A Details of Models

Most of the models we targeted in attacks come from Hugging Face Model Hub (`https://huggingface.co/models`). Since no BERT-small model was available, we imported the corresponding publically-available TensorFlow model. When a fine-tuned model for an architecture/task combination was not available, we trained one ourselves using standard parameters using a version of the `run_glue.py` script from HuggingFace Transformers library extended to train models on AG News. Concretely, we fine-tuned the pre-trained model for 3 epochs using an initial learning rate of 2×10^{-5} using AdamW with a linear learning rate scheduler. We used a batch size of 8 for RoBERTa-base and BART-large and 32 for the rest.

Architecture (params)	Pre-trained model	Fine-tuned model	Accuracy
BERT-small (15M)	From TensorFlow	Own trained	0.875
BERT-base (110M)	bert-base-uncased	textattack/bert-base-uncased-SST-2	0.924
RoBERTa-base (125M)	roberta-base	textattack/roberta-base-SST-2	0.940
ALBERT-base-v2 (11M)	albert-base-v2	textattack/albert-base-v2-SST-2	0.925
XLNet-base (110M)	xlnet-base-cased	textattack/xlnet-base-cased-SST-2	0.944
DistilBERT (66M)	distilbert-base-uncased	Own trained	0.909
BART-large (406M)	facebook/bart-large	textattack/facebook-bart-large-SST-2	0.953

Table 1: Details of target models for SST-2 task. Named models were obtained from Hugging Face Model Hub.

Architecture (params)	Pre-trained model	Fine-tuned model	Accuracy
BERT-small (15M)	From TensorFlow	Own trained	0.777
BERT-base (110M)	bert-base-uncased	textattack/bert-base-uncased-SST-2	0.846
RoBERTa-base (125M)	roberta-base	textattack/roberta-base-MNLI	0.881
ALBERT-base-v2 (11M)	albert-base-v2	Own trained	0.849
XLNet-base (110M)	xlnet-base-cased	textattack/xlnet-base-cased-MNLI	0.871
DistilBERT (66M)	distilbert-base-uncased	Own trained	0.812
BART-large (406M)	facebook/bart-large	textattack/facebook-bart-large-MNLI	0.889

Table 2: Details of target models for MNLI task. Named models were obtained from Hugging Face Model Hub.

Architecture (params)	Pre-trained model	Fine-tuned model	Accuracy
BERT-small (15M)	From TensorFlow	Own trained	0.939
BERT-base (110M)	bert-base-uncased	textattack/bert-base-uncased-ag-news	0.955
RoBERTa-base (125M)	roberta-base	textattack/roberta-base-ag-news	0.948
ALBERT-base-v2 (11M)	albert-base-v2	textattack/albert-base-v2-ag-news	0.948
XLNet-base (110M)	xlnet-base-cased	Own trained	0.951
DistilBERT (66M)	distilbert-base-uncased	textattack/distilbert-base-uncased-ag-news	0.951
BART-large (406M)	facebook/bart-large	Own trained	0.952

Table 3: Details of target models for AG News task. Named models were obtained from Hugging Face Model Hub.

B Results with BERT-Small

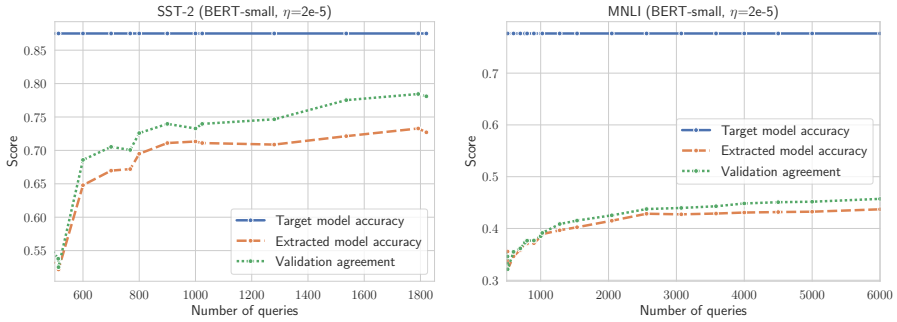

Figure 1: Effect of number of inputs on extracted model task accuracy and agreement of extracted model. Baseline accuracy with a random guess is 50% for SST-2 and 33% for MNLI.

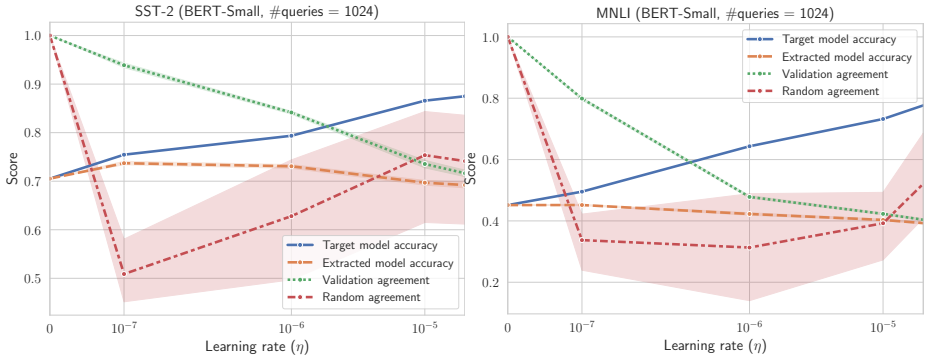

Figure 2: Effect of learning rate on original model task accuracy, extracted model task accuracy, and fidelity of extracted model.

C Full Experimental Results Varying Learning Rate

Effect of learning rate on task accuracy of extracted models, and agreement with target model on in-distribution (Fig. 3) and random (Fig. 4) queries.

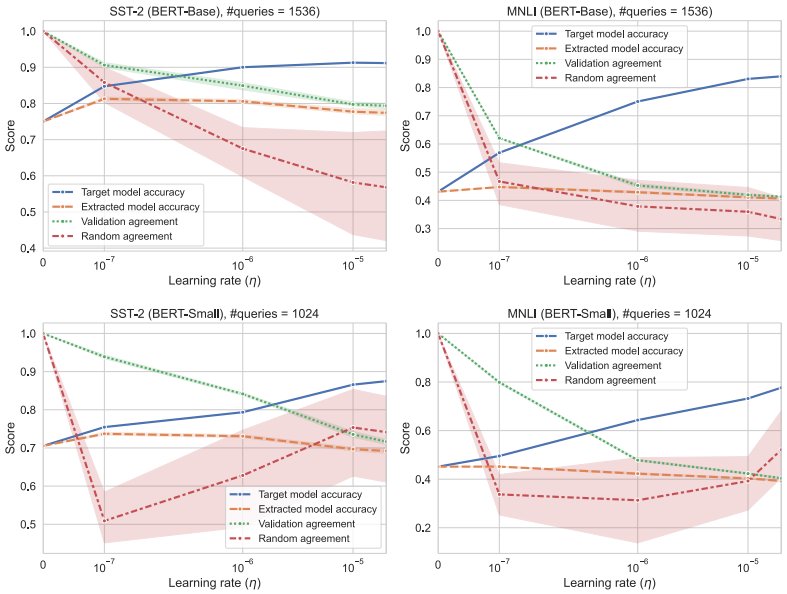

Figure 3: Extraction with in-distribution queries

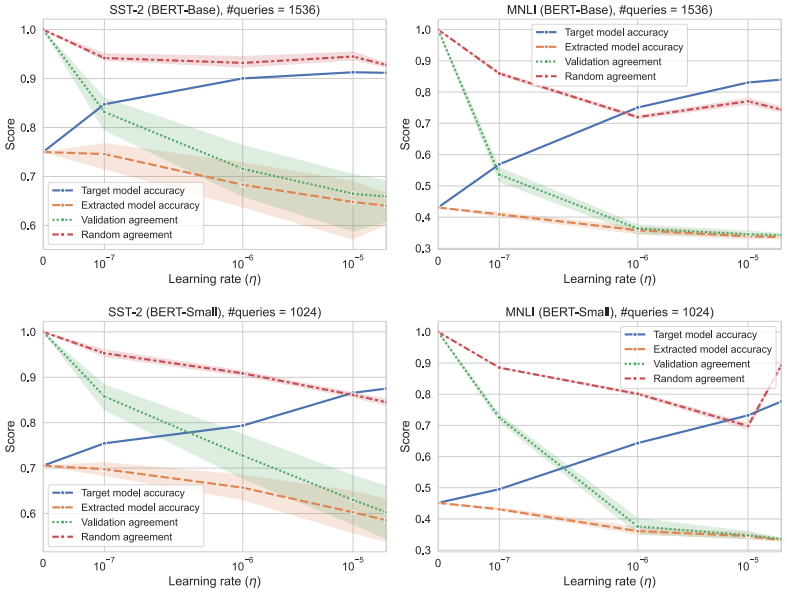

Figure 4: Extraction with random queries

D Full Experimental Results Varying Number of Queries

D.1 BERT-Base and SST-2

Effect of number of queries on task accuracy of extracted model and agreement with target model, for in-distribution (Fig. 5) and random (Fig. 6) queries. Baseline accuracy of random guess: 50%.

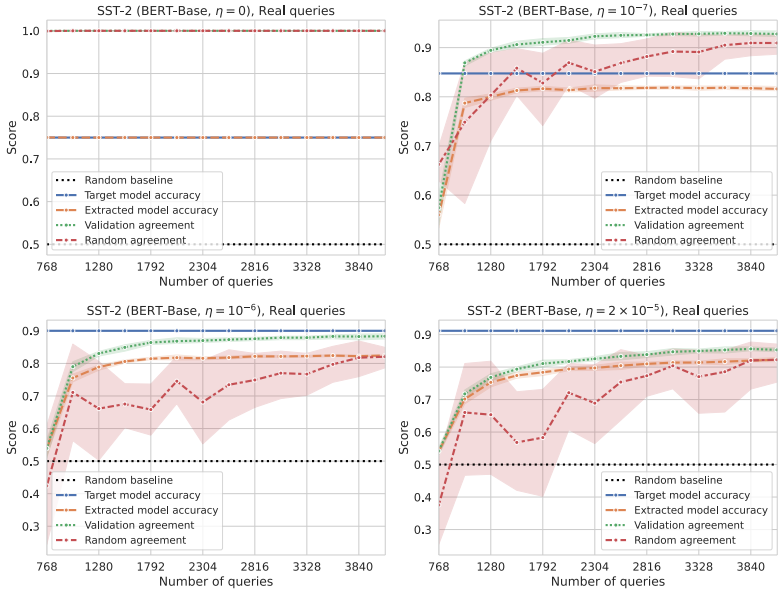

Figure 5: Extraction with in-distribution queries

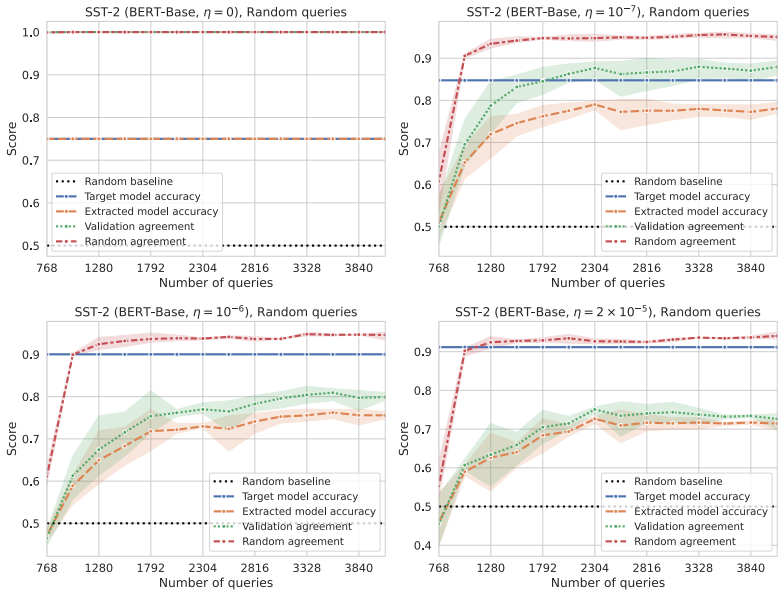

Figure 6: Extraction with random queries

D.2 BERT-Base and MNLI

Effect of number of queries on task accuracy of extracted model and agreement with target model, for in-distribution (Fig. 7) and random (Fig. 8) queries. Baseline accuracy of random guess: 33%.

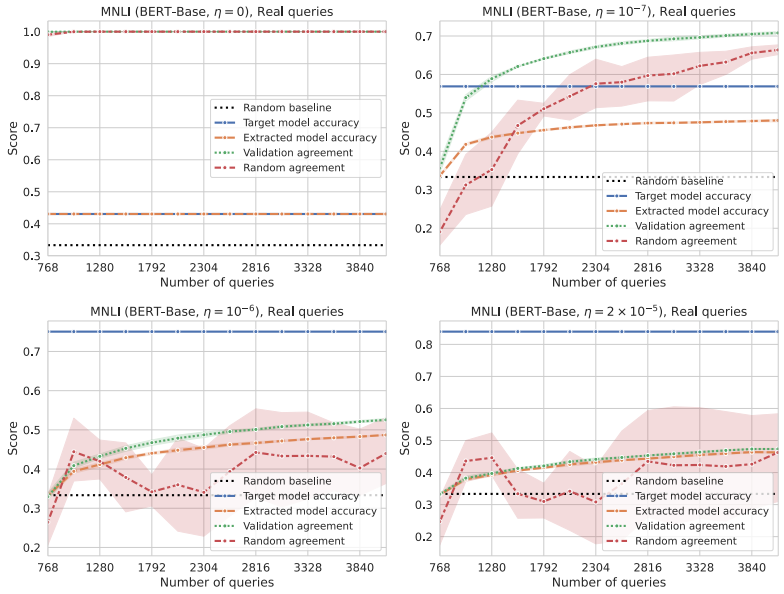

Figure 7: Extraction with in-distribution queries

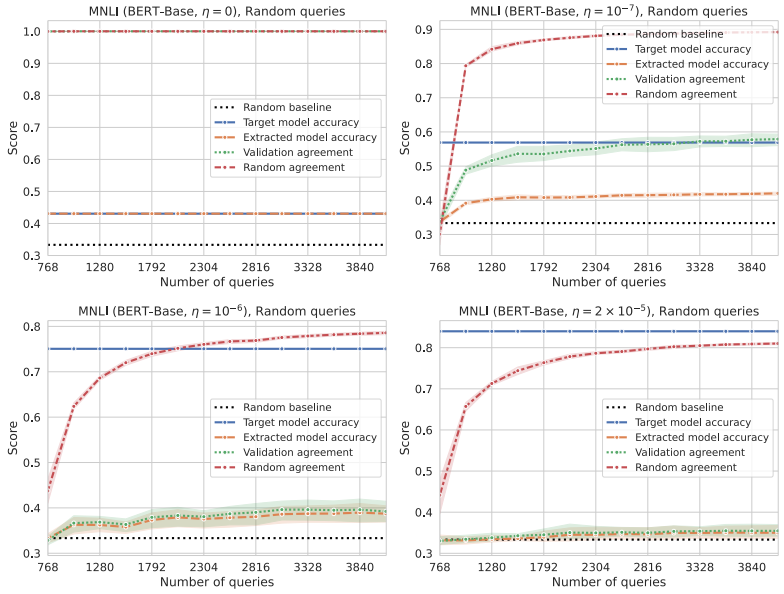

Figure 8: Extraction with random queries

D.3 BERT-Small and SST-2

Effect of number of queries on task accuracy of extracted model and agreement with target model, for in-distribution (Fig. 9) and random (Fig. 10) queries. Baseline accuracy of random guess: 50%.

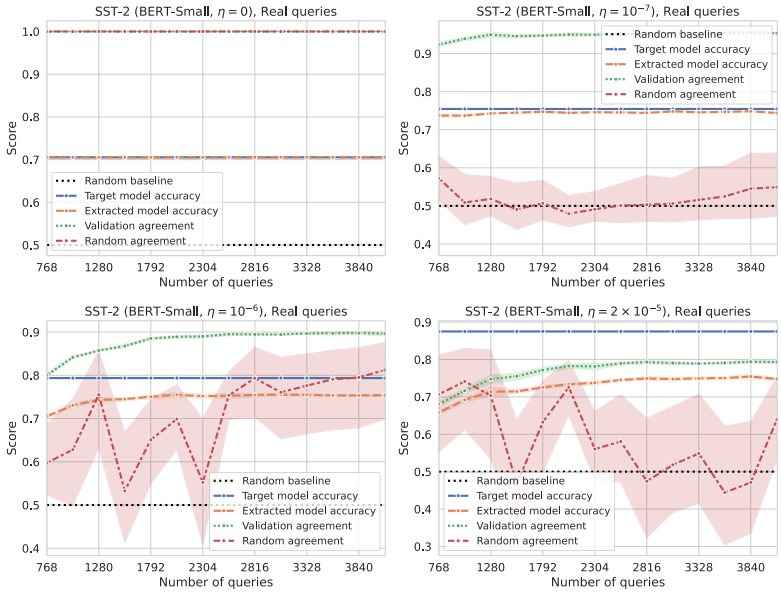

Figure 9: Extraction with in-distribution queries

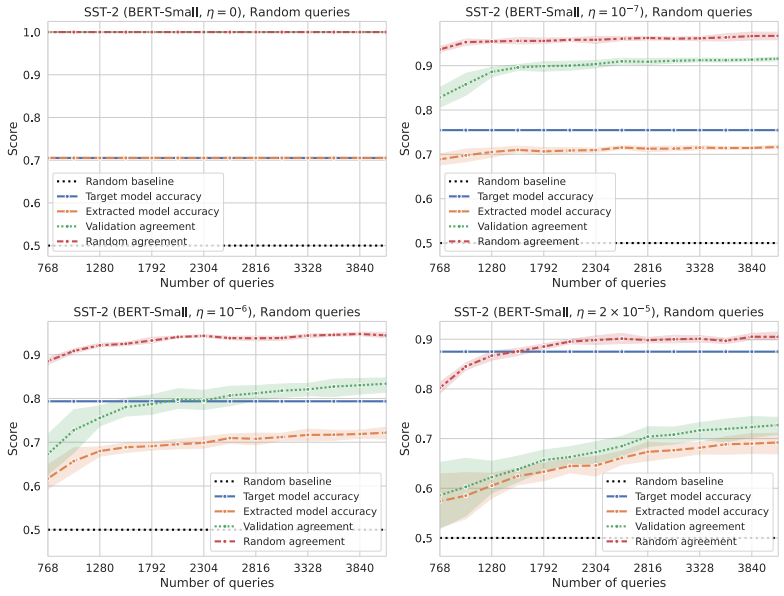

Figure 10: Extraction with random queries

D.4 BERT-Small and MNLI

Effect of number of queries on task accuracy of extracted model and agreement with target model, for in-distribution (Fig. 11) and random (Fig. 12) queries. Baseline accuracy of random guess: 33%.

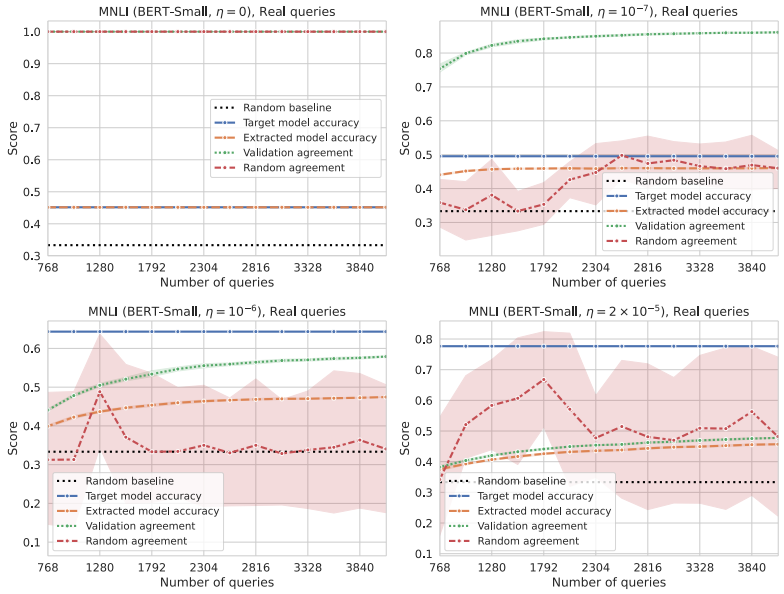

Figure 11: Extraction with in-distribution queries

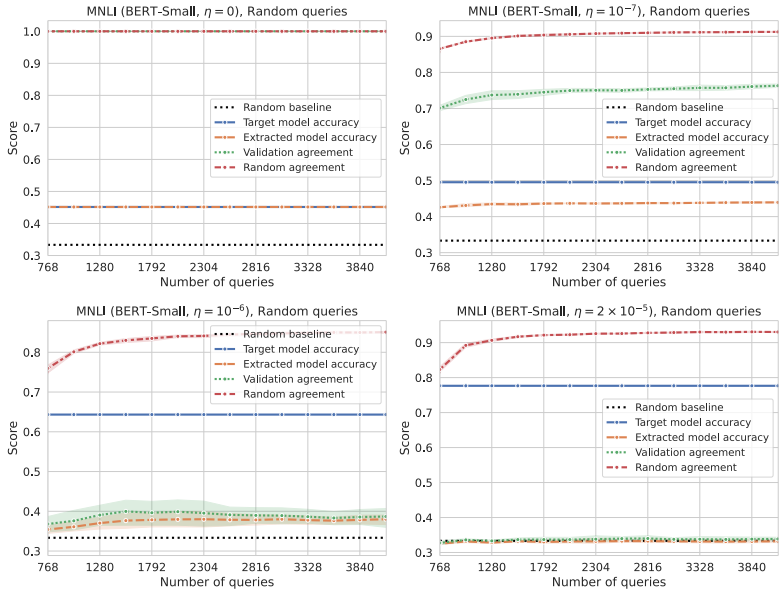

Figure 12: Extraction with random queries