-2 model
model = GPT2LMHeadModel.from_pretrained(model_name)

# Freeze the initial layers of the model
for param in model.transformer.h[:6].parameters():
    param.requires_grad = False

# Set up training arguments with adjusted logging frequency
training_args = TrainingArguments(
    output_dir='./results',
    overwrite_output_dir=True,
    num_train_epochs=3,
    per_device_train_batch_size=2,
```

Creating our own GPT-2 Model

```python
import pandas as pd
from transformers import GPT2LMHeadModel, GPT2Tokenizer
from tqdm import tqdm
import torch
import random

model_name = './fine-medical-tuned-gpt2'
tokenizer = GPT2Tokenizer.from_pretrained(model_name)
model = GPT2LMHeadModel.from_pretrained(model_name).to('cuda')

def generate_symptom_description(disease, max_length=100, temperature=1.0, top_p=0.9):
    prompt_variations = [
        f"As a doctor, a patient came to me telling me a description of his symtoms for {disease}, which he did not know and did not mention explictly before.
        He/She mentioned demographics and his lifestyle randomly in the explanation:"
    ]
    prompt = random.choice(prompt_variations)
    inputs = tokenizer.encode(prompt, return_tensors='pt').to('cuda')
    attention_mask = inputs.ne(tokenizer.pad_token_id).to('cuda')
    outputs = model.generate(
        inputs,
        max_length=max_length,
        num_return_sequences=1,
        no_repeat_ngram_size=3,
        num_beams=30,
        do_sample=True,
        num_beam_groups=1,
        early_stopping=False,
        temperature=temperature,
        top_k=50,
        top_p=top_p,
        attention_mask=attention_mask,
        pad_token_id=tokenizer.eos_token_id
    )
    description = tokenizer.decode(outputs[0], skip_special_tokens=True)
    return description
```

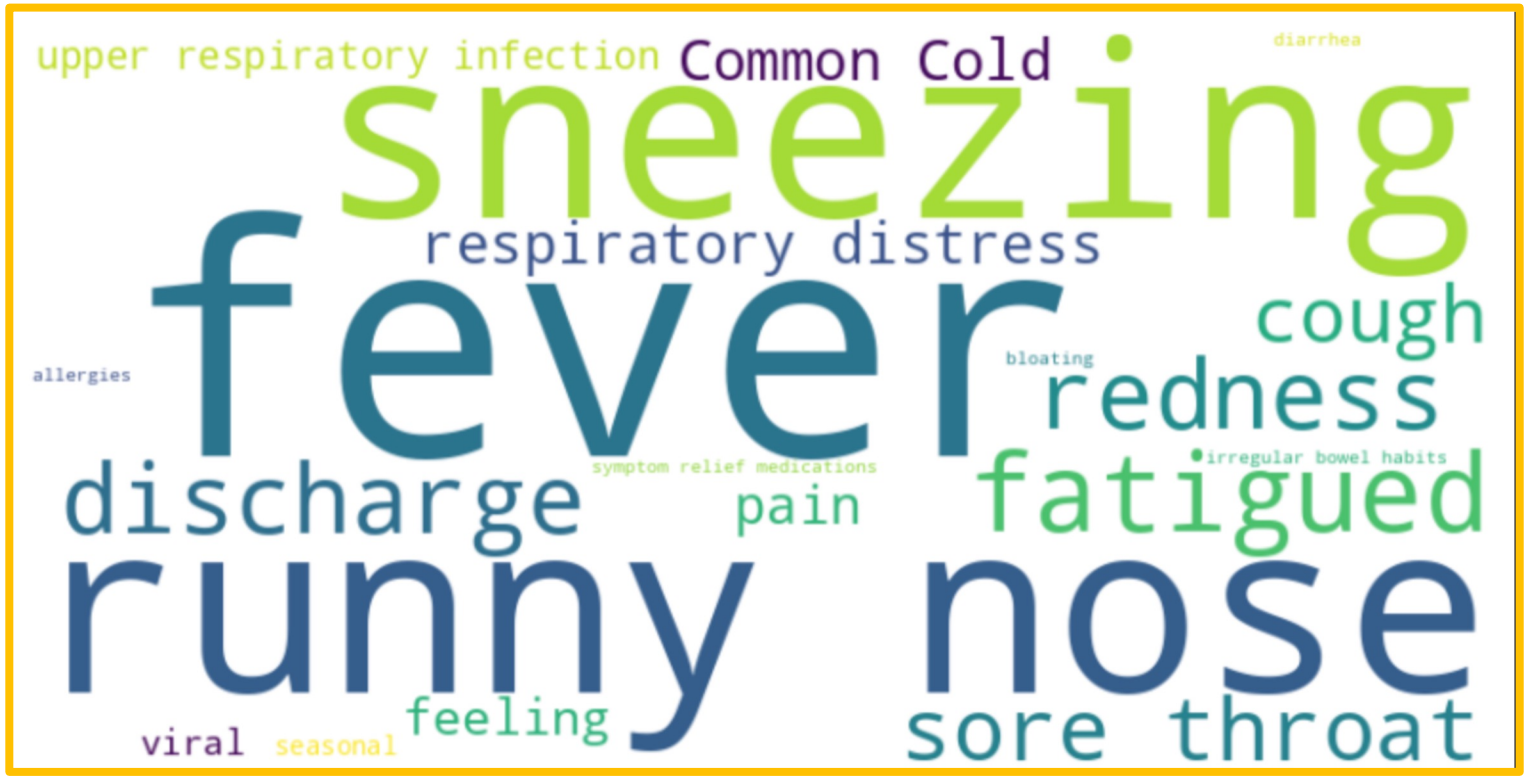
Btw, this is how your specific task BERT, GPTs and other specialized transformers are created!

Medical Entity Recognition

Unfolding Trends

Final Dataset

Hugging Face × Clinical - AI - Apollo



Medical Entity Recognition

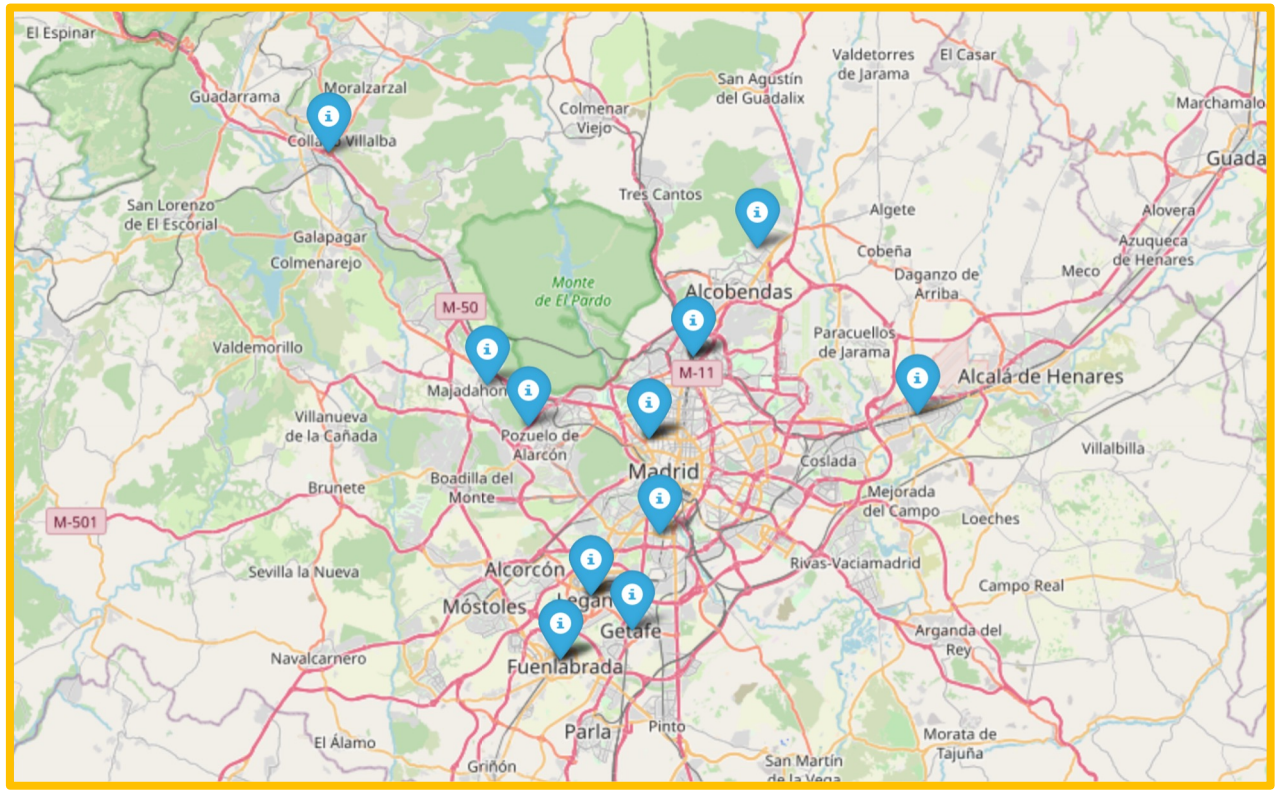
Geo-Patterns

Final Dataset

Hugging Face

Clinical - AI - Apollo

"FEVER"



Based on the words that appear in the patient_comment, we use NLP Entity Recognition domain-specific transformers to make a most important terms count and display where those terms have been reported.

- This allows to connect specific diseases and symptoms to geographical areas.
- Allows to predict a sudden increase in the number of cases and contain it before it is too late.

Application Showcase

Logistic Regression

👩‍⚕️ Medical Trends Through Entity Recognition & Synthetic Datasets

Enter your symptoms to explore medical entities and visualize symptom frequencies.

Describe your symptoms:

I am 24 years old and I have been coughing. I feel my lungs are not working properly. → Predicted Disease: Bronchitis

I am 24 years old and I have been coughing and fever → Predicted Disease: Common Cold

Analyze Symptoms

TF-IDF

https://medicalinsights.streamlit.app/

Future Lines of Work

1. **Increase Datasets disease categories**: Expand the synthetic dataset to include a wider variety of diseases, symptoms. We only have 24 for this demo.

2. **Using Real-Time Data**: Incorporate real-time data feeds to track disease trends!

3. **Predictive Analytics Tools**: Creating more tools to predict disease trends and patterns based on both historical and real-time data, aiding in proactive public health measures.

4. **Collaborate with Healthcare Providers**: Partner with healthcare institutions to validate synthetic datasets and models against real-world data, ensuring practical applicability and reliability.

Team Members

Contact Information

Ignacio Alonso López-Linares

IE Sci-Tech – MBD Student

i.alonso.linares@student.ie.edu

Morten Aas-Lyngby

IE Sci-Tech – MBD Student

morten.aas.lyngby@student.ie.edu

Harel Ben David

IE Sci-Tech – MBD Student

harel@student.ie.edu

Kangjie Yu

IE Sci-Tech – MBD Student

carlos.yu@student.ie.edu

Andreu Artigues Martínez

IE Sci-Tech – MBD Student

andreuartigues@student.ie.edu

Guillermo Brun Rubio

IE Sci-Tech – MBD Student

guillermo.brun@student.ie.edu

Any Questions?