1 Project summary

I chose to fine-tune a transformer model for text summarization.

I found a dataset on Hugging Face that summarizes transcribed clinical dialogue. It's a modified version of this dataset. The team that uploaded the modified dataset to Hugging Face added structure to the summaries, converting them from paragraph format to bullet-point format with 4 bullets: symptoms, diagnosis, patient history, and plan of action. The team only uploaded the training and validation splits.

The original dataset was intended for training both summarization and classification into medical categories, but for simplicity, I only attempted summarization.

The team that uploaded the dataset fine-tuned "facebook/bart-large-cnn", but I chose to fine-tune "facebook/bart-base" for efficiency and because the dataset is very small – only 1301 samples.

These are my key discoveries from fine-tuning on this dataset:

- For accurate summaries, mean cross-entropy loss needs to be < 0.4. The model hallucinates at higher losses, which can drastically alter the meaning of the clinical note.
    - With the language modeling head unfrozen or 38M trainable parameters, training loss is 0.07 after 15 epochs
    - With all layers unfrozen or 140M trainable parameters, training loss is 0.07 after 15 epochs (almost no difference from unfreezing only the language modeling head)
    - With the base model frozen and LoRA applied with rank=512 or 28M trainable parameters, training loss is 0.29 after 30 epochs
- In all cases, the model overfits. Validation loss increases as training progresses and never falls below 1. The dataset is too small and the task is too complex for the model to generalize well – applying dropout and weight decay degrades both training and validation loss. The model generates summaries that are in the correct format and may even contain correct information, but the model often hallucinates, either contradicting the ground truth or contradicting itself.
- The dataset is too small and too noisy for the complexity of the task. For a given sample, there aren't enough similar samples; some samples contain almost no information; and some samples have inaccurate ground truth summaries.

    - Training across 5-fold training/validation splits are nearly identical, which means the problem isn't splitting
    - Encoding the dialogue and summary using a SentenceTransformer and then plotting the encodings using t-SNE shows that the dataset is quite scattered. I break down a specific example.
    - I generated summaries using dataset team's model, and their model performs poorly

I made a lot of mistakes since this was my first NLP project. Coming into the project, I was confident in my understanding of transformers, but I realized how much I didn't understand once I started reading through Hugging Face's API documentation. I spent a lot of time looking for a dataset, but I didn't know what to look for. I had to rush and didn't spend enough time vetting the dataset.

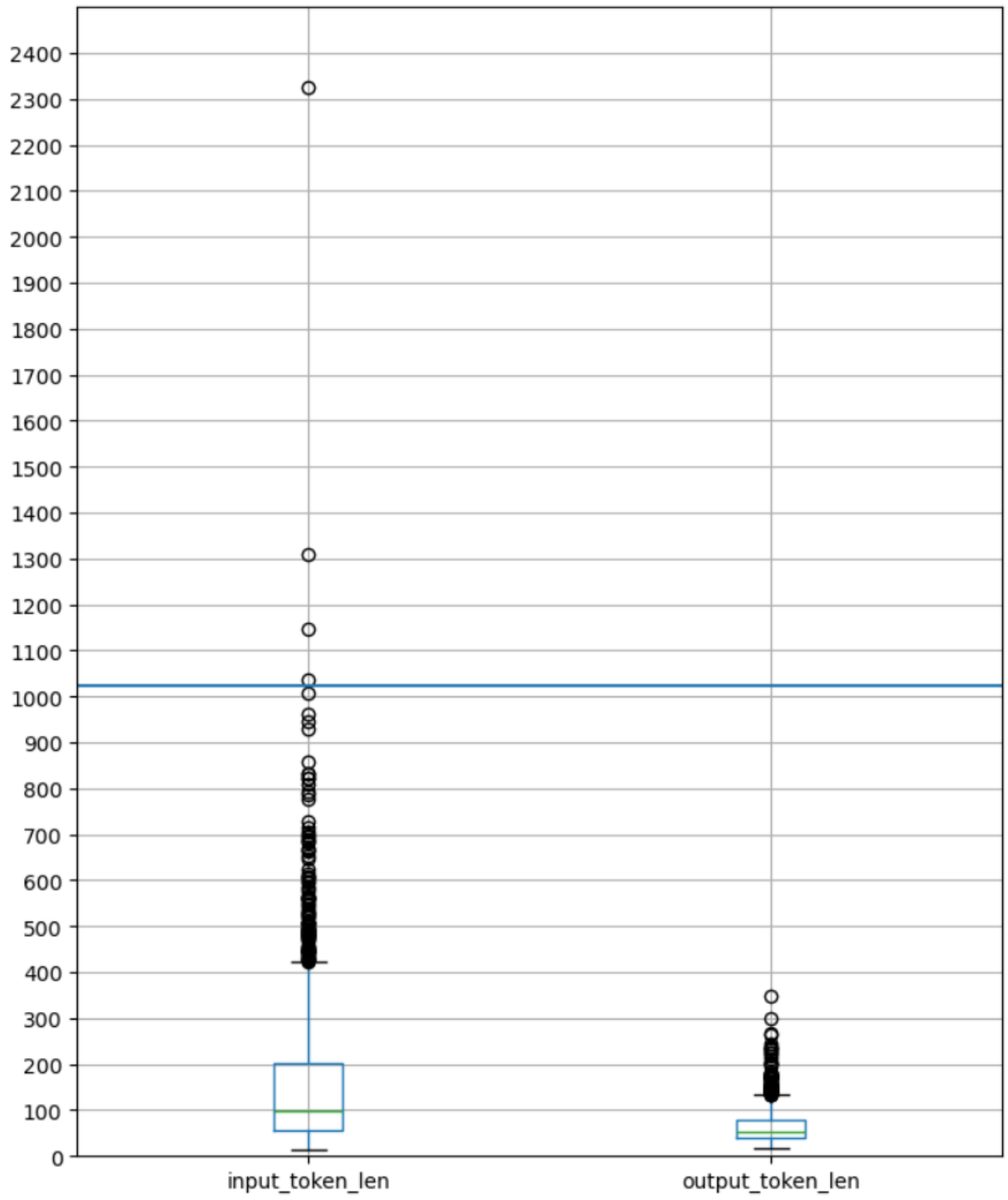
Miscellaneous notes:

- Generation strategy is greedy decoding with max generation length of 500 tokens

2 Dataset preprocessing

The dataset consists of 1301 samples. Each sample contains an ID, a section header that categorizes the clinical note and can be used for classification training, dialogue, and the dialogue summary.

After tokenizing with "facebook/bart-base", there are 4 samples where the encoder input sequence exceeds the model's maximum sequence length of 1024 tokens:



input_token_len = encoder input length; output_token_len = decoder target output length.

I removed the sample with the longest encoder input length and kept the others. I applied truncation from the left since the dialogue usually contains less information at the beginning (hi, how are you?, etc.).

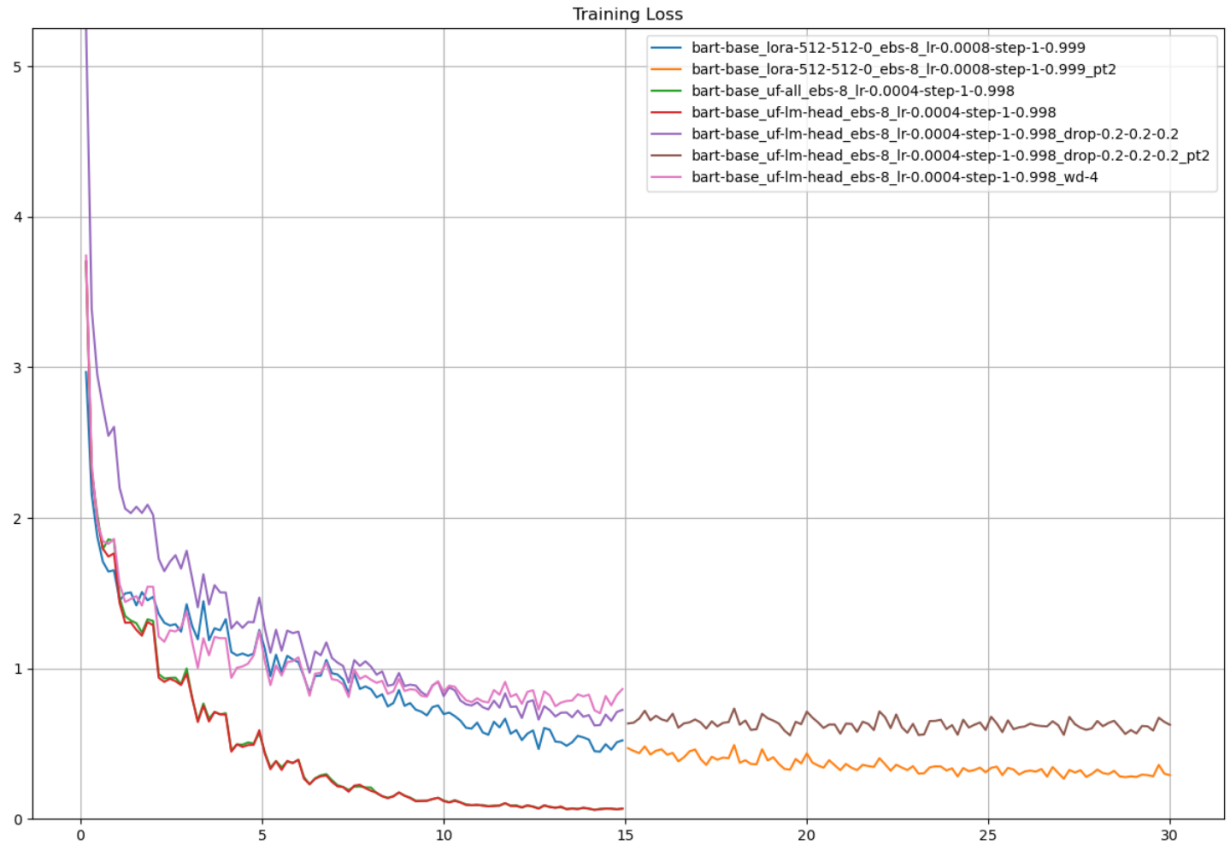
I split the dataset into train/val/test 80-10-10, splitting along the section header labels.

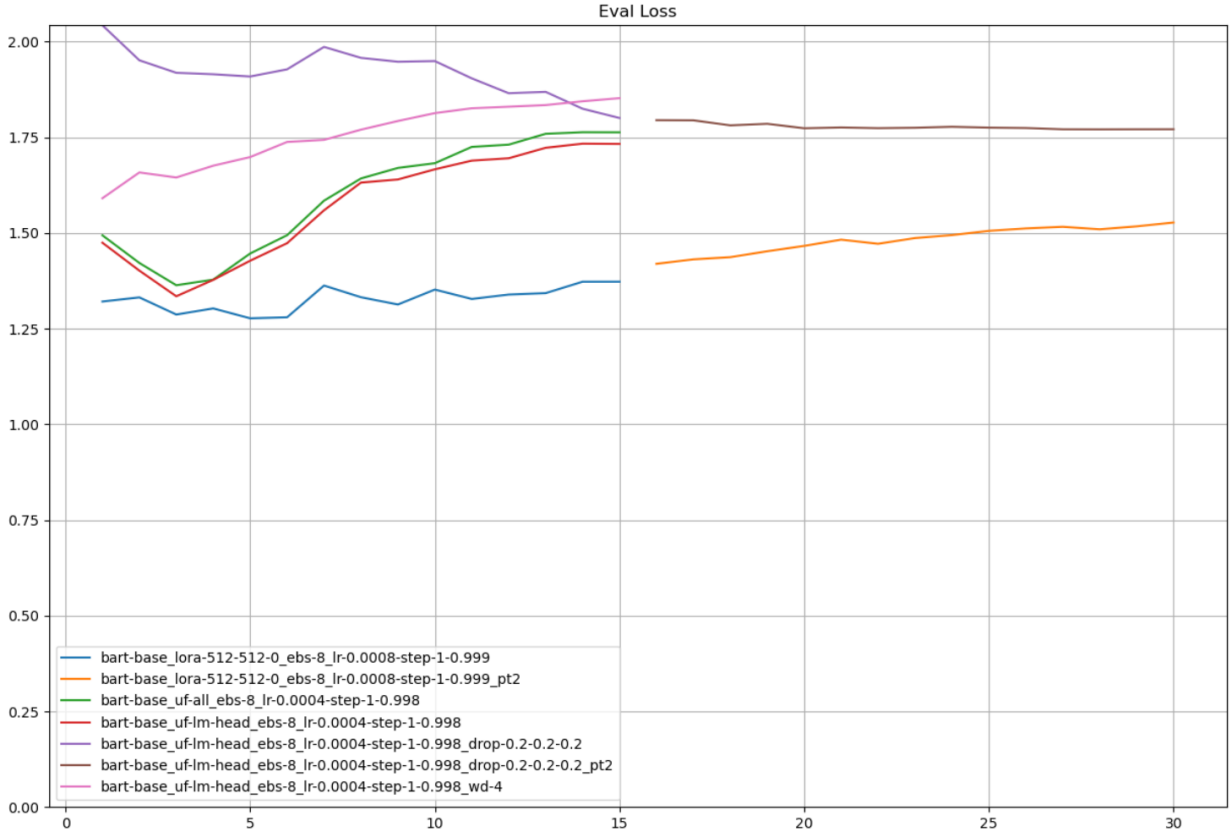
3 Training and validation

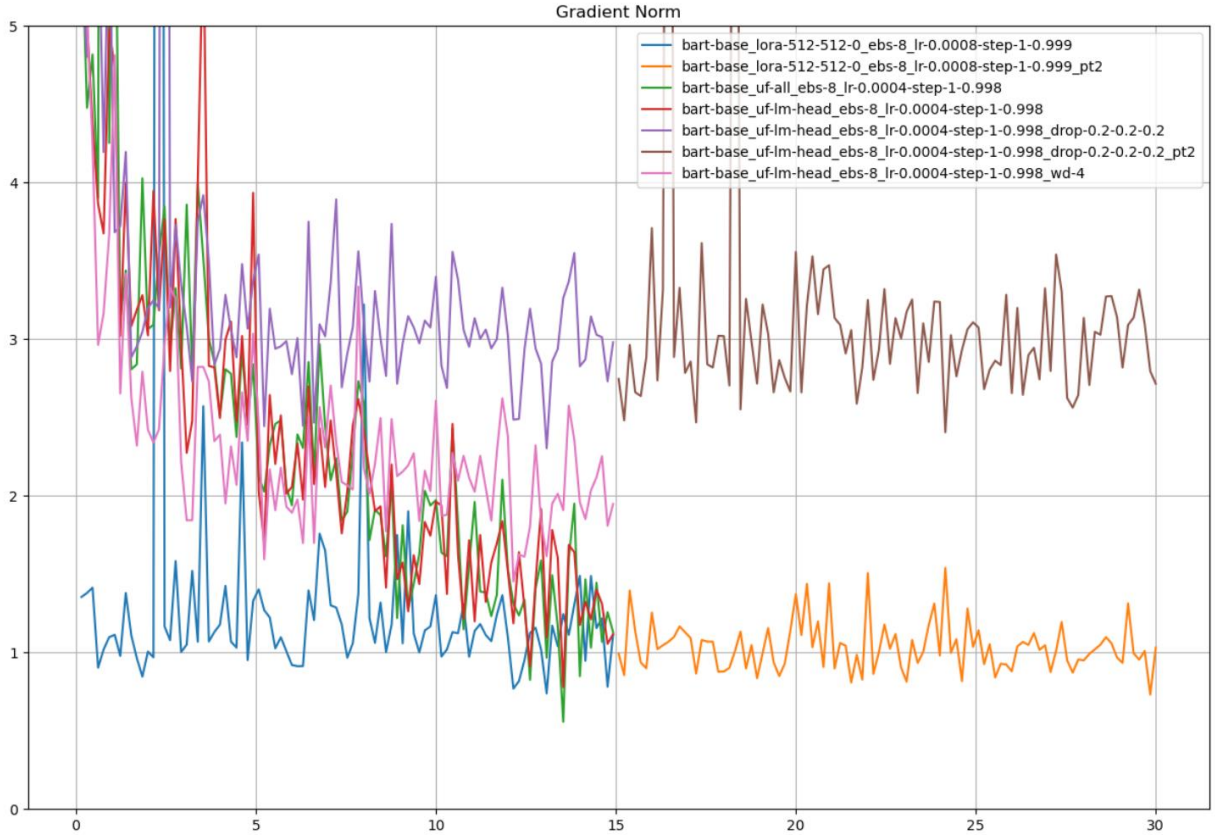
The plots below summarize the training results: I can achieve good training performance by either unfreezing layers in the base model or applying high-rank LoRA adaptation, but regularization degrades both training and validation performance.

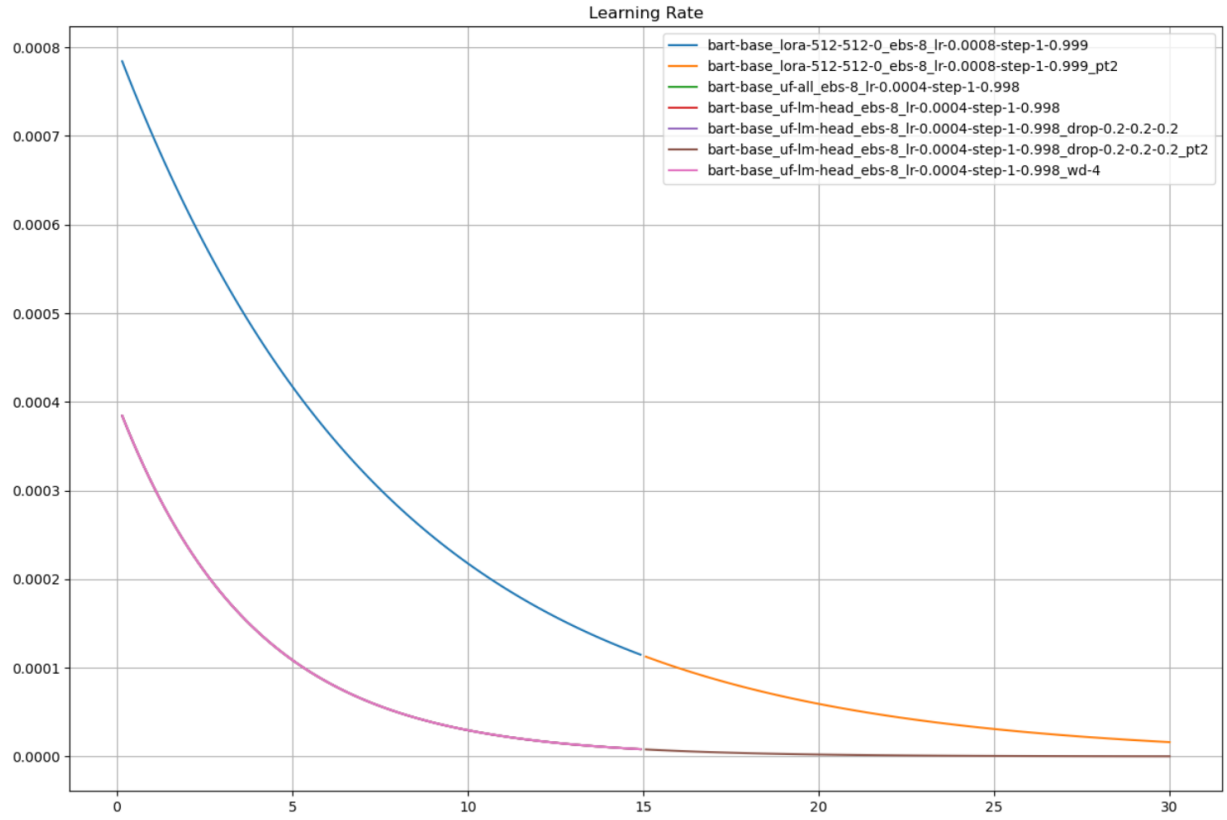
Model configurations (top to bottom in the legend):

1. LoRA (r = 512, alpha = 512, dropout = 0), effective batch size = 8, initial learning rate = 8e-4, StepLR(step_size = 1, gamma = 0.999)
2. Continuation of training from #1
3. Unfreeze all layers, effective batch size = 8, initial learning rate = 4e-4, StepLR(step_size = 1, gamma = 0.998)
4. Unfreeze language modeling head, effective batch size = 8, initial learning rate = 4e-4, StepLR(step_size = 1, gamma = 0.998)
5. Unfreeze language modeling head, effective batch size = 8, initial learning rate = 4e-4, StepLR(step_size = 1, gamma = 0.998), dropout = 0.2 (default 0.1)
6. Continuation of training from #5
7. Unfreeze language modeling head, effective batch size = 8, initial learning rate = 4e-4, StepLR(step_size = 1, gamma = 0.998), weight decay = 4 (default 0)

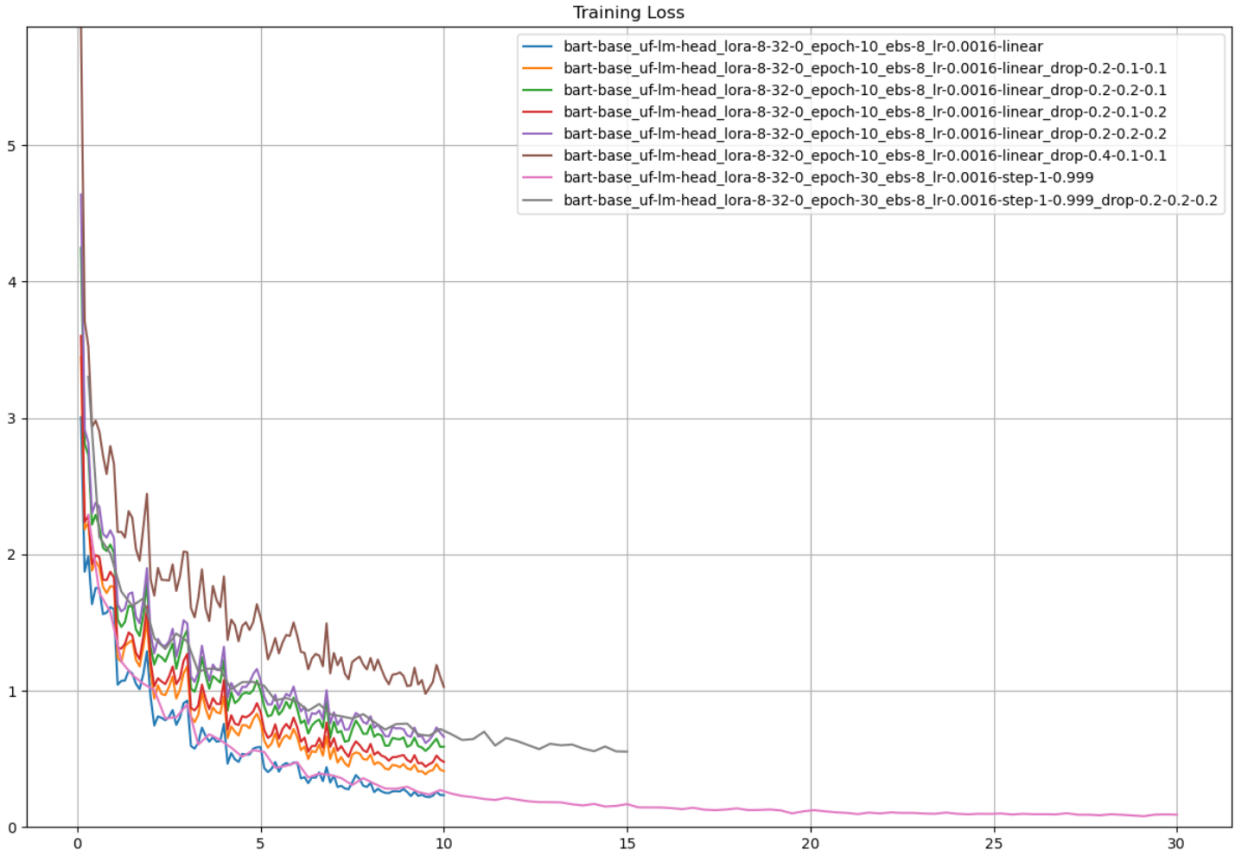

Training Loss
bart-base_lora-512-512-0_ebs-8_lr-0.0008-step-1-0.999
bart-base_lora-512-512-0_ebs-8_lr-0.0008-step-1-0.999_pt2
bart-base_uf-all_ebs-8_lr-0.0004-step-1-0.998
bart-base_uf-lm-head_ebs-8_lr-0.0004-step-1-0.998
bart-base_uf-lm-head_ebs-8_lr-0.0004-step-1-0.998_drop-0.2-0.2-0.2
bart-base_uf-lm-head_ebs-8_lr-0.0004-step-1-0.998_drop-0.2-0.2-0.2_pt2
bart-base_uf-lm-head_ebs-8_lr-0.0004-step-1-0.998_wd-4
0
1
2
3
4
5
0
5
10
15
20
25
30

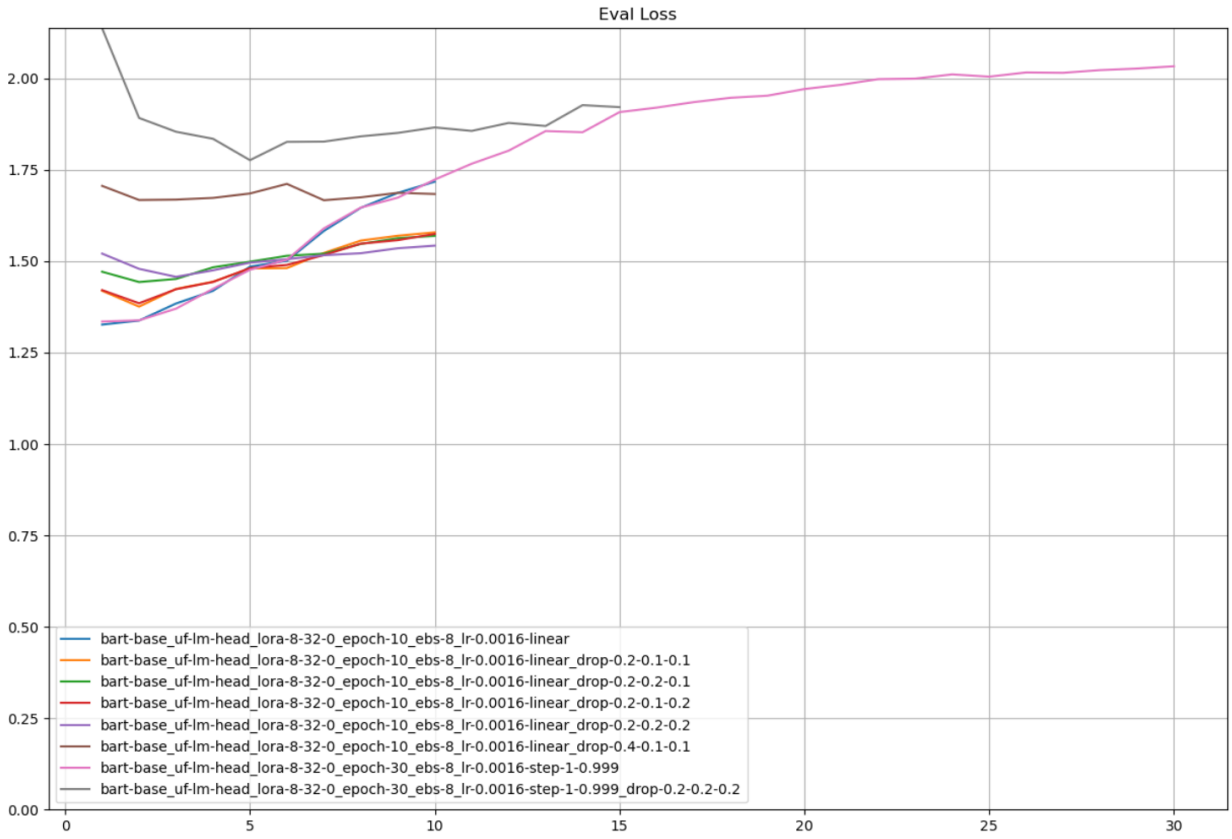

Eval Loss
bart-base_lora-512-512-0_ebs-8_lr-0.0008-step-1-0.999
bart-base_lora-512-512-0_ebs-8_lr-0.0008-step-1-0.999_pt2
bart-base_uf-all_ebs-8_lr-0.0004-step-1-0.998
bart-base_uf-lm-head_ebs-8_lr-0.0004-step-1-0.998
bart-base_uf-lm-head_ebs-8_lr-0.0004-step-1-0.998_drop-0.2-0.2-0.2
bart-base_uf-lm-head_ebs-8_lr-0.0004-step-1-0.998_drop-0.2-0.2-0.2_pt2
bart-base_uf-lm-head_ebs-8_lr-0.0004-step-1-0.998_wd-4
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
0
5
10
15
20
25
30

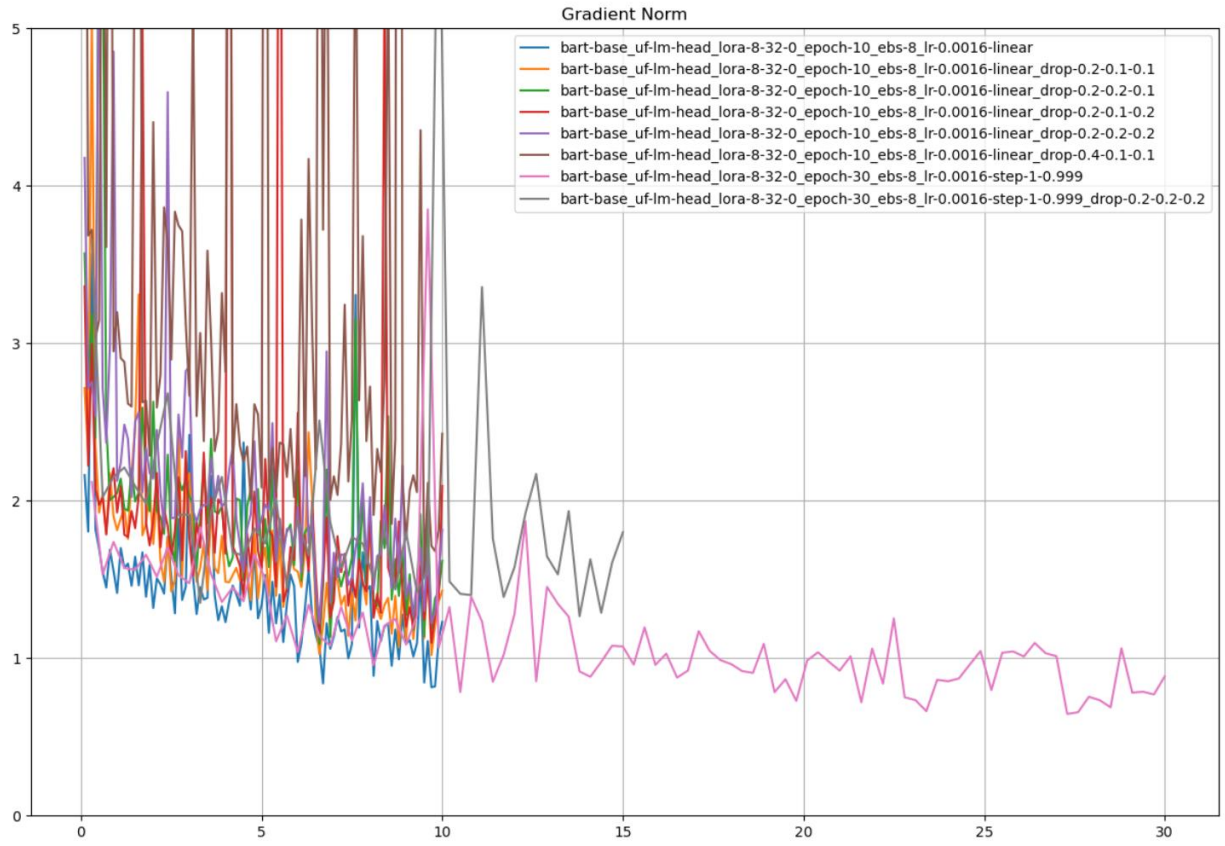

Gradient Norm
bart-base_lora-512-512-0_ebs-8_lr-0.0008-step-1-0.999
bart-base_lora-512-512-0_ebs-8_lr-0.0008-step-1-0.999_pt2
bart-base_uf-all_ebs-8_lr-0.0004-step-1-0.998
bart-base_uf-lm-head_ebs-8_lr-0.0004-step-1-0.998
bart-base_uf-lm-head_ebs-8_lr-0.0004-step-1-0.998_drop-0.2-0.2-0.2
bart-base_uf-lm-head_ebs-8_lr-0.0004-step-1-0.998_drop-0.2-0.2-0.2_pt2
bart-base_uf-lm-head_ebs-8_lr-0.0004-step-1-0.998_wd-4
0
1
2
3
4
5
0
5
10
15
20
25
30

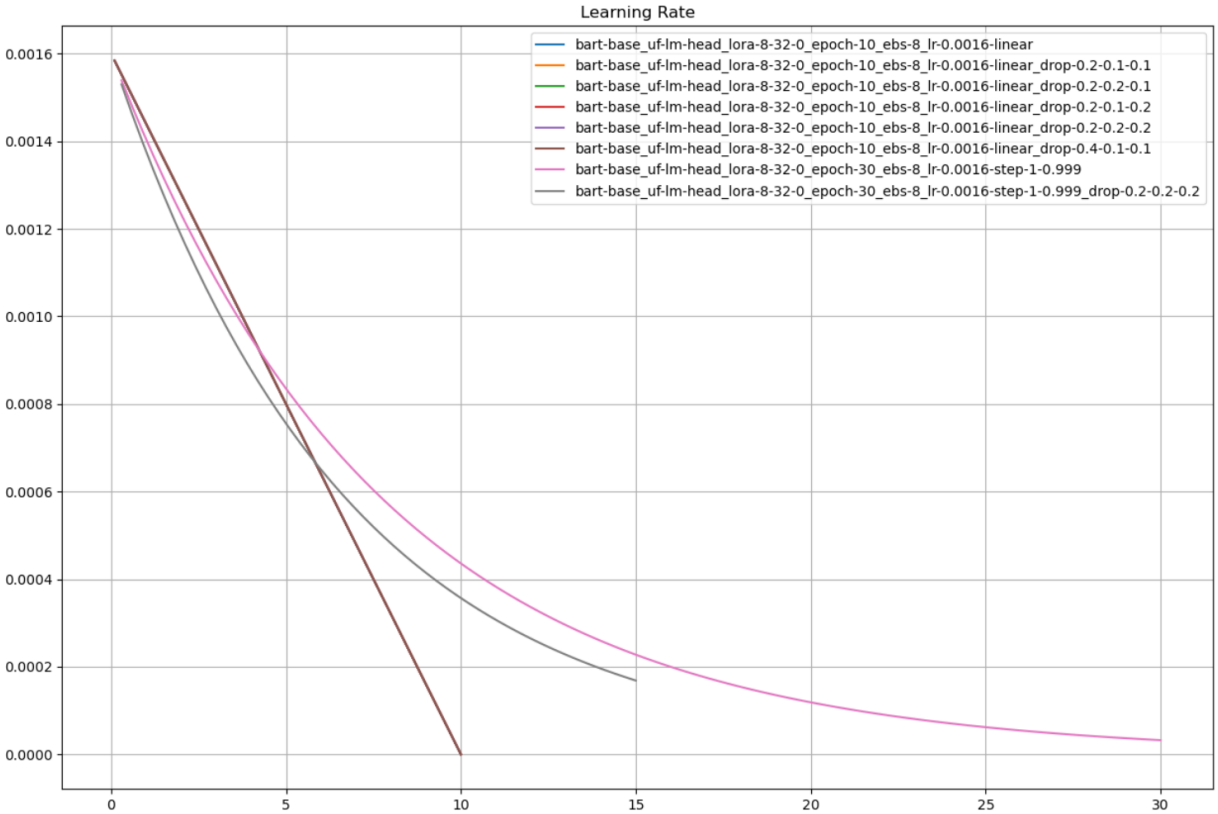

Learning Rate
bart-base_lora-512-512-0_ebs-8_lr-0.0008-step-1-0.999
bart-base_lora-512-512-0_ebs-8_lr-0.0008-step-1-0.999_pt2
bart-base_uf-all_ebs-8_lr-0.0004-step-1-0.998
bart-base_uf-lm-head_ebs-8_lr-0.0004-step-1-0.998
bart-base_uf-lm-head_ebs-8_lr-0.0004-step-1-0.998_drop-0.2-0.2-0.2
bart-base_uf-lm-head_ebs-8_lr-0.0004-step-1-0.998_drop-0.2-0.2-0.2_pt2
bart-base_uf-lm-head_ebs-8_lr-0.0004-step-1-0.998_wd-4
0.0000
0.0001
0.0002
0.0003
0.0004
0.0005
0.0006
0.0007
0.0008
0
5
10
15
20
25
30

3.1 Dropout experiments

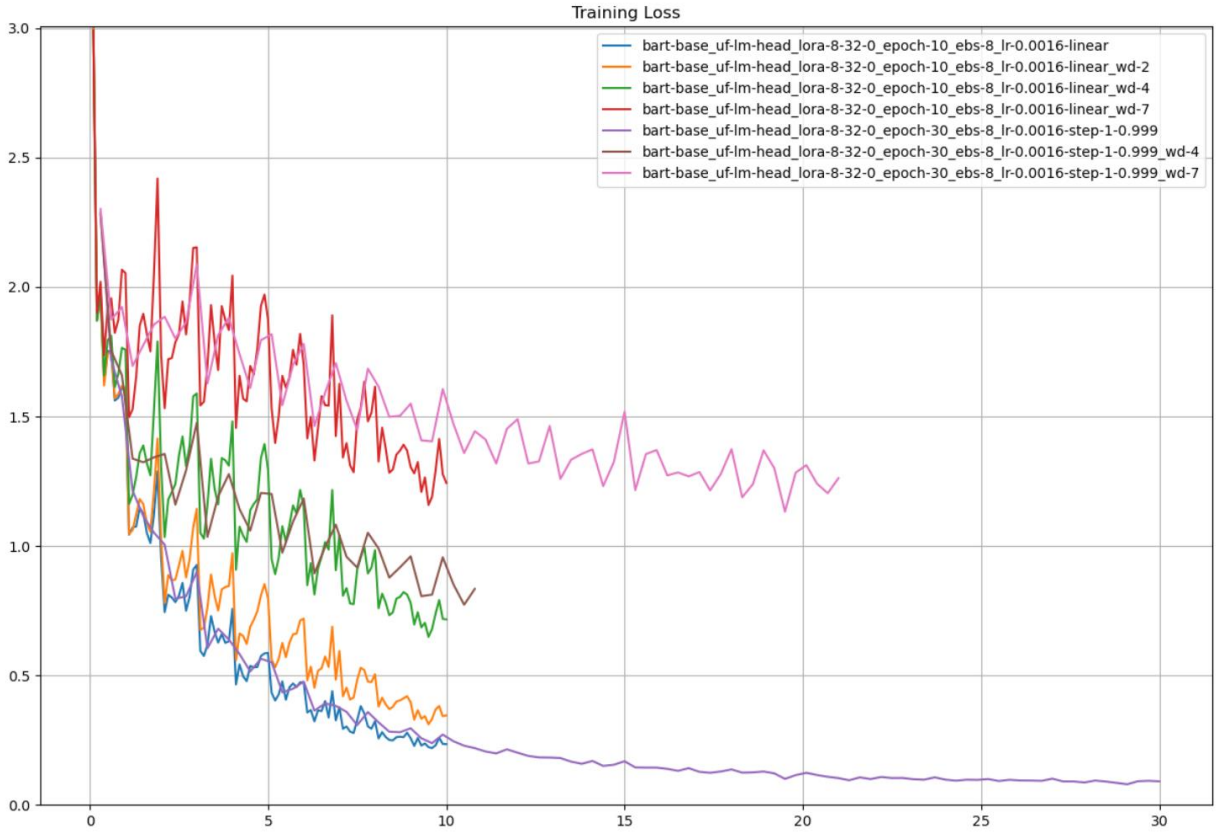


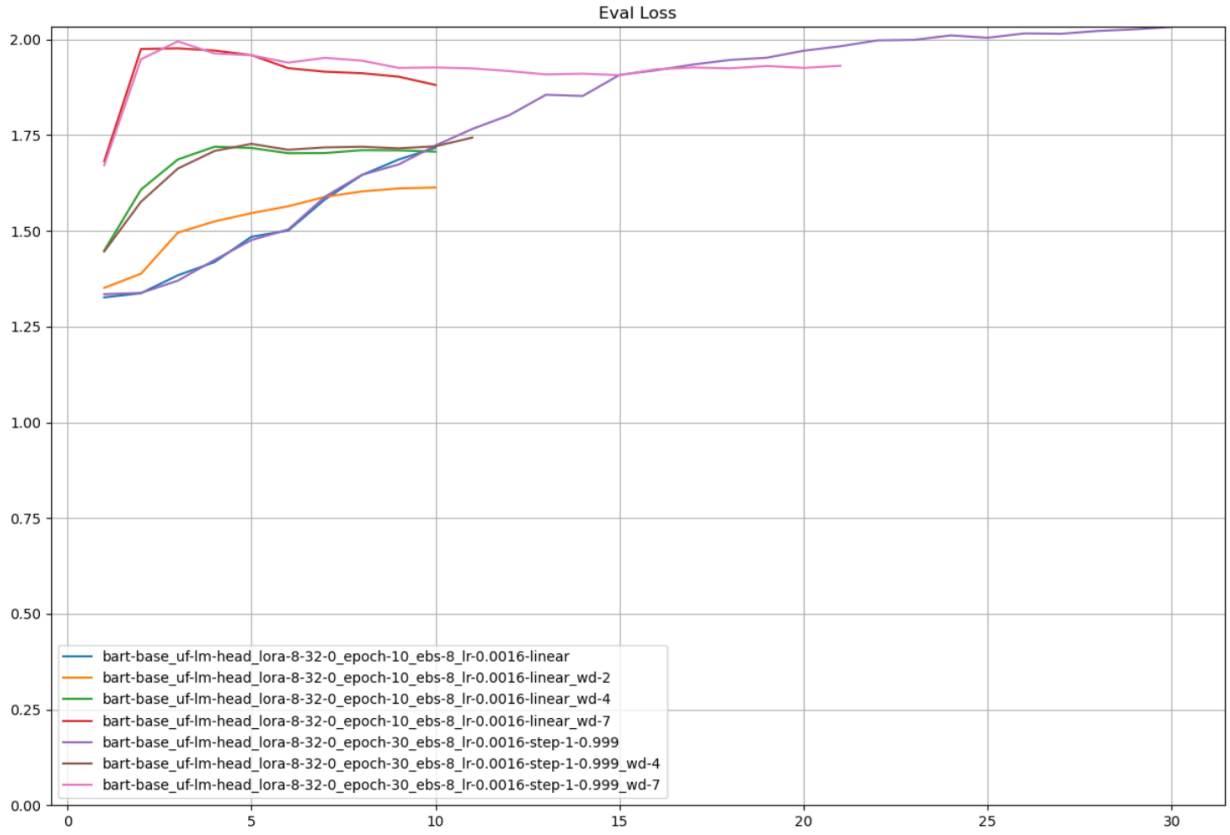

Eval Loss
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0
5
10
15
20
25
30
bart-base_uf-lm-head_lora-8-32-0_epoch-10_ebs-8_lr-0.0016-linear
bart-base_uf-lm-head_lora-8-32-0_epoch-10_ebs-8_lr-0.0016-linear_drop-0.2-0.1-0.1
bart-base_uf-lm-head_lora-8-32-0_epoch-10_ebs-8_lr-0.0016-linear_drop-0.2-0.2-0.1
bart-base_uf-lm-head_lora-8-32-0_epoch-10_ebs-8_lr-0.0016-linear_drop-0.2-0.1-0.2
bart-base_uf-lm-head_lora-8-32-0_epoch-10_ebs-8_lr-0.0016-linear_drop-0.2-0.2-0.2
bart-base_uf-lm-head_lora-8-32-0_epoch-10_ebs-8_lr-0.0016-linear_drop-0.4-0.1-0.1
bart-base_uf-lm-head_lora-8-32-0_epoch-30_ebs-8_lr-0.0016-step-1-0.999
bart-base_uf-lm-head_lora-8-32-0_epoch-30_ebs-8_lr-0.0016-step-1-0.999_drop-0.2-0.2-0.2

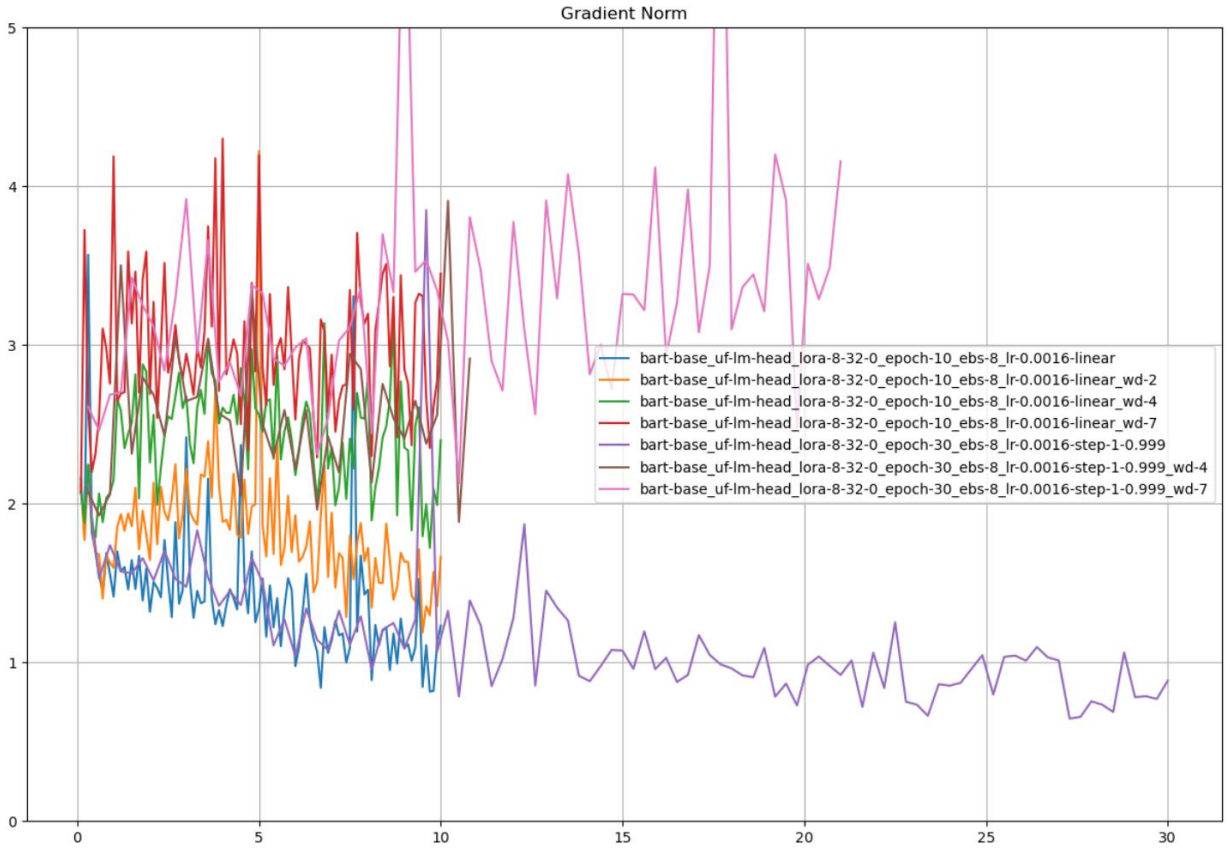

Gradient Norm
bart-base_uf-lm-head_lora-8-32-0_epoch-10_ebs-8_lr-0.0016-linear
bart-base_uf-lm-head_lora-8-32-0_epoch-10_ebs-8_lr-0.0016-linear_drop-0.2-0.1-0.1
bart-base_uf-lm-head_lora-8-32-0_epoch-10_ebs-8_lr-0.0016-linear_drop-0.2-0.2-0.1
bart-base_uf-lm-head_lora-8-32-0_epoch-10_ebs-8_lr-0.0016-linear_drop-0.2-0.1-0.2
bart-base_uf-lm-head_lora-8-32-0_epoch-10_ebs-8_lr-0.0016-linear_drop-0.2-0.2-0.2
bart-base_uf-lm-head_lora-8-32-0_epoch-10_ebs-8_lr-0.0016-linear_drop-0.4-0.1-0.1
bart-base_uf-lm-head_lora-8-32-0_epoch-30_ebs-8_lr-0.0016-step-1-0.999
bart-base_uf-lm-head_lora-8-32-0_epoch-30_ebs-8_lr-0.0016-step-1-0.999_drop-0.2-0.2-0.2
0
1
2
3
4
5
0
5
10
15
20
25
30

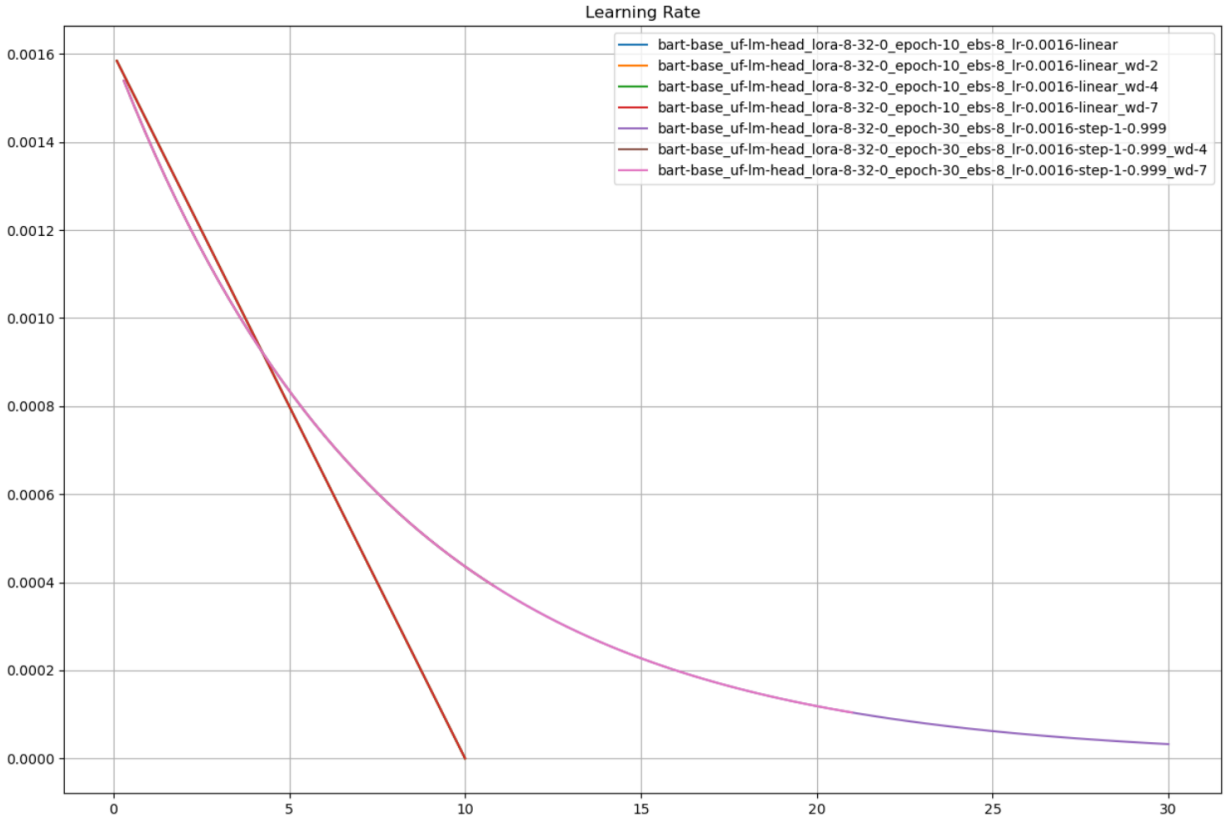

Learning Rate
bart-base_uf-lm-head_lora-8-32-0_epoch-10_ebs-8_lr-0.0016-linear
bart-base_uf-lm-head_lora-8-32-0_epoch-10_ebs-8_lr-0.0016-linear_drop-0.2-0.1-0.1
bart-base_uf-lm-head_lora-8-32-0_epoch-10_ebs-8_lr-0.0016-linear_drop-0.2-0.2-0.1
bart-base_uf-lm-head_lora-8-32-0_epoch-10_ebs-8_lr-0.0016-linear_drop-0.2-0.1-0.2
bart-base_uf-lm-head_lora-8-32-0_epoch-10_ebs-8_lr-0.0016-linear_drop-0.2-0.2-0.2
bart-base_uf-lm-head_lora-8-32-0_epoch-10_ebs-8_lr-0.0016-linear_drop-0.4-0.1-0.1
bart-base_uf-lm-head_lora-8-32-0_epoch-30_ebs-8_lr-0.0016-step-1-0.999
bart-base_uf-lm-head_lora-8-32-0_epoch-30_ebs-8_lr-0.0016-step-1-0.999_drop-0.2-0.2-0.2
0.0000
0.0002
0.0004
0.0006
0.0008
0.0010
0.0012
0.0014
0.0016
0
5
10
15
20
25
30

3.2 Weight decay experiments

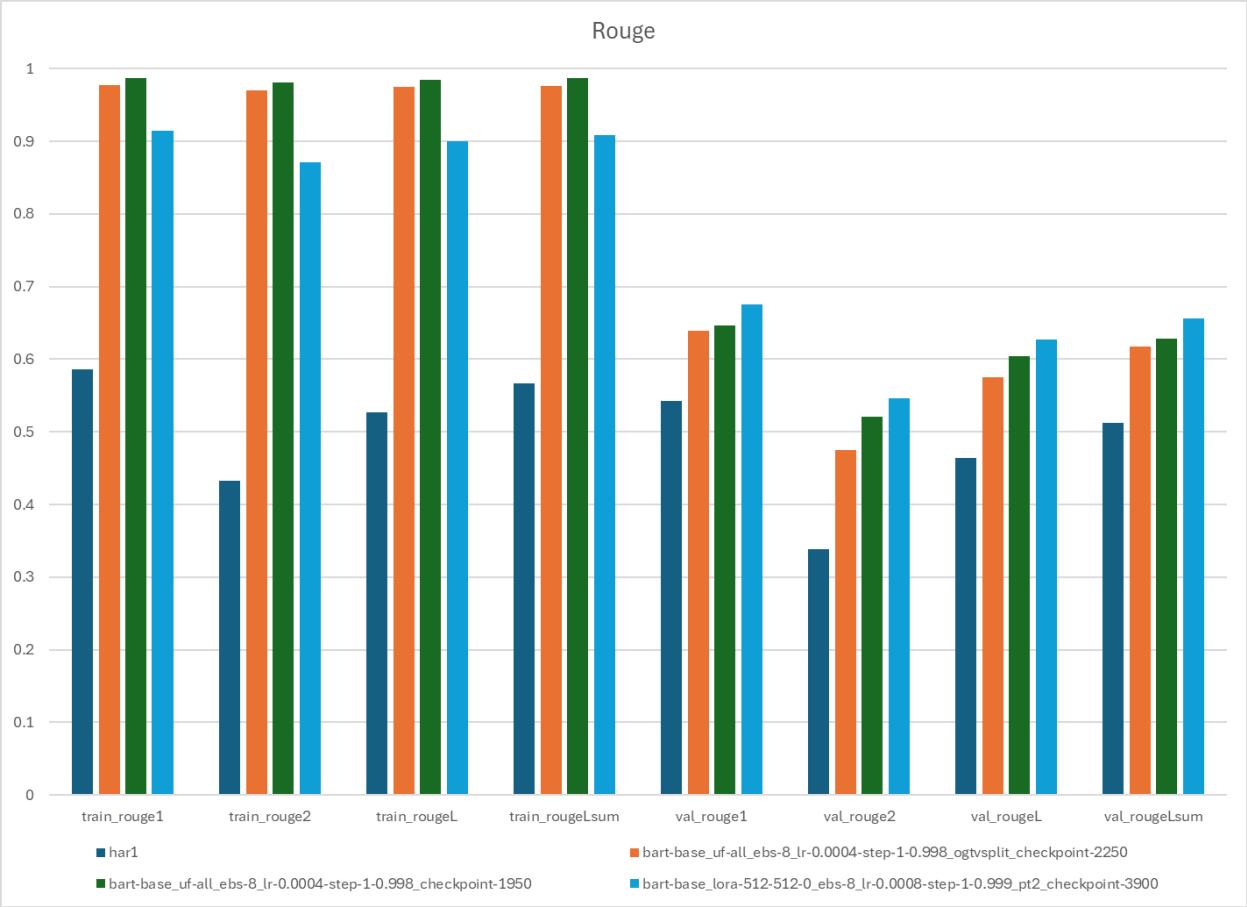


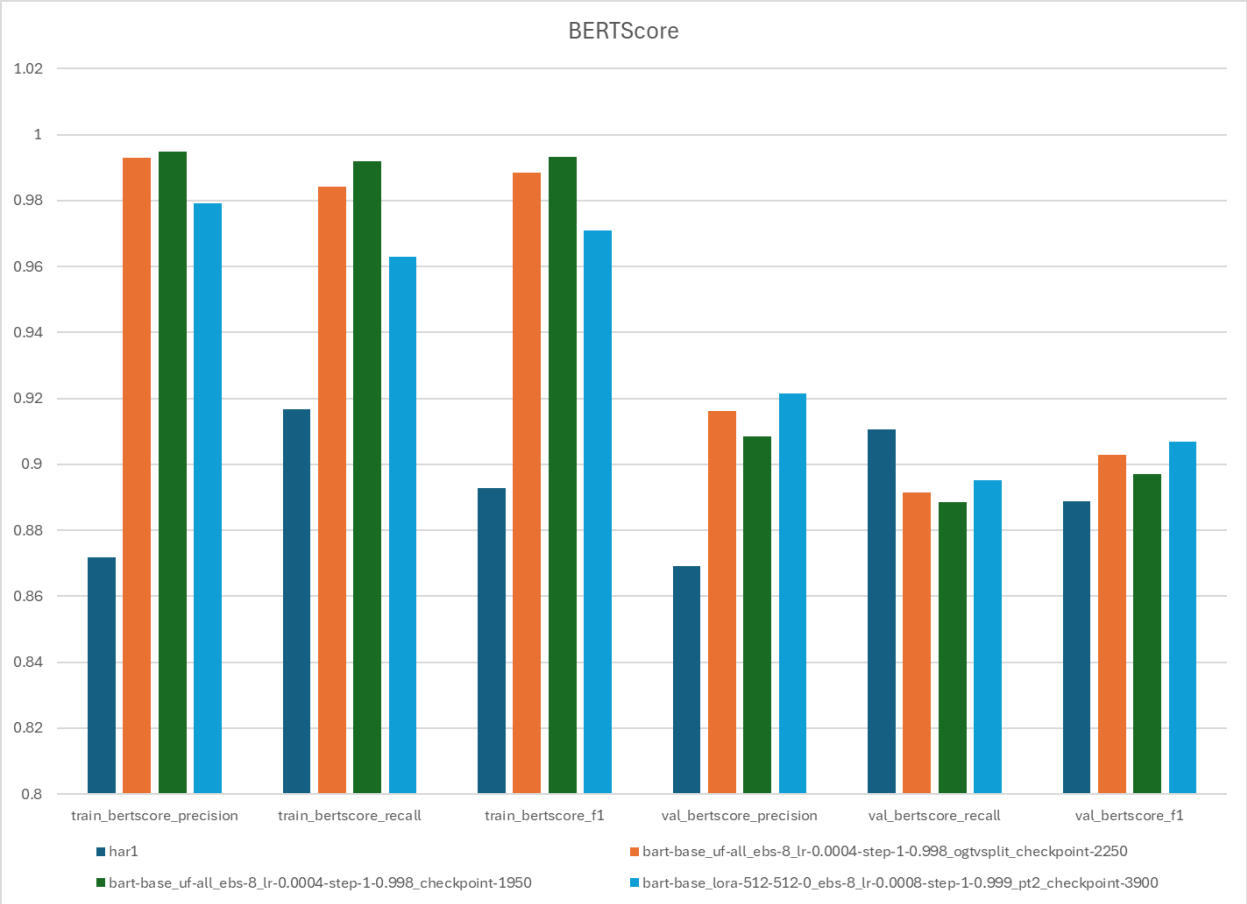

Eval Loss
bart-base_uf-lm-head_lora-8-32-0_epoch-10_ebs-8_lr-0.0016-linear
bart-base_uf-lm-head_lora-8-32-0_epoch-10_ebs-8_lr-0.0016-linear_wd-2
bart-base_uf-lm-head_lora-8-32-0_epoch-10_ebs-8_lr-0.0016-linear_wd-4
bart-base_uf-lm-head_lora-8-32-0_epoch-10_ebs-8_lr-0.0016-linear_wd-7
bart-base_uf-lm-head_lora-8-32-0_epoch-30_ebs-8_lr-0.0016-step-1-0.999
bart-base_uf-lm-head_lora-8-32-0_epoch-30_ebs-8_lr-0.0016-step-1-0.999_wd-4
bart-base_uf-lm-head_lora-8-32-0_epoch-30_ebs-8_lr-0.0016-step-1-0.999_wd-7


Gradient Norm
bart-base_uf-lm-head_lora-8-32-0_epoch-10_ebs-8_lr-0.0016-linear
bart-base_uf-lm-head_lora-8-32-0_epoch-10_ebs-8_lr-0.0016-linear_wd-2
bart-base_uf-lm-head_lora-8-32-0_epoch-10_ebs-8_lr-0.0016-linear_wd-4
bart-base_uf-lm-head_lora-8-32-0_epoch-10_ebs-8_lr-0.0016-linear_wd-7
bart-base_uf-lm-head_lora-8-32-0_epoch-30_ebs-8_lr-0.0016-step-1-0.999
bart-base_uf-lm-head_lora-8-32-0_epoch-30_ebs-8_lr-0.0016-step-1-0.999_wd-4
bart-base_uf-lm-head_lora-8-32-0_epoch-30_ebs-8_lr-0.0016-step-1-0.999_wd-7


Learning Rate
bart-base_uf-lm-head_lora-8-32-0_epoch-10_ebs-8_lr-0.0016-linear
bart-base_uf-lm-head_lora-8-32-0_epoch-10_ebs-8_lr-0.0016-linear_wd-2
bart-base_uf-lm-head_lora-8-32-0_epoch-10_ebs-8_lr-0.0016-linear_wd-4
bart-base_uf-lm-head_lora-8-32-0_epoch-10_ebs-8_lr-0.0016-linear_wd-7
bart-base_uf-lm-head_lora-8-32-0_epoch-30_ebs-8_lr-0.0016-step-1-0.999
bart-base_uf-lm-head_lora-8-32-0_epoch-30_ebs-8_lr-0.0016-step-1-0.999_wd-4
bart-base_uf-lm-head_lora-8-32-0_epoch-30_ebs-8_lr-0.0016-step-1-0.999_wd-7
0.0016
0.0014
0.0012
0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
0
5
10
15
20
25
30

4 Inference


Rouge
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
train_rouge1
train_rouge2
train_rougeL
train_rougeLsum
val_rouge1
val_rouge2
val_rougeL
val_rougeLsum
har1
bart-base_uf-all_ebs-8_lr-0.0004-step-1-0.998_ogtvsplit_checkpoint-2250
bart-base_uf-all_ebs-8_lr-0.0004-step-1-0.998_checkpoint-1950
bart-base_lora-512-512-0_ebs-8_lr-0.0008-step-1-0.999_pt2_checkpoint-3900



I ran inference on the training and validation samples for 4 different models:

1. har1: this is the model trained by the team that prepared and uploaded the dataset to Hugging Face
2. bart-base_uf-all_ebs-8_lr-0.0004-step-1-0.998_checkpoint-1950: BART, unfreeze all layers (training loss = 0.07, validation loss = 1.76)
3. bart-base_uf-all_ebs-8_lr-0.0004-step-1-0.998_ogtvsplit_checkpoint-2250: same model as #2 but different training/validation split (training loss = 0.08, validation loss = 1.99)
4. bart-base_lora-512-512-0_ebs-8_lr-0.0008-step-1-0.999_pt2_checkpoint-3900: BART (frozen), LoRA (rank = 512, alpha = 512, dropout = 0), training loss = 0.29, validation loss = 1.53

Performance:

1. har1 performs poorly on both training and validation samples
2. BART with all layers unfrozen performs very well on training and poorly on validation

3. Compared to BART with all layers unfrozen, BART with LoRA performs slightly worse on training and slightly better on validation, which tracks the training and validation loss

What do “good” and “poor” performance mean? Let’s look at validation sample ID 209:

Doctor: Good morning, young man. Are these your parents?
Patient: Yes.
Doctor: Good, can you tell me more about your son, please?
Guest_family_1: Well, he's five now, and **he fell onto his right arm** on December fifth two thousand seven.
Doctor: After he fell, how was he treated?
Guest_family_1: We went to the E D right after he fell, and they said he had **complete fractures of both bones in the arm.**
Doctor: Yes, I see that here, he also has shortening bayonet apposition.
Guest_family_1: What can we do for this?
Doctor: There's actually a few options here. First we can cast it and see how he heals, generally, children heal up very well from fractures.
Guest_family_1: That's good, we like that option more than any kind of surgery.
Doctor: However, **surgery is also an option here as well.**
Guest_family_1: Yeah, to be completely sure we fix this, I think we should opt for the surgery, what do you think, honey?
Guest_family_2: Yes, I agree. What are the risks of infection for this surgery?
Doctor: The **risk of infection is very low**, generally less than one percent. We use antibiotics to control for infection.
Guest_family_1: Will he be asleep for the surgery?
Doctor: Absolutely, he won't feel a thing. **Other risks include bleeding, changes in sensation and motion of the extremity, hardware failure, and need for later hardware removal, and cast tightness. I would not worry about these risks. We have great results with these surgeries.**
Guest_family_1: Then **yes, we'd like to do the surgery.**

Symptoms: **refracture** of right forearm
Diagnosis: complete fractures of both bones in right arm, shortening bayonet apposition
History of Patient: fell onto right arm on December 5, 2007
Plan of Action: surgery for **closed reduction and pinning**, with risks including anesthesia, infection, bleeding, changes in sensation and motion of extremity, hardware failure, need for later hardware removal, cast tightness

The model’s generated summary is

Symptoms: fell onto right arm on December 5, 2007, **incomplete fractures** of both bones in right arm, shortening bayonet apposition, possible need for later hardware removal and possible continuous nerve Symptoms
Diagnosis: N/A
History of Patient: fell onto right arm on December 5, 2007, **incomplete fractures of both bones, received physical therapy and bandaging, no other surgeries recommended due to risk of infection**
Plan of Action: Surgery recommended due to **non-healing extremity injuries**, risks of infection discussed and patient agreed to surgery after informed consent

There are a couple of problems here. First, the model is hallucinating, and second, the ground truth summary does not completely follow from the dialogue. There is information in the summary that isn't present in the dialogue.

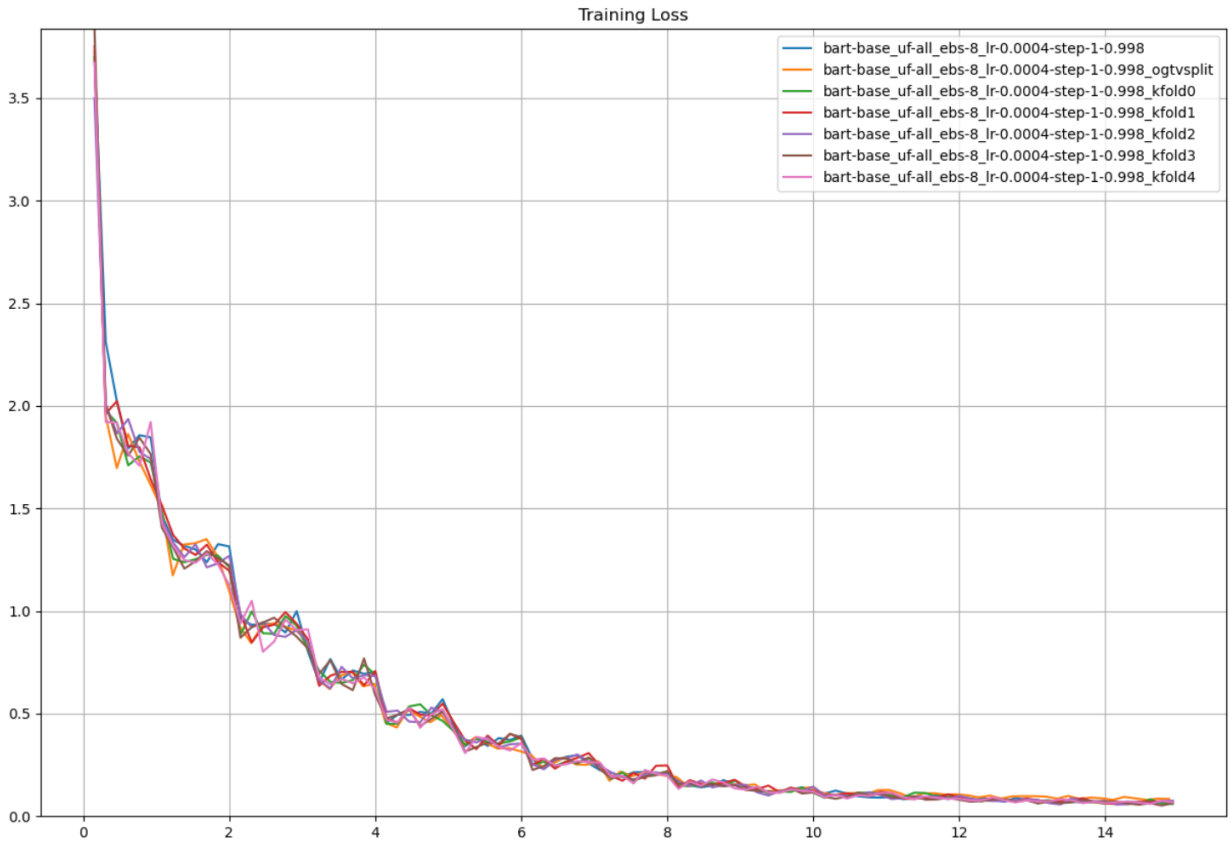
5 Dataset investigation

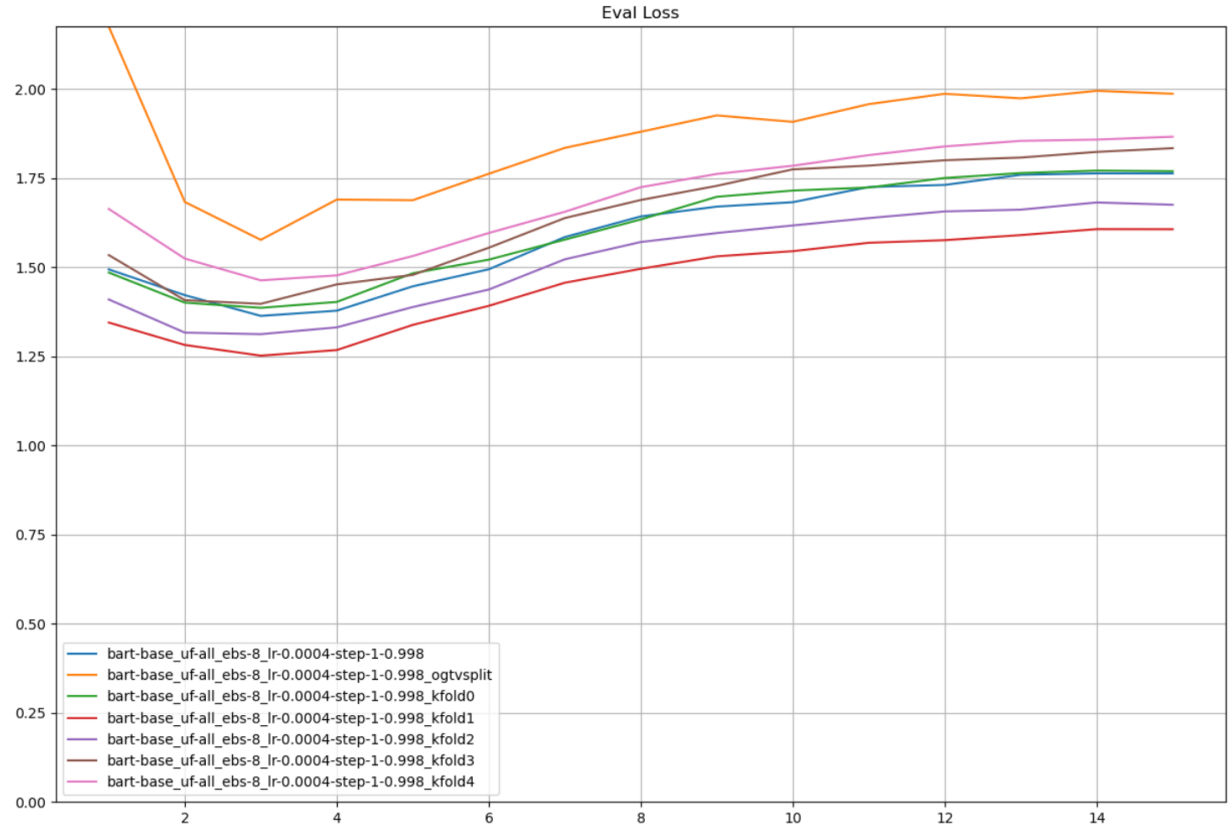
Since adding regularization to the training process does not improve validation performance, I suspected that the dataset is the problem. I tried a couple of experiments:

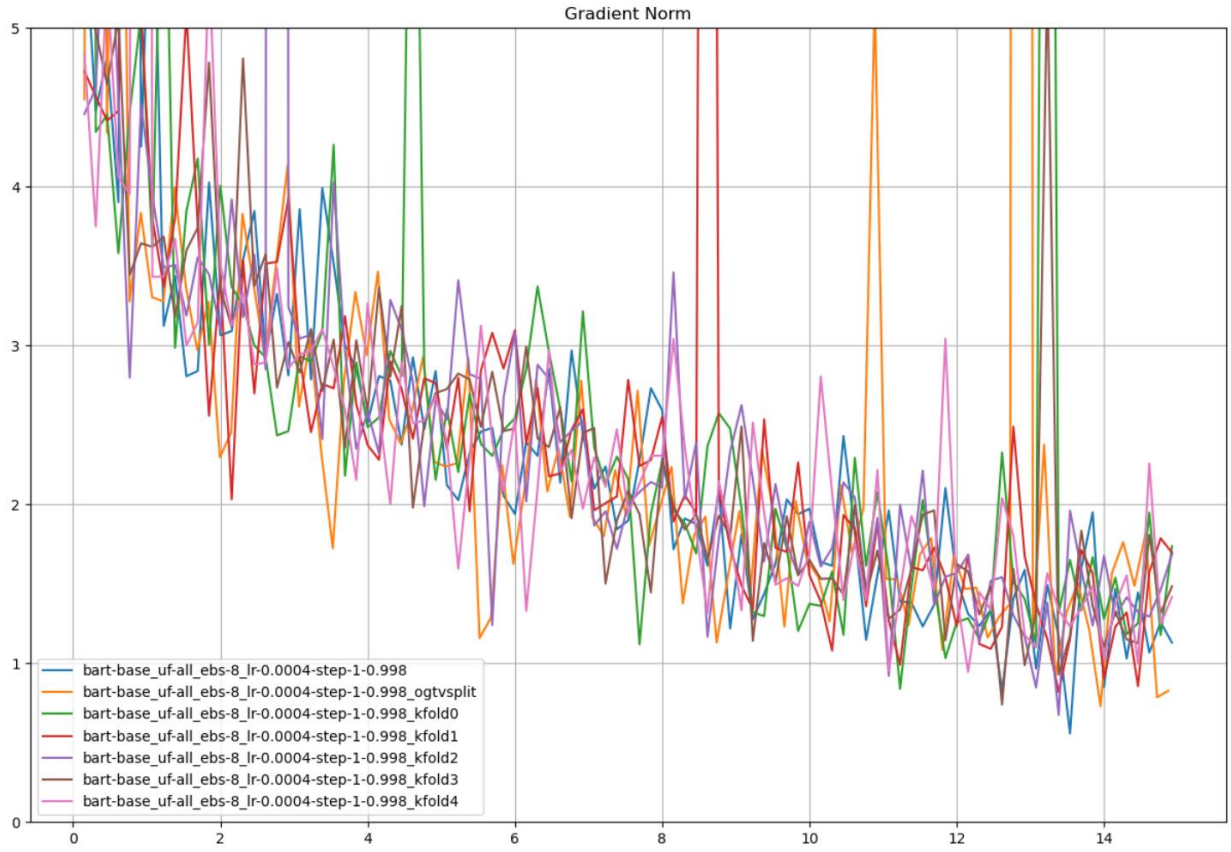
1. Run training on 5 different train/val folds to see if it's a problem in the distribution of samples between training and validation. Training and validation loss is nearly identical across all folds, so splitting isn't the problem.
2. Encode the dialogue and summaries using a SentenceTransformer and plot the dataset in (x, y) coordinate space using t-SNE to visualize the distribution of samples. For nearly all validation and test samples, there are similar training samples to learn from, but the problem is that the number of training samples is too low for the complexity of the task. In the example outlined below, there are only 8 samples related to fractures and surgery. Additionally, I found noisy samples and samples that contain almost no information.

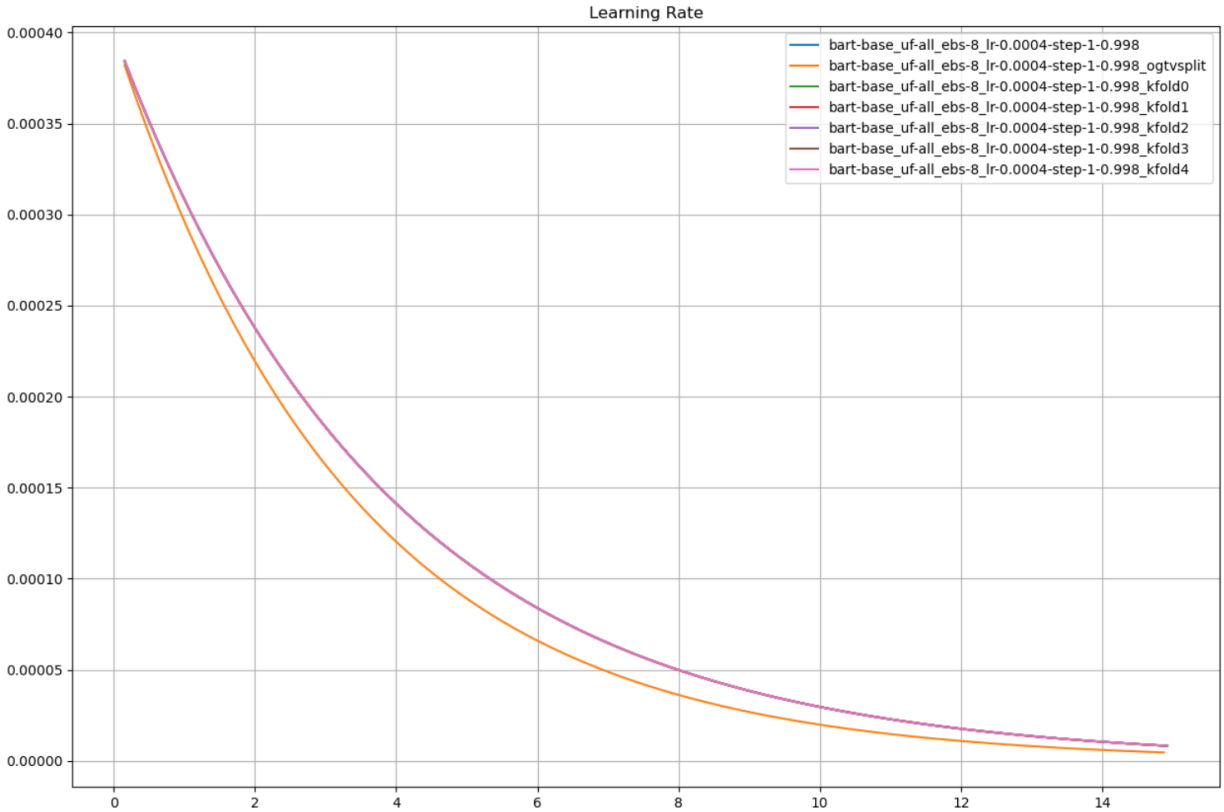
5.1 5-fold cross validation

I trained for 7 different splits, including the 5-fold splits.


Training Loss
bart-base_uf-all_ebs-8_lr-0.0004-step-1-0.998
bart-base_uf-all_ebs-8_lr-0.0004-step-1-0.998_ogtvsplit
bart-base_uf-all_ebs-8_lr-0.0004-step-1-0.998_kfold0
bart-base_uf-all_ebs-8_lr-0.0004-step-1-0.998_kfold1
bart-base_uf-all_ebs-8_lr-0.0004-step-1-0.998_kfold2
bart-base_uf-all_ebs-8_lr-0.0004-step-1-0.998_kfold3
bart-base_uf-all_ebs-8_lr-0.0004-step-1-0.998_kfold4
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0
2
4
6
8
10
12
14


Eval Loss
bart-base_uf-all_ebs-8_lr-0.0004-step-1-0.998
bart-base_uf-all_ebs-8_lr-0.0004-step-1-0.998_ogtvsplit
bart-base_uf-all_ebs-8_lr-0.0004-step-1-0.998_kfold0
bart-base_uf-all_ebs-8_lr-0.0004-step-1-0.998_kfold1
bart-base_uf-all_ebs-8_lr-0.0004-step-1-0.998_kfold2
bart-base_uf-all_ebs-8_lr-0.0004-step-1-0.998_kfold3
bart-base_uf-all_ebs-8_lr-0.0004-step-1-0.998_kfold4
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
2
4
6
8
10
12
14

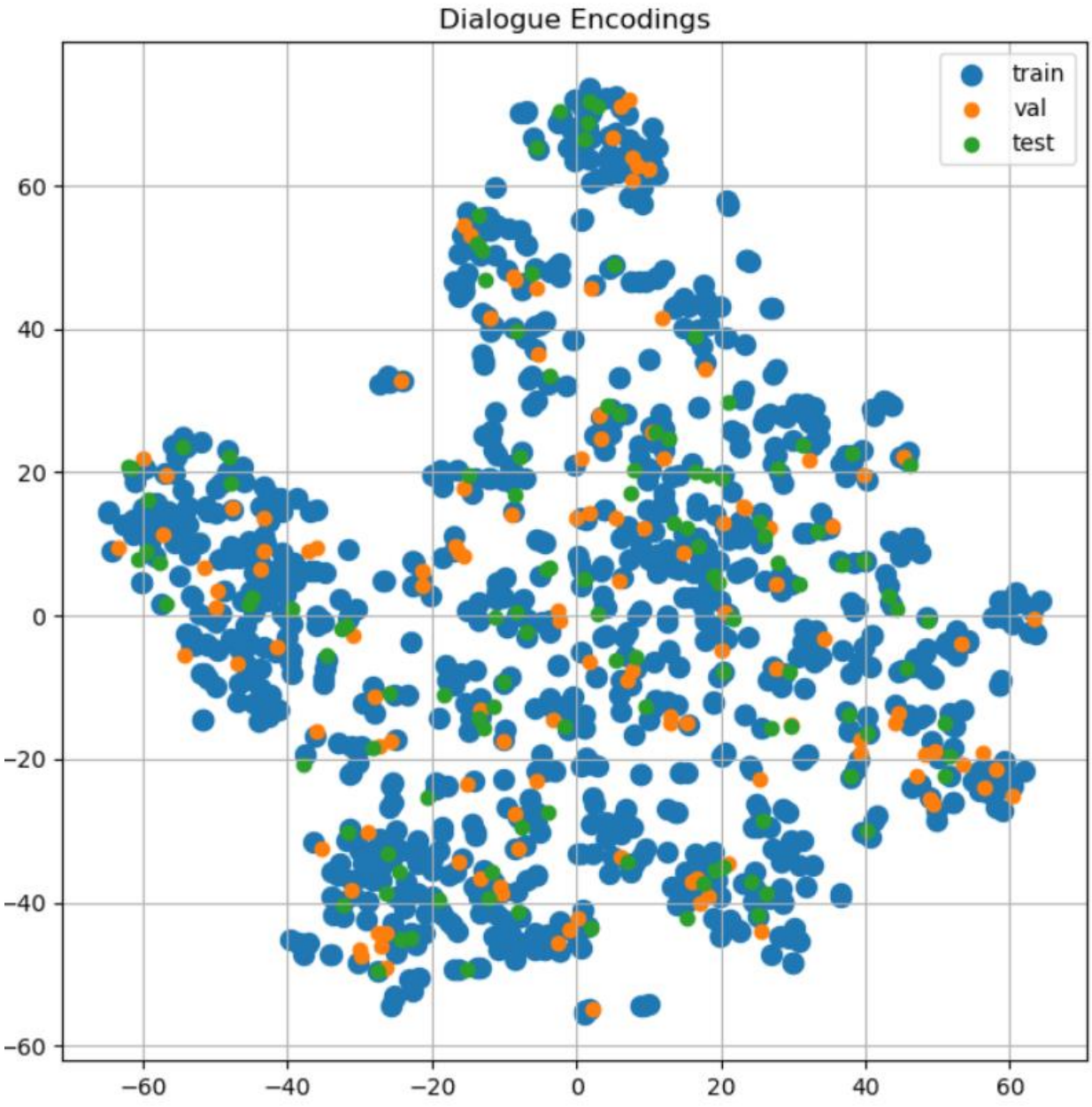

Gradient Norm
bart-base_uf-all_ebs-8_lr-0.0004-step-1-0.998
bart-base_uf-all_ebs-8_lr-0.0004-step-1-0.998_ogtvsplit
bart-base_uf-all_ebs-8_lr-0.0004-step-1-0.998_kfold0
bart-base_uf-all_ebs-8_lr-0.0004-step-1-0.998_kfold1
bart-base_uf-all_ebs-8_lr-0.0004-step-1-0.998_kfold2
bart-base_uf-all_ebs-8_lr-0.0004-step-1-0.998_kfold3
bart-base_uf-all_ebs-8_lr-0.0004-step-1-0.998_kfold4

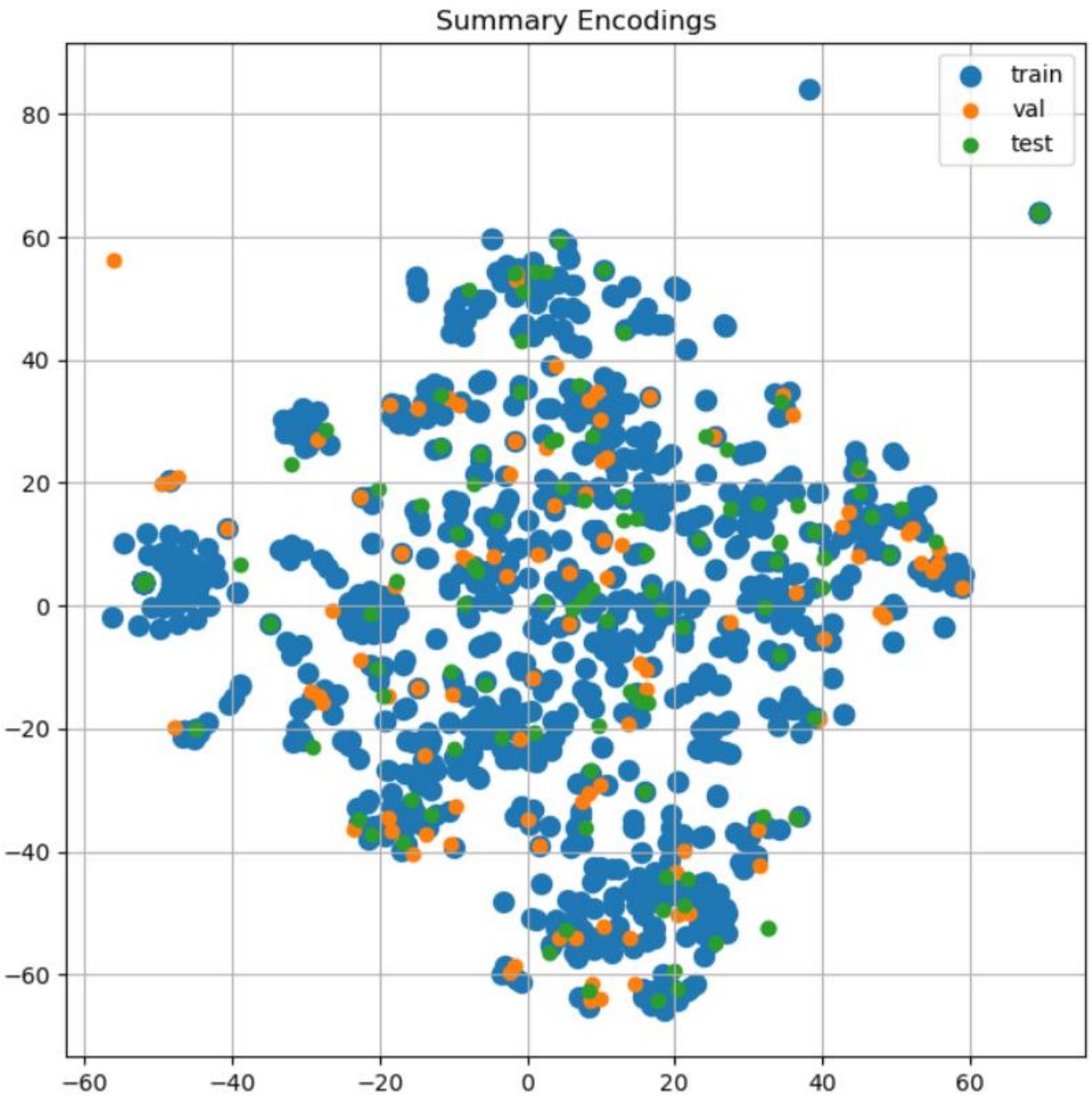

Learning Rate
bart-base_uf-all_ebs-8_lr-0.0004-step-1-0.998
bart-base_uf-all_ebs-8_lr-0.0004-step-1-0.998_ogtvsplit
bart-base_uf-all_ebs-8_lr-0.0004-step-1-0.998_kfold0
bart-base_uf-all_ebs-8_lr-0.0004-step-1-0.998_kfold1
bart-base_uf-all_ebs-8_lr-0.0004-step-1-0.998_kfold2
bart-base_uf-all_ebs-8_lr-0.0004-step-1-0.998_kfold3
bart-base_uf-all_ebs-8_lr-0.0004-step-1-0.998_kfold4

5.2 SentenceTransformer + t-SNE

I used the SentenceTransformer "all-distilroberta-v1" since it supports the longest input sequence among Hugging Face models (512 tokens). I removed the 55 samples with dialogue sequence length greater than 512 tokens.

I applied t-SNE with perplexity=10 and plotted the resulting 2D dialogue and summary encodings:





When I calculated the pairwise distances between validation and training samples, validation sample ID 209 had the closest match to a training sample, so I analyzed it to try and understand the training results.

To understand the overall dataset, I also calculated the distances for 209 against the validation and test samples and then looked at the top 5 closest matches from each split for both dialogue and summary.

209 talks about fractures and surgery, so as a sanity check, I also searched for mentions of fracture/fractures/fractured in the dialogues.

After reading through the text samples, I identified the samples below as being the closest in subject matter and then fed them into the model for inference:

Notes	Split	ID	Dialogue	Summary	Prediction
Reference sample	Val	209	Patient fell onto right arm and completely fractured both arm bones; doctor suggests surgery and says that risk of infection is less than one percent with antibiotics; other surgical risks include bleeding, etc.	Symptoms: refracture of right forearm Diagnosis: complete fractures of both bones in right arm, shortening bayonet apposition History of Patient: fell onto right arm on December 5, 2007 Plan of Action: surgery for closed reduction and pinning, with risks including anesthesia, infection, bleeding, changes in sensation and motion of extremity, hardware failure, need for later hardware removal, cast tightness	Symptoms: fell onto right arm on December 5, 2007, incomplete fractures of both bones in right arm, shortening bayonet apposition, possible need for later hardware removal and possible continuous nerve Symptoms Diagnosis: N/A History of Patient: fell onto right arm on December 5, 2007, incomplete fractures of both bones, received physical therapy and bandaging, no other surgeries recommended due to risk of infection Plan of Action: Surgery recommended due to non-healing extremity injuries, risks of infection discussed and patient agreed to surgery after informed consent
	Train	727	Patient fractured left elbow and underwent ORIF surgery; doctor says fracture has healed well; patient has been going to physical therapy; doctor says another surgery is required to remove hardware and says that risk of infection is less than one percent with antibiotics; other surgical risks include bleeding, etc.	Symptoms: N/A Diagnosis: N/A History of Patient: 10-year-old right-hand dominant male threw himself off a quad on 10/10/2007, underwent open reduction and internal fixation of left elbow fracture dislocation, sustained nondisplaced right glenoid neck fracture, experiencing postoperative stiffness treated with physical therapy and Dynasplint, neurologically intact distally Plan of Action: Surgery recommended for hardware removal from left elbow to decrease irritation with elbow extension, risks discussed include anesthesia, infection, bleeding, changes in sensation and motion of extremities, failure to remove hardware, failure to relieve pain, continued postoperative stiffness; parents agreed to the plan	Symptoms: N/A Diagnosis: N/A History of Patient: 10-year-old right-hand dominant male threw himself off a quad on 10/10/2007, underwent open reduction and internal fixation of left elbow fracture dislocation, sustained nondisplaced right glenoid neck fracture, experiencing postoperative stiffness treated with physical therapy and Dynasplint, neurologically intact distally Plan of Action: Surgery recommended for hardware removal from left elbow to decrease irritation with elbow extension, risks discussed include anesthesia, infection, bleeding, changes in sensation and motion of extremities, failure to remove hardware, failure to relieve pain, continued postoperative stiffness; parents agreed to the plan
	Val	649	Patient underwent screw compression to fix distracted left lateral condyle fracture; patient is coming in to remove the hardware; doctor says surgical risks include anesthesia, infection, bleeding, etc.	Symptoms: N/A Diagnosis: distracted left lateral condyle fracture History of Patient: Underwent screw compression for the fracture in October 2007, now presents for hardware removal Plan of Action: Risks and benefits of surgery discussed, including risk of anesthesia, infection, bleeding, changes in sensation and motion of extremity, failure of removal of hardware, failure to relieve pain or improve range of motion; family agreed to the plan	Symptoms: N/A Diagnosis: N/A History of Patient: distracted left lateral condyle fracture (PUD) with screw compression to prevent fracture Plan of Action: N/A
	Train	902	Patient fell off liner ten feet and landed on left foot; no deformity in ankle but X-ray shows fractured talus (heel); doctor recommends surgery to repair ankle; risk of infection is low	Symptoms: left ankle pain, disfigurement Diagnosis: grade IV Hawkins fracture of left talus History of Patient: 50-year-old male, fell from approximately 10 feet onto left foot, no other injuries reported, distal neurovascularly intact Plan of Action: surgery recommended due to risk of avascular necrosis, risks of infection discussed and antibiotics planned	Symptoms: left ankle pain, disfigurement Diagnosis: grade IV Hawkins fracture of left talus History of Patient: 50-year-old male, fell from approximately 10 feet onto left foot, no other injuries reported, distal neurovascularly intact Plan of Action: surgery recommended due to risk of avascular necrosis, risks of infection discussed and antibiotics planned
	Train	1097	Patient is undergoing conservative treatment for lower back fracture, but there's been no improvement; doctor recommends bileral L five kyphoplasty to fix incomplete healing of L five compression fracture; doctor says risk of infection is less than one percent with antibiotics	Symptoms: back and buttock pain Diagnosis: L5 compression fracture with sclerosis, incomplete healing History of Patient: 86-year-old female with history of back and buttock pain, conservative treatment unsuccessful, CT scan shows incomplete healing of L5 compression fracture Plan of Action: bilateral L5 kyphoplasty scheduled, patient denies bowel or bladder incontinence, wears back brace and corset, no weakness reported	Symptoms: back and buttock pain Diagnosis: L5 compression fracture with sclerosis, incomplete healing History of Patient: 86-year-old female with history of back and buttock pain, conservative treatment unsuccessful, CT scan shows incomplete healing of L5 compression fracture Plan of Action: bilateral L5 kyphoplasty scheduled, patient denies bowel or bladder incontinence, wears back brace and corset, no weakness reported
	Train	182	Patient came in for surgery to fix acute on chronic right slipped capital femoral epiphysis; X-ray shows screw is going into the hip joint; doctor says another operation is required to remove the screw and replace it with a shorter one; doctor says risk of infection is less than one percent; other surgical risks include bleeding, etc.	Symptoms: N/A Diagnosis: acute on chronic right slipped capital femoral epiphysis History of Patient: presented in November Plan of Action: underwent in situ pinning, screw exchange discussed	Symptoms: N/A Diagnosis: acute on chronic right slipped capital femoral epiphysis History of Patient: presented in November Plan of Action: underwent in situ pinning, screw exchange discussed
	Val	640	Following ORIF surgery to fix right tibial plateau fracture, patient had surgery one week ago to remove Ex Fix from the right knee with MUA to break up scar tissue; patient does not have much pain, flu symptoms, numbness, or tingling; patient is doing range of motion exercises	Symptoms: N/A Diagnosis: N/A History of Patient: Approximately one week status post removal of Ex-Fix from the right knee with manipulation under anesthesia (MUA) following open reduction internal fixation (ORIF) of right tibial plateau fracture, well-controlled pain, mild drainage from pin sites, just started range of motion exercises for the right knee, no fevers, chills, or night sweats, no numbness or tingling Plan of Action: Continue range of motion exercises, monitor pin sites for drainage, follow-up as scheduled	Symptoms: N/A Diagnosis: right tibial plateau fracture History of Patient: surgery performed approximately a week ago for right tibial plateau fracture, previous surgery for right tibial plateau fracture, unsuccessful interventional management strategies
	Test	606	Patient underwent percutaneous screw fixation of a Schatzker I V tibial plateau fracture; patient has been going to physical therapy; patient says there are no signs of infection; patient has no fever symptoms but has some tingling in feet	Symptoms: Tingling in both feet. Diagnosis: N/A. History of Patient: 59-year-old male, 4 months post percutaneous screw fixation of Schatzker IV tibial plateau fracture and nonoperative management of second through fifth metatarsal head fractures. Currently at home after leaving nursing home facility. Pain well controlled. Working with physical therapy 2-3 times a week. No drainage or fever. History of spinal stenosis with lower extremity neuropathy. Plan of Action: N/A.	Symptoms: percutaneous screw fixation of a Schatzker I V tibial plateau fracture, toe fractures nonoperatively, no pain, no injuries reported, has a history of spinal stenosis, severe cough, runny nose, sore throat, fatigue, and tingling in both feet. Diagnosis: percutaneous screw fixation of a Schatzker I V tibial plateau fracture, toe fractures nonoperatively, no other associated injuries. History of Patient: toe fractures and tingling in both feet, history of spinal stenosis, and neuropathy, post-surgery, normal side effects, no prior history of spinal stenosis, and neuropathy. Plan of Action: N/A

Split	ID	Loss	Bertscore (F1)	Rouge1	Rouge2	RougeL	RougeLsum
Val	209	1.78	0.89	0.55	0.36	0.41	0.52
Train	727	0.02	1.00	1.00	1.00	1.00	1.00
Val	649	1.89	0.85	0.41	0.26	0.32	0.41
Train	902	0.00	1.00	1.00	1.00	1.00	1.00
Train	1097	0.01	1.00	1.00	1.00	1.00	1.00
Train	182	0.00	1.00	1.00	1.00	1.00	1.00
Val	640	2.89	0.84	0.31	0.14	0.26	0.29
Test	606	2.92	0.83	0.42	0.25	0.33	0.40

As expected, the generated summaries for the training samples exactly match the ground truth. The generated summaries for the validation and test samples look plausible at first glance, but upon closer inspection, the model is hallucinating. See below.

5.2.1 Validation sample ID 209

Dialogue:

```
Doctor: Good morning, young man. Are these your parents?
Patient: Yes.
Doctor: Good, can you tell me more about your son, please?
Guest_family_1: Well, he's five now, and he fell onto his right arm on
December fifth two thousand seven.
Doctor: After he fell, how was he treated?
```

Guest_family_1: We went to the E D right after he fell, and they said he had **complete fractures of both bones in the arm.**
Doctor: Yes, I see that here, he also has shortening bayonet apposition.
Guest_family_1: What can we do for this?
Doctor: There's actually a few options here. First we can cast it and see how he heals, generally, children heal up very well from fractures.
Guest_family_1: That's good, we like that option more than any kind of surgery.
Doctor: However, **surgery is also an option here as well.**
Guest_family_1: Yeah, to be completely sure we fix this, I think we should opt for the surgery, what do you think, honey?
Guest_family_2: Yes, I agree. What are the risks of infection for this surgery?
Doctor: The **risk of infection is very low**, generally less than one percent. We use antibiotics to control for infection.
Guest_family_1: Will he be asleep for the surgery?
Doctor: Absolutely, he won't feel a thing. **Other risks include bleeding, changes in sensation and motion of the extremity, hardware failure, and need for later hardware removal, and cast tightness. I would not worry about these risks. We have great results with these surgeries.**
Guest_family_1: Then **yes, we'd like to do the surgery.**

Summary:

Symptoms: **refracture** of right forearm
Diagnosis: complete fractures of both bones in right arm, shortening bayonet apposition
History of Patient: fell onto right arm on December 5, 2007
Plan of Action: surgery for **closed reduction and pinning**, with risks including anesthesia, infection, bleeding, changes in sensation and motion of extremity, hardware failure, need for later hardware removal, cast tightness

Generated summary:

Symptoms: fell onto right arm on December 5, 2007, **incomplete fractures** of both bones in right arm, shortening bayonet apposition, possible need for later hardware removal and possible continuous nerve Symptoms
Diagnosis: N/A
History of Patient: fell onto right arm on December 5, 2007, **incomplete fractures of both bones, received physical therapy and bandaging, no other surgeries recommended due to risk of infection**
Plan of Action: **Surgery recommended due to non-healing extremity injuries,** risks of infection discussed and patient agreed to surgery after informed consent

The model is hallucinating – it thinks the fractures are incomplete fractures and contradicts itself by saying that surgery is both recommended and not recommended due to risk of infection.

The ground truth summary contains information not present in the dialogue (noisy dataset).

5.2.2 Validation sample ID 649

Dialogue:

Doctor: Hello. How are you both doing today?
Guest_family: We're doing great. The E D told us to come here. We're here to see if we could get the stuff in his leg taken out.
Doctor: I see. Did he have a fracture before?
Guest_family: Yeah, he did. Here's the report from the surgery from October two thousand and seven.
Doctor: Great. Let me take a look. **It looks like he had a distracted left lateral condyle fracture. He then underwent screw compression to fix the fracture, correct?**
Guest_family: That's correct.
Doctor: Perfect. **So, now we just have to remove the hardware that was used to fix the fracture, right?**
Guest_family: Yeah, that's correct. **What are the risks of the surgery?**
Doctor: **The risks include the risk of anesthesia, infection, bleeding, changes in sensation and motion of extremity, failure of removal of hardware, failure to relieve pain or improved range of motion.**
Guest_family: Oh, wow. That's a lot. Could you go over the benefits?
Doctor: Of course. Before I do, I just need to confirm your son's age.
Guest_family: He can tell you.
Patient: Hi. I'm five years old. My birthday is in four months.
Doctor: Thank you. Happy almost birthday.
Guest_family: He gets excited about birthdays. I just have a few more questions, but it looks like we're going to go through with the surgery.

Summary:

Symptoms: N/A
Diagnosis: distracted left lateral condyle fracture
History of Patient: Underwent screw compression for the fracture in October 2007, now presents for hardware removal
Plan of Action: Risks and benefits of surgery discussed, including risk of anesthesia, infection, bleeding, changes in sensation and motion of extremity, failure of removal of hardware, failure to relieve pain or improve range of motion; family agreed to the plan

Generated summary:

Symptoms: N/A
Diagnosis: N/A
History of Patient: distracted left lateral condyle fracture (PUD) with screw compression to prevent fracture
Plan of Action: N/A

The model hallucinates the previous operation prevented, rather than fixed, the fracture. The ground truth summary's "Plan of Action" should include hardware removal, but it's under "History of Patient" (noisy dataset).

5.2.3 Validation sample ID 640

Dialogue:

Doctor: Good morning, sir. Remind me please, **when was your surgery?**
Patient: Hello, doctor. **It was about a week ago now.**
Doctor: Great, and just to remind you, **you had a removal of Ex Fix from the right knee with an M U A following O R I F of a right tibial plateau fracture.**
Patient: Yes, that all sounds familiar. What's M U A?
Doctor: It stands for manipulation under anesthesia. We bent the knee in order to forcibly break up some scar tissue.
Patient: Oh, wow.
Doctor: Yeah, it's great, eliminates some need for P T in fact. How are you doing today?
Patient: **Overall, my pain is under control.**
Doctor: Are you having any flu like, or infection symptoms?
Patient: No, nothing like that at all.
Doctor: **So, no fever, or chills, or sweating at night.**
Patient: **Yeah, nothing like that.**
Doctor: Great. **I see you're having some drainage from the pin sites, but that doesn't bother me too much.**
Patient: Okay, good. I was getting worried.
Doctor: No need to worry. **Are you doing your range of motion exercises? Those are imperative.**
Patient: **Yes, doctor. I am.**
Doctor: **Good, and do you have any numbness or tingling?**
Patient: **No, doctor. I don't have any of that.**

Summary:

Symptoms: N/A
Diagnosis: N/A
History of Patient: Approximately one week status post removal of Ex-Fix from the right knee with manipulation under anesthesia (MUA) following open reduction internal fixation (ORIF) of right tibial plateau fracture, well-controlled pain, mild drainage from pin sites, just started range of motion exercises for the right knee, no fevers, chills, or night sweats, no numbness or tingling
Plan of Action: Continue range of motion exercises, monitor pin sites for drainage, follow-up as scheduled

Generated summary:

Symptoms: N/A
Diagnosis: right tibial plateau fracture
History of Patient: surgery performed approximately a week ago for right tibial plateau fracture, **previous surgery for right tibial plateau fracture, unsuccessful interventional management strategies**

The model hallucinates about “unsuccessful interventional management strategies.”

5.2.4 Test sample ID 606

Dialogue:

Doctor: Hello, sir. So, remind me, how long ago was your procedure?
Patient: Let me think. Now, it's been about four months. I just turned fifty nine.
Doctor: Okay, and just to refresh your memory, **we did a percutaneous screw fixation of a Schatzker I V tibial plateau fracture, and we've managed those toe fractures nonoperatively.**
Patient: Which toes did I break?
Doctor: **You broke your second and fifth toes.**
Patient: Do you have the x rays? Can I see where the breaks are?
Doctor: Absolutely, if you look here, right here is where the breaks are. These are called the metatarsal heads.
Patient: Wow, that's right where it hurts.
Doctor: Yep, and where are you living now?
Patient: Thankfully, I left the nursing home, and I'm back in my own house now.
Doctor: That's really good. **Are you in any pain today?**
Patient: **No, not today.**
Doctor: **Great, have you been going to P T?**
Patient: **Yeah, I've been going about two or three times a week.**
Doctor: Are there any signs of infection, like drainage from the incision?
Patient: Nope, **the wound healed up nicely.** I'll be able to wear shorts again.
Doctor: Great, **are you having any fever symptoms, like cough, runny nose, sore throat, fatigue, anything like that?**
Patient: **No, but I have had some tingling in both my feet.** I don't know if its related to surgery. If I'm remembering the names right, um, I have a history of spinal stenosis and neuropathy in my legs.
Doctor: With your history of spinal stenosis, the broken toes, and the surgery, that's a pretty normal side effect.

Summary:

Symptoms: Tingling in both feet.
Diagnosis: N/A.
History of Patient: 59-year-old male, 4 months post percutaneous screw fixation of Schatzker IV tibial plateau fracture and nonoperative management of second through fifth metatarsal head fractures. Currently at home after leaving nursing home facility. Pain well controlled. Working with physical therapy 2-3 times a week. No drainage or fever. History of spinal stenosis with lower extremity neuropathy.
Plan of Action: N/A.

Generated summary:

Symptoms: percutaneous screw fixation of a Schatzker I V tibial plateau fracture, **toe fractures nonoperatively**, no pain, no injuries reported, has a history of spinal stenosis, **severe cough**, **runny nose**, **sore throat**, **fatigue**, and tingling in both feet.
Diagnosis: percutaneous screw fixation of a Schatzker I V tibial plateau fracture, toe fractures nonoperatively, no other associated injuries.
History of Patient: toe fractures and tingling in both feet, history of spinal stenosis, and neuropathy, post-surgery, normal side effects, **no prior history of spinal stenosis**, and neuropathy.

```
Plan of Action: N/A
```

The model hallucinates that the patient has flu symptoms and contradicts itself on the patient's history of spinal stenosis.

5.2.5 All other samples

The other samples I found weren't similar to the reference validation sample ID 209. Some samples contain almost no information, and some are noisy (training sample ID 931 and test sample ID 308).

Notes	Split	ID	Dialogue	Summary
Noise	Train	931	Patient jumped off swing and suffered a closed type three supracondylar fracture of left distal humerus	Symptoms: N/A Diagnosis: N/A History of Patient: patient had cardioversion for atrial fibrillation, taking Coumadin, history of smoking but quit several years ago, denies COPD or emphysema, no family members are sick Plan of Action: N/A
	Train	219	Doctor says X-ray does not show any open fracture or bone abnormality	Symptoms: N/A Diagnosis: N/A History of Patient: N/A Plan of Action: N/A .
	Train	194	Doctor asks patient when the surgery happened	Symptoms: N/A Diagnosis: slipped capital femoral epiphysis (SCFE) bilaterally History of Patient: post-surgery, 2-1/2 months Plan of Action: N/A
	Train	401	Patient hit left elbow against a railing trying to do a new trick on skateboard; patient has iced the elbow, but the pain has only gotten worse; patient has not taken Advil or Tylenol	Symptoms: Left elbow pain. Diagnosis: N/A. History of Patient: Injured left elbow by hitting it against a railing while attempting a new trick on a skateboard about a week ago, pain came on gradually, no other body parts injured, hasn't taken any medication for the pain. Plan Of Action: N/A.
	Train	1155	Patient injured elbow during a fight in juvenile hall; patient has left ankle pain; patient hit head against the floor during fight but does not have headache, nausea, blurry vision, or chest or abdominal pain	Symptoms: Pain in the left elbow and left ankle. Diagnosis: N/A History of Patient: The 17-year-old male sustained an elbow injury during a fight with other kids in Juvenile Hall, experiencing sudden pain in his left elbow. He also reports pain in his left ankle, with previous left knee pain. He denies passing out, neck pain, chest pain, or abdominal pain. No weapons were involved in the fight. Plan of Action: N/A
	Train	1290	Patient injured left elbow during fight in juvie; patient has left ankle pain; patient got hit in the head; patient does not have chest or abdominal pain	Symptoms: Pain in left elbow and left ankle, previous pain in left knee Diagnosis: N/A History of Complaint: 17-year-old male experienced pain in left elbow and left ankle during a fight in Juvenile Hall, denies neck pain despite being hit in the head, no chest or abdominal pain, no weapons involved Plan of Action: N/A
	Train	1240	Patient splashed hot oil onto his arm while working in coffee shop kitchen and was burned from elbow to wrist	Symptoms: Burn on the arm from elbow to wrist, mainly on the medial aspect Diagnosis: N/A History of Complaint: Workers' Compensation injury, hot oil splashed onto arm while cooking in coffee shop kitchen Plan of Action: Provide care for burn injury
	Val	1160	Doctor looks at patient's ultrasound	Symptoms: N/A. Diagnosis: N/A. History of Patient: N/A. Plan of Action: The only significant finding in the ultrasound of the area is that it shows this to be related to bone.
	Val	616	Patient twisted right ankle while running	Symptoms: right ankle pain on the lateral aspect Diagnosis: N/A History of Patient: twisted right ankle while running, no other injuries, brought in by mother, primary care physician is Dr. Brown Plan of Action: N/A
	Val	1009	Patient plays basketball for University of Houston; patient landed on another player's foot, was taped up, and kept playing; patient's foot is swollen and can't put weight on it; patient is in a walking boot	Symptoms: Swelling, pain onset immediately after injury, pain with weightbearing activities, limping Diagnosis: Inversion injury, tenderness around the navicular History of Patient: Injured foot during basketball game while traveling to Duke, landed on another player's foot, taped by trainer John Houston, continued playing after injury Plan of Action: Patient has been in a walking boot, foot taped firmly, advised to continue wearing the tape, continue monitoring for pain and swelling
	Val	1181	Patient has a right side shoulder strain and possibly some nerve compression	Symptoms: Right shoulder pain, most likely secondary to muscular strain Diagnosis: very mild evidence of impingement History of Patient: N/A Plan of Action: Further evaluation and treatment will be done.
	Val	273	Patient had knee replacement surgery for both knees three years ago and is now having pain in both knees	Symptoms: pain in both knees Diagnosis: N/A History of Patient: bilateral knee replacement three years ago, experiencing pain in both knees Plan of Action: N/A
	Val	877	Patient injured left knee after slipping in a grocery store; patient's primary care provider gave him a knee brace and referred him to physical therapy	Symptoms: left knee pain Diagnosis: N/A History of Patient: fell in grocery store on 10/02/08, slipped on grape, went to ED then followed up with PCP, referred to Physical Therapy, received knee brace Plan of Action: N/A
	Test	540	Two doctors talking about patient - multiple areas with hypergranulation tissue on the left leg	Symptoms: N/A Diagnosis: N/A History of Patient: N/A Plan of Action: N/A
	Test	257	Patient had circumcision and had minor post-op bleeding	Symptoms: N/A Diagnosis: phimosis (resolved) History of Patient: circumcision performed on 09/16/2007 at Children's Hospital, minor bleeding post-operation requiring additional sutures, pain managed with oral analgesics for a couple of days, normal urination and bowel movements Plan of Action: N/A
	Test	521	Patient has pain in navicular bone, and wearing shoes makes it worse; patient has had surgery for osteochondroma in the foot; patient requests surgery	Symptoms: extreme pain over the navicular bone with shoe gear. Diagnosis: hereditary osteochondromas, previous dissection of osteochondromas. History of Patient: pain in the foot, multiple osteochondromas of unknown origin, desire for surgical treatment. Plan of Action: surgical treatment for pain in the foot caused by osteochondromas.
	Test	89	When was the last time you got a tetanus shot?	Symptoms: N/A. Diagnosis: N/A. History of Patient: N/A. Plan of Action: N/A.
Noise	Test	308	Patient has stomach pain	Symptoms: pain in my tummy. Diagnosis: left patellar chondromalacia. History of Patient: started suddenly around noon. Plan of Action: N/A
	Test	88	Patient with history of falling does not want chair and bed monitor	Symptoms: N/A. Diagnosis: N/A. History of Patient: History of multiple falls. Plan of Action: Recommending chair and bed monitor for fall detection, patient declined.
	Test	678	Patient has history of osteomyelitis in right fifth toe and diabetes mellitus	Symptoms: N/A Diagnosis: N/A History of Patient: Diabetes mellitus, history of osteomyelitis of the right fifth toe treated with IV antibiotics therapy 5 years ago for 6 weeks
	Test	367	Patient fell onto right hip and lower back while bringing in groceries from car	Symptoms: back pain Diagnosis: N/A History of Patient: fell on steps three nights ago while bringing in groceries, landed on right hip and hit low back on railing Plan of Action: doctor wants to get imaging done
	Train	814	Patient recounts past surgeries: craniotomy for brain hemorrhage, leg surgery to fix fracture, stomach surgery as a child	Symptoms: N/A Diagnosis: N/A History of Patient: 1. Surgery on stomach as a child. Type unknown, common procedure with no complications. 2. Surgery for leg fracture with pins inserted. 3. Craniotomy seven years ago for intracranial hemorrhage/subdural hematoma. Plan of Action: N/A
	Train	398	Two doctors talking about patient history - surgeries and right hip fracture	Symptoms: N/A. Diagnosis: N/A. History of Patient: Appendectomy, right hip fracture from a fall in 2005, total abdominal hysterectomy (TAH) with bilateral salpingo-oophorectomy (BSO). Plan of Action: N/A.
	Train	514	Patient hurt back of left thigh; patient hurt kneecap while boat fishing and got surgery; patient has external fixation on knee while the fracture heals	Symptoms: pain and injury to the back of left thigh, knee injury from a boat accident, hypergranulation tissue around graft site, drainage from areas with hypergranulation tissue. Diagnosis: N/A. History of Patient: traumatic injury to left posterior thigh, surgery for large defect in left posterior thigh, ongoing external fixation for healing fractures in leg, grafting and full thickness skin grafting for closure of defect, nearly healed in gluteal fold area. Plan of Action: referred to clinic for management of hypergranulation tissue and drainage from graft site.