Comparison of Metrics across finetuned (N=408) vs vanilla (N=209) large models