Automatic Quality Estimation for Data Selection and Curriculum Learning

Hiep Nguyen and **Lynn Yip** and **Justin DeBenedetto**
Department of Computing Sciences
Villanova University

Abstract

The size of neural models within natural language processing has increased at a rapid pace in recent years. With this increase in model size comes an increase in the amount of training data required for training. While these larger models have shown strong performance, their use comes with added training and data costs, can be resource-prohibitive for many researchers, and uses an amount of language data that is not always available for all languages. This work focuses on exploring quality estimation as a method of data selection or filtering. The aim is to provide models with higher quality data as compared to larger amounts of data. This approach was applied to machine translation models with varying data sizes as well as to the BabyLM Challenge. Given the 100M word dataset provided in the BabyLM Challenge, we test out various strategies for selecting 10M words for pretraining and use a curriculum learning approach based on the quality estimation scoring. We find small improvements in certain data settings.

1 Introduction

In recent years, there has been a dramatic rise in the size of neural network models used for natural language processing tasks. To train these larger models, there has been a similar rise in the size of datasets used for training or pretraining. While these models have been quite successful, this trend comes with several downsides including the cost of creating these larger systems which also inhibits the ability of many researchers who lack access to the large scale computing resources required. By contrast, human language development occurs in children with exposure to far fewer words of training data. Inspired by this, the BabyLM Challenge (Choshen et al., 2024) focuses on "sample-efficient pretraining on a developmentally plausible corpus."

One approach to improve model performance in data-limited settings is curriculum learning (Elman, 1993). Just as human language learners are typically exposed to simpler language before building up to more complex utterances, curriculum learning involves increasing the difficulty of training examples over the course of model training. In order to do this, there must be some measure of "difficulty" in order to assign an ordering to training examples. In this work, we apply quality estimation scoring as an estimation of difficulty. These scores are used to train models for the BabyLM Challenge, specifically restricted to 10 million words or less of training data.

Quality estimation (QE) in machine translation scores the quality of translation output without the need for a reference translation (Specia et al., 2018). Through a series of experiments, we explore the effects of using QE to filter data for machine translation systems for both initial model training and fine-tuning, as well as the result of training on different quantities of data for each model (see Section 3). Prior work has shown that data filtering through QE can increase model performance (Batheja and Bhattacharyya, 2023). We explore that further in this work for both machine translation and language modeling in data restricted settings.

Since quality estimation scores the quality of the output of a machine translation system, it is likely that higher QE scores correspond to sentences which the system has an easier time translating. Motivated by this, we experiment with using QE scores as an estimation of the difficulty of a given sentence for an NLP system. In particular, we use this for data selection as well as for difficulty scoring for curriculum learning training of a "baby" language models as part of the 2024 BabyLM Challenge (see Section 4).

2 Related work

2.1 BabyLM Challenge and Curriculum Learning

As this is the second year of the BabyLM Challenge, there is a body of existing work which relates directly to our BabyLM experiments (Warstadt et al., 2023). There were many submissions (41.9% of teams) in last year's iteration which made use of curriculum learning. Using a curriculum to make training difficulty scale up during training is known as curriculum learning (Bengio et al., 2009). It has been shown that reordering input data during training can have a large effect on model performance across tasks such as natural language inference (NLI) (Schluter and Varab, 2018) and neural machine translation (NMT) (Liu et al., 2020). The most similar approaches from last year's BabyLM Challenge to this current work were by Chobey et al. and by Hong et al.. In those works, a teacher language model was trained first and used to determine the curriculum for training a new model. We similarly are using another model to inform the curriculum, though the model and curriculum forming is done differently.

2.2 Quality Estimation

QE has been used to assist with both automated post-editing (APE) (Chatterjee et al., 2018) and human post-editing tasks (Béchara et al., 2021). QE can be used in tandem with APE to determine which sentences from a machine translation system need to be corrected (Chatterjee et al., 2018). In contrast, we use QE in this work to filter out data to be used for fine-tuning the machine translation model.

QE has also been used to extract high-quality data from both parallel and pseudo-parallel data for training machine translation systems (Batheja and Bhattacharyya, 2022, 2023). We take this work one step further by fine-tuning machine translation systems on the model's own output which was also filtered using QE. The results from fine-tuning on both high and low-quality data were evaluated.

3 Quality Estimation for Machine Translation

3.1 Methodology

3.1.1 Dataset

We used the German-English IWSLT 2017 dataset (Cettolo et al., 2017) for all machine translation experiments described in this section. The original dataset was initially divided into eight sets of different sizes ranging from approximately 1500 sentences to the full-sized set of 198669 sentences as shown in Table 1. The full dataset was first halved to create the next smallest sized dataset. This smaller dataset was then also halved to create the next smallest size and so on for all eight. Sentences that were removed during this process did not reappear in smaller sets. This ensured that each smaller set of sentences consisted solely of sentences from the larger set.

The smallest dataset split of roughly 1500 sentences was dropped due to the BLEU scores being too low to be meaningful after initial model training. All experiments listed were completed with the remaining seven splits of data.

3.1.2 Model Training

The fairseq (Ott et al., 2019) sequence modeling toolkit was used to train machine translation models from German to English. A new model was trained on each dataset split. The results from initial model training resulted in BLEU scores ranging from 0.04 to 36.82, with the largest dataset split corresponding to the highest BLEU score.

3.1.3 Quality Estimation Filtering

TransQuest is a framework for machine translation quality estimation that can be used to rate translations at either the word or sentence level (Ranasinghe et al., 2020). The SiameseTransQuest sentence-level quality estimation model was used throughout these experiments[1].

The quality estimation threshold to separate high-quality and low-quality sentences was determined by selecting the threshold that gave the widest range of filtered sentence quantity across all seven split datasets.

3.1.4 Model Fine-Tuning

Using the sentences that were filtered out using TransQuest quality estimation, the original fairseq translation models for the specified dataset split were fine-tuned on the filtered sentences. The BLEU scores were recorded after fine-tuning to see if any improvements had been made as a result of the fine-tuning. To replicate any of our results, please see our GitHub repository[2].

3.2 Experiments

Experiments 1 through 3 start with fairseq translation models trained on the original IWSLT 2017 German-English dataset splits. See Table 5 for the full results. In experiments 4 through 6, the original dataset is first filtered with QE and only the high-quality data is used for initial model training. QE is then used to filter the model output for fine-tuning (see Table 6).

3.2.1 Experiment 1

Seven fairseq models were trained on the original IWSLT 2017 German-English dataset which had been split into varying sizes. Each model was then used to translate the test set, which introduced new data to the model. The output translations went through TransQuest quality estimation. The low-quality sentences as rated by TransQuest were used to fine-tune the models.

For the models initially trained on the smallest splits, excluding the eighth split, fine-tuning resulted in BLEU score improvements from 0.46 and 0.76. The most significant score improvement was

[1] https://huggingface.co/TransQuest/siamesetransquest-da-en_de-wiki

[2] https://github.com/lsyip/mt-qe-filtering

Dataset Split	Number of Sentences	BLEU
1 (Full set)	198669	36.82
2	99335	33.15
3	49668	29.14
4	24834	23.00
5	12417	16.63
6	6209	11.75
7	3105	9.40
8	1553	0.04†

Table 1: Initial Model BLEU Scores for Experiments 1-3. Model trained on unfiltered data, fine-tuned on high or low quality data.
†Not used in experiments

seen in split 7, where the initial model had been trained on the smallest amount of data.

3.2.2 Experiment 2

The seven base translation models trained on the dataset splits remain the starting point for this experiment. This time, the models were used to translate the training sentences that they had been trained on, thus re-introducing the same data the model was trained on. The translation output went through TransQuest quality estimation and the low-quality sentences were used to fine-tune the models.

For this experiment, splits 5 and 6 had BLEU score improvements of over 0.6 points. The remaining splits did not show significant improvement after fine-tuning.

3.2.3 Experiment 3

Starting again with the seven base translation models, the models were again used to translate the training set. This translation output then went through TransQuest quality estimation and the high quality sentences were used for fine-tuning.

The highest BLEU score improvement for this pipeline was on split 5, which showed an increase of 0.41 points after fine-tuning on the high-quality sentences. The remaining splits did not show significant improvement in BLEU scores after fine-tuning.

3.2.4 Experiment 4

For this experiment, we first filtered each of the IWSLT 2017 splits through TransQuest quality estimation. See Table 2 for details. Next, new fairseq translation models were trained on the sentences that were rated to be of high quality. These models serve as the starting point for the following two pipelines. This setup mirrors the parallel corpus filtering via quality estimation previously done by Batheja and Bhattacharyya.

Dataset Split	Number of Sentences	BLEU
1	75898	30.56
2	36269	25.00
3	17776	19.13
4	8882	15.20
5	4286	6.90
6	2149	0.05
7	1082	0.07

Table 2: Initial Model BLEU Scores for Experiments 4-6. Model trained on filtered data, fine-tuned on high or low quality data.

3.2.5 Experiment 5

Using the new fairseq models that were trained in experiment 4, the model was asked to translate all sentences from their respective training set, which did not introduce new data to the model. The new translations were sent through TransQuest quality estimation and the sentences that were rated high-quality were used for fine-tuning.

After fine-tuning, the models trained on the smallest three splits did not show significant improvement to their BLEU scores. However, some improvement was made with the larger splits. The models initially trained on splits 1, 2, 3 and 4 showed BLEU score improvements of 0.38, 0.6, 0.58, and 0.46, respectively. It is important to note that with each larger split, the number of sentences in the fine-tuning set also increases as more translations were sent through quality estimation.

3.2.6 Experiment 6

Starting again with the fairseq models that were previously trained in experiment 4, these models were again used to translate sentences from their respective original training sets and the new translations were sent through TransQuest for quality estimation. The sentences that were rated to be of low-quality were then used to fine-tune the models.

After fine-tuning, smallest 3 splits did not show any improvements in BLEU score. Splits 1, 2, and 4 showed an increase in BLEU score between 0.21 and 0.28. Split 3 had the highest BLEU score increase of 0.51.

3.3 Machine Translation Results

In experiments 1-3, we observed that some models trained on smaller datasets saw improvements in BLEU score after fine-tuning on training data that had been filtered through quality estimation. The differences between using low-quality and high-quality data to fine-tune, however, were marginal. This suggests that the quality of the data may not matter as much as the quantity that is available. For the smaller datasets, improvements could be seen after fine-tuning in both the low and high-quality instances.

For experiments 4-6, which used TransQuest quality estimation to filter both the original dataset and the data for fine-tuning, the initial model BLEU scores were lower than the first three experiments due to having fewer training sentences. We observed that some improvements in BLEU score can be made after fine-tuning on the filtered high-quality on the larger dataset splits. The most significant differences after fine-tuning were seen in splits 2, 3, and 4.

4 BabyLM Challenge

4.1 Methodology

4.1.1 Dataset

The data for the BabyLM Challenge provided by the challenge organizers consists of text from six sources and was selected to represent language data that a human child may be exposed to when developing their language skills. The provided dataset contains 100 million words of text data. From this, we could form training datasets containing up to 10 million words to train models for the strict-small track.

4.1.2 Model and Training

The data preprocessing involved removing blank lines and special characters, with a focus on dialogue-related elements. The sentences within each dataset were then rearranged based on length to streamline the training process. After preprocessing, the sentences were translated from English to German using base translation models from fairseq. See Table 3 for metadata of processed datasets.

The quality of resulting pairs of German-English sentences was assessed using TransQuest and xCOMET frameworks. COMET, which stands for Crosslingual Optimized Metric for Evaluation of Translation, is a neural framework to predict human judgments on machine translation quality from source and target language samples (Rei et al., 2020). Specifically, the wmt23-cometkiwi-da-xl[3] model was chosen for xComet, and its results were compared to those from TransQuest. However, we found the scores from both models to be inconsistent with each other. In the end, the xComet scores were selected to rank and filter the data for training, since the TransQuest scores were heavily influenced by sentence length.

Our model is a RoBERTa (Liu et al., 2019) model. RoBERTa is a modification of the BERT (Devlin et al., 2018) model, which showed improved performance across several benchmarks.

We conducted several experiments to explore different strategies for selecting a training subset with a budget of 10 million words from the original 100 million words. The experiments varied based primarily on:

- The order of training: ascending or descending (original curriculum learning) order of quality estimation scores (equivalently, reversed machine comprehension level),
- The separation or combination of datasets from various sources,
- The number of hidden layers and heads during model training.

After filtering and training, models were evaluated using a standardized evaluation pipeline provided by the organizers to compute their scores on the BLiMP and EWoK benchmarks.

Code to train our models can be found on GitHub[4].

4.2 Experiments

4.2.1 Experiment 1

For this experiment, all sources were combined and rearranged based on their xComet scores. The 10 million words in sentences with highest scores were kept and divided into 3 files, namely easy (top 2 million), medium (next 4 million), and hard (next 4 million). Sentences with higher QE scores are considered to be easier sentences. Those files are trained in order from easy to hard, following the curriculum learning approach.

4.2.2 Experiments 2 and 3

For Experiment 2, 1.8 million words from the highest scored sentences of each sources were selected, except for Switchboard from which all 0.8 million words were taken. For Experiment 3, we did the opposite, by selecting the 1.8 million words from the lowest scored sentences from each source. This means we followed the typical order for curriculum learning in Experiment 2 and the reversed order in Experiment 3.

In both experiments, the following file order was used for training: CHILDES, OpenSubtitles, Switchboard, BNC_spoken, Simple_wiki, and Gutenberg.

4.2.3 Experiments 4-6

For Experiment 4, we tried to replicate Experiment 1 in the way word data were selected, starting by combining all sources into one stream for score ranking. However, we divided the word budget into 5 files grouped by QE score with each file containing around 2 million words. The model is trained on these files from easiest to hardest in Experiment 4 and reversed order in Experiment 6.

For Experiment 5, instead of selecting 10 million words from highest scored sentences, we filtered

[3] https://huggingface.co/Unbabel/wmt23-cometkiwi-da-xl

[4] https://github.com/jdebened/BabyLM2024

Dataset	**Description**	**# Words (Original)**	**# Words (Processed)**
CHILDES	Child-directed speech	28.9M	15.6M
British National Corpus (BNC)	Dialogue	7.7M	5.3M
Standardized Project Gutenberg Corpus	Written English	26.3M	21.7M
OpenSubtitles	Movie subtitles	20.0M	13.5M
Simple Wikipedia	Wikipedia (Simple English)	14.7M	11.5M
Switchboard Dialog Act Corpus	Dialogue	1.3M	0.9M
Total		**99M**	**68.5M**

Table 3: Original and Processed Dataset provided for the strict track of the BabyLM Challenge. Dataset names, domain descriptions, and word counts

those from lowest scored ones and trained resulting files in the order of hardest to easiest files.

4.2.4 Experiments 7-10

For these experiments, only one source was used in each experiment, namely either CHILDES or Gutenberg. In Experiments 7 and 8, the 10 million words were selected from lowest scores of CHILDES and Gutenberg datasets respectively. The order of training is from sentences with lowest scores to those with higher scores, opposite of expected curriculum learning order.

In Experiment 9, we filtered down to the 10 million words from highest scoring sentences of CHILDES to compare with the result from Experiment 7. It is noted that this comparison is based on data selection of highest and lowest scored sentences as well as training order of increasing and decreasing complexity.

In Experiment 10, we used the same subset of 10 million words from Experiment 10. However, the number of hidden layers and heads were doubled for further comparison.

The motivation behind choosing these sources rather than others is because we wanted to test the opposition between child-directed speech and written texts.

4.2.5 Experiments 11 and 12

For these experiments, we tried to replicate Experiments 4 and 5 respectively. However, we decided to split into smaller files, each with 1 million words.

4.2.6 Experiments 13 and 14

For these experiments, the mixture of 5 million words from highest scored sentences and 5 million words from lowest scored ones were used.

The primary difference between these experiments are based on their order of training. While Experiment 13's order followed curriculum training, Experiment 14 did the opposite.

Full experiment descriptions, mainly in how data was selected for model training, can be found at Table 4.

4.3 Results

Full experiment results including BLiMP and EWoK scores from the evaluation pipeline can be found at Table 4.

Evaluation pipeline provided by BabyLM Challenge 2024 included zero shot evaluation on tasks from the BLiMP benchmark and hidden evaluation tasks from the Ewok benchmark (Warstadt et al., 2020; Ivanova et al., 2024).

BLiMP is made up of tasks designed to test how well language models adhere to the structure of English. Each task presents a pair of sentences, where one is grammatically correct, and the other is incorrect, with the two sentences differing as little as possible. A model is considered accurate for a given example if it assigns a higher probability to the correct sentence in the pair (Warstadt et al., 2023).

Elements of World Knowledge (EWoK) framework evaluates world modeling in language models by testing their ability to use knowledge of concepts across physical and social domains to determine plausible or implausible contexts. It flexibly constructs multi-step scenarios, targets specific cognitive concepts, and generates controlled evaluation items using a template-based approach. This framework focuses on how well language models can productively apply concept knowledge, rather than just matching individual sentences or facts (Ivanova et al., 2024).

From the table of results, we found several patterns in the varied BLiMP and EWoK scores:

- Models with order of training from harder to easier, opposite to expected order from curriculum learning (decreasing complexity) had slightly higher BLiMP_complement scores compared to others with/without same datasets such as models 3, 5, 10, 12, 14. The exception in this case is Model 6, compared to Model 4. However, the BLiMP_filtered and EWoK_filtered scores did not experience the similar pattern with no noticeable improvement for any order. This inconsistency may stem from our assumption of the relationship

#	Data setup	BLiMP complement	BLiMP filtered	EWoK filtered	Details
1	All data sources Sources combined CL training order	58.01	60.69	49.99	10M highest QE scores separated into 2M highest, next 4M, next 4M
2	All data sources Sources kept separate CL training order	54.98	60.87	49.15	1.8M words of each source by highest QE score (Switchboard max 0.8M)
3	All data sources Sources kept separate Reversed CL order	60.25	60.17	50.47	1.8M words of each source by lowest QE score (Switchboard max 0.8M)
4	All data sources Sources combined CL training order	58.92	61.01	50.00	10M highest QE scores separated into 5 equal-sized files
5	All data sources Sources combined Reversed CL order	61.41	60.45	50.10	10M lowest QE scores separated into 5 equal-sized files
6	All data sources Sources combined Reversed CL order	56.83	60.31	49.71	10M highest QE scores separated into 5 equal-sized files, train in the reverse order (compared to experiment 4)
7	CHILDES data only Reversed CL order	59.34	57.80	50.21	10M lowest QE scores separated into 5 equal-sized files
8	Gutenberg data only Reversed CL order	58.27	61.94	50.46	10M lowest QE scores separated into 5 equal-sized files
9	CHILDES data only CL training order	55.41	57.99	50.30	10M highest QE scores separated into 5 equal-sized files
10	Gutenberg data only Reversed CL order	59.73	**62.39**	49.66	10M lowest QE scores separated into 5 equal-sized files, double the number of hidden layers and heads (compared to experiment 8)
11	All data sources Sources combined CL training order	56.92	61.29	**50.97**	10M highest QE scores separated into 10 equal-sized files
12	All data sources Sources combined Reversed CL order	59.80	60.11	50.55	10M lowest QE scores separated into 10 equal-sized files
13	All data sources Sources combined CL training order	59.74	60.35	50.31	5M highest QE scores separated into 5 equal-sized files, then 5M lowest QE scores separated into 5 equal-sized files
14	All data sources Sources combined Reversed CL order	**63.02**	60.66	50.18	5M highest QE scores separated into 5 equal-sized files, then 5M lowest QE scores separated into 5 equal-sized files; train in reverse order (compared to experiment 13)

Table 4: Experiments setups and results (%). Comparison between models trained on 10 million word budget filtered from original 100 million words provided in the 2024 BabyLM Challenge. **Bolded** values show best in column. Strategies to filter the data to form training datasets containing up to 10 million words to train models.

between quality estimation and machine comprehension level.

- Models with combined sources did not show superior results in BLiMP_complement scores compared to separated ones. Models with multiple sources also did not outperform single-source models. However, regarding the BLiMP_filtered scores, single-source model using Gutenberg showed better performance compared to multiple-source models or other single-source models, especially derived from CHILDES data. Additionally, this may also relate to the fact that Gutenberg's written style and higher quality can improve the performance.

- Models' performance and the number of files to train were not proportional in terms of BLiMP_complement scores, but showed a clear positive correlation in EWoK_filtered. Models using 5 training files get the highest BLiMP_complement in comparison to 3 or 10 files. Meanwhile, in the case of implementing the curriculum learning order (models 1, 4, 11), the BLiMP_filtered accuracy positively correlated with the number of files.

- Models using doubled number of heads and hidden layers took more time to train and had better BLiMP_complement and BLiMP_filtered scores (models 8 and 10), but not EWoK_filtered scores.

5 Conclusion

This work explored quality estimation for data filtering and curriculum learning on both machine translation systems and language models. As shown in our machine translation experiments (see Section 3.2), modest improvements can be obtained through finetuning on filtered data. This benefit largely went away as the data size scaled up to the full IWSLT17 dataset, suggesting that this method has more use for certain data limited settings rather than for general model use. Furthermore, base model performance went up more noticeably when additional data was added, showing that more data made a larger difference than higher quality data in this setting.

For the BabyLM Challenge strict-small track, teams could form datasets consisting of up to 10 million words to train their language models. We explored several options for data selection from the provided 100 million word dataset. Each model was then trained using a curriculum learning approach based on quality estimation scoring. Overall, data source made a bigger difference to model performance than curriculum choice. In particular, models trained using the Project Gutenberg data generally had higher scores on downstream tasks. This suggests that while the other data sources are useful for human children learning language, the higher quality data available in the Gutenberg dataset produced a better language model.

Acknowledgments

This work used the Augie High-Performance Computing cluster, funded by award NSF 2018933, at Villanova University.

References

Akshay Batheja and Pushpak Bhattacharyya. 2022. Improving machine translation with phrase pair injection and corpus filtering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5395–5400.

Akshay Batheja and Pushpak Bhattacharyya. 2023. A little is enough: Few-shot quality estimation based corpus filtering improves machine translation. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Hannah Béchara, Constantin Orăsan, Carla Parra Escartín, Marcos Zampieri, and William Lowe. 2021. The role of machine translation quality estimation in the post-editing workflow. *Informatics*, 8(3).

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuhito Sudoh, Koichiro Yoshino, and Christian Federmann. 2017. Overview of the IWSLT 2017 evaluation campaign. In *Proceedings of the 14th International Conference on Spoken Language Translation*, pages 2–14, Tokyo, Japan. International Workshop on Spoken Language Translation.

Rajen Chatterjee, Matteo Negri, Marco Turchi, Frédéric Blain, and Lucia Specia. 2018. Combining quality estimation and automatic post-editing to enhance machine translation output. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 26–38, Boston, MA. Association for Machine Translation in the Americas.

Aryaman Chobey, Oliver Smith, Anzi Wang, and Grusha Prasad. 2023. Can training neural language models on a curriculum with developmentally plausible data improve alignment with human reading behavior? In *Proceedings of the BabyLM Challenge at the 27th Conference* *on Computational Natural Language Learning*, pages 98–111.

Leshem Choshen, Ryan Cotterell, Michael Y Hu, Tal Linzen, Aaron Mueller, Candace Ross, Alex Warstadt, Ethan Wilcox, Adina Williams, and Chengxu Zhuang. 2024. [call for papers] the 2nd babylm challenge: Sample-efficient pretraining on a developmentally plausible corpus. *arXiv preprint arXiv:2404.06214*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jeffrey L Elman. 1993. Learning and development in neural networks: The importance of starting small. *Cognition*, 48(1):71–99.

Xudong Hong, Sharid Loáiciga, and Asad Sayeed. 2023. A surprisal oracle for active curriculum language modeling. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 259–268.

Anna A Ivanova, Aalok Sathe, Benjamin Lipkin, Unnathi Kumar, Setayesh Radkani, Thomas H Clark, Carina Kauf, Jennifer Hu, RT Pramod, Gabriel Grand, et al. 2024. Elements of world knowledge (ewok): A cognition-inspired framework for evaluating basic world knowledge in language models. *arXiv preprint arXiv:2405.09605*.

Xuebo Liu, Houtim Lai, Derek F. Wong, and Lidia S. Chao. 2020. Norm-based curriculum learning for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 427–436, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2020. Transquest: Translation quality estimation with cross-lingual transformers. In *Proceedings of the 28th International Conference on Computational Linguistics*.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702.

Natalie Schluter and Daniel Varab. 2018. When data permutations are pathological: the case of neural natural language inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4935–4939, Brussels, Belgium. Association for Computational Linguistics.

Lucia Specia, Frédéric Blain, Varvara Logacheva, Ramón F. Astudillo, and André F. T. Martins. 2018. Findings of the WMT 2018 shared task on quality estimation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 689–709, Belgium, Brussels. Association for Computational Linguistics.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34, Singapore. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

A Appendix

			Experiment 1		Experiment 2		Experiment 3	
Split	Train Sents	Initial BLEU	FT† Sents	BLEU	FT Sents	BLEU	FT Sents	BLEU
1	198669	36.82	1884	36.81 (-0.01)	128246	36.94 (+0.12)	70423	36.84 (+0.02)
2	99335	33.15	1862	33.17 (+0.02)	66261	33.30 (+0.15)	33074	33.29 (+0.14)
3	49668	29.14	1962	29.43 (+0.29)	33540	29.45 (+0.31)	16128	29.49 (+0.35)
4	24834	23.00	2110	23.15 (+0.15)	16934	23.28 (+0.28)	7900	23.1 (+0.10)
5	12417	16.63	2232	**17.09 (+0.46)**	8670	**17.24 (+0.61)**	3747	**17.04 (+0.41)**
6	6209	11.75	2682	**12.21 (+0.46)**	4419	**12.41 (+0.66)**	1790	11.86 (+0.11)
7	3105	9.40	2928	**10.14 (+0.74)**	2185	9.42 (+0.02)	920	9.41 (+0.01)

Table 5: Experiment 1-3 Results. Model trained on unfiltered IWSLT17 dataset, fine-tuned on high or low quality data.
†Fine-tune

	Experiment 4		Experiment 5		Experiment 6	
Split	Train Sents	Initial BLEU	FT† Sents	BLEU	FT† Sents	BLEU
1	75898	30.56	60187	30.94 (+0.38)	15711	30.77 (+0.21)
2	36269	25.00	29657	25.6 (+0.60)	6612	25.28 (+0.28)
3	17776	19.13	14433	**19.71 (+0.58)**	3343	**19.64 (+0.51)**
4	8882	15.20	7728	**15.66 (+0.46)**	1154	15.44 (+0.24)
5	4286	6.90	1402	6.87 (-0.03)	2884	6.72 (-0.18)
6	2149	0.05	178	0.06 (+0.01)	1971	0.05
7	1082	0.07	83	0.07	999	0.07

Table 6: Experiment 4-6 Results. Model trained on filtered IWSLT17 dataset, fine-tuned on high or low quality data.
†Fine-tune