
Internship @ TLab

— Anusha Madapura —

The Rise of Synthetic Data

- Data that is artificially generated
- Supposed to resemble the patterns of original data - while also having enough “noise”
- Used to fill any gaps in real world data, mitigate privacy and accessibility concerns.

Synthetic Social Media Data: Why?

- People are more likely to share their experiences with peers than with professionals [when it comes to side effects](#)
- Real-time signal detection for pharmacovigilance
- Novel methods for signal detection are constantly being developed
- Lack of training data for such models

Performance on Unseen Data: SMM4H Competition Results

	F1-Norm	P-Norm	R-Norm	Unseen: F1-Norm	Unseen: P-Norm	Unseen: R-Norm
Mean	28.264	29.244	33.388	20.936	20.524	28.676
Median	29.3	33.9	32.6	14.1	14.4	36.5

Task: Normalizing natural language ADEs to MedDRA PTs

Project: Making realistic synthetic data

- Used 1824 annotated tweets dataset from SMM4H
- What defines realistic synthetic data in this context?
 - Diversity in phrasing, word choice, style etc
- We came up with three different ways of prompting the GPT-4 model, and wanted to evaluate which method produced the most diverse dataset:
 - Zero-Shot, Selected-One-Shot and Sampled One-Shot
- (version=gpt-4-1106-preview, temperature=0.7 max new tokens=512),

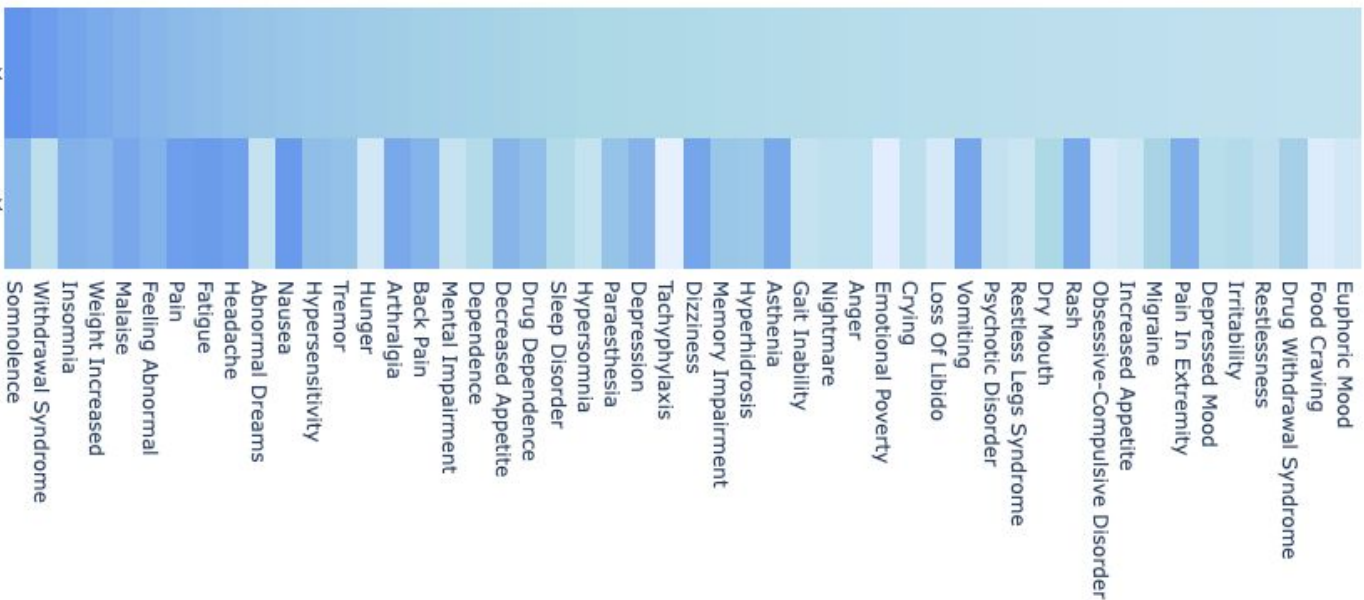
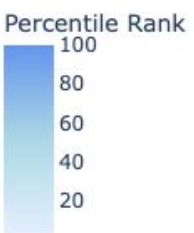
Hypothesis:

Providing the LLM with an example (1-shot) will help generate a more diverse dataset. Specifically, selected one-shot will be more varied than sampled one-shot.

Category

SMM4H Percentile Rank

FAERS Percentile Rank

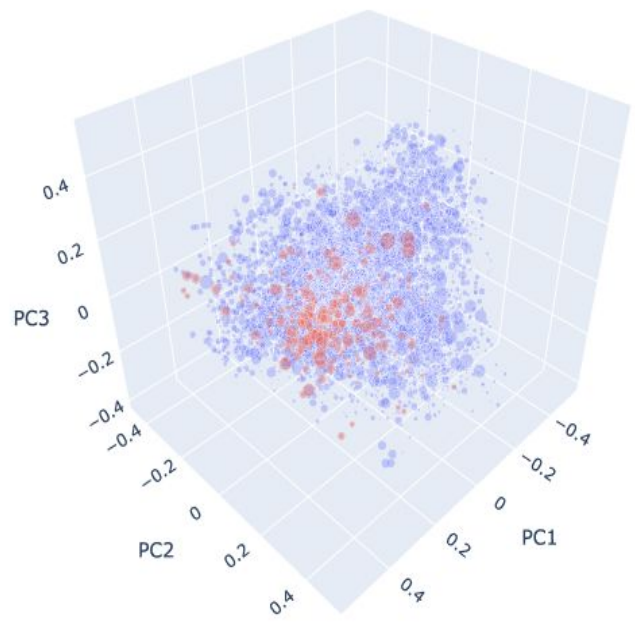


Name

Label

faers

smm4h



Zero-shot

```
{
```

```
    "role": "system",
```

```
    "content": "You are an assistant that, given an Adverse Drug Event (ADE), writes a tweet containing the specified ADE. Ensure that the tweet does not mention any other symptoms besides the prompted ADE. The tweet should be concise, realistic, and reflect a genuine user experience. Return nothing besides the Tweet."
```

```
},
```

```
{
```

```
    "role": "user",
```

```
    "content": f'The ADE to incorporate is: {row["Faers_ADE"]}'
```

```
}
```

Sampled One-Shot

```
"role": "system",
```

```
"content": "You are an assistant that, given an Adverse Drug Event (ADE), writes a tweet containing the specified ADE. Ensure that the tweet does not mention any other symptoms besides the prompted ADE. The tweet should be concise, realistic, and reflect a genuine user experience. The user will provide an example and you should incorporate its structure, tone, word, and language choices in your new Tweet. Return nothing besides the Tweet."
```

```
},
```

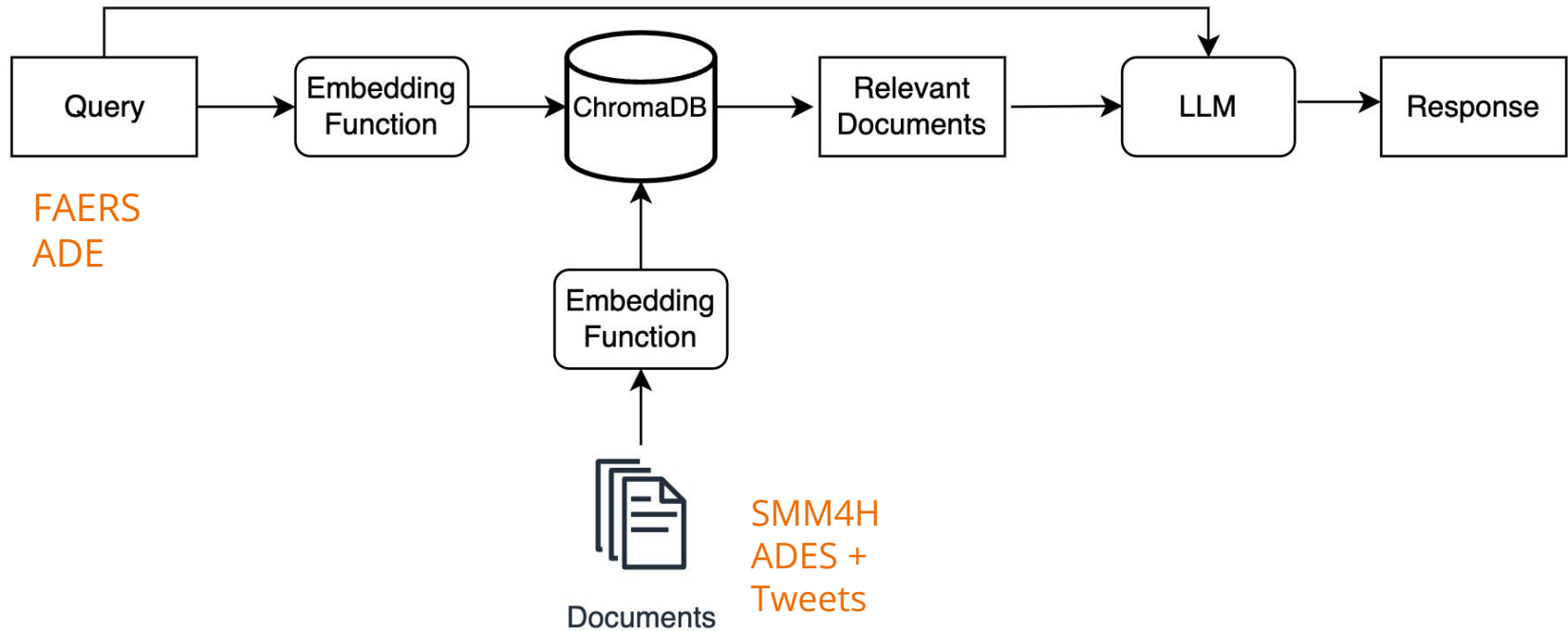
```
{
```

```
"role": "user",
```

```
"content": f'The ADE to incorporate is: {row["Faers_ADE"]}. The example to go off of is:  
ADE: {row["Text1"]}, Tweet: {row["Tweet1"]}'
```

```
}
```

ChromaDB for Generating Selected One-Shot Tweets



Selected One-Shot

```
"role": "system",
```

```
  "content": "You are an assistant that, given an Adverse Drug Event (ADE), writes a tweet containing the specified ADE. Ensure that the tweet does not mention any other symptoms besides the prompted ADE. The tweet should be concise, realistic, and reflect a genuine user experience. The user will provide an example and you should incorporate its structure, tone, word, and language choices in your new tweet. Return nothing besides the tweet."
```

```
},
```

```
{
```

```
  "role": "user",
```

```
  "content": f"The ADE to incorporate is: {row['Faers_ADE']}. The example to go off of is:  
ADE: {row['Text1']}, Tweet: {row['Tweet1']}"
```

Previous Synthesis Approaches

Graph networks

Diffusion models

Transformer models

Why:

- Talk about SMM4H problem - extraction and normalization from social media
- Lack of training data - LLMs -
- Not being able to recognize terms that haven't been introduced previously
- Introduce synthetic data
- Hypothesis:
- Sampled one-shot v/s selected one-shot
-
-

Zero-Shot

Aplasia Pure Red Cell	Just found out that my anemia is actually due to Aplasia Pure Red Cell after starting a new medication. Feeling overwhelmed and hoping for a treatment plan that works. #health #anemia #aplasia #medicationawareness
Endophthalmitis	Just got diagnosed with endophthalmitis after my eye surgery. Never thought I'd have to deal with something like this. If you've had eye procedures, be vigilant about any signs of infection! #EyeHealth #PostSurgeryCare
Bacterial Sepsis	Just got diagnosed with bacterial sepsis after starting a new medication. Please be careful and monitor your symptoms when on new meds! #health #sepsisawareness
Erythema	Just noticed a weird redness on my skin after starting my new meds. Is this erythema a common side effect? #medication #sideeffects #erythema

Selected Few-Shot

Aplasia Pure Red Cell	Has anyone with aplasia pure red cell been able to maintain normal hemoglobin levels? Just curious. #aplasia #medication #healthquestions
Endophthalmitis	Just got diagnosed with endophthalmitis after my eye injection. Scary stuff, but getting treatment ASAP. Take your eye symptoms seriously, folks! #EyeHealth #Endophthalmitis
Bacterial Sepsis	@USER_____ after switching to the new antibiotic, I ended up with bacterial sepsis. Back to square one with treatments. #frustrated
Erythema	Just noticed some redness on my arms after starting the new meds. No biggie, slapped on some hydrocortisone cream. #ErythemaSquad

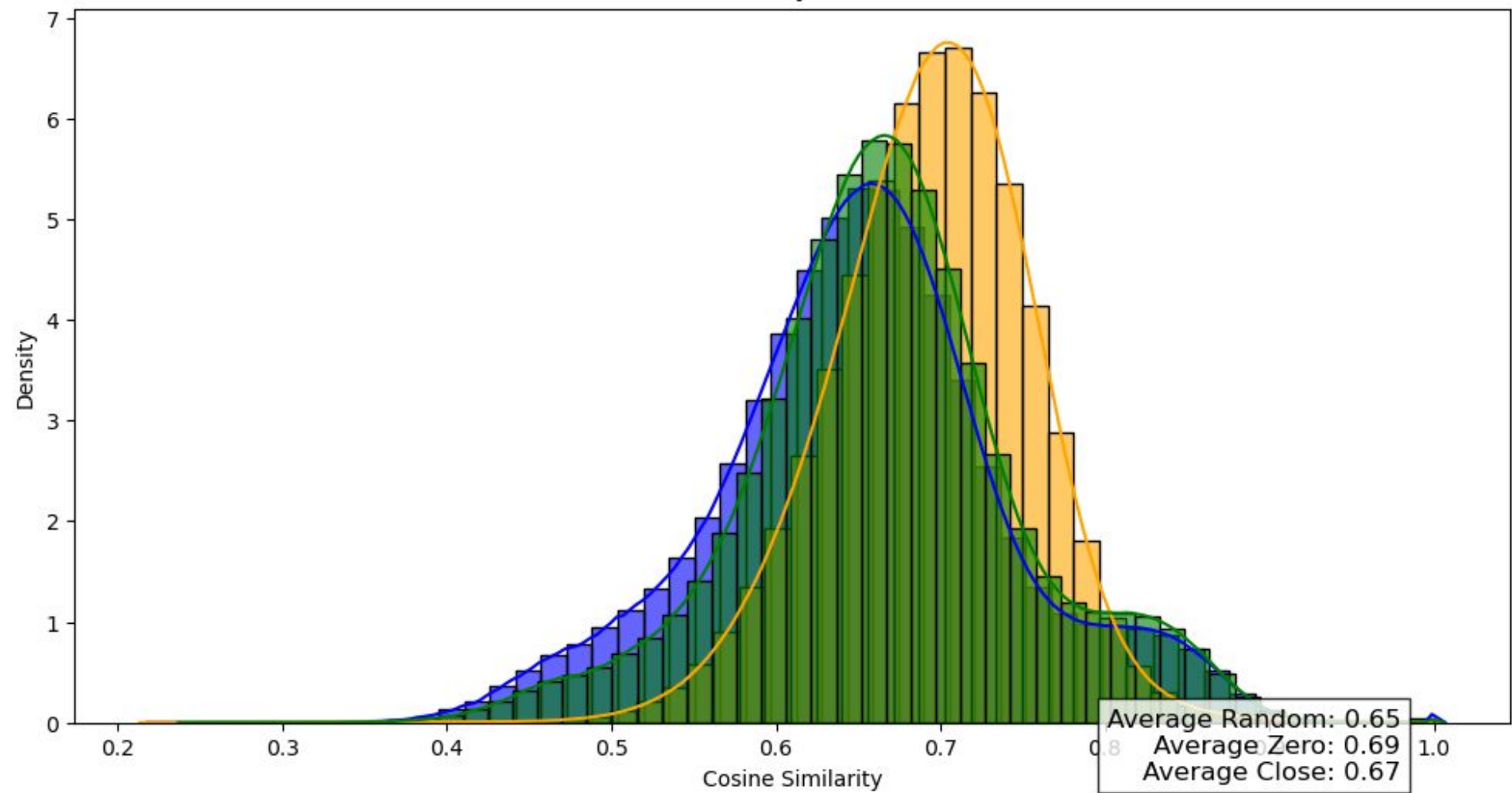
Sampled Few-Shot

Aplasia Pure Red Cell	Been dealing with aplasia pure red cell for a while now. Not easy, but I'm pushing through. No energy is tough, but I'm tougher. #healthchallenges #aplasia
Endophthalmitis	Just got diagnosed with endophthalmitis after my eye injection. Inflammation, pain, and vision issues - not what I signed up for. 🙄 #eyehealth #postinjection #whythis #visualnightmare
Bacterial Sepsis	@USER_____ never expected an antibiotic could lead to this. Bacterial sepsis is no joke. Felt like my body was shutting down. Recovery's tough. Stay vigilant with meds, folks. #healthscare #sepsis
Erythema	@USER_____ if you're dealing with redness after taking a new med, you're not alone. Erythema hit me out of nowhere. Stay vigilant with these meds!

Metrics for Evaluating Diversity in Synthetic Data

- Measuring semantic similarity within that particular dataset
- Used “encode” method from SentenceTransformers Library to embed all the tweets
- Calculated cosine similarity between every tweet (measures how semantically similar the tweets are)
- Still need method for comparing selected one-shot and sampled one-shot

Cosine Similarity of 3 Datasets



Results

	Zero-shot	Sampled one-shot	Selected one-shot
Mean	0.694	0.653	0.666
Variance	0.00392	0.00825	0.00724
Kurtosis	1.19	0.334	0.526

Conclusions

- Providing one example (one-shot) leads to a more varied dataset
- Providing a sampled example does cause a more varied dataset than a selected example
- Variation within a dataset doesn't necessarily need to be the only marker of quality

Future Directions

- Compare cosine similarity distribution of synthetic datasets to that of SMM4H data
- Develop method to assess and compare quality of the different one-shot methods
- Use Wasserstein Distance to compare semantic similarity
- Recreating drug-effect relationships/drug-drug interactions in synthetic data

A Potential Use Case

Wanted to explore more unusual sources of data for extracting ADEs

Found a few papers on using online health forum data

Could the synthetic data be used for fine-tuning a model that extracts ADEs from any form of social media?

18	3	deleted_user	11/12/2011	Original Post	Hello....i am a 38 male with (achon) c
19	3	deleted_user	01/31/2012	Comment	Hey Ric, I'm no stranger to pain either
20	3	4evertryingtoimprove	01/04/2017	Comment	I hope you're okay now.
21	4	d212ewi	09/21/2015	Original Post	Hi,...i am a mother of 14th yrs daugh
22	5	deleted_user	05/31/2014	Original Post	I do not have Achondroplasia. I am h
23	5	deleted_user	05/31/2014	Comment	I want to add to this by saying that I
24	5	deleted_user	06/02/2014	Comment	Just discovered the Short Stature gro
25	6	deleted_user	10/03/2012	Original Post	I have moderate to severe acne but I
26	6	deleted_user	10/03/2012	Comment	Maybe your doctor can prescribe som
27	6	deleted_user	06/01/2014	Comment	Totally with you on this one! Can't ex
28	7	deleted_user	09/03/2012	Original Post	Anyone with any experience with hyp
29	7	deleted_user	09/04/2012	Comment	Hi CrayCray,My son was diagnosed v
30	7	deleted_user	09/04/2012	Comment	Hey, maybe one of these links will he
31	7	deleted_user	09/04/2012	Comment	Thank you for your response! When
32	7	deleted_user	09/04/2012	Comment	Thank you lilstubby! I've found some
33	7	deleted_user	09/05/2012	Comment	No problem...blessings to you and yo
34	7	deleted_user	05/11/2014	Comment	Hil have Acondroplasia and while I c
35	8	deleted_user	06/19/2011	Original Post	i came out as gay 5 months ago and
36	8	PinterestFan	12/31/2013	Comment	People fear change of every kind. Wl
37	8	PinterestFan	12/31/2013	Comment	Hmmm - I have no idea how I ended

Thank you for an amazing summer!

