

# HPO - reader only - v1

Training a reader model using the gold document.

Valentin Liévin

## Setup

Searching for the best hyperparameters using N=636 trials with HyperOpt search alg. and Async HyperBand scheduler. Each model is trained using 4 GPUs and a training batch size of 24. The reader model encodes the question and the document together as a  $d$ -dimensional vector  $\mathbf{h}_{qd}$  and each answer candidate as  $\mathbf{h}_{a_i}$  using BERT. The score for each answer candidate is:  $\mathbf{S}_I(\mathbf{q}, \mathbf{d}) := \langle \mathbf{h}_{qd}, \mathbf{h}_{a_i} \rangle$ .

## Results

BioBERT seems to be the best backbone model, dropout=0 and lr=0.0001 also seem to be important factors. Nonetheless, the effect of the random seed remains unclear.

The best model scores 69% on the test set and uses the following hyperparameters





















- hidden\_size=128
- bert=BioBERT
- lr=1e-4
- weight\_decay=1e-3
- dropout=0

## Discussion

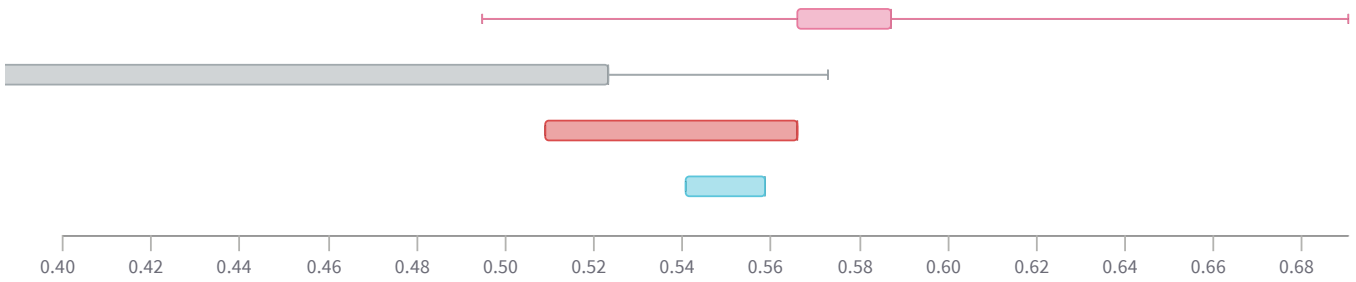
The best performances are much higher than the original MedQA model (vs 36.7% on the **whole** USMLE dataset for the MedQA paper). We need to analyse the best model in order to exclude any implementation issue. The performances could be explained by a model bug, test set contamination or because we are using a subset of the USMLE dataset.

The architecture of the reader model is rather simple, a more complete model that concatenates  $[\mathbf{q}, \mathbf{d}, \mathbf{a}_i]$  shall be tested in the future.

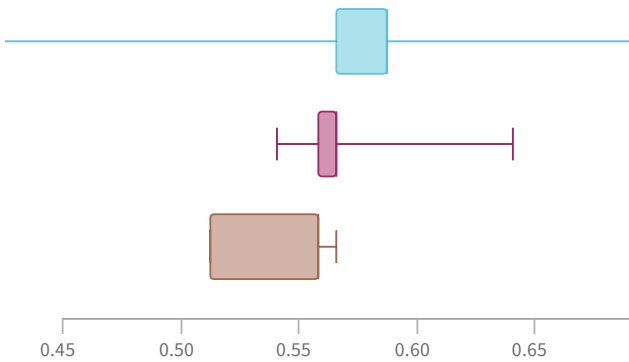
# Data

<div><input checked="" type="checkbox"/></div> <div>Run set 683</div> <div></div>						
Name (683 visualized)	Runtime	model/hidd	model/lr	model/weig	model/bert	model/drop
  neat-tree-1022	46m 31s	128	0.0001	0.001	dmis-lab/l	0
  dashing-darkness-981	39m 53s	128	0.0001	0.001	dmis-lab/l	0
  fallen-bee-820	34m 19s	256	0.0001	0.00001	dmis-lab/l	0
  lucky-plasma-1456	59m 19s	32	0.0001	0.00001	dmis-lab/l	0
  treasured-river-1573	40m 13s	64	0.0001	0.00001	dmis-lab/l	0
  autumn-wave-875	36m 4s	256	0.0001	0.001	dmis-lab/l	0
  resilient-paper-558	40m 47s	256	0.0001	0.00001	dmis-lab/l	0
  gallant-plant-980	33m 48s	256	0.0001	0.001	dmis-lab/l	0
  fast-shape-653	45m 47s	128	0.0001	0.001	dmis-lab/l	0
  olive-night-859	1h 16m 13	256	0.0001	0.00001	dmis-lab/l	0
1-10 ▾ of 683 < >						

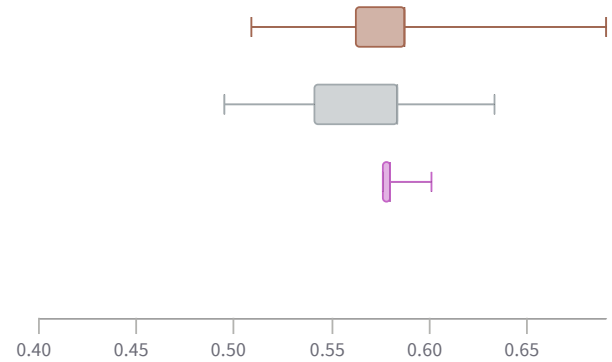
Accuracy vs. BERT type



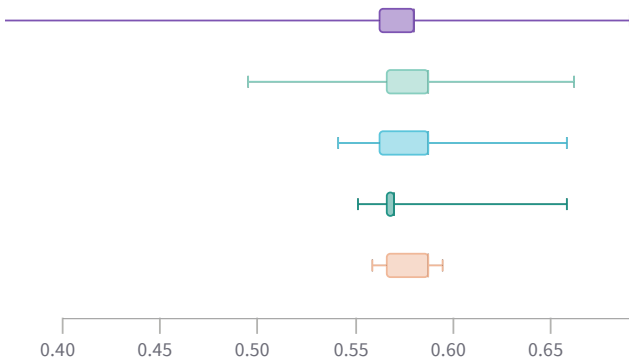
Accuracy vs. learning rate



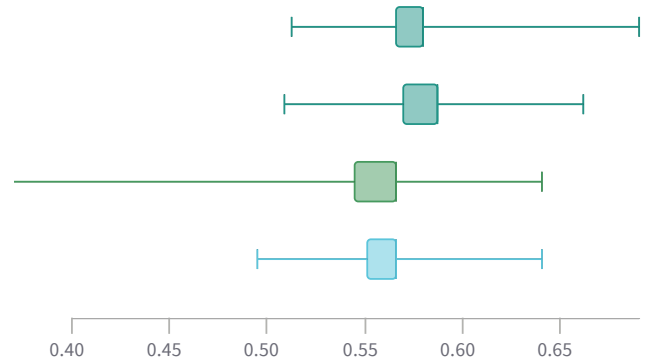
Accuracy vs. dropout



Accuracy vs. hidden\_size



Accuracy vs. weight decay



Created with  on Weights & Biases.

<https://wandb.ai/vlievin/findzebra-qa/reports/HPO-reader-only-v1--VmIldzo4ODIxMTg>