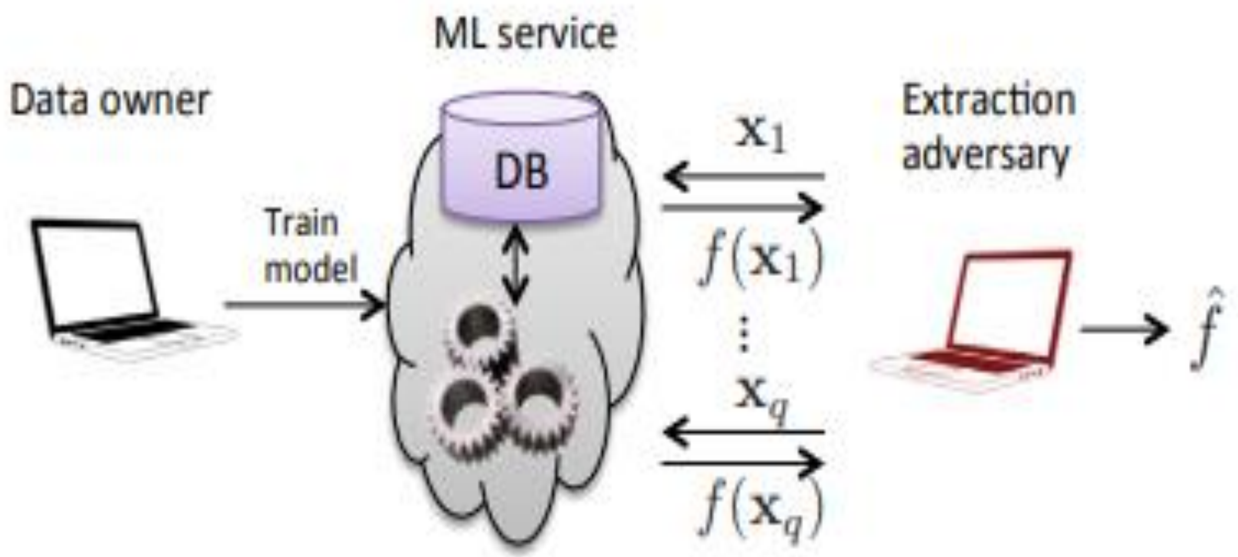
Actively Fake it Until you Make it in Neural Machine Translation

- Frank Kelly
- R00044319
- MSc Artificial Intelligence

Related Work:

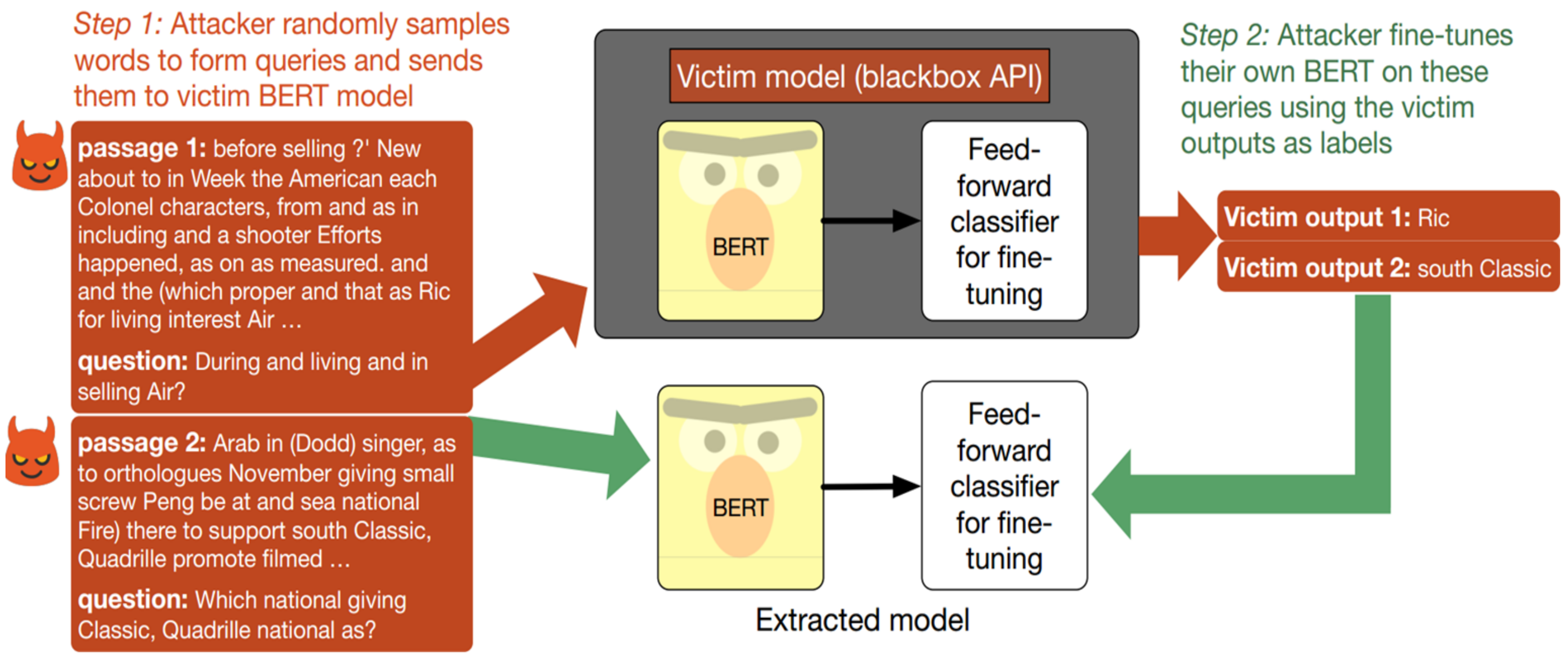
Create functionally equivalent models given only query access to a victim model



Source: Stealing Machine Learning Models via Prediction APIs, (Tramer et al 2016)

Related Work

Function Approximation of Neural Network in NLP domain



Krishna, K., et al. (2019). "Thieves on sesame street! model extraction of bert-based apis." arXiv preprint arXiv:1910.12366.

Model Extraction Attacks

- Differential Extraction Attack (Carlini et al 2020)
- Bus snooping Architecture Extraction (Hu, Liang et al. 2020)

Frank Kelly
R00044319

Why is Model Extraction A Problem?

Undermines:

- Valuable Pay for Prediction API Business Models
- Intellectual Property
- Data Privacy
- Models' security

DeepL

Research Aim:

Evaluate how effective a synthetic parallel corpus created via active learning model extraction is at training a substitute neural machine translation model.

Research Objectives:

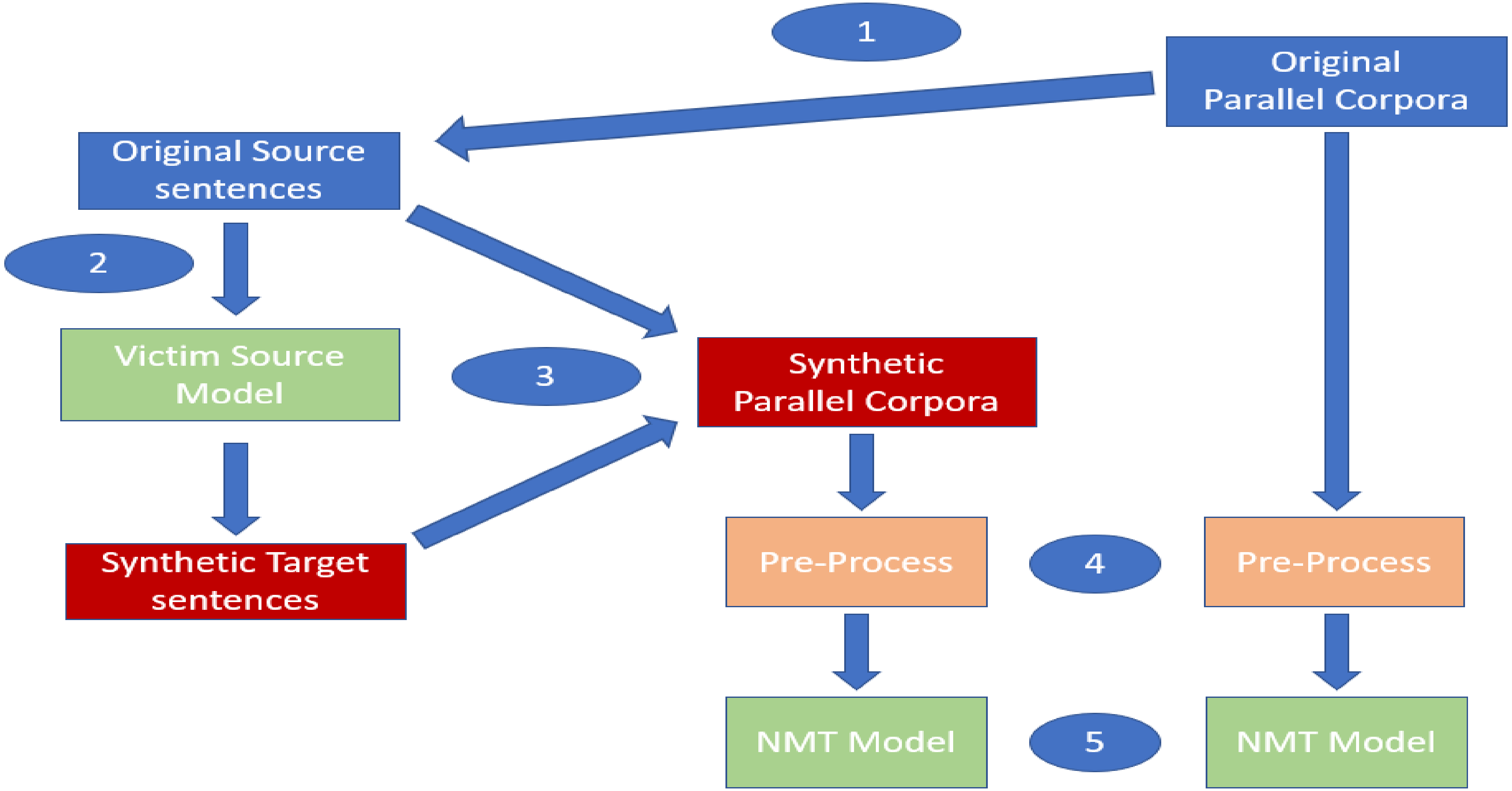
- Compare performance between NMT models trained on synthetic data and associated original data.
- Analyse performance obtained from training on synthetic datasets of various sample sizes.
- Analyse NMT model performance with various translation evaluation metrics.
- Perform statistical significance test on evaluation metric results.

Research Contribution

- Determine threat posed by ALME to the monetization of NMT APIs.
- Evaluate how effective Active Learning is with modern NMTs
- Analyse NMT model performance with various translation evaluation metrics.

Experiment Methodology

Active Learning Model Extraction Experiment Pipeline



Victim Source Models

Victim Source Model Experiment 1

OpenNMT

An open source neural machine translation system.

Pretrained transformer models from OpenNMT-tf
(Klein, Hernandez, Nguyen, & Senellart, 2020)
Source: https://opennmt.net/Models-tf/

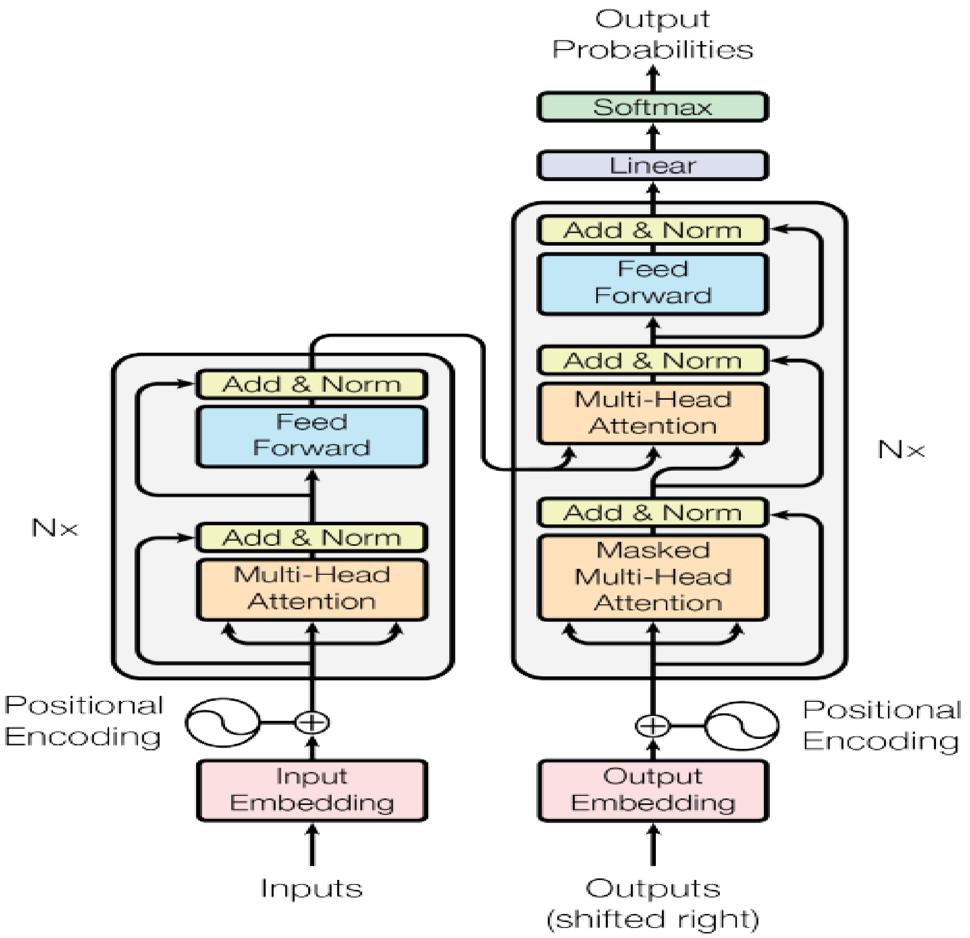
Victim Source Model Experiment 2 to 7

FAIRSEQ

Facebook FAIR's WMT19 News Translation Task Submission
(Ng et al., 2019) https://opennmt.net/Models-tf/
Source: https://github.com/pytorch/fairseq/blob/master/examples/wmt19/

Adversary NMT Model Selection

Transformer Model



Ref: (Vaswani et al., 2017)

Datasets Used

EMNLP 2017
SECOND CONFERENCE ON
MACHINE TRANSLATION (WMT17)

http://data.statmt.org/wmt17/translation-task/

ParaCrawl

https://s3.amazonaws.com/web-language-models/paracrawl/release5.1/

OPUS

the open parallel corpus

https://opus.nlpl.eu/download.php?f=CCAligned/v1/

UFAL Medical Corpus

ÚFAL

https://ufal.mff.cuni.cz/ufal_medical_corpus

Pre-Processing

- Tokenise
- Normalise
- Remove Long sentences
- Train Subword Tokeniser
- Apply Subword Tokeniser

Evaluation Metrics

- Human Relative Ranking (Callison-Burch et al., 2008)
- Bilingual Evaluation Understudy (BLEU) metric (Papineni et al., 2002)
- Word and character n-gram F-scores (chrF++) metric (Popović, 2017)

Statistical Significance Tests

- **Wilcoxon signed rank test**
- **Paired Bootstrap resampling**
- **Approximate randomization exchanges**

Statistical Significance Tests

- **Wilcoxon signed rank test**
- **Paired Bootstrap resampling**
- **Approximate randomization exchanges**

Hardware

Google Compute Engine

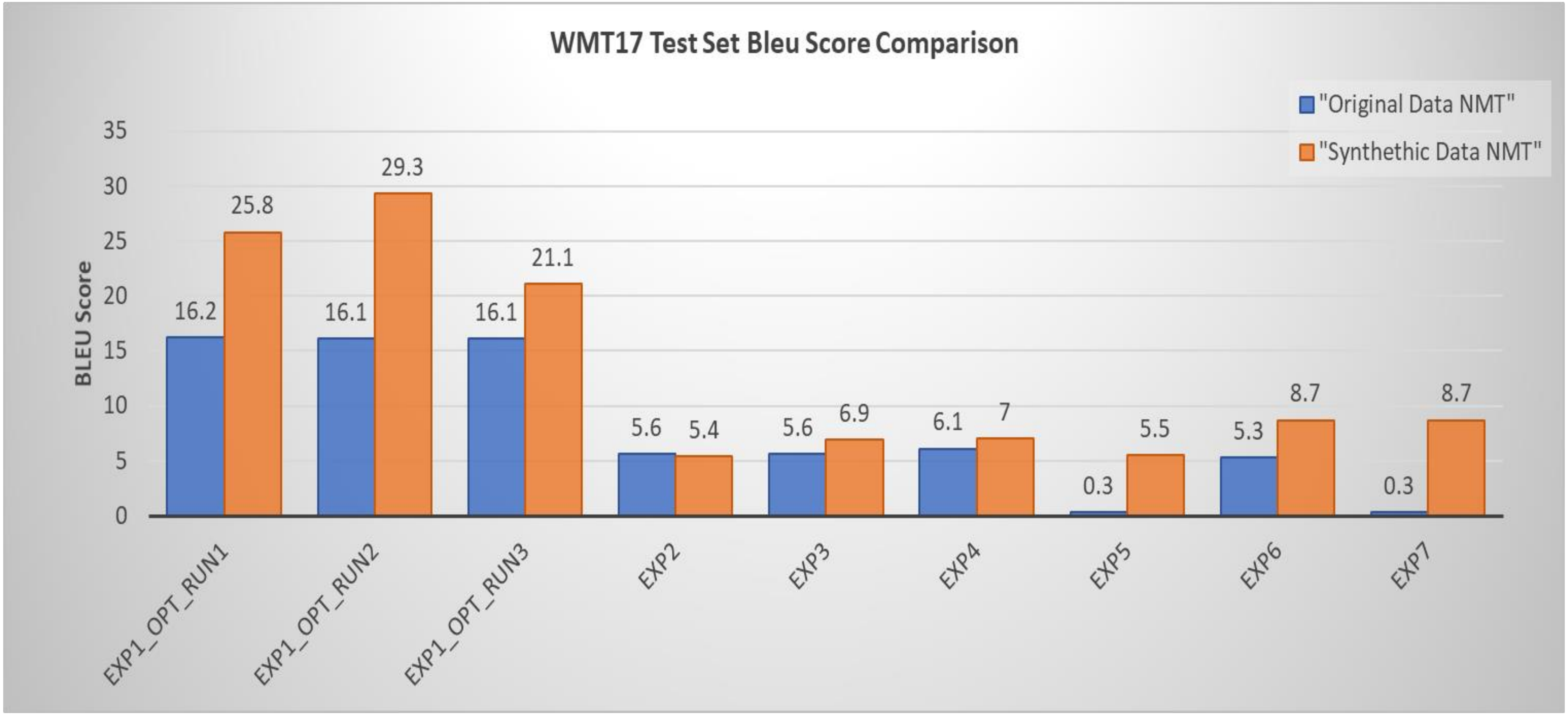
1 x NVIDIA TESLA V100 GPU

Implementation

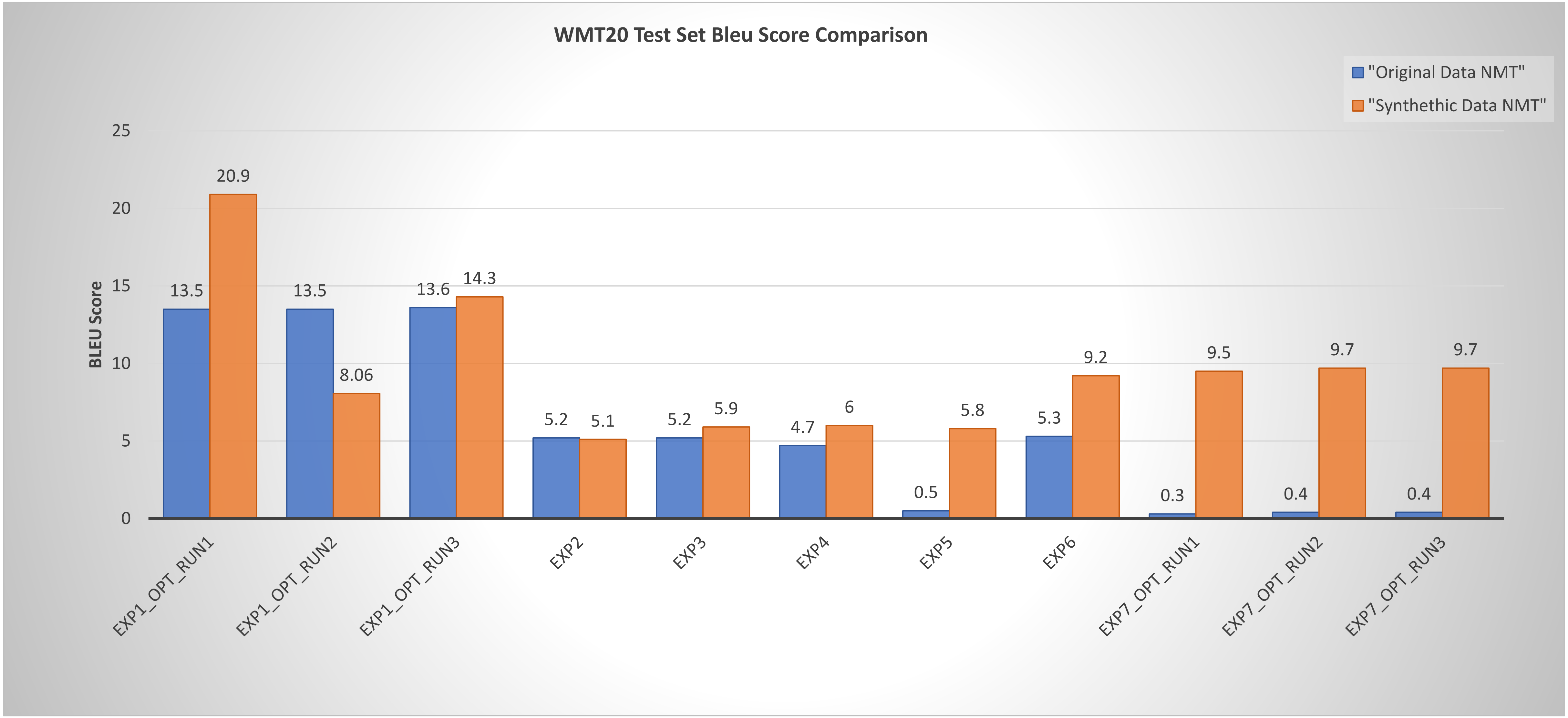
Experiment	Corpus	Framework	Source Model Supplied by	Source Model BLUE Score on *	Optimizer Runs	Encoding	Training Data Size	Min-Max Sentence Length tokens	Updates during Training	Aprox Training Time (hrs)
1	WMT2017	Tensorflow	ONMT-TF	28	3	Sentece Piece	100,000	2 to 200	50k	17
2	Paracrawl	Pytorch	Fairseq	30.9	1	BPE	100,000	2 to 200	22K	2.5
3	OPUS	Pytorch	Fairseq	30.9	1	BPE	100,000	2 to 200	20k	2
4	OPUS	Pytorch	Fairseq	30.9	1	BPE	100,000	2 to 80	20k	2
5	UFAL	Pytorch	Fairseq	30.9	1	BPE	100,000	2 to 80	20k	2
6	UFAL	Pytorch	Fairseq	30.9	1	BPE	3,500,000	2 to 80	50k	4.5
7	UFAL	Pytorch	Fairseq	30.9	3	BPE	11,500,000	2 to 80	100k	12

*Stated BLEU score was achieved on WMT2018 New test set.

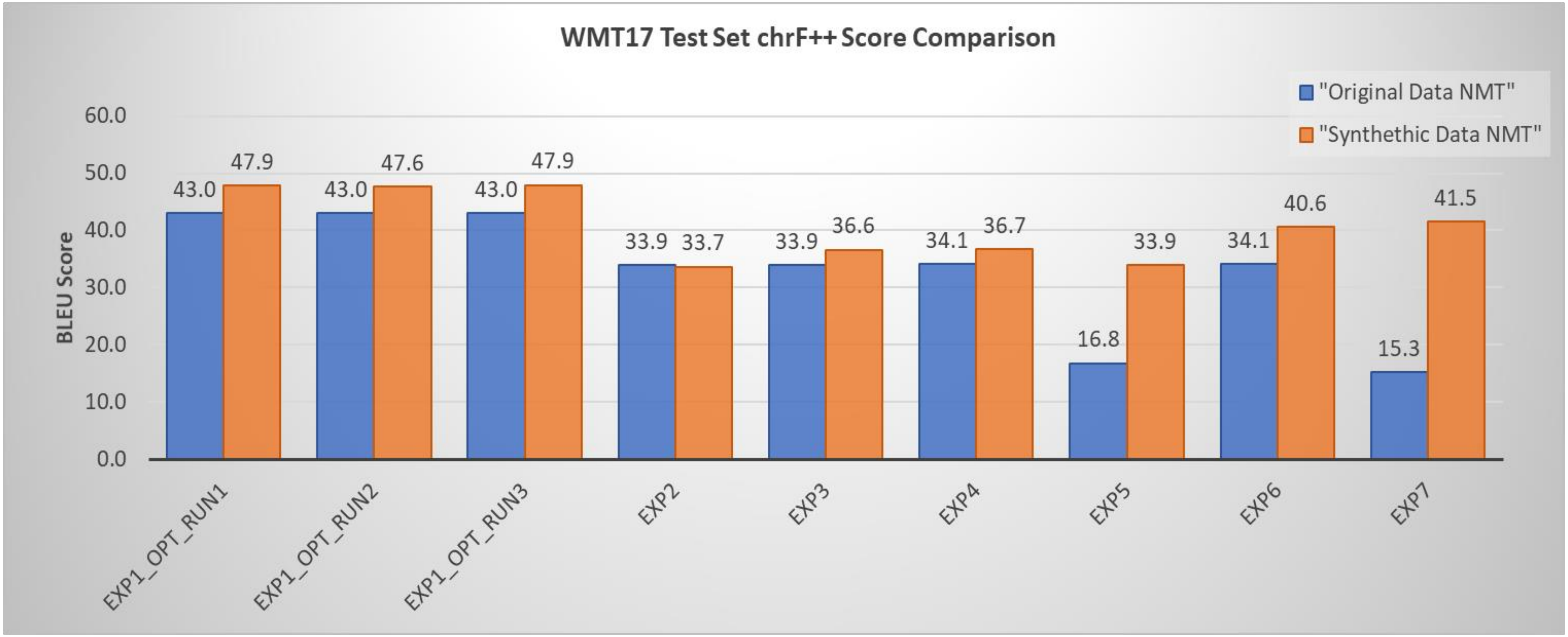
Experiment Results



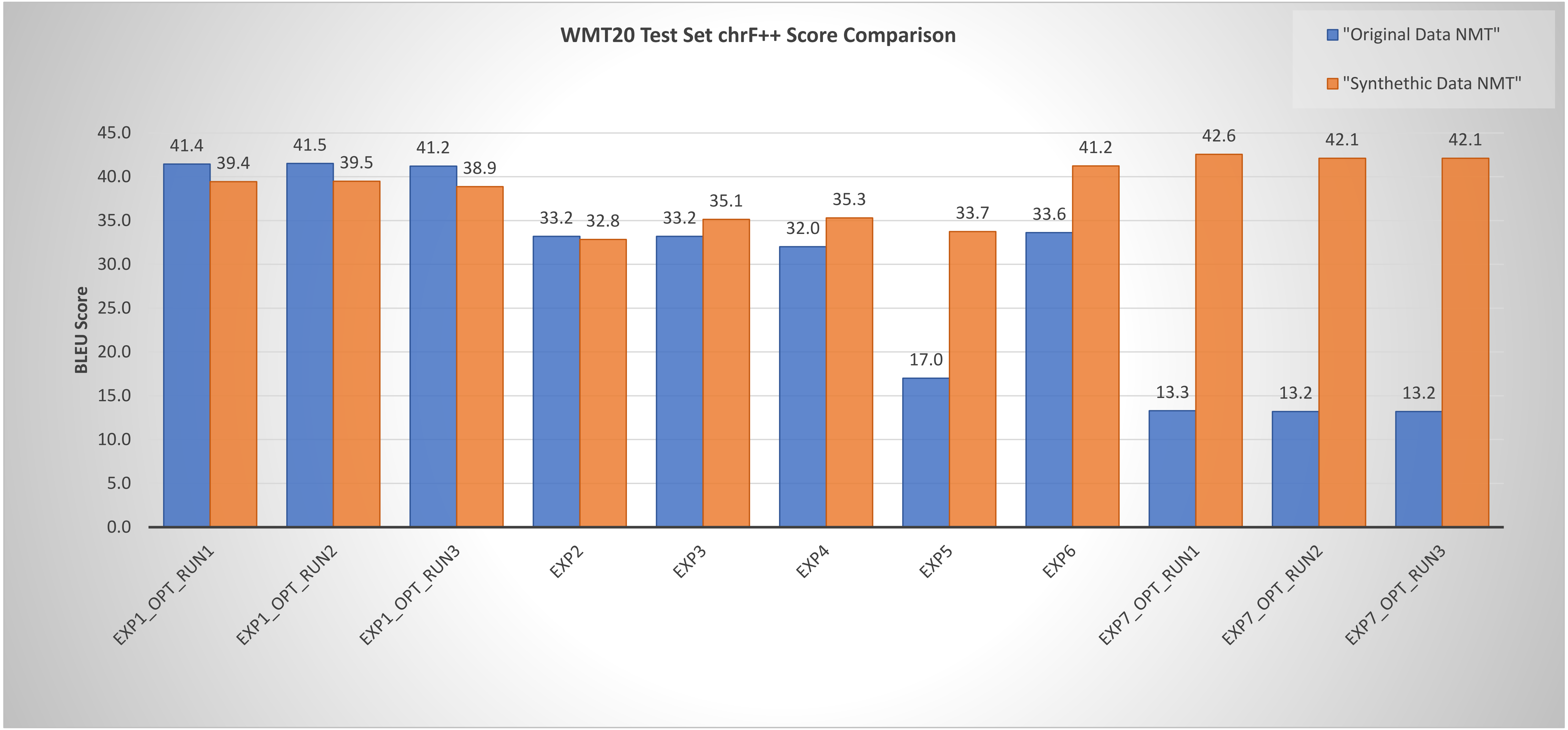
Experiment Results



Experiment Results

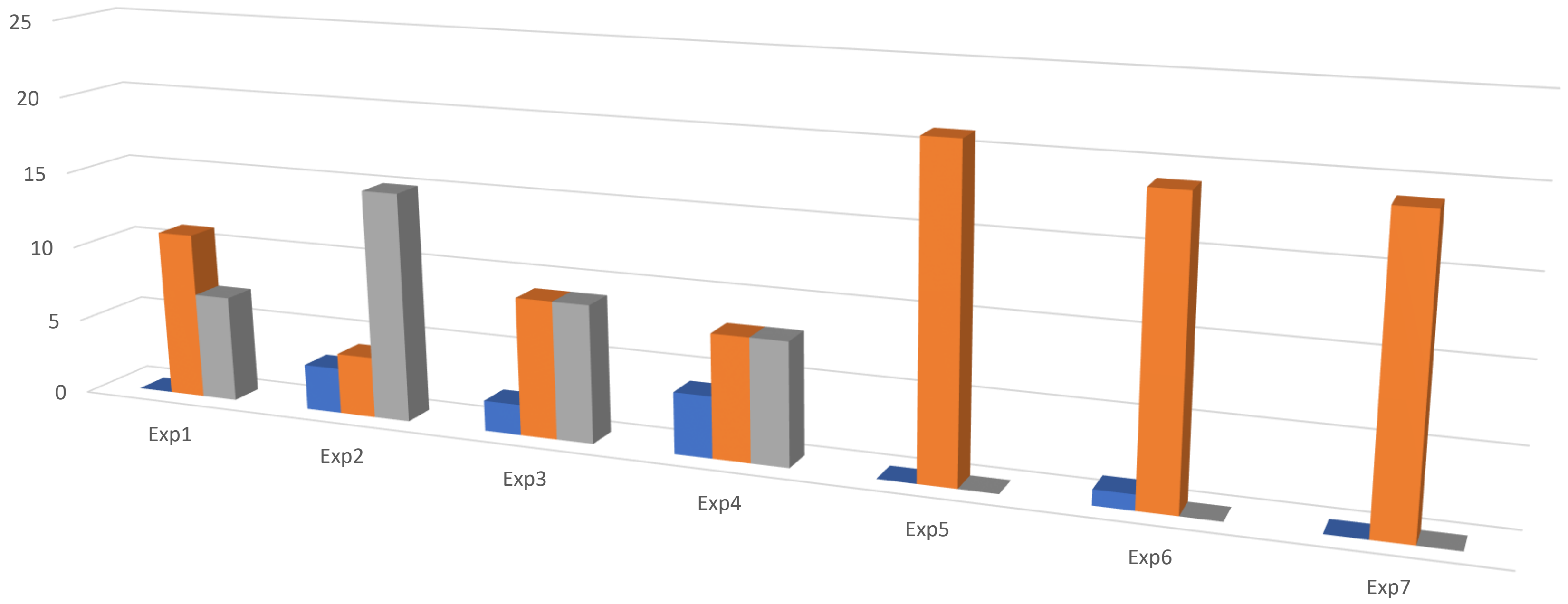


Experiment Results



Experiment Results

Comparison of Human Relative Ranking Scores Across All experiments



	Exp1	Exp2	Exp3	Exp4	Exp5	Exp6	Exp7
Original Model Score	0	3	2	4	0	1	0
Synthetic Model Score	11	4	9	8	21	19	19
Tied Score	7	15	9	8	0	0	0

Analysis

Synthetic NMT models outperformed Original NMT models

Trained Model Data	BLUE Score Mean	BLEU Score Standard Deviation	c6+w2-F2 Score Mean	c6+w2-F2 Score Standard Deviation
Original Data	14.05	10.25	41.18	8.07
Synthetic Data	16.74	11.09	44.12	8.57

Descriptive Statistics from analysis of all experiments 1 to experiment 4

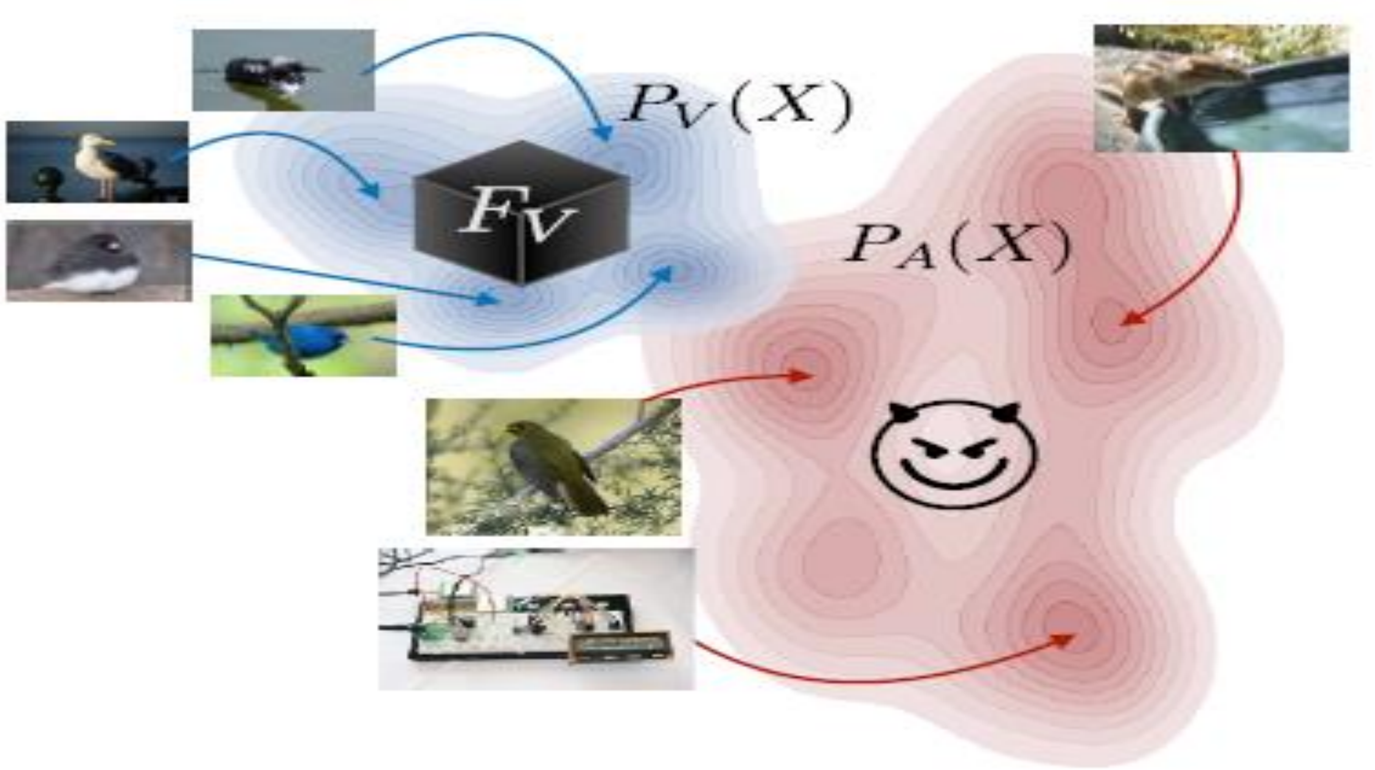
- Poorly paired sentences in original dataset
- Good translations provided by source model

Conclusion

- Active Learning Model Extraction are a genuine threat to the monetization of NMT models
- Active learning viable approach for data augmentation
- Results vary between evaluation metrics

Future Work

Sample Selection



Orekondy, T., et al. (2019). Knockoff nets: Stealing functionality of black-box models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Future Work

Higher Quality Data

QUALITY
TRANSLATIONS

source: http://www.rl-translations.com/index_english.html

https://commons.wikimedia.org/wiki/File:Thank-you-word-cloud.jpg

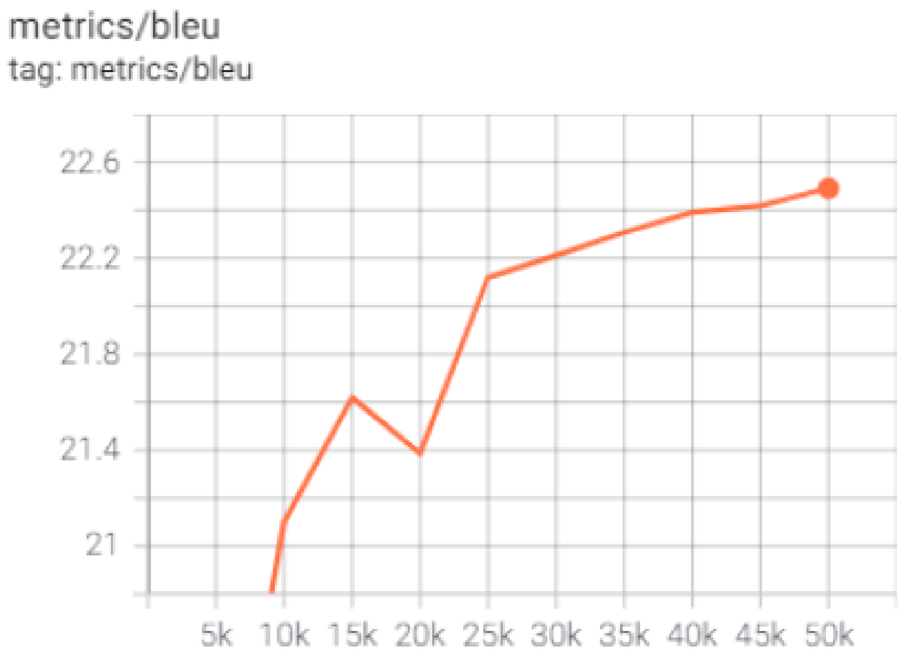
Samples from Experiment 1

Model A = NMT trained on Synthetic data.

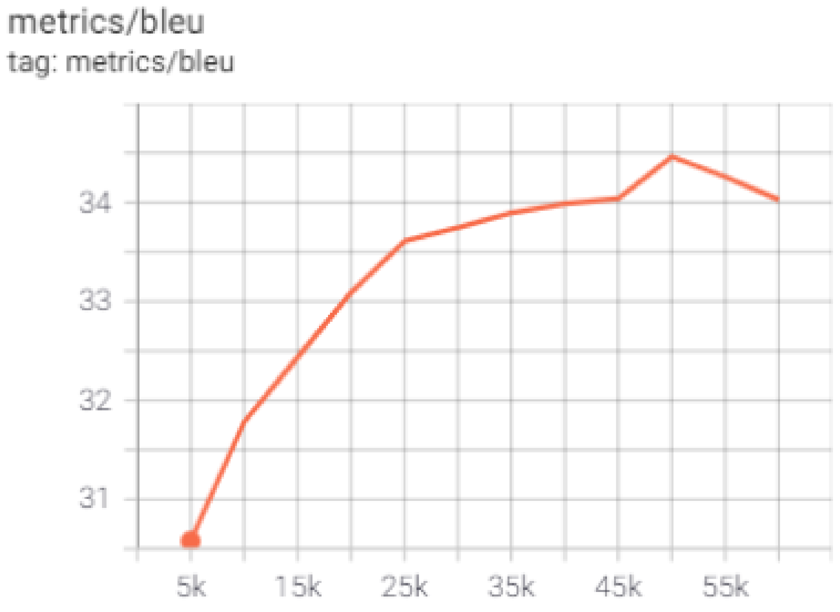
Model B = NMT trained on Original data.

Source_Sentence	Reference_Sentence	Model_A_Hypothesis	Model_B_Hypothesis	Better Translation (Model A or Model B)
but they are not sufficient .	aber sie reichen nicht aus .	aber sie reichen nicht aus.	sie sind aber nicht ausreichend.	same
but will it work ?	aber wird es funktionieren ?	aber wird es funktionieren?	aber wird es arbeiten?	a
a lot has changed since 2005 .	seit 2005 hat sich viel verändert .	Seit 2005 hat sich viel verändert.	Vieles hat sich seit 2005 verändert.	same
paris - who would have thought it ?	paris - wer hätte das gedacht ?	Paris - wer hätte das gedacht?	paris - wer hätte das gedacht?	same
one could now imagine much more clearly what might happen if a nuclear bomb exploded .	man konnte sich jetzt viel deutlicher vorstellen , was passieren könnte , wenn eine atombombe explodierte .	Man konnte sich jetzt viel deutlicher vorstellen, was passieren könnte, wenn eine Atombombe explodierte.	Man kann sich jetzt viel klarer vorstellen, was möglicherweise ein Atombomben explodierte.	a

Experiment 1 Training Charts

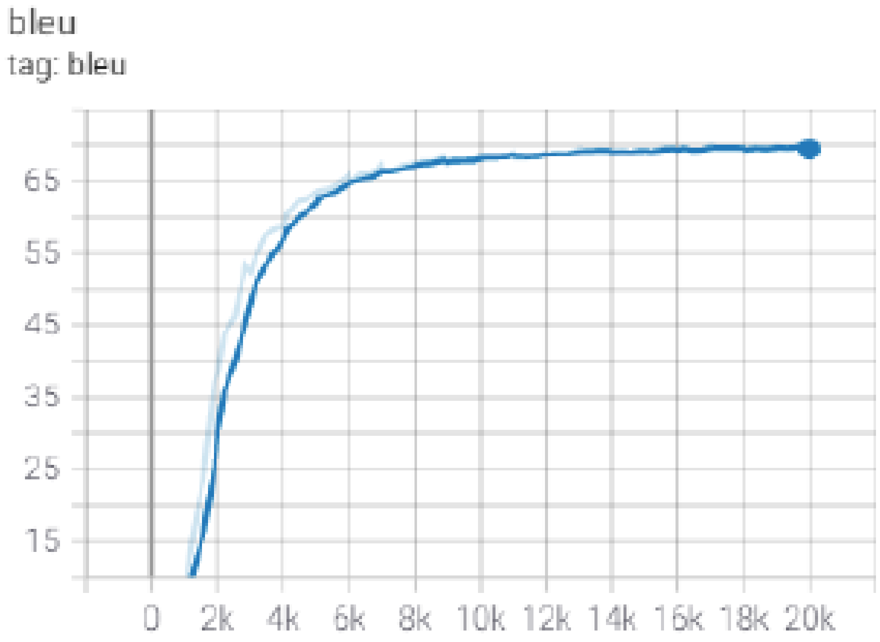


Experiment 1 Synthetic Dataset NMT
validation BLEU score during training optimizer run 1

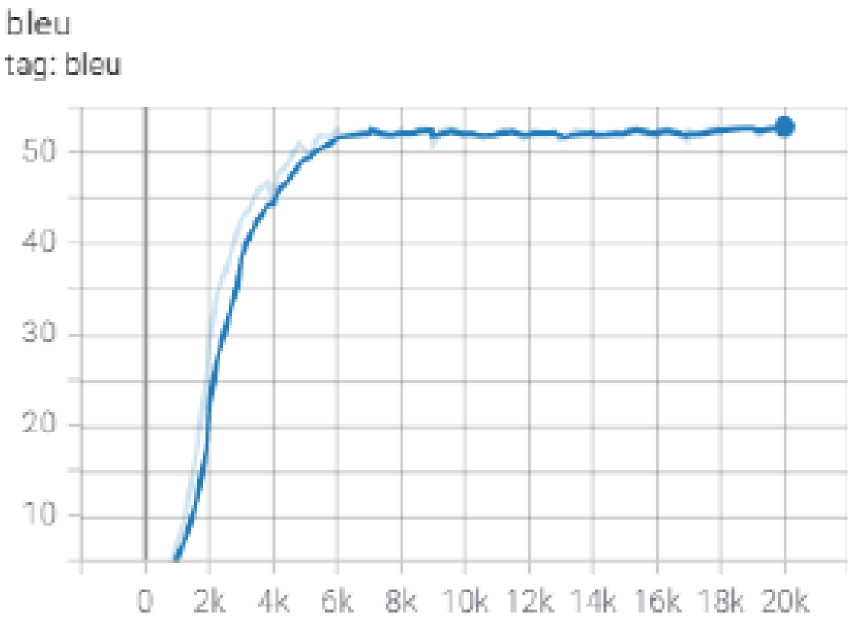


Experiment 1 Original Dataset NMT
validation BLEU score during training

Experiment 3 Training Charts

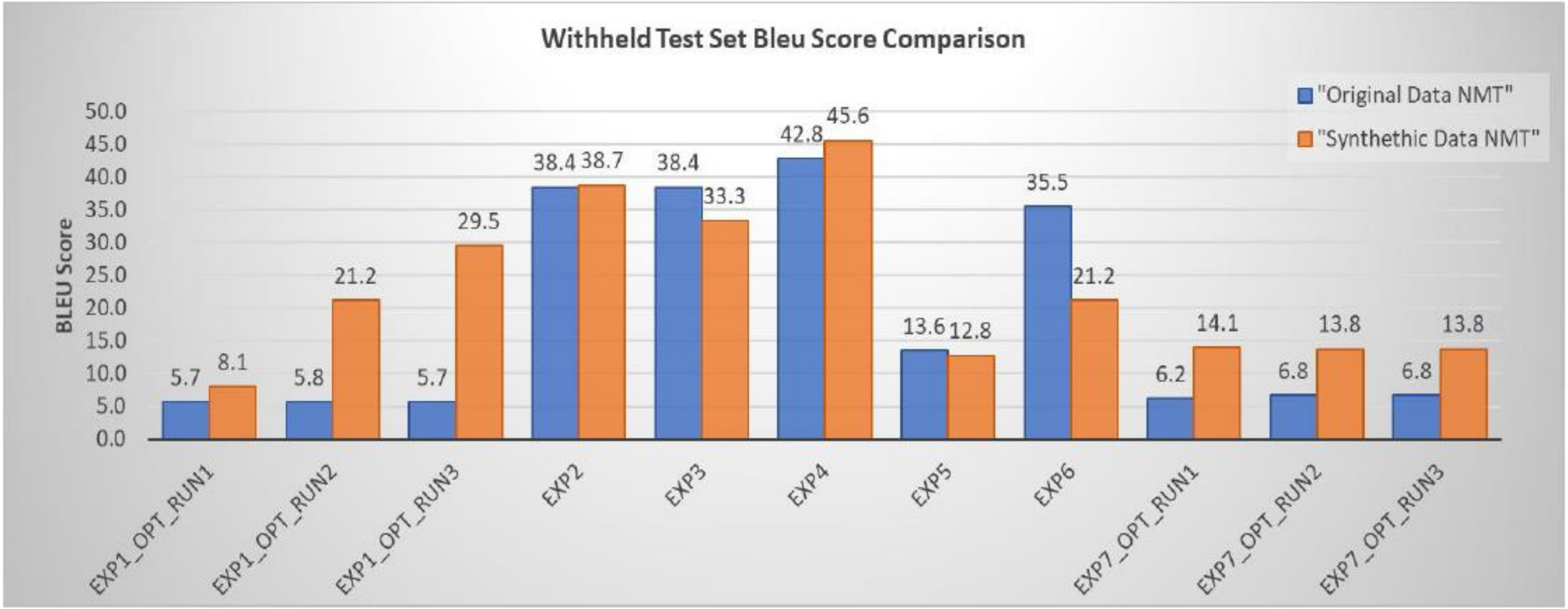


Experiment 3 Synthetic Dataset NMT validation BLEU score during training single run



Experiment 3 Original Dataset NMT validation BLEU score during training single run

Withheld Data Training Results



Experiment 1 Statistical Significance Test

Withheld Corpus Test Dataset Descriptive Statistics and P-values

n=3	BLEU (s_sel/s_opt/p)	METEOR (s_sel/s_opt/p)	TER (s_sel/s_opt/p)	Length (s_sel/s_opt/p)
baseline	12.1 (0.1/0.2/-)	19.0 (0.0/0.1/-)	72.2 (0.1/0.1/-)	96.9 (0.1/0.5/-)
system 1	17.2 (0.1/0.1/0.0001)	22.0 (0.1/0.0/0.0001)	66.4 (0.1/0.0/0.0001)	96.4 (0.1/0.3/0.0001)

WMT17 Test Dataset Descriptive Statistics and P-values

n=3	BLEU (s_sel/s_opt/p)	METEOR (s_sel/s_opt/p)	TER (s_sel/s_opt/p)	Length (s_sel/s_opt/p)
baseline	16.5 (0.3/0.0/-)	22.0 (0.1/0.1/-)	66.0 (0.3/0.1/-)	97.6 (0.3/0.2/-)
system 1	21.5 (0.3/0.2/0.0001)	25.0 (0.2/0.1/0.0001)	60.3 (0.4/0.1/0.0001)	98.9 (0.3/0.4/0.0001)

WMT18 Test Dataset Descriptive Statistics and P-values

n=3	BLEU (s_sel/s_opt/p)	METEOR (s_sel/s_opt/p)	TER (s_sel/s_opt/p)	Length (s_sel/s_opt/p)
baseline	21.8 (0.3/0.1/-)	25.3 (0.1/0.1/-)	58.4 (0.3/0.3/-)	96.5 (0.3/0.1/-)
system 1	29.7 (0.3/0.2/0.0001)	29.4 (0.2/0.1/0.0001)	50.3 (0.3/0.1/0.0001)	97.4 (0.3/0.3/0.0001)

WMT19 Test Dataset Descriptive Statistics and P-values

n=3	BLEU (s_sel/s_opt/p)	METEOR (s_sel/s_opt/p)	TER (s_sel/s_opt/p)	Length (s_sel/s_opt/p)
baseline	18.9 (0.3/0.1/-)	23.4 (0.2/0.2/-)	61.0 (0.4/0.3/-)	91.7 (0.4/0.3/-)
system 1	26.8 (0.4/0.1/0.0001)	27.6 (0.2/0.1/0.0001)	53.2 (0.4/0.1/0.0001)	94.0 (0.4/0.5/0.0001)

WMT20 Test Dataset Descriptive Statistics and P-values

n=3	BLEU (s_sel/s_opt/p)	METEOR (s_sel/s_opt/p)	TER (s_sel/s_opt/p)	Length (s_sel/s_opt/p)
baseline	14.5 (0.3/0.1/-)	20.9 (0.2/0.0/-)	65.7 (0.4/0.1/-)	86.1 (0.5/0.5/-)
system 1	15.8 (0.4/0.3/0.0001)	20.4 (0.3/0.2/0.0001)	66.7 (0.5/0.3/0.0001)	72.4 (1.0/0.8/0.0001)

References

Tramèr, Florian, et al. "Stealing machine learning models via prediction apis." *25th {USENIX} Security Symposium ({USENIX} Security 16)*. 2016.

Papernot, Nicolas, et al. "Practical black-box attacks against machine learning." *Proceedings of the 2017 ACM on Asia conference on computer and communications security*. 2017.

Krishna, Kalpesh, et al. "Thieves on sesame street! model extraction of bert-based apis." *arXiv preprint arXiv:1910.12366* (2019).

Jagielski, Matthew, et al. "High accuracy and high fidelity extraction of neural networks." *29th {USENIX} Security Symposium ({USENIX} Security 20)*. 2020.

Carlini, Nicholas, Matthew Jagielski, and Ilya Mironov. "Cryptanalytic extraction of neural network models." *Annual International Cryptology Conference*. Springer, Cham, 2020.

Hu, X., L. Liang, S. Li, L. Deng, P. Zuo, Y. Ji, X. Xie, Y. Ding, C. Liu and T. Sherwood (2020). Deepsniffer: A dnn model extraction framework based on learning architectural hints. Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems.

References

Orekondy, T., Schiele, B., & Fritz, M. (2019). *Knockoff nets: Stealing functionality of black-box models.* Paper presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Callison-Burch, C., Fordyce, C. S., Koehn, P., Monz, C., & Schroeder, J. (2007). *(Meta-) evaluation of machine translation.* Paper presented at the Proceedings of the Second Workshop on Statistical Machine Translation

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). *Bleu: a method for automatic evaluation of machine translation.* Paper presented at the Proceedings of the 40th annual meeting of the Association for Computational Linguistics.

Popović, M. (2017). *chrF++: words helping character n-grams.* Paper presented at the Proceedings of the second conference on machine translation.

Ng, N., Yee, K., Baevski, A., Ott, M., Auli, M., & Edunov, S. (2019). Facebook FAIR's WMT19 News Translation Task Submission. *arXiv preprint arXiv:1907.06616*.

Klein, G., Hernandez, F., Nguyen, V., & Senellart, J. (2020). *The OpenNMT neural machine translation toolkit: 2020 edition.* Paper presented at the Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (AMTA 2020).