



Holden

Office visit at ONE MEDICAL GROUP INC

\$527.75

Partial **Denied**

Fighting Health Insurance w/LLMs

Holden Karau

Timbit's claim (#11020282) is complete

\$393.00

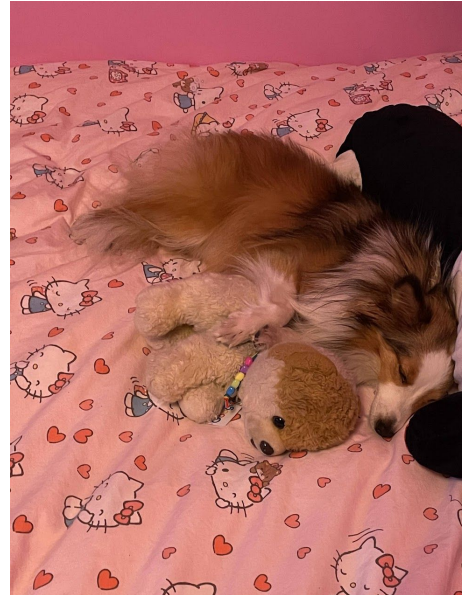
Denied

Oct 2024



Content Warning

- References to surgery, hospitals, and medical insurance
- Probably some trans references
- Maybe some references to broken bones



Who am I?



- My name is Holden Karau
- Pronouns are she/her
- Apache Spark PMC (think committer with tenure)
- previously Apple, IBM, Alpine, Databricks, Google, Foursquare & Amazon
- co-author of High Performance Spark, Learning Spark, and Kubeflow for Machine Learning, Scaling Python With {Ray, Dask}
- Twitter: [@holdenkarau](https://twitter.com/holdenkarau), bluesky holdenkarau.com, mastodon @holden@tech.lgbt
- OOS Livestreams: <https://youtube.com/user/holdenkarau>
- Github <https://github.com/holdenk>
- My employer is hiring a manager for my team – but I don't represent them
 - Especially in this talk
- Dog mom





What's the problem?

- Health Insurance in America frequently denies medical care
- Some insurers are (allegedly) [using AI to deny claims](#)
- Appealing these denials is hard

▼ [Anthem Blue Cross:](#)

Individual Market

2016 Total Claims Received: 7,379,867

2016 Total Claims Denied: 2,422,282

2017 Total Claims Received: 5,634,611

2017 Total Claims Denied: 1,731,253

2018 Total Claims Received: 2,331,417

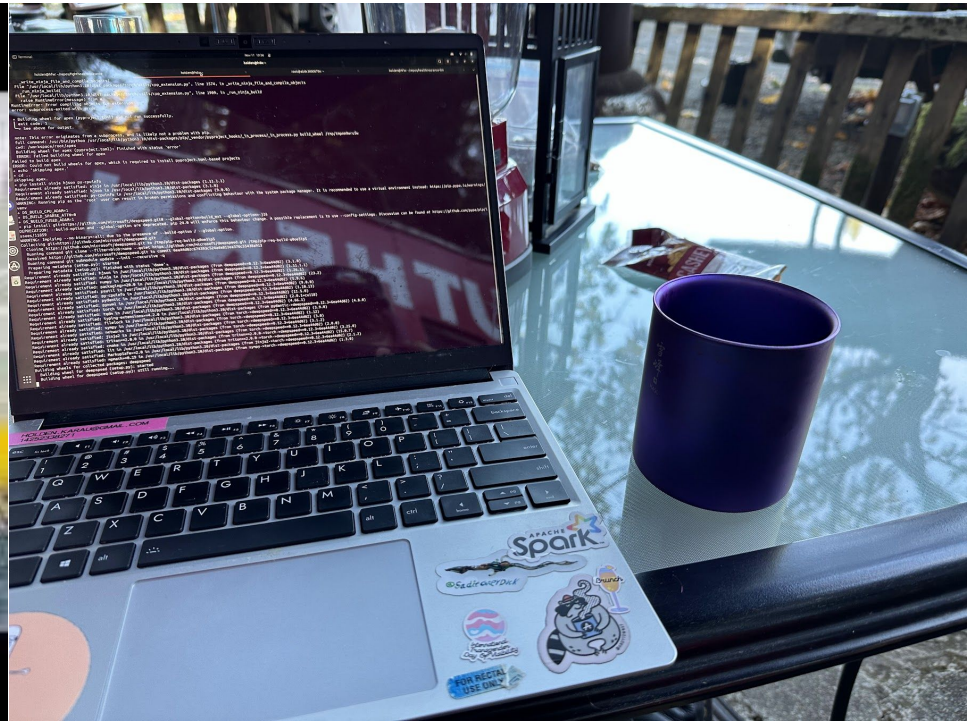
2018 Total Claims Denied: 629,010

And some personal motivation



I have a social problem

- Let's use computers to fix it



Ok so "how" do we use computers to fix that problem?

- An ML model to take denials and produce appeals
- Frontend Interface for people to access said model
- Library to remove and put the personal data back in

To be clear "not one true way"



What do we need to tune a language model?

- Training data
 - Lots of training data
- Evaluation Data
- Computers (especially with GPUs)
- An existing base model
- Some software to put it all together



So training data: how?

- Request them from someone who has them
 - That's insurance companies & doctors offices
- Google/Reddit search for denials
 - Then write some appeals
- Find some other data we can produce appeals/denials from
 - Oooh! Insurance commissioners



Independent Medical Review Boards

CalHHS Open Data – <https://wpso.dmhca.ca.gov/imr/>

Texas: Has the data at the state level but couldn't find a quick open version (also california was "enough" to start)

Example CA IMR Record

Findings: The physician reviewer found that Nature of Statutory Criteria/Case Summary: The patient has requested authorization and coverage for Fiasp injection 100 units/mL. Multiple daily injections, which involve several shots of short-acting prandial insulin, along with one to two doses of a long-acting basal preparation, are oftentimes recommended for management of type 1 diabetes. One study noted that randomised controlled trials have shown <93>an advantage of continuous subcutaneous insulin infusion (CSII) over multiple daily injections (MDI) in the general type 1 diabetes mellitus (T1DM) population<94>. Various options for prandial insulins are currently available. Researchers found that fast-acting insulin aspart (Fiasp), is insulin aspart in a new formulation aiming to mimic the physiologic prandial insulin release more closely than available rapid-acting insulin products. The records indicate that this patient has responded well to Fiasp, unlike the suboptimal response to other prandial insulins. For these reasons, the requested Fiasp injection 100 units/mL is medically necessary for treatment of this patient. Final Result: The reviewer determined that the requested medication is medically necessary for treatment of the patient<92>s medical condition. Therefore, the Health Plan<92>s denial should be overturned. Credentials/Qualifications: The reviewer is board certified in internal medicine with subspecialty certification in endocrinology, diabetes, and metabolism and is actively practicing. The reviewer is an expert in the treatment of the enrollee<92>s medical condition and knowledgeable about the proposed treatment through recent or current actual clinical experience treating those with the same or a similar medical condition.

The sketchy solution ask an LLM to use the findings :)

f""The independent medical review findings were {findings} and grounds were {grounds}. In your response instead of independent say internal in reference to any review or reviewers. Use this information to write the original insurance denial.""",

Putting it together (1/3):

```
for r in 1:

    print(r[1])

    model_index = 0

    for m in models:

        # For the first model we don't add an index but subsequent ones we do.

        mistr = ""
```



Putting it together (2/3)

```
if model_index > 0:
    mistr = f"{model_index}-"
for response_type in r[1].keys():
    i = 0
    for v in r[1][response_type]:
        idx = r[0]
        target_file = join(gen_loc,
f"{idx}MAGIC{mistr}{i}{response_type}.txt")
        i = i + 1
```



Putting it together (3/3):

```
        if not os.path.exists(target_file) or not  
check_for_bad_file(response_type, target_file):  
  
        response = make_request(m, v)  
  
        if not check_for_bad(response_type, response):  
  
            if not check_for_bad(response_type, response):  
  
                print(f"Writing out to {target_file}")  
  
                with open(target_file, "w") as f:  
  
                    f.write(response)
```

What's the downsides of generated data?

- Costs money to generate the data
 - But generally less than people
- The generated data might not be very good
 - And filtering it becomes a fair amount of work on top of that
- Possible license complications if not careful



Ok so how do we pick a model to fine tune?

- Make sure the model will fit in memory
- Check the license (again)
- Ideally otherwise performs well



Ok how do we do the fine tuning?

- Structure the data in the "correct" way
- Pile of shell scripts
- Axolotl simplifies this so much

What models did we make?

<https://huggingface.co/TotallyLegitCo/appeal-alpaca>

https://huggingface.co/TotallyLegitCo/fighthealthinsurance_model_v0.2

https://huggingface.co/TotallyLegitCo/fighthealthinsurance_model_v0.3
(now up to v0.5)

Waiting to see if I can get permission to use latest Mistral base model for 0.6

How did I make the latest model

- Inference on big model to produce synthetic output
- Fine tuned smaller model (did dolly first last year, then Mistral 7B using axolotl)



Ok so... how exactly?

`run_remote.sh` – copies everything over to your cloud computer of chose

`Setup.sh` – install everything

`Run.sh` – execute actual magic

Some config files:



Model Fine Tune Config (1/3)

`base_model: mistralai/Mistral-7B-Instruct-v0.2`

`model_type: MistralForCausalLM`

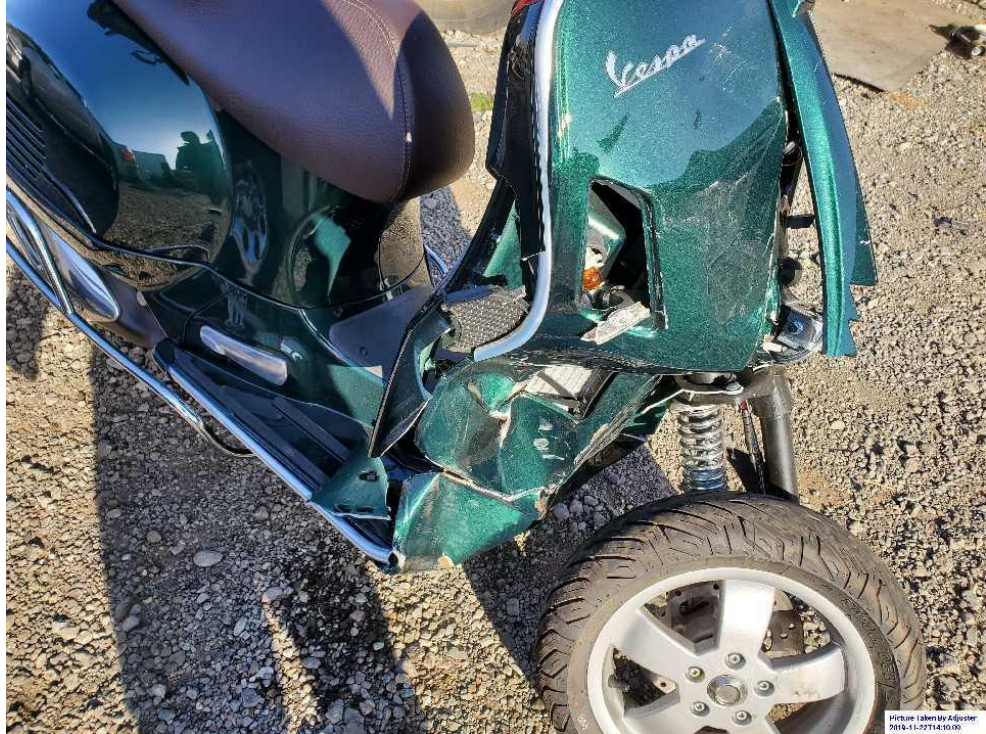
`tokenizer_type: LlamaTokenizer`

`is_mistral_derived_model: true`

`load_in_8bit: false`

`load_in_4bit: false`

`strict: false`



Model Fine Tune Config (2/3)

datasets:

- path: out/train_alpaca.jsonl

type:

alpaca

dataset_prepared_path:

val_set_size: 0.05

output_dir: ./mistral_fine_out

sequence_len: 8192

sample_packing: true

pad_to_sequence_len: true

Model Fine Tune Config (le fin)

.....

special_tokens:

 bos_token: "<s>"

 eos_token: "</s>"

 unk_token: "<unk>"

model_config:

 sliding_window: 4096



Ok and... cost?

USAGE	RATE	SPEND
0.8 hr	\$2.49 / hr	\$1.99
45.11 hr	\$2.49 / hr	\$112.32



Model Serving

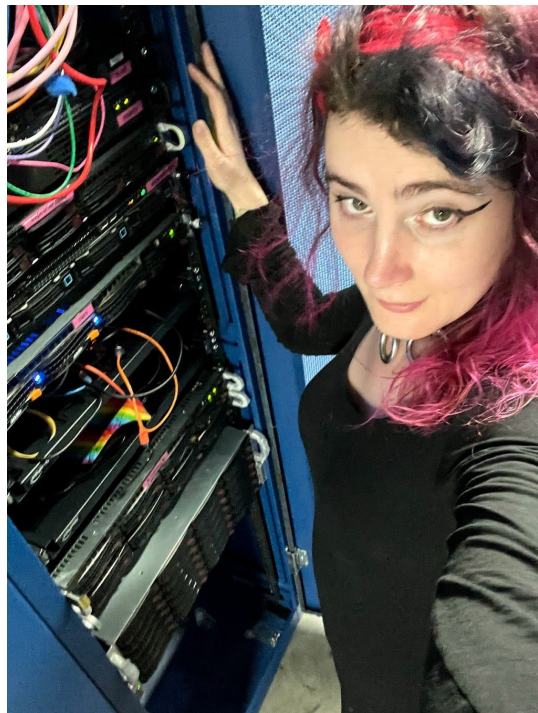
vLLM very fancy

Still need GPUs:

My ARM+GPUs don't party well

RTX 3090 does not fit well in a 2U server :(

Also uses power? And it costs money?



Model Serving (1/3):

kind: Deployment

metadata:

name: vllm-health

labels:

app: vllm-health

spec:

replicas: **1**

selector:

matchLabels:

app: vllm-health



Model Serving (2/3):

template:

metadata:

labels:

app: vllm-health

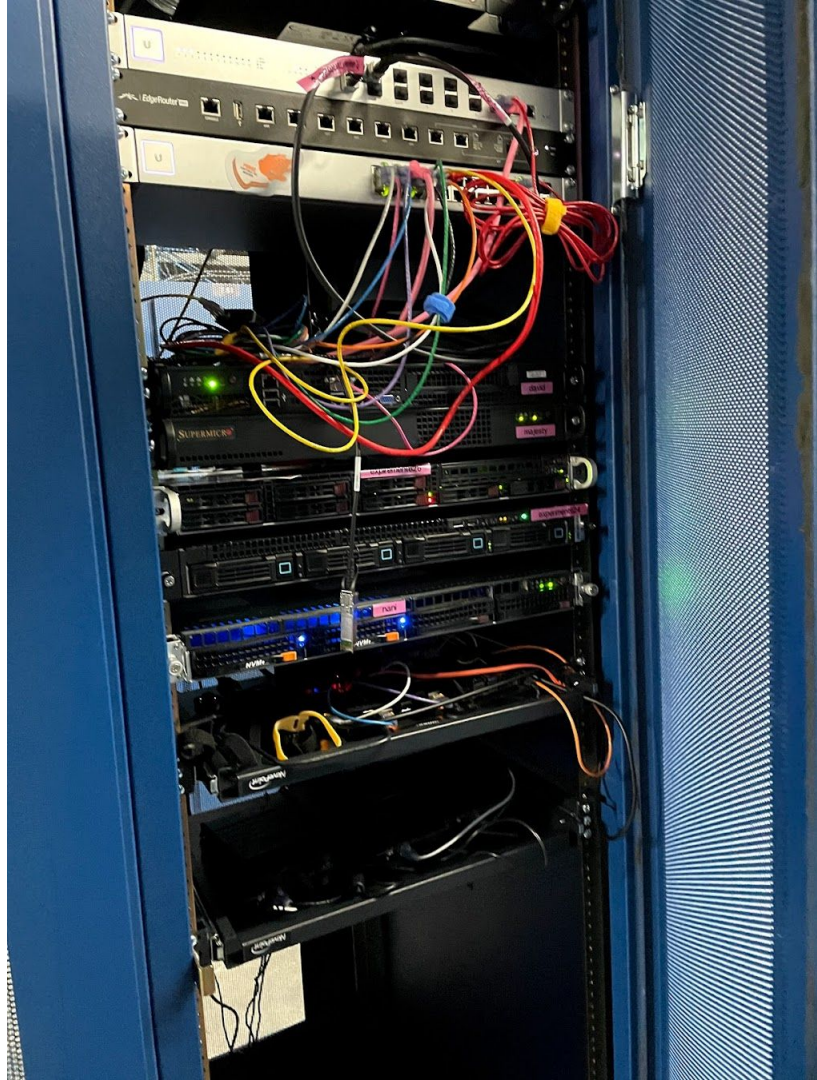
spec:

nodeSelector:

node.kubernetes.io/gpu: gpu

kubernetes.io/arch: amd64

runtimeClassName: nvidia



Model Serving (3/3):

containers:

- name: vllm-container

image: vllm/vllm-openai:latest

imagePullPolicy: Always

args: ["--model=TotallyLegitCo/fighthealthinsurance_model_v0.5", "--host",
"0.0.0.0", "--gpu-memory-utilization=0.9", "--max-model-len=26280",
"--enforce-eager"]

resources:

requests:

nvidia.com/gpu: 1

ephemeral-storage: "60Gi"



Ok ... and cost?

ZOTAC GAMING GeForce RTX 3090 Trinity OC 24GB GDDR6X ...



Total: US \$748.46

Order number: 14-10810-53797

Item ID: 386277557751

Seller: [ki-9851](#)

[Give feedback](#) →



Frontend?

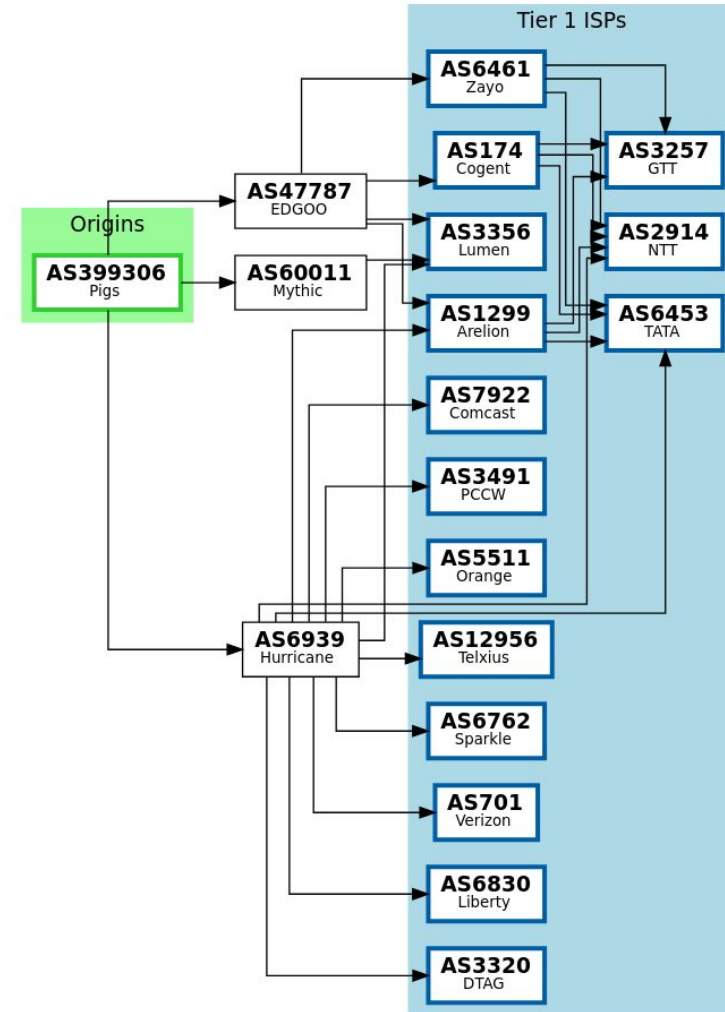
- I chose Django because I work in Python
- Also LLMs do a pretty good job of generating Python
- Needs computers (but less picky)



And internet access

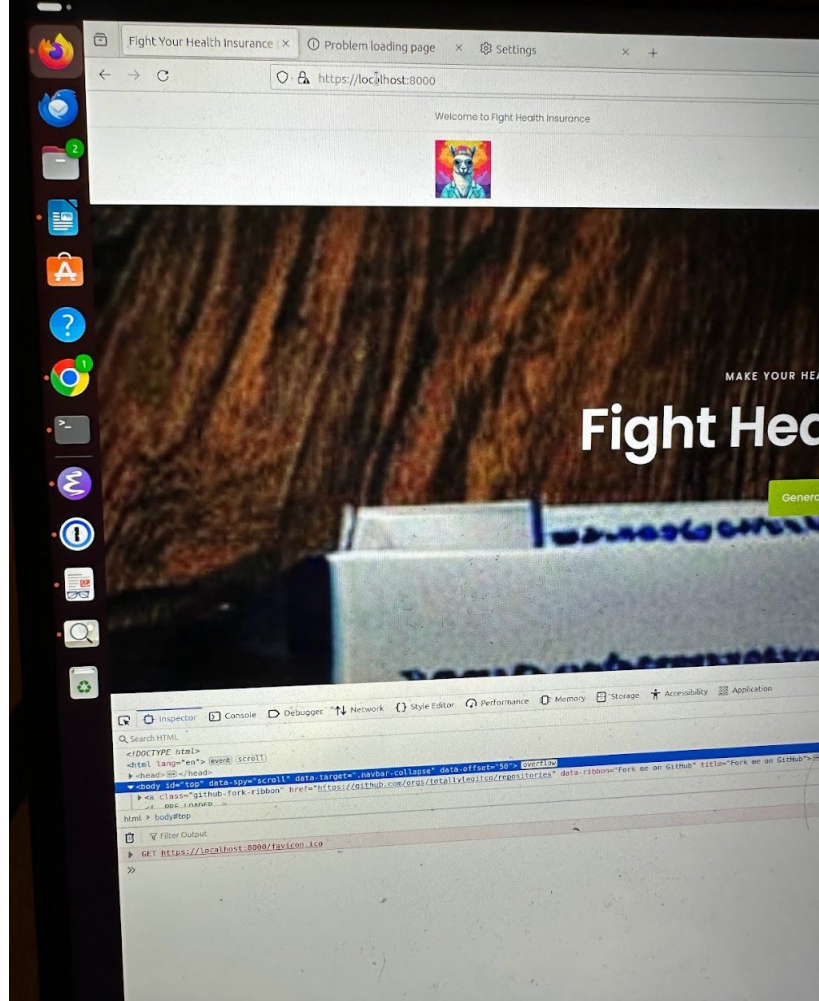
To be fair BGP was very optional

Also if you want internet at FMT2...



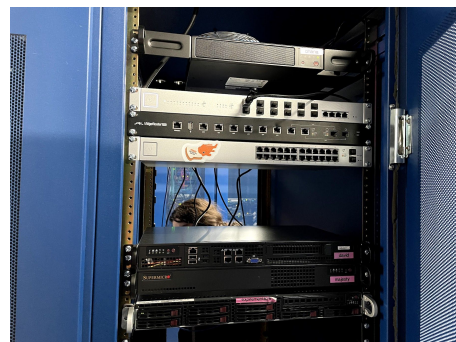
Now lets do a demo!

www.fightthehealthinsurance.com



Ok how do we do better for our next model?

- Use "real" data
 - Takes time to collect feedback etc.
- Use a better model to generate our synthetic data
 - Mistral 8x7 perhaps or Llama-3.1-Nemotron-70B-Instruct
- Filter out the bad generated data
 - Humans & regexes
- Have humans fix up the generated data
- Use Retrieval augmented generation
 - WIP using PubMed data



Links

<https://github.com/orgs/totallylegitco/repositories>

<https://www.youtube.com/user/holdenkarau>

<https://www.fightthehealthinsurance.com/>

<https://www.linkedin.com/in/holdenkarau/>

<https://bit.ly/hkHealthInsuranceVideos>

