

# Results

Stefan Reifberger

September 26, 2024

## 1 Appendix

The following tables show the results of the experiment. They are to be read as follows:

- D: Number of decoder layers.
- E: Number of encoder layers.
- M: Mean value range of epochs.
- MD: 95% confidence interval distance between means.
- **Bold text**: The best value for given epochs or the current model performs better for the range of epochs up to the 95% confidence intervals of the difference of means not overlapping.

Model	Ep.: 1	5	M 1–5	MD 1–5	M 40–50	MD 40–50
LSTM, D=1	.829	<b>.513</b>	.661	.000	<b>.274</b>	.000
LSTM, D=2	<b>.808</b>	.675	.750	<b>.089<math>\pm</math>.042</b>	.385	<b>.111<math>\pm</math>.017</b>
LSTM, D=3	.812	.519	<b>.653</b>	$-.008 \pm .008$	.340	<b>.066<math>\pm</math>.006</b>
LSTM, D=4	.899	.682	.760	<b>.099<math>\pm</math>.027</b>	.382	<b>.108<math>\pm</math>.008</b>
TrFo, D=1	.635	.251	.365	.000	.145	.000
TrFo, D=2	.465	.220	.303	$-.061 \pm .034$	.138	$-.008 \pm .002$
TrFo, D=3	.441	.187	.273	$-.092 \pm .032$	.134	$-.011 \pm .001$
TrFo, D=4	<b>.429</b>	<b>.175</b>	<b>.254</b>	$-.110 \pm .030$	<b>.129</b>	$-.017 \pm .001$

(a) Test loss results.

Model	Ep.: 1	5	M 1–5	MD 1–5	M 40–50	MD 40–50
LSTM, D=1	.731	.793	.783	.000	<b>.894</b>	.000
LSTM, D=2	<b>.745</b>	.794	.760	$-.023 \pm .018$	.856	$-.038 \pm .010$
LSTM, D=3	.729	<b>.826</b>	<b>.794</b>	<b>.011<math>\pm</math>.010</b>	.870	$-.025 \pm .015$
LSTM, D=4	.724	.763	.753	$-.030 \pm .017$	.840	$-.054 \pm .017$
TrFo, D=1	.733	.909	.860	.000	.942	.000
TrFo, D=2	.833	<b>.941</b>	.896	<b>.036<math>\pm</math>.022</b>	.944	.002 $\pm$ .012
TrFo, D=3	.839	<b>.941</b>	.908	<b>.048<math>\pm</math>.023</b>	<b>.946</b>	.004 $\pm$ .011
TrFo, D=4	<b>.844</b>	.929	<b>.911</b>	<b>.051<math>\pm</math>.022</b>	.939	$-.003 \pm .013$

(b) Test accuracy results.

Figure 1: Test data results for varying number of decoder layers with one encoder layer.

Model	Ep.: 1	5	M 1-5	MD 1-5	M 40-50	MD 40-50
LSTM, E=1	.812	.519	<b>.653</b>	.000	.340	.000
LSTM, E=2	<b>.803</b>	<b>.502</b>	.658	.005 $\pm$ .010	<b>.320</b>	-.020 $\pm$ .003
LSTM, E=3	.904	.879	.890	<b>.237<math>\pm</math>.064</b>	.418	<b>.078<math>\pm</math>.024</b>
TrFo, E=1	<b>.441</b>	.187	<b>.273</b>	.000	.134	.000
TrFo, E=2	.522	.189	.290	.017 $\pm$ .020	.126	-.008 $\pm$ .001
TrFo, E=3	.606	<b>.181</b>	.302	.029 $\pm$ .042	<b>.122</b>	-.012 $\pm$ .001

(a) Test loss results.

Model	Ep.: 1	5	M 1-5	MD 1-5	M 40-50	MD 40-50
LSTM, E=1	.729	<b>.826</b>	<b>.794</b>	.000	.870	.000
LSTM, E=2	<b>.748</b>	.817	.774	-.020 $\pm$ .014	<b>.880</b>	.010 $\pm$ .010
LSTM, E=3	.694	.733	.728	-.066 $\pm$ .014	.836	-.034 $\pm$ .012
TrFo, E=1	.839	<b>.941</b>	<b>.908</b>	.000	.946	.000
TrFo, E=2	<b>.841</b>	.919	.896	-.012 $\pm$ .005	.952	.005 $\pm$ .008
TrFo, E=3	.774	.927	.897	-.011 $\pm$ .019	<b>.954</b>	.008 $\pm$ .016

(b) Test accuracy results.

Figure 2: LSTM and Transformer results for different numbers of encoder layers, with decoder layers fixed to one.

Model	Ep.: 1	5	M 1-5	MD 1-5	M 40-50	MD 40-50
TrFo 1H, E=1	.449	.230	.313	.000	.144	.000
TrFo 2H, E=2	<b>.437</b>	.216	.292	-.020 $\pm$ .005	.138	-.006 $\pm$ .001
TrFo 4H, E=3	.441	.187	<b>.273</b>	-.040 $\pm$ .011	.134	-.010 $\pm$ .001
TrFo 8H, E=3	.507	<b>.181</b>	.277	-, 035 $\pm$ .029	<b>.127</b>	-.017 $\pm$ .001

(a) Test loss results.

Model	Ep.: 1	5	M 1-5	MD 1-5	M 40-50	MD 40-50
TrFo 1H	.850	.928	.894	.000	.941	.000
TrFo 2H	<b>.875</b>	.931	.896	.002 $\pm$ .012	.948	.007 $\pm$ .008
TrFo 4H	.839	.941	<b>.908</b>	.015 $\pm$ .017	.946	.005 $\pm$ .010
TrFo 8H	.803	<b>.953</b>	.900	.002 $\pm$ .020	<b>.950</b>	.008 $\pm$ .010

(b) Test accuracy results.

Figure 3: Transformer results for different numbers of attention heads.

Model	Ep.: 1	5	M 1-5	MD 1-5	M 40-50	MD 40-50
TrFo 1H, E=1	.449	.230	.313	.000	.144	.000
TrFo 2H, E=2	<b>.437</b>	.216	.292	— <b>.020<math>\pm</math>.005</b>	.138	— <b>.006<math>\pm</math>.001</b>
TrFo 4H, E=3	.441	.187	<b>.273</b>	— <b>.040<math>\pm</math>.011</b>	.134	— <b>.010<math>\pm</math>.001</b>
TrFo 8H, E=3	.507	<b>.181</b>	.277	— <b>.035<math>\pm</math>.029</b>	<b>.127</b>	— <b>.017<math>\pm</math>.001</b>

(a) Test loss results.

Model	Ep.: 1	5	M 1-5	MD 1-5	M 40-50	MD 40-50
TrFo 1H	.850	.928	.894	.000	.941	.000
TrFo 2H	<b>.875</b>	.931	.896	.002 $\pm$ .012	.948	.007 $\pm$ .008
TrFo 4H	.839	.941	<b>.908</b>	.015 $\pm$ .017	.946	.005 $\pm$ .010
TrFo 8H	.803	<b>.953</b>	.900	.002 $\pm$ .020	<b>.950</b>	.008 $\pm$ .010

(b) Test accuracy results.

Figure 4: Transformer results for different numbers of attention heads.

Model	Ep.: 1	5	M 1-5	MD 1-5	M 40-50	MD 40-50
LSTM/TrFo, D=1	.829/. <b>.635</b>	.513/. <b>.251</b>	.661/. <b>.365</b>	<b>.296<math>\pm</math>.041</b>	.274/. <b>.145</b>	<b>.129<math>\pm</math>.006</b>
LSTM/TrFo, D=2	.808/. <b>.465</b>	.675/. <b>.220</b>	.750/. <b>.303</b>	<b>.447<math>\pm</math>.033</b>	.385/. <b>.138</b>	<b>.247<math>\pm</math>.017</b>
LSTM/TrFo, D=3	.812/. <b>.441</b>	.519/. <b>.187</b>	.653/. <b>.273</b>	<b>.380<math>\pm</math>.025</b>	.340/. <b>.134</b>	<b>.206<math>\pm</math>.002</b>
LSTM/TrFo, D=4	.899/. <b>.429</b>	.682/. <b>.175</b>	.760/. <b>.254</b>	<b>.506<math>\pm</math>.012</b>	.382/. <b>.129</b>	<b>.254<math>\pm</math>.005</b>

(a) Test loss results.

Model	Ep.: 1	5	M 1-5	MD 1-5	M 40-50	MD 40-50
LSTM/TrFo, D=1	<b>.731/.733</b>	.793/. <b>.909</b>	.783/. <b>.860</b>	— <b>.077<math>\pm</math>.028</b>	.894/. <b>.942</b>	— <b>.048<math>\pm</math>.011</b>
LSTM/TrFo, D=2	<b>.745/.833</b>	.794/. <b>.941</b>	.760/. <b>.896</b>	— <b>.137<math>\pm</math>.018</b>	.856/. <b>.944</b>	— <b>.088<math>\pm</math>.010</b>
LSTM/TrFo, D=3	<b>.729/.839</b>	.826/. <b>.941</b>	.794/. <b>.908</b>	— <b>.114<math>\pm</math>.008</b>	.870/. <b>.946</b>	— <b>.077<math>\pm</math>.011</b>
LSTM/TrFo, D=4	<b>.724/.844</b>	.763/. <b>.929</b>	.753/. <b>.911</b>	— <b>.159<math>\pm</math>.016</b>	.840/. <b>.939</b>	— <b>.099<math>\pm</math>.015</b>

(b) Test accuracy results.

Figure 5: LSTM and Transformer results compared for different numbers of decoder layers, with encoder layers fixed to one.

<b>Model</b>	<b>Ep.: 1</b>	<b>5</b>	<b>M 1–5</b>	<b>MD 1–5</b>	<b>M 40–50</b>	<b>MD 40–50</b>
LSTM/TrFo, E=1	.812/. <b>441</b>	.519/. <b>187</b>	.653/. <b>273</b>	.380 $\pm$ .025	.340/. <b>134</b>	.206 $\pm$ .002
LSTM/TrFo, E=2	.803/. <b>522</b>	.502/. <b>189</b>	.658/. <b>290</b>	.368 $\pm$ .039	.320/. <b>126</b>	.195 $\pm$ .004
LSTM/TrFo, E=3	.904/. <b>606</b>	.879/. <b>181</b>	.890/. <b>302</b>	.588 $\pm$ .092	.418/. <b>122</b>	.296 $\pm$ .024

(a) Test loss results.

<b>Model</b>	<b>Ep.: 1</b>	<b>5</b>	<b>M 1–5</b>	<b>MD 1–5</b>	<b>M 40–50</b>	<b>MD 40–50</b>
LSTM/TrFo, E=1	.729/. <b>839</b>	.826/. <b>941</b>	.794/. <b>908</b>	–.114 $\pm$ .008	.870/. <b>946</b>	–.077 $\pm$ .011
LSTM/TrFo, E=2	.748/. <b>841</b>	.817/. <b>919</b>	.774/. <b>896</b>	–.122 $\pm$ .015	.880/. <b>952</b>	–.072 $\pm$ .015
LSTM/TrFo, E=3	.694/. <b>774</b>	.733/. <b>927</b>	.728/. <b>897</b>	–.169 $\pm$ .030	.836/. <b>954</b>	–.118 $\pm$ .017

(b) Test accuracy results.

Figure 6: LSTM and Transformer results compared for different numbers of encoder layers, with decoder layers fixed to three.