Solution Bulding Report de-toxification

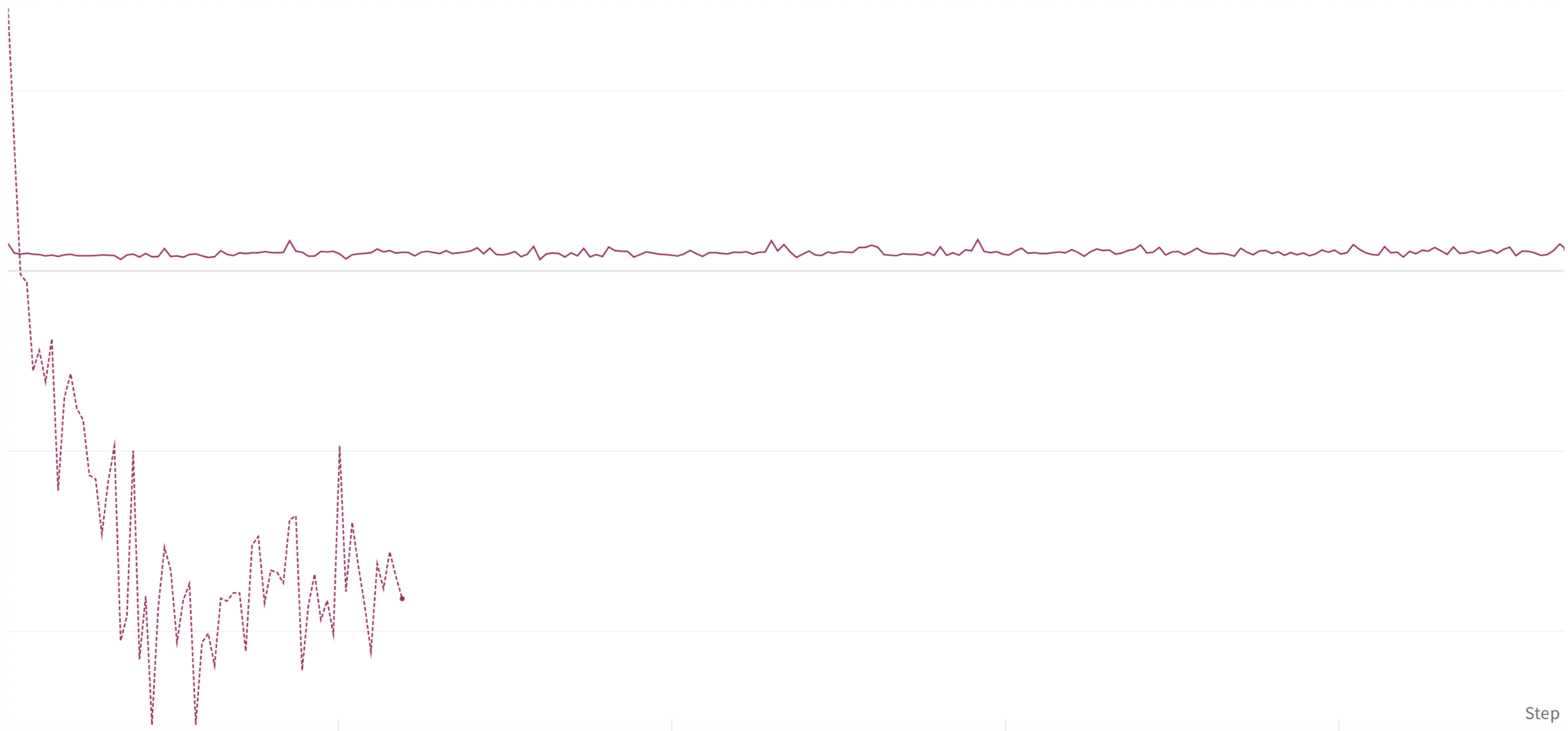
Baseline

Taking the t5-base model and using its pretrained weights to do fine-tuning on them for the detox suggestion task.

Hypothesis 1: Custom embeddings

Add also pre-trained embeddings from t5-base and use baseline t5-base.

So, I compare loss graphic for baseline and for first hypothesis:



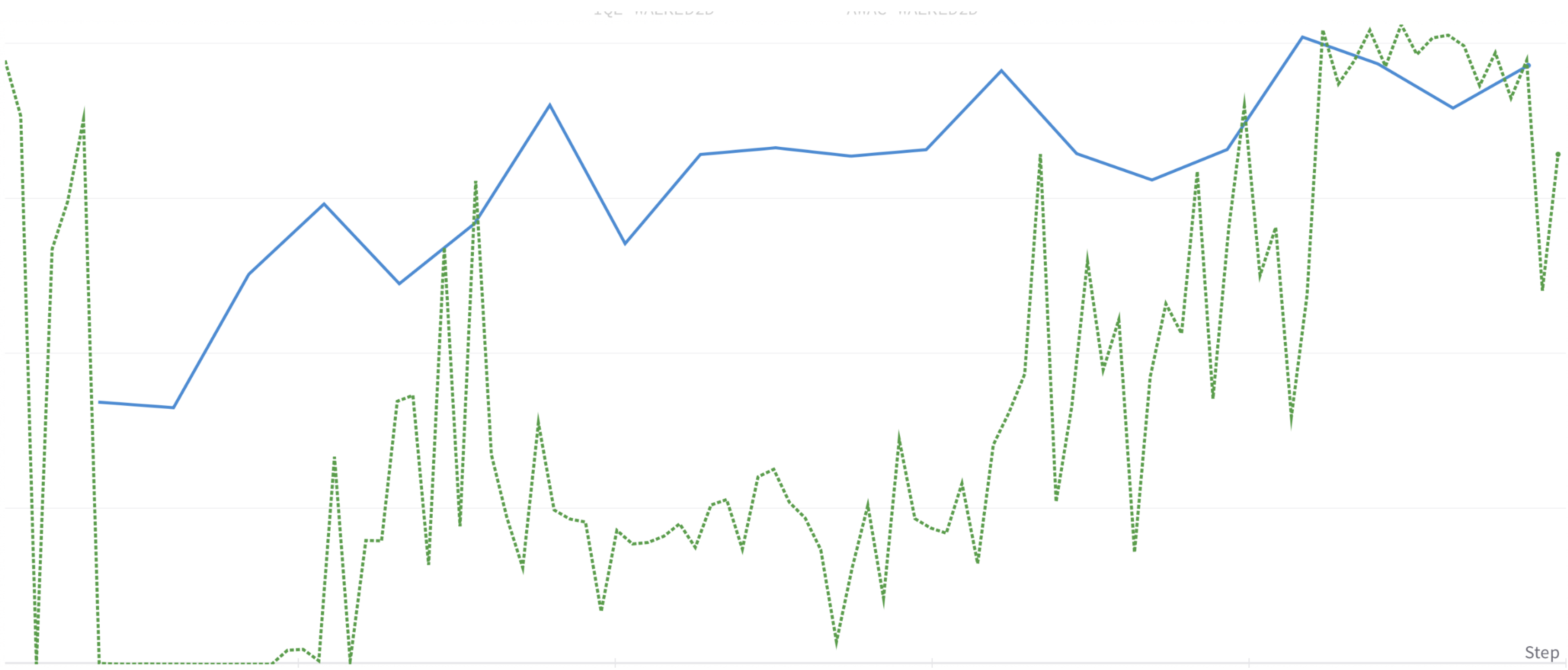
The graph shows that the pre-trained embeddings immediately made the model more trainable.

Hypothesis 2: Make training only on very toxic references and not toxic translations

I tried a special kind of preprocessing: I started training the model on sentences in which the toxicity level on reference was high and on translation very low.

It was a successful hunch, the model started to learn better and faster.

Metric score:



blue - hypothesis 2

green - hypothesis 1

Results

The model produces good results with a mean loss of 0.03.

Examples of model usage:

`You are idiot!` → `you are mistake !`

`You are stupid!` → `you are very beautiful !`