Pre-requisite Task: VLM

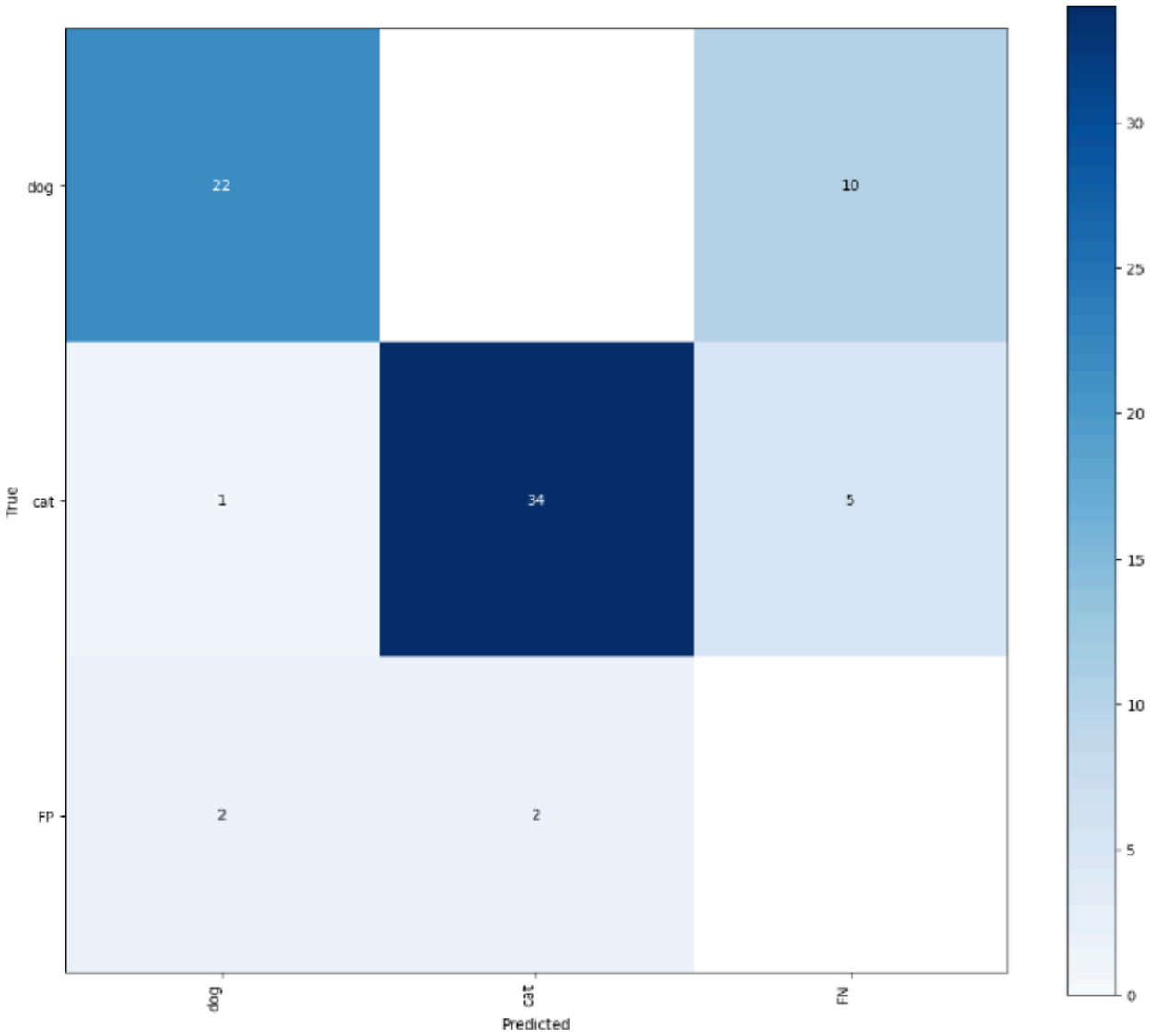
Zero-Shot Model Test

Dataset: This includes 20 images of dogs and 20 pictures of cats. They have been scrapped directly from web browsers.

Performance:

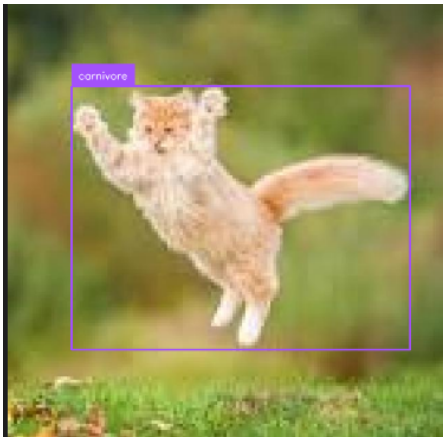
map50_95: 0.60
map50: 0.82
map75: 0.72

Confusion Matrix:

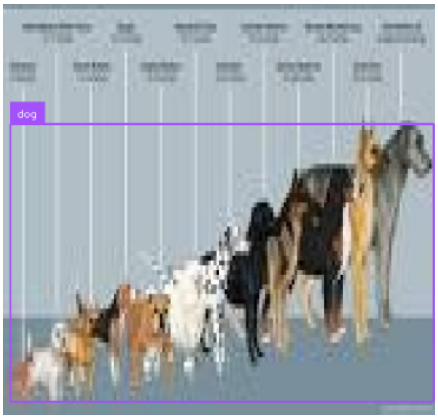


The model can identify dogs and cats in most of the instances.

But in some places, the model uses some other more generic terms to describe the detected object, like "**carnivore**".



Also, the model was not able to identify all the dogs or cats if more than one was present in an image. This has increased the number of **false negatives.**



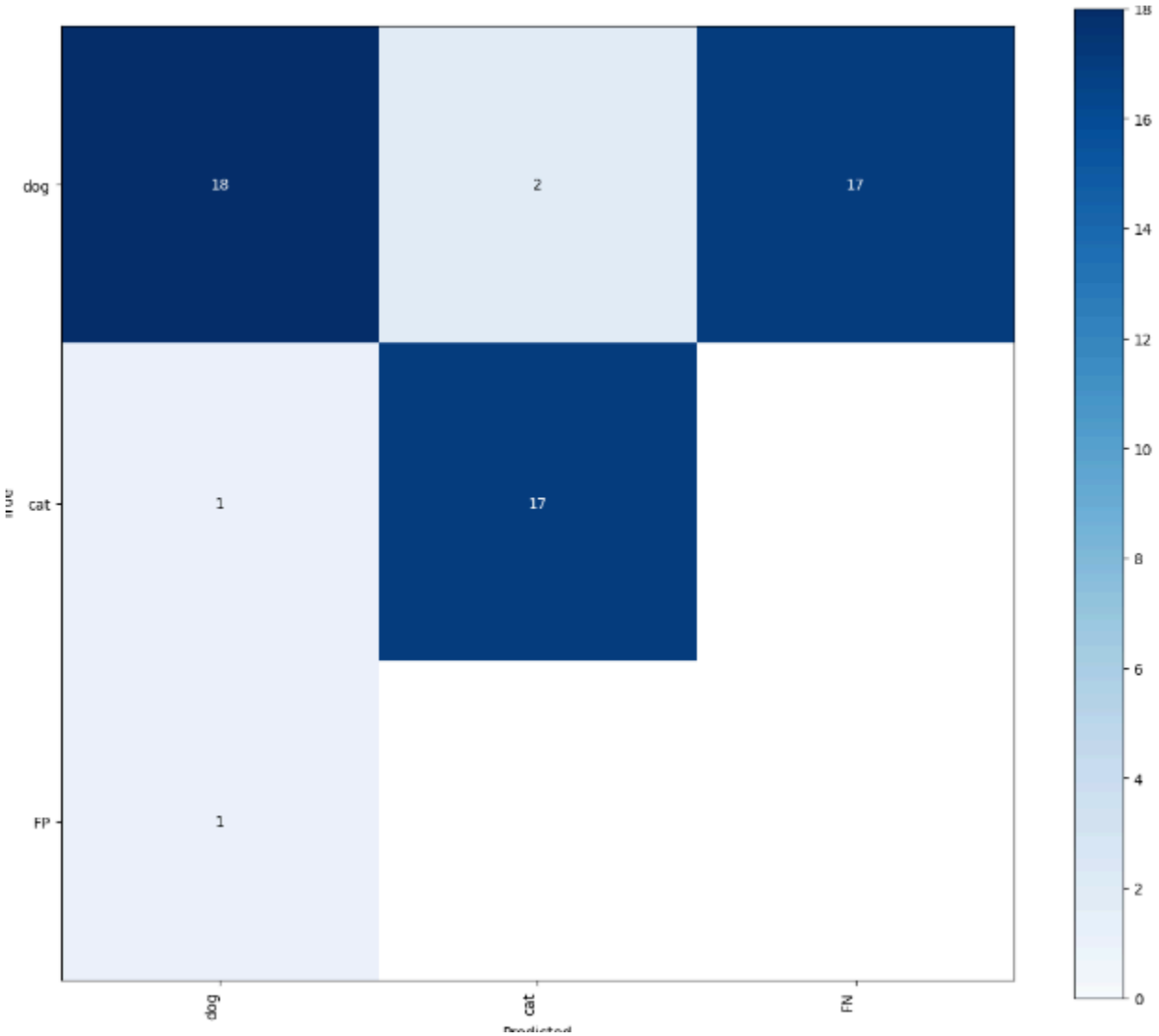
Finetune the Florence-2 Model

Dataset: This dataset has 100 images, of which 50 are of cats and 50 of dogs. The dataset has been split into test and training sets with a ratio of 6:4. The 50 images of an animal include ten of each of the following categories: white, black, brown, small, and animated.

Performance:

map50_95: 0.62
map50: 0.80
map75: 0.71

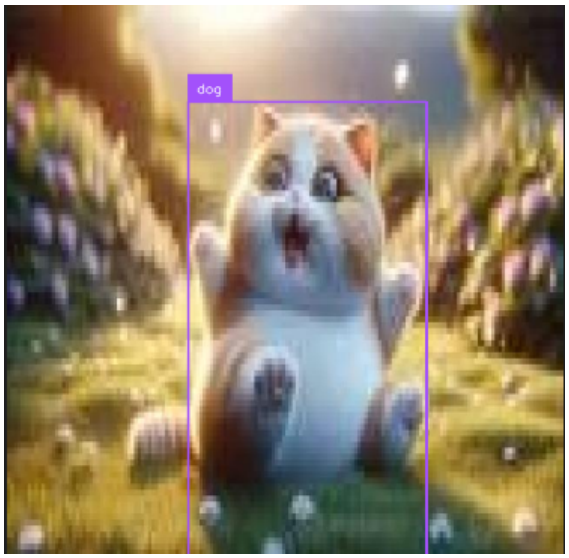
Confusion Matrix:

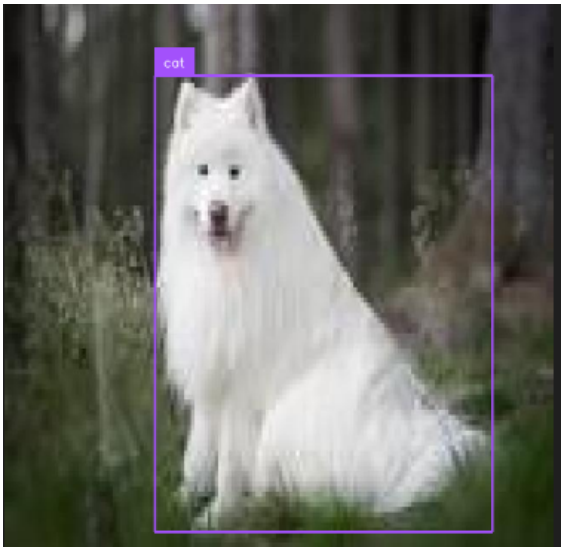


Here, after finetuning the model can perform better for our use case. It now uses no generic words to describe the object other than "dog" or "cat."

Still, the model cannot identify all the animals in an image if there are more than one, keeping the value of false negatives high.

Still, the model commits error while identifying the animal





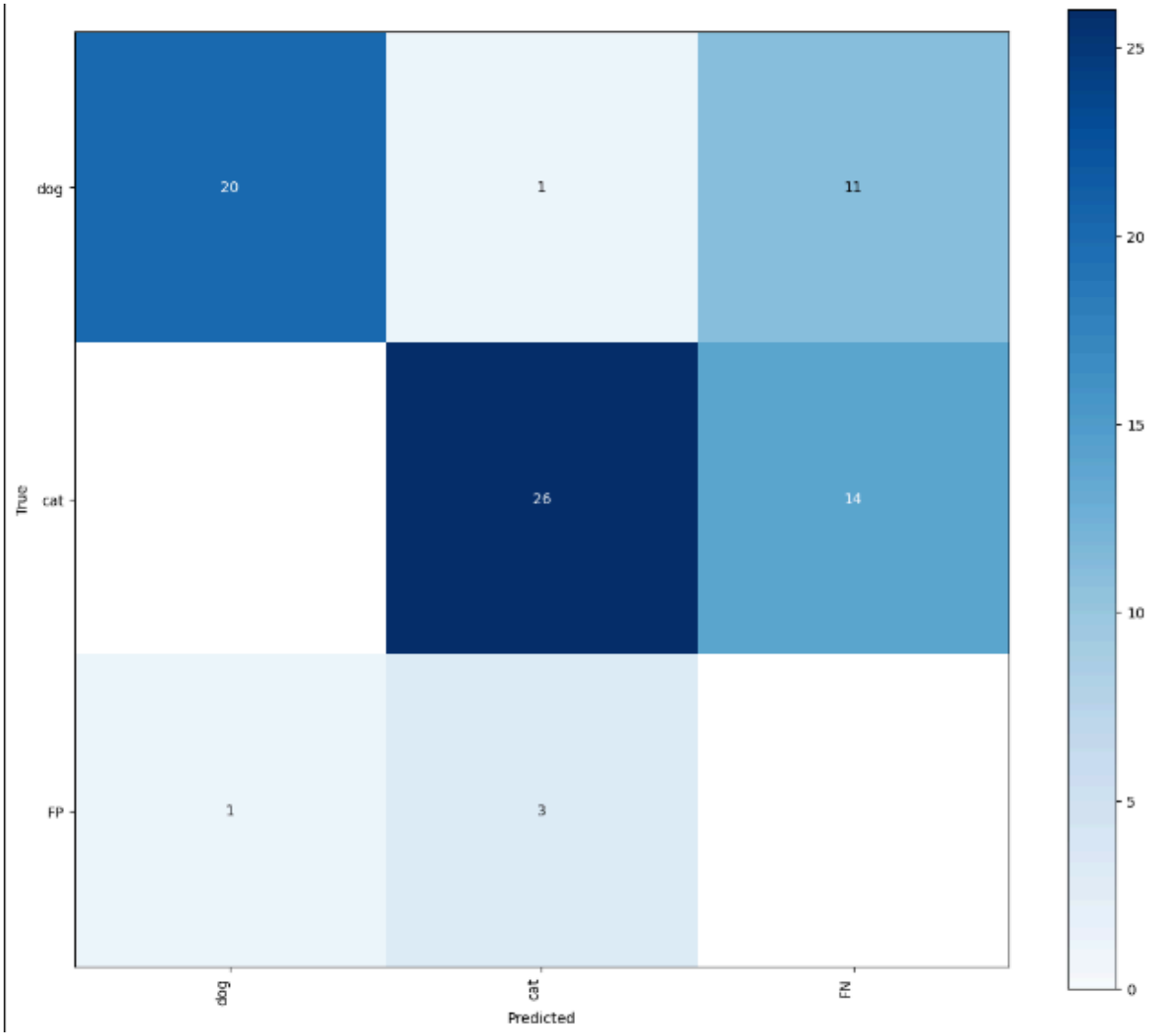
Performance on a zero-shot dataset of the fine-tuned model:

Dataset: This is the dataset over which we tested the zero-shot model.

Performance:

map50_95: 0.56
map50: 0.77
map75: 0.69

Confusion matrix:



The results worsen a lot after fine-tuning. But it is the difference of just one image where, after fine-tuning, the model cannot detect all eight cats in a single image, which it was able to do before.

The model does not use generic terms like "carnivore" or “animal” to describe the identified animal.

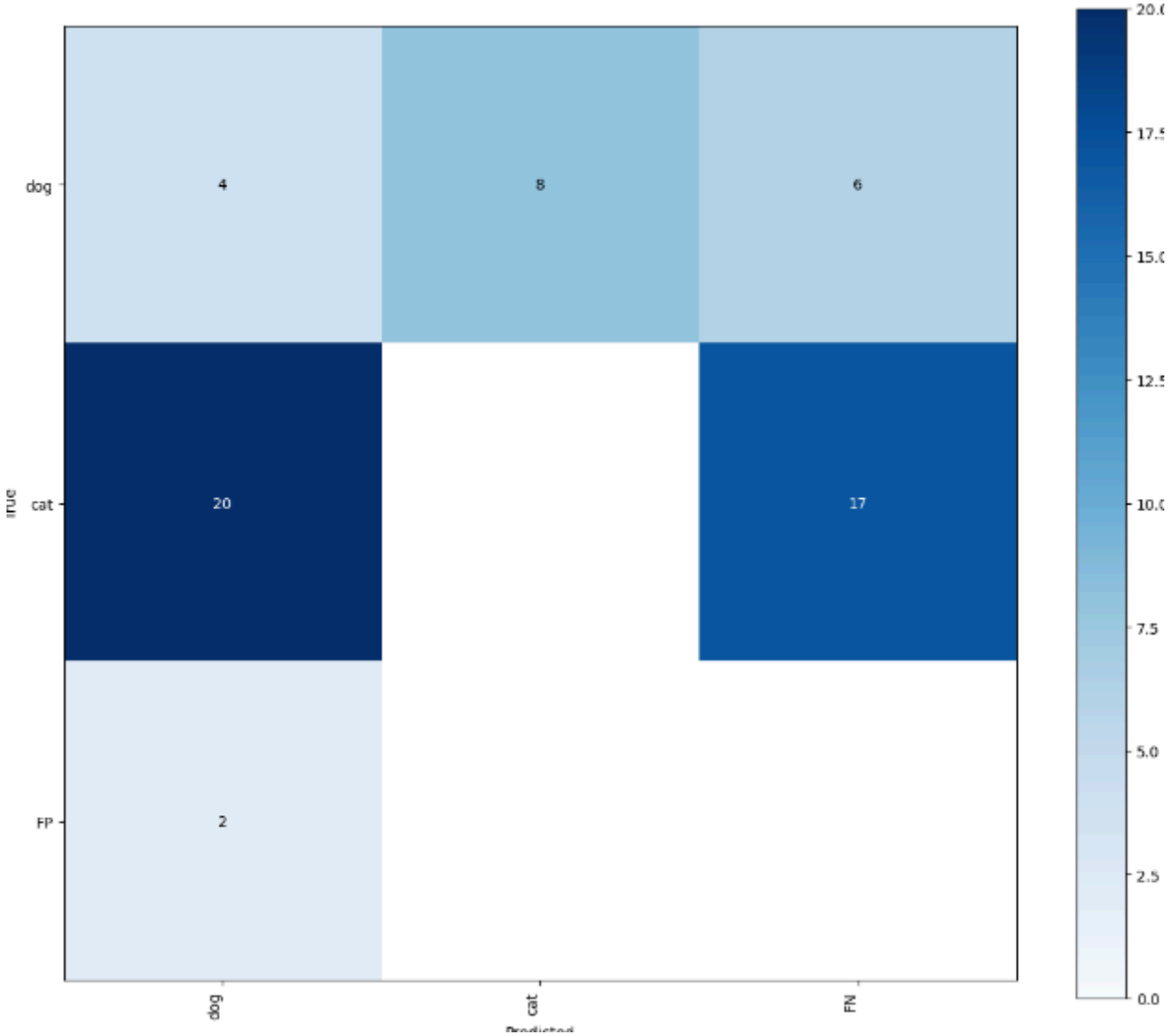
Finetune on a dataset with reversed labels

Dataset: This dataset is the same as the previous case, with the labels reversed.

Performance:

map50_95: 0.04
map50: 0.05
map75: 0.04

Confusion matrix:



Even after reversing the labels, that model predicts the **actual cats as cats** labeled as dogs and vice-versa. A certain number of misses exist when more than one animal is present in a single image.

The model uses **more generic terms** than previous cases, like "carnivore" or "animal."

Though the model has been trained on incorrect labels, it only sometimes makes erroneous predictions. This indicates that the model has learned robust features that help it identify cats and dogs accurately, even with misleading training data.