Appendix

A Hyper-Parameters for Knowledge Injection

The hyper-parameters for knowledge injection for three closed-book question answering datasets are summarized in Table 9. We train fewer steps for HEAD-QA since the size of its retrieved knowledgeable corpus is relatively small.

Hyper-Parameters	**Natural Questions**	**WebQuestions**	**HEAD-QA**
# NKB Slot	3072	3072	3072
NKB Position	the last decoder layer	the last decoder layer	the last decoder layer
Dropout for NKB	0	0	0
Sequence Length	512 tokens	512 tokens	512 tokens
Batch Size	256	256	256
Optimizer	AdaFactor	AdaFactor	AdaFactor
Maximum Learning Rate	1e-3	1e-3	1e-2
Learning Rate Scheduler	constant with warmup	constant with warmup	constant with warmup
Total Steps	30K	30K	3K
Warm-up Steps	5K	5K	0.5K
Gradient Clip Norm	1.0	1.0	1.0
Random Seed	1234	1234	1234

Table 9. Hyper-parameters for knowledge injection for three closed-book question answering datasets.

B Hyper-Parameters for Finetuning

The hyper-parameters for closed-book question answering are summarized in Table 10. For all the reported models, we use the same hyper-parameters.

The hyper-parameters for summarization and machine translation are summarized in Table 11. For all the reported models except for the vanilla Transformer, we use the same hyper-parameters. For the vanilla Transformer, we train 200K steps for summarization and machine translation since it converges more slowly than pretrained models.

Hyper-Parameters	**Natural Questions**	**WebQuestions**	**HEAD-QA**
Maximum Sequence Length	256 tokens	256 tokens	256 tokens
Batch Size	768	256	256
Optimizer	AdaFactor	AdaFactor	AdaFactor
Maximum Learning Rate	1e-3	1e-3	1e-3
Learning Rate Scheduler	constant	constant	constant
Total Steps	10K	5K	1K
Gradient Clip Norm	1.0	1.0	1.0
Dropout	0.1	0.2	0.2
Dropout for NKB	0	0	0
Random Seed	1234	1234	1234

Table 10. Hyper-parameters for finetuning on closed-book question answering.

Hyper-Parameters	**Xsum**	**WMT-En-De**	**WMT-En-Ro**
Maximum Source Sequence Length	512 tokens	128 tokens	128 tokens
Maximum Target Sequence Length	128 tokens	128 tokens	128 tokens
Batch Size	16	96	96
Optimizer	Adam	Adam	Adam
Maximum Learning Rate	3e-5	3e-5	3e-5
Learning Rate Scheduler	linear with warmup	linear with warmup	linear with warmup
Total Steps	50K	50K	50K
Warm-up Steps	10K	10K	10K
Gradient Clip Norm	1.0	1.0	1.0
Label Smoothing	0.1	0.1	0.1
Beam Size	4	4	4
Dropout	0.1	0.1	0.1
Dropout for NKB	0	0	0
Random Seed	1234	1234	1234

Table 11. Hyper-parameters for finetuning on summarization and machine translation.