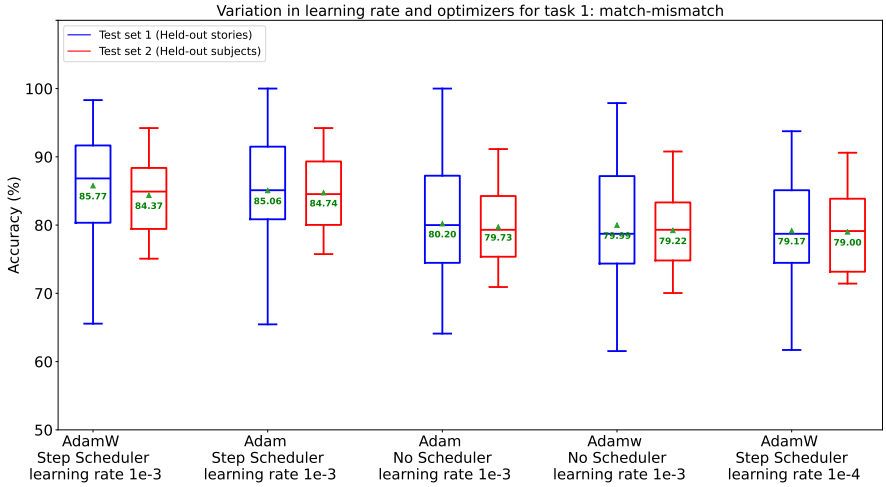
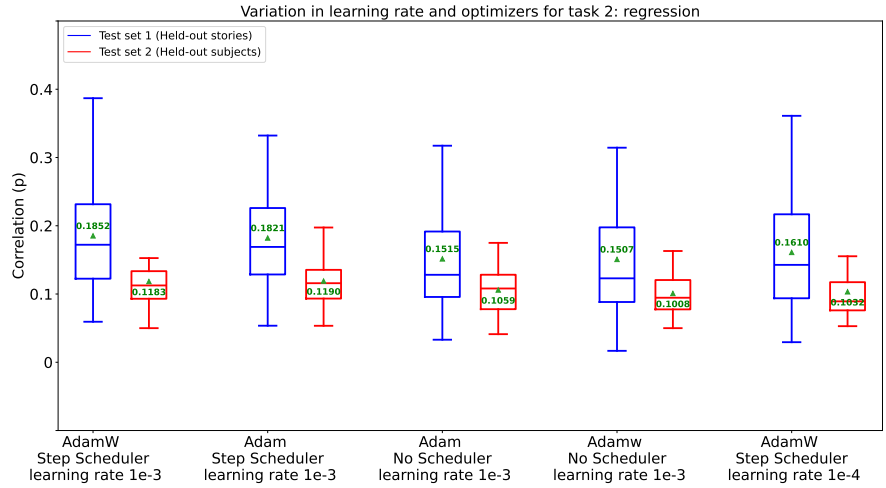
A Variation in the optimizer and learning rate

In our experiments, we noticed that the combination of using the Adam optimizer with weight decay, together with a step scheduler, consistently yielded the most stable and best results. In figure 1, this is clearly illustrated. On the left, we train a model using the Adam optimizer with weight decay (AdamW), a step scheduler, which decays the learning rate of each parameter group of the model by a factor 0.1 every 10 epochs and an initial learning rate of 0.001. In the other models, we systematically vary the optimizer (Adam versus Adam with weight decay), the learning rate scheduler (Step scheduler versus no scheduler) and the initial learning rate (0.001 versus 0.0001).



(a)



(b)

Figure 1. Results for task 1 and task 2 when varying the optimizer, learning rate scheduler and initial learning rate. **(a)** evaluated on task 1: match-mismatch. Each point in the boxplot corresponds to the average accuracy of one subject. The green points denote the mean accuracy across all subjects. **(b)** evaluated on task 2: regression. Each point in the boxplot corresponds to the average correlation of one subject. The green points denote the mean correlation across all subjects.