Improving Domain Adaptation of Transformer Models For Generating Reddit Comments

Fan Pu Zeng
Department of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
fzeng@andrew.cmu.edu

Joseph Salmento
Department of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
jsalment@andrew.cmu.edu

Abstract

We improve upon the recent success of large language models based on the transformer architecture by investigating and showing several methods that have empirically improved its performance in domain adaptation. We use a pre-trained GPT-2 model and perform fine-tuning on 5 different subreddits, and use different methods of ordering the training data based on our priors about the input to see how this affects the prediction quality of the trained model. We propose a new metric for evaluating causal language modeling tasks called APES (Average Perplexity Evaluation for Sentences) to address the limitations of existing metrics, and apply them to our results. Our results are evaluated against both LSTM and GPT-2 baselines.

1 Introduction

There have been many exciting breakthroughts in language generation models in recent years. From the simple n-gram model that has been studied since the early 20th century, to the introduction of neural language models that utilizes word embeddings [4] at the turn of the century (2001), in the past few years we saw the development of powerful language models such as Word2Vec (2013), Transformer (2017), BERT (2018), GPT (2018), GPT-3 (2020). Such models have already beaten humans in accuracy in tasks such as reading comprehension [6], and have displayed high levels of fluency in language-generation tasks. These successes can be attributed to the great strides taken in Deep Learning, the increase in computational resources, and the proliferation of publicly available datasets and benchmarks [11].

We aim to replicate and build upon existing work in causal language models by investigating and answering the following research questions in this paper:

1. Can we improve the performance of domain adaptation of transformer language models by various methods of ordering the inputs seen at training time, based on our priors about the input?
2. Can we successfully perform domain adaptation using just commodity hardware on unstructured internet discourse that can contain new vocabulary, carry unique speech patterns, and require expert domain knowledge in order to provide a cogent response to a prompt?
3. Can we successfully perform domain adaptation on large language models with billions of parameters even with very small datasets?
4. What is a suitable evaluation metric to determine the success of domain adaptation for text generation, given the limitations of existing metrics for evaluating causal language models?

10617 Class Project (2022).

Our paper shows mild results for point 1, and answers points 2 and 3 affirmatively empirically. We propose a new metric to address point 4 which we call APES (Average Perplexity Evaluation for Sentences), and use it to evaluate our results.

We achieve our results by fine-tuning a pre-trained GPT-2 language model using data from 5 subreddits with distinct topics and talking styles, and use various methods of ordering the input data to improve the performance of fine-tuning as measured by the Bilingual Evaluation Understudy (BLEU) score and a qualitative evaluation of the results given a prompt from different subreddits. Finally, we use the APES metric to manually score 200 results from the test set from each dataset from 1 to 5. Our results are evaluated against a LSTM baseline and a GPT-2 model without fine-tuning.

2 Background

2.1 Recurrent Neural Networks

Recurrent neural networks (RNNs) are a family of deep learning methods that aim to process sequential data. They are called recurrent because of the hidden state that is passed through the network, which allows it to store a low-dimensional representation of the history of inputs [9].

RNNs are especially suited for language tasks, as words depends on the context in which they appear in. However, there are several major limitations of RNNs: it suffers from the vanishing gradient and exploding gradient problem due to the presence of self-loops, it is difficult to parallelize training due to its inherent sequential nature as evident in Equation **??**, and it is difficult to capture long-term dependencies in the input data.

2.2 LSTMs

The Long Short-Term Memory (LSTM) sequence model aims to address some of the limitations of basic RNNs. LSTM is based on the gated recurrent unit (GRU), which addresses the problem of vanishing or exploding gradients in RNNs. Their design also allows them to better capture long-term dependencies [5].

There are three main types of gates in an LSTM cell: the input gate, the forget gate, and the output gate. The input gate controls which information from the current input is added to the cell state. The forget gate controls which information from the previous cell state is retained in the current cell state. Finally, he output gate controls which information from the current cell state is output as the LSTM's prediction.

2.3 Transformers

The Transformer architecture is based on an encoder and decoder, and attention layers [6]. The encoder is used for building a latent representation of various inputs, and the decoder is a generative model that can generate outputs based on its learned feature. It makes use of two types of attention functions: scaled dot-product attention, and multi-head attention. The former allows it to perform computations in parallel, and the latter allows it to access different representations learned at different places in the input. [12].

Transformers are more powerful than RNNs because they use a self-attention mechanism to compute relationships between input elements, whereas RNNs use a recurrent connection that passes information from one time step to the next. This allows transformers to capture longer-term dependencies in the input data, and to process input sequences of arbitrary length, whereas RNNs are limited by the fixed-length unrolling of the recurrent connection.

2.4 BLEU

Evaluating natural language generation (NLG) systems automatically is challenging due to the difficulty of capturing all the nuances of the task at hand [11]. On the other hand, human evaluation is resource-intensive and impractical. We use the Bilingual Evaluation Understudy (BLEU) score to compare our performance to the baseline despite its limitations due to the lack of better alternatives.

BLEU works by multiplying two components: the brevity penalty term and the N-gram overlap term. The brevity penalty term penalizes outputs that are too short compared to the reference text with an exponential decay. The N-gram overlap is a precision metric that counts how many N-grams match their N-gram counterparts in the reference text [7]. BLEU scores ranges from 0 to 1. A score below 0.2 is poor, and a score above 0.6 is considered to be better than human performance [1].

2.5 Reddit

Reddit[1] is a website with user-run communities where users can create posts, make comments on them, and upvote/downvote posts and comments. Each community is known as a subreddit, and they cover different interest groups. This makes subreddits suitable for evaluating domain adaptation techniques as each subreddit has distinct topics and styles of speech.

3 Related Work

Domain adaptation of language models have been applied to sentiment classification [8] and domain adaption through changing the tokenizer [10]. Our work distinguishes itself from prior works by focusing on text generation without changing the tokenizer.

4 Methods

4.1 Dataset

We obtain comments from each of the subreddits from the Reddit Corpus, which is part of the ConvoKit dataset. The Reddit Corpus contains posts and comments from 948,169 subreddits from its inception until Oct 2018. [3]

The subreddits that we will be training and analyzing our results on are: /r/wallstreetbets ($\sim$250MB), /r/cmu ($\sim$4MB), /r/AskScienceFiction ($\sim$250MB), /r/piano ($\sim$77MB), /r/poker ($\sim$206MB).

4.1.1 Choice of Subreddits

We intentionally chose subreddits that vary significantly in their topic of discourse and style of speech. This allows the result of fine-tuning to be more evident. We discuss what each of the subreddits are about and why they are interesting from the perspective of our research question.

/r/wallstreetbets is a subreddit meant for discussing stock trading and investing, with a particular focus on risky or unconventional strategies. The subreddit has also developed its own set of unique lingo and colloquialism, which is very interesting for our experiments as we would like to see if the language model is able to pick up such lingo that it has never seen before.

/r/cmu is a subreddit for Carnegie Mellon University students, as well as prospective students, to talk about various aspects of student life. Common topics of discussion include course recommendations, rants about classes and being stressed, finding housing, admissions advice, and so on. It is a relatively small and inactive subreddit, but is still interesting for us to see whether a small dataset is also able to successfully fine-tune a large language model.

/r/AskScienceFiction is a subreddit for people to ask and answer questions related to science fiction. Threads of the subreddit often discuss and debate the feasibility of the technologies and scenarios that are depicted in science fiction works. This subreddit is interesting as the comments tend to be very scientific and well-crafted.

/r/piano is dedicated to the piano and other keyboard instrument, and provides a platform for piano enthusiasts to share their experiences, learn from each other, and discuss a wide range of topics related to playing the piano. It is interesting as the style of speech tends to be very supportive and wholesome.

/r/poker is a subreddit for the card game poker. Members of the subreddit often share tips and advice for improving as a poker player, and also post anecdotes about their games. It is interesting as comments tend to make use of a lot of technical game-specific terminology

[1] https://reddit.com

4.2 Data Pre-Processing

We had to remove empty comments, comments that were removed due to community guideline violations, and those that were deleted by the poster. These comments show up as "", "[removed]", "[deleted]" respectively.

Since we need all inputs to have the same length during training, we had a maximum context length of 64, and split up long comments into multiple new inputs, padding the last one if necessary. We performed padding with an end-of-sentence (EOS) token.

We shuffled the data, and split our dataset into 90% train data and 10% test data.

4.3 LSTM Baseline

We will use a LSTM network with four hidden layers as the baseline. This was trained on tensorflow for a five epoch with a batch size of 256 and a initial learning rate 0.01. We used the /r/wallstreetbets dataset for training the LSTM because it was the largest dataset. The RNN baseline was evaluated by its BLEU score and manually scored.

4.4 Domain Adaptation on a Transformer Model

Instead of training a model from scratch, we fine-tune a pre-trained GPT-2 model with 124.4M parameters from HuggingFace [2]. This helps to reduce both our computation costs and the amount of data required to produce good results, as compared to a model trained from scratch [2]. We used the same tokenizer used for the GPT-2 model when preparing inputs for training.

We used the following hyperparameters for training, which are common among all training runs: Learning rate: 0.0005; Batch size: 64; Seed: 42; Gradient accumulation steps: 8; Optimizer: Adam with $\beta_1, \beta_2 = 0.9, 0.999$, $\epsilon = 10^{-8}$; Learning rate scheduler: cosine with 1000 warmup steps.

Training and evaluation was performed on a Tesla T4 GPU, taking around 150 hours in total across all our experiments.

4.5 Training Input Ordering Analysis

This experiment aims to answer our question on whether it is possible to improve the performance of domain adaptation, by making use of our prior knowledge about the train input to re-order it before training.

In our case where our data are Reddit comments, we have access to metadata such as the score (i.e number of upvotes), anonymized speaker ID, and time posted of each comment. The most relevant metadata is the score, which motivates ordering the inputs for training in the following set of ways:

1. Inputs in the original order
2. Inputs sorted by score in ascending order, from least upvotes to most upvotes
3. Inputs sorted by score in descending order, from most upvotes to least upvotes
4. Using only inputs with score at least 5, sorted in descending order from most upvotes to least upvotes

Note that the original order is shuffled, as we always shuffle our data with the same seed before making the train/test split.

We chose to use the /r/wallstreetbets subreddit to perform input ordering analysis due to the large size of the dataset, and the prevalence of both popular and controversial opinions being voiced, which indicates that a large proportion of comments tend to receive either a lot of upvotes or downvotes. The BLEU score on the test set is used to evaluate the results for each of the experiments.

4.5.1 Domain Adaptation Analysis

To understand whether it is feasible to perform domain adaptation on commodity hardware on internet discourse with unseen vocabulary and potentially limited amounts of data, we train a model for each of the chosen subreddits.

We evaluate each of the trained models quantitatively using both the BLEU score and manual scoring via APES. In addition, we also take one random prompt from each of the subreddits, and have all models generate predictions. This allows us to qualitatively compare how each model has learned the unique styles of speech from their respective subreddits. We call this approach cross-prompt analysis.

4.6 BLEU Score Computation

To compute the BLEU score, we take each comment in the test set that has length at least 16 characters (to avoid high entropy in the amount of possible outputs for extremely short comments that will severely penalize the BLEU score), and split it into word tokens. If there are n such tokens, we concatenate the first $\lfloor n/2 \rfloor$ tokens to produce the prompt, and then sample from the model for up to $\max(3n/5, 5)$ new tokens. We then compare this predicted output with the ground truth via the BLEU metric.

4.7 APES Score Computation

We propose a new metric called APES (Average Perplexity Evaluation for Sentences) to address the limitations of BLEU scores. APES is scored by manually assigning a score between 1-5 for each predicted example, and taking the average. We applied APES by manually scoring the predictions from 200 randomly chosen test prompts from each subreddit.

A 1 means that the predicted result is unexpected or nonsensical given the prompt. For example, the /r/piano prompt of:

```
Absolutely, sorry, re-reading my post it does come off like "there's nothing you
can do to improve your ear playing," and that's not what I meant to say. You're
100% right, you can train your ear, and
```

produced a predicted result of:

```
till improve it. Best of! \n Have to check her stuff out tho :) \n Anyways? :)
\n Regards Bhatn :) \n Your Dad.
```

We scored it a 1 due to the response randomly switching topics and emojis. A 5 means that the response is understandable given the prompt and matches the style of the subreddit. We believed we could see the response on a given subreddit. For example, the /r/AskScienceFiction prompt of:

```
So, i finnaly found time to watch the old trilogy again and yes, it is exactly
in this moment. Luke says that there is Vader on that
```

produced a predicted result of:

```
ship which brings back the old Vader. So Luke calls off all the hunting after
he's seen the film. At this point he is probably the only
```

We scored it as a 5 due to it referencing both Luke and Vader from Star Wars. While it does end in the middle of the sentence, this is due to the restriction on the maximum number of tokens that we allow it to output.

5 Results

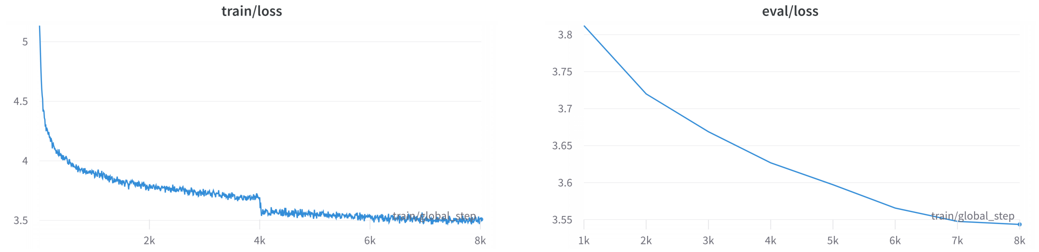
5.1 Training Input Ordering Analysis Results

The results for our training input ordering analysis is summarized in Table 1. Each of the models took around 8 hours to train. Note that a higher BLEU score is better. The brevity penalty measures how much penalization the computed BLEU score received as a result of the generated text being shorter than the reference text. The length ratio is the overall ratio of the sum of the lengths of all predicted outputs to that of the reference texts. The n-gram precision values measure how well the predictions matched the reference n-grams, where higher values are better.

The training and evaluation loss for the model trained on the original input order is given in Figure 1. We note that there is no overfitting based on the evaluation loss, and that the drop in loss has also plateaued near the end of training which indicates that 2 epochs is likely sufficient.

Input Processing	**BLEU Score**	**Brevity Penalty**	**Length Ratio**	**Precisions**			
				2-gram	**3-gram**	**4-gram**	**5-gram**
LSTM	0.001498	-	-	0.059026	0.000953	0.000370	0.000242
GPT-2, no training	0.004406	0.82769	0.84096	0.13910	0.00681	0.0013199	0.0006417
Original	0.029650	0.73153	0.76184	0.16716	0.030444	0.0237893	0.0222940
Sorted (Asc)	**0.031639**	0.70931	0.74434	**0.17250**	**0.033916**	**0.0268437**	0.0252075
Sorted (Desc)	0.029475	0.72489	0.75658	0.16729	0.03005	0.0240943	**0.0225727**
Positive, Sorted (Desc)	0.012917	0.71241	0.74677	0.15220	0.014769	0.0077368	0.0062138

Table 1: BLEU scores on /r/wallstreetbets evaluated against various input processing techniques.



(a) Training Loss

(b) Evaluation Loss

Figure 1: Loss over steps when performing training on /r/wallstreetbets corpus on the original input order for 2 epochs

5.2 Domain Adaptation Analysis Results

Subreddit	**BLEU**	**APES**	**Epochs**	**# Examples**	**Epochs × # Ex.**	**Time (hrs)**
LSTM (/r/wallstreetbets)	0.001498	2.337	5	-	-	-
/r/wallstreetbets	0.029650	3.550	2	2053180	4106360	8.13
/r/cmu	0.008762	3.685	20	14336	286720	0.58
/r/AskScienceFiction	0.005563	3.490	5	849363	4246815	8.98
/r/piano	0.008547	3.160	5	233936	1169680	2.11
/r/poker	0.005274	3.380	2	957910	1915820	3.89

Table 2: The BLEU and APES scores of our LSTM baseline trained on /r/wallstreetbets, and the subreddits that we fine-tuned on, evaluated against unseen test examples from their own subreddits. The number of epochs, number of training examples, product of epochs and training examples, and total training time are also given for comparison.

Table 2 summarizes the BLEU and APES scores of each of the models fine-tuned on their respective subreddits. We could not train all of them for a comparable number of examples due to overfitting for subreddits with smaller datasets, such as /r/cmu. The BLEU scores obtained are not directly comparable between each of them as their training and test sets are all completely different, and we include them only for reference. The APES scores are directly comparable as it is based on expert human evaluation.

Tables 3 and 4 shows the results of cross-prompt analysis for a prompt taken from /r/wallstreetbets and /r/cmu respectively. The results for other subreddits are given in the appendix, which we encourage the reader to check out as they are also very interesting.

Prompt Subreddit:	/r/wallstreetbets
Prompt:	I'm up 30% on my Nov 17 $170 calls. I'm gonna wait until the last 30 minutes of the day
Reference:	to decide whether I'm gonna sell or not. I wouldn't be scared to hold tho. COST has a high ceiling
Training Subreddit	**Predictions**
GPT-2	for a call, for the rest of the afternoon for another
LSTM	value off before, only TSLA on to tomorrow pay sales. **Thoughts, maybe similar and I share
/r/wallstreetbets	for more tendies and then double down after a solid report. Oh but, my portfolio!
/r/cmu	before finally making my decision. Last time I tried this was
/r/AskScienceFiction	to call and see how much money they earn by phoning
/r/piano	when the rest is more available. :) Thank you!
/r/poker	before calling with A hi, hoping it goes through.

Table 3: Predictions for the chosen subreddits using an unseen prompt taken from the test set /r/wallstreetbets. The actual ground truth reference is also provided.

Prompt Subreddit:	/r/cmu
Prompt:	Do you know how many people generally try and fail? (Out of the people who fulfill the requirements of completing the courses 15-122, 15-150, 15-210, 15-213,
Reference:	15-251 and 21-127 and has a 3.6 QPA in the set of required courses listed above and have an overall 3.0 QPA?) Thanks for the response!
Training Subreddit	**Predictions**
GPT-2	15-220, 15-230, 12-243, 12-248
LSTM	known your don't and go for in I bullshit most You 10 and. be ONTX $SCON COULDN'T less I've, probably a report to I
/r/wallstreetbets	22, 23?) And here you are with 500K in debt
/r/cmu	15-251, and 21-269) They might even make a "do it" list, and then only take 15-150 with 15
/r/AskScienceFiction	15-220) Are they simply unaware of their powers?
/r/piano	etc.) Are you sure he's doing his PhD in composition
/r/poker	etc.) Most people who try to work hard to make $25

Table 4: Predictions for the chosen subreddits using an unseen prompt taken from the test set /r/cmu. The actual ground truth reference is also provided.

6 Discussion and Analysis

6.1 Training Input Ordering Results Analysis

We noticed that the BLEU scores improved slightly when the input was ordering in ascending order. Since there were only 2 epochs of training, this was likely to be helpful because of the inherent recency bias in training updates, allowing it to learn from the best examples when it finished training. However, we note that the BLEU score improvements are not significant.

On the other hand, when the ordering was in descending order, the BLEU score decreased very slightly, which could be explained by the same reasoning, noting that now it is finishing training with the worst examples.

When the ordering was done in descending order and using only comments with at least 5 upvotes, the BLEU score decreased significantly. This was initially surprising, as we expected it to perform better when we only gave it good training examples. However, we noticed that the number of comments that received at least 5 upvotes was significantly smaller than the total number of comments, meaning that the amount of data available for training was severely restricted and thus it could not generalize as well. Furthermore, since the test set contains comments of arbitrary scores, it could be possible it also did not learn how to properly continue prompts from downvoted comments, contributing to its low score.

6.2 Cross-Prompt Results Analysis

The qualitative results from cross-prompt analysis given in Tables 3, 4, and Tables 5, 6, 7 in the Appendix, is highly promising.

For instance, given the /r/wallstreetbets prompt in Table 3, the /r/wallstreetbets model mentions "tendies", which is a common figure of speech used by members of the subreddit to refer to their stock gains, which would otherwise refer to chicken nuggets in a context outside of the subreddit. Similarly, the response from /r/cmu talks about "making my decision", which is a common phrase in threads in /r/cmu by prospective students about making a decision on which school they would like to go to. Similarly, we notice how /r/piano adds a smiley and a "Thank you!" to the end of the message, which is similar to the nice and wholesome manner that members of /r/piano tend to speak in. It is also interesting to note how /r/poker interprets "calls" as a poker call instead of a stock option call, and responds accordingly by mentioning that one should call as well with an ace high. This is an excellent demonstration of how the trained models have specialized their understanding of probability distributions of words to the unique context that they were trained in.

6.2.1 Domain Adaptation APES Scores Analysis

The manual scoring of the domain adapted results reveals that each model adapted to the style and culture of their respective subreddits. Each model scored above 3 on average when manually scored, indicating that the response was generally comprehensible given the prompt. Many of the test prompts scored a 5 and could be reasonably be expected to be seen on their subreddits. Less technical subreddits like /r/cmu, /r/wallstreetbets, and /r/AskScienceFiction seemed to have adapted better to fine-tuning. This could have been the case because the style of speech of these subreddits are simpler, and the model has to learn fewer new technical words and phrases to engage in meaningful discussion. All of the models outperformed the baseline LSTM model.

6.3 Limitations of BLEU Scores

Our results also show the limitation of BLEU scores for evaluating text generation tasks. BLEU scores are dependent on exact matches of n-grams between the reference text and the predicted text. This works well for translation tasks where there is a more limited set of possible translations for a given input. However, it fails to capture the open-ended nature of text generation in causal language modeling tasks, where there are many reasonable outputs for a given prompt. This can be seen by the very low BLEU scores resulting in Table 1, but the relatively high APES scores. In addition, BLEU is not comparable between different datasets. Since the BLEU score is dependant on the reference text provided, it is not possible to compare BLEU scores across datasets because they will have different reference texts. Therefore, we propose our APES metric as a way to evaluate such tasks.

7 Conclusion

We successfully adapted a GPT-2 model to generate Reddit comments in the style of a specific subreddit. Each model was able to generate comments that a user could expect to see on a particular subreddit. Our trained models were able to learn both the colloquialisms and style of speech of their respective subreddits. In addition, we showed that the models can better adapt to the new style when the training data is sorted in ascending order of upvotes. This is due to the most recent training examples of the highest scoring comments being given the most weight in the model. We demonstrate the limitations of BLEU scores for open-ended text generation tasks, and show that our proposed APES metric is a strong contender as a new standard metric.

References

[1] Evaluating models | automl translation documentation ; google cloud. URL https://cloud.google.com/translate/automl/docs/evaluate.

[2] URL https://huggingface.co/docs/transformers/training.

[3] Reddit corpus (by subreddit). URL https://convokit.cornell.edu/documentation/subreddit.html.

[4] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3(null):1137–1155, mar 2003. ISSN 1532-4435.

[5] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9:1735–80, 12 1997. doi: 10.1162/neco.1997.9.8.1735.

[6] H. Li. Language models: Past, present, and future. *Commun. ACM*, 65(7):56–63, jun 2022. ISSN 0001-0782. doi: 10.1145/3490443. URL https://doi.org/10.1145/3490443.

[7] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu. Bleu: a method for automatic evaluation of machine translation. 10 2002. doi: 10.3115/1073083.1073135.

[8] A. Rietzler, S. Stabinger, P. Opitz, and S. Engl. Adapt or get left behind: Domain adaptation through BERT language model finetuning for aspect-target sentiment classification. *CoRR*, abs/1908.11860, 2019. URL http://arxiv.org/abs/1908.11860.

[9] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. 1986.

[10] V. Sachidananda, J. S. Kessler, and Y. Lai. Efficient domain adaptation of language models via adaptive tokenization. *CoRR*, abs/2109.07460, 2021. URL https://arxiv.org/abs/2109.07460.

[11] A. B. Sai, A. K. Mohankumar, and M. M. Khapra. A survey of evaluation metrics used for NLG systems. *CoRR*, abs/2008.12009, 2020. URL https://arxiv.org/abs/2008.12009.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL http://arxiv.org/abs/1706.03762.

Appendices

A Cross-Prompt Analysis

The results of cross-prompt analysis for prompts taken from /r/AskScienceFiction, /r/piano, and /r/poker are given below.

Prompt Subreddit:	/r/AskScienceFicition
Prompt:	So, i finnaly found time to watch the old trilogy again and yes, it is exactly in this moment. Luke says that there is Vader on that
Reference:	ship while they're waiting for the codes to work and allow them safe pass to the Endor and right after Han's like: "there are alota command ships"
Training Subreddit	**Predictions**
GPT-2	ship, which, to him, is clearly not the most ideal choice for
LSTM	in. constante from do than it's are above, the and go what- the a death between But but the a But it people years
/r/wallstreetbets	timeline. Who'll take Vader's job? They just never come
/r/cmu	old one, but this time he is talking about a missing Malaysian exchange student
/r/AskScienceFiction	ship which brings back the old Vader. So Luke calls off all the hunting after he's seen the film. At this point he is probably the only
/r/piano	storm and Leia finds him... but I could be wrong xD thanks!!
/r/poker	able and he asks him a question because of Luke's story. So what

Table 5: Predictions for the chosen subreddits using an unseen prompt taken from the test set /r/AskScienceFiction. The actual ground truth reference is also provided.

Prompt Subreddit:	/r/piano
Prompt:	Work at it still. Ravel wrote an entire piano concerto for just left hand for a friend that lost his right arm during the
Reference:	war. You can probably get used to using the index finger in place of the thumb but you won't be able to extend very far.
Training Subreddit	**Predictions**
GPT-2	fight on the street. I was shocked that they just gave him the
LSTM	an stocks on options : so prices based, the wrong too where currently either a start 500 price what they've year and were
/r/wallstreetbets	'08 recession and lost half of his accts after the
/r/cmu	second class together. He was so enthusiastic about working with students and faculty
/r/AskScienceFiction	tragic accident. Its not like Batman was planning to use
/r/piano	Revolutionary. Those left hand notes make my butt hurt and want it to heal. Play your hardest repertoire and see where the numbers line up with
/r/poker	WSOP. He couldn't play it anymore.

Table 6: Predictions for the chosen subreddits using an unseen prompt taken from the test set /r/piano. The actual ground truth reference is also provided.

Prompt Subreddit:	/r/poker
Prompt:	If he’s not calling the turn, he’s rarely calling the river. All you do by checking there is
Reference:	give these hands infinite odds to beat you. Make your opponent make a mistake. Which you did. WP.
Training Subreddit	**Predictions**
GPT-2	his car," he says. But most
LSTM	still option : 04 we &gt go 2 anything will the in? should another gold to,
/r/wallstreetbets	give-aways, and you do you!
/r/cmu	a temporary stopping point so you don’t have issues
/r/AskScienceFiction	water flow but that’s rarely part of the game
/r/piano	turn it into a bassoon.
/r/poker	fold out draws. I wouldn’t necessarily advise that because V can possibly be capable of barrelling Q

Table 7: Predictions for the chosen subreddits using an unseen prompt taken from the test set /r/poker. The actual ground truth reference is also provided.