1 Question answering task on the SQUADv2 dataset

	SQUADv2 (Exact Match)	SQUADv2 (F1)	SQUADv2 (Train loss)
Adam	48.41 ± 0.57	49.99 ± 0.54	2.73 ± 0.01
M-FAC	49.80 ± 0.43	52.18 ± 0.20	2.44 ± 0.02

Table 1: Comparing M-FAC optimizer (without weight decay) against HuggingFace's Adam baseline on the **bert-tiny** model.

	SQUADv2 (Exact Match)	SQUADv2 (F1)	SQUADv2 (Train loss)
Adam	54.80 ± 0.47	58.13 ± 0.31	1.86 ± 0.02
M-FAC	58.02 ± 0.39	61.35 ± 0.24	1.75 ± 0.01

Table 2: Comparing M-FAC optimizer (without weight decay) against HuggingFace's Adam baseline on the **bert-mini** model.

2 Text classification on a subset of GLUE tasks

	SST-2 (Acc.)	SST-2 (Train loss)	MRPC (F1)	MRPC (Acc.)	MRPC (Train loss)
Adam	80.11 ± 0.65	0.41 ± 0.01	81.68 ± 0.33	69.90 ± 0.32	0.61 ± 0.01
M-FAC	81.86 ± 0.76	0.32 ± 0.01	82.77 ± 0.22	72.94 ± 0.37	0.55 ± 0.01

	STS-B (Pearson)	STS-B (Spearman)	STS-B (Train loss)	QNLI (Acc.)	QNLI (Train loss)
Adam	64.39 ± 5.02	66.52 ± 5.67	4.04 ± 0.45	77.85 ± 0.15	0.50 ± 0.01
M-FAC	80.15 ± 0.52	80.62 ± 0.43	1.09 ± 0.02	81.17 ± 0.43	0.48 ± 0.01

	QQP (F1)	QQP (Acc.)	QQP (Train loss)
Adam	77.58 ± 0.08	81.09 ± 0.15	0.42 ± 0.01
M-FAC	79.71 ± 0.13	84.29 ± 0.08	0.40 ± 0.01

	MNLI-m (Acc.)	MNLI-mm (Acc.)	MNLI (Train loss)
Adam	65.36 ± 0.13	66.78 ± 0.15	0.85 ± 0.01
M-FAC	68.28 ± 3.29	68.98 ± 3.05	0.81 ± 0.05

Table 3: Comparing M-FAC optimizer (without weight decay) against HuggingFace's Adam baselines on the **bert-tiny** model.

	SST-2 (Acc.)	SST-2 (Train loss)	MRPC (F1)	MRPC (Acc.)	MRPC (Train loss)
Adam	85.46 ± 0.58	0.31 ± 0.01	84.57 ± 0.36	76.57 ± 0.80	0.54 ± 0.01
M-FAC	84.20 ± 0.58	0.29 ± 0.01	85.06 ± 1.63	78.87 ± 2.33	0.46 ± 0.01

	STS-B (Pearson)	STS-B (Spearman)	STS-B (Train loss)	QNLI (Acc.)	QNLI (Train loss)
Adam	82.09 ± 0.54	82.64 ± 0.71	1.58 ± 0.10	83.85 ± 0.10	0.41 ± 0.01
M-FAC	84.66 ± 0.30	84.65 ± 0.30	0.85 ± 0.03	83.70 ± 0.13	0.42 ± 0.01

	QQP (F1)	QQP (Acc.)	QQP (Train loss)
Adam	82.43 ± 0.10	86.45 ± 0.12	0.34 ± 0.01
M-FAC	82.67 ± 0.23	86.75 ± 0.20	0.35 ± 0.01

	MNLI-m (Acc.)	MNLI-mm (Acc.)	MNLI (Train loss)
Adam	73.30 ± 0.20	74.85 ± 0.09	0.70 ± 0.01
M-FAC	74.59 ± 0.41	75.95 ± 0.14	0.68 ± 0.01

Table 4: Comparing M-FAC optimizer (without weight decay) against HuggingFace's Adam baselines on the **bert-mini** model.