AI Song Lyric Generation

Casey Hahn & Michael Lu
DATASCI 266, Fall 2023

Abstract

This project investigates how song lyricists can leverage generative artificial intelligence (AI) as a collaborative tool to help inspire and augment their creative process. Leveraging pre-trained natural language processing (NLP) models as a starting point, the project experiments with model fine tuning and generation parameter adjustments to make the models better suited for the generation of song lyrics. Unlike previous efforts, this project's experimental models simplify the user experience by eliminating the need to specify song lyric-specific constraints. The experiments showed that fine tuning pre-trained NLP models can produce results that are comparable or better than the results produced by prior work, as measured by widely accepted NLP evaluation metrics.

Introduction

Generative AI has attracted substantial attention following the introduction of OpenAI's ChatGPT and Google's Bard. AI has the potential to become a great productivity tool helping humans complete tasks with greater efficiency. Unlike the long form text on which NLP models are generally trained, song lyrics do not follow common sentence structures. Additionally, songwriters' use of rhymes, alliteration, repetition, and other expressive techniques make it difficult for general purpose NLP models to generate lyrics of good quality. These intricacies prompted our exploration into building an AI solution tailored specifically to songwriting.

Prior work in this space often relied on users to input multiple constraints, including genre specifications, rhyme schemes, syllable counts, and desired tones. In this paper, we propose a genre-agnostic model harnessing state-of-the-art pre-trained NLP models to suggest lyrics, aiming to overcome these limitations by autonomously inferring genre, rhyme schemes, tone, and other characteristics, without requiring users to input predefined constraints. This innovative approach offers a more versatile solution for generating song lyrics.

Related Work

Previous work has been done on developing generated lyric systems for song writing support. For instance, DeepBeat (Malmi et al., 2016) is a rap lyric generation tool that works by combining lines sourced from existing songs, ensuring both rhyme and meaningfulness. Similarly, LyriSys (Watanabe et al., 2017) emerged as a pioneering lyric generation tool infused with semantic themes, such as darkness or love. Both these tools are tailored to specific genres and require users to pre-define constraints such as rhyming patterns, semantic themes, and syllable counts.

The current landscape lacks a tool harnessing the latest large language NLP models capable of generating genre-agnostic lyrics that possess coherence, rhyme, and proficiency without the need for predefined constraints. The objective of this study is to fine-tune a pre-trained model to

generate lyrics when prompted with a small sample of lyrics. Specifically, we aim for our fine-tuned NLP model to infer genre, rhyme scheme, tone, and other pertinent characteristics without relying on manual user input of these constraints.

Methods

Overview

The project uses two pre-trained NLP models as baseline starting points. Next, we fine tune the models by training them on a song lyric dataset. This allows the models to learn features that are specific to song lyrics, which should improve the quality of the generated results. Additionally, we experiment with adjusting generation parameters, such as number of beams and temperature. Finally, we evaluate the quality of the output using common NLP metrics.

Model Architectures

Songwriters typically put pen to paper and write each lyric of a song. Songwriting is naturally a text generation task. Thus, a generative pre-trained transformer (GPT) model would be a suitable architecture for this task. For that reason, we chose OpenAI's GPT-2 as one of the base models. GPT-2 is trained on a very large corpus of publicly available English-language text scraped from outbound links found on Reddit. It was trained specifically to predict the next word in sentences. For our experiments, we used the small and medium model checkpoints which encompass 124 million and 355 million parameters, respectively.

Alternatively, assisting a songwriter in their creative process can be thought of as a text-to-text task: given some lyrics of a song, generate the next lyric. A sequence-to-sequence model with an encoder-decoder architecture is best suited for this type of task; so we chose Google's Text-to-Text Transfer Transformer (T5) model as a second base model. T5 is pre-trained on a number of different task types. We focused on T5's summarization task as a method for generating song lyrics. For summarization, T5 was trained on the Colossal Clean Crawled Corpus (C4) and Wiki-DPR datasets, both of which are large corpuses of English-language text scraped from the web. For our experiments, we used the small and large model checkpoints, which encompass 60 million and 770 million parameters, respectively.

Fine Tuning

Song lyrics are unique in that they do not adhere to the same sentence structures as and may incorporate linguistic techniques not commonly found in long form text. Both GPT-2 and T5 were trained on corpuses made up largely of long form text. As a result, the pre-trained models will likely generate lyrics that do not match the style of the human-written lyrics. Fine tuning the pre-trained models by training them on a song lyric dataset will allow the models to learn features unique to song lyrics. Training the models with additional examples of song lyrics should make them better suited for the task of lyric generation and produce lyrics that more closely resemble the style of human-written lyrics.

Data Pre-Processing

We leveraged a Kaggle song lyric dataset that was originally scraped from Genius.com, a website that crowdsources song lyrics and related annotations. Consisting of over 5 million songs, the dataset includes titles, artists, genres, and lyrics for each song. User-submitted annotations include production credits, part of song markers (e.g. intro, chorus, bridge, outro, etc.), and indicators of spoken passages and instrumental interludes. Since our task is lyric generation, we scrubbed the song lyrics of these annotations.

Given limited computing and scarce GPU resources for fine tuning our models, we narrowed our training dataset by focusing only on English-language songs in the rap, rhythm & blues, rock, and country genres. We performed stratified sampling to create a smaller dataset of 100,000 songs equally weighted across the four genres. Finally, we performed a 60/20/20 split of the dataset for training, validation, and testing.

Each base model requires a specific format for the training dataset to be used for fine tuning. T5 is a text-to-text model that requires (i) an input string with a task-specific prefix and (ii) a corresponding output string. We constructed input strings consisting of a task prefix (e.g. “write the next line for:”) plus the first eight lyrics from each song. The corresponding output string is the song’s next (9th) lyric. GPT-2 similarly requires training data structured as an input string and an output string. Since it is trained to guess the next word in a sentence, the output string is simply the input string shifted to the right by one word.

Generation

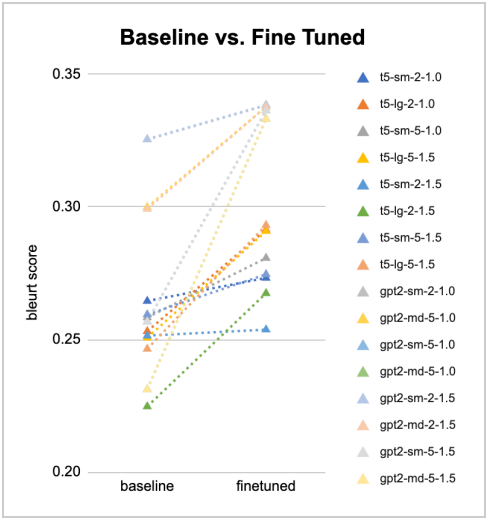
Besides the model architecture (2), model size (2), whether the model was fine tuned or not (2), we also varied two generation hyperparameters: number of beams (2) and temperature (2). The number of beams controls how many options the model considers when choosing the “best” output at each generation step. Temperature controls the randomness of the generated text and is often associated with the creativity of the output. For each song in our test dataset, we generated lyrics a total of 32 times across the combination of variables ($2^5 = 32$).

From our 20,000 song test dataset, we took a stratified sample of 250 songs from each of the 4 musical genres. For each song, we tokenized the first eight lyric lines of each song (and if needed, the task-specific prefix) and passed them as inputs to the models for lyric generation. To evaluate the quality of the output strings, we used each song’s next (ninth) lyric line as the reference string for calculating the evaluation metrics. Table 1 in the Appendix shows examples of the input strings, the output strings (generated lyrics), and the reference string.

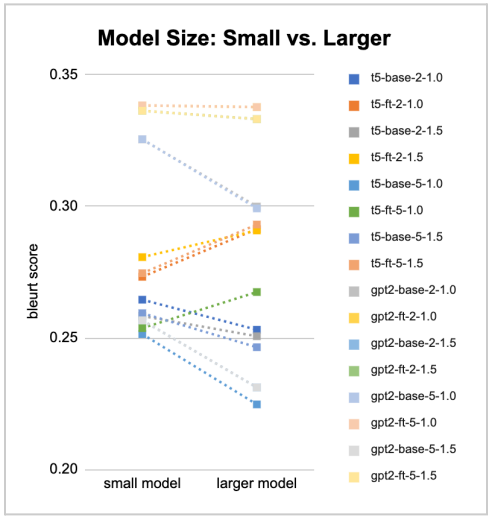
Evaluation

To measure the quality of the generated lyrics, we utilized BLEU, BLEURT, ROUGE-1, ROUGE-2, and ROUGE-L as our evaluation metrics. BLEURT evaluates semantic coherence, tone consistency, and overall meaning conveyed in the generated lyrics. BLEU and ROUGE gauged each model's proficiency in incorporating specific terms and the overall utilization of words in the generated lyrics. Collectively these metrics provide a comprehensive framework for

evaluating the linguistic quality and contextual relevance of the model's generated lyrics compared to the human-written reference lyrics.

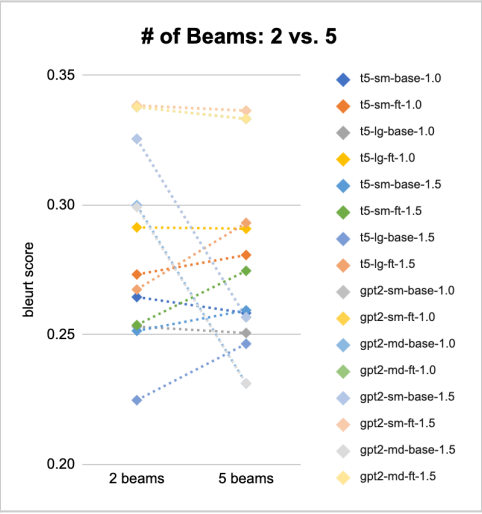


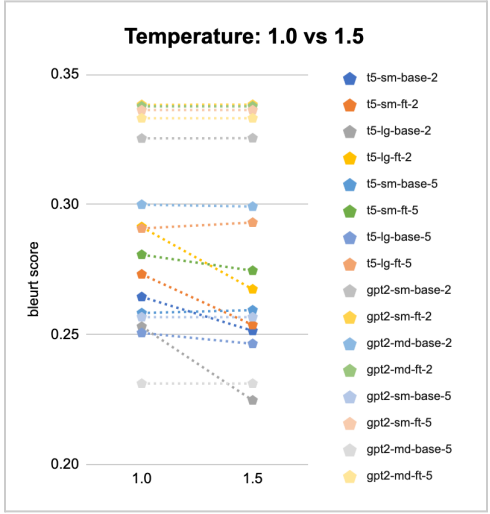
Fine tuning the pre-trained models improved the evaluation metrics for GPT-2 and T5 models across the board (holding all other variables constant). This observation seems to confirm that the additional training on the song lyric dataset allowed the models to learn features present in the song lyrics that were not present in the long-form text corpuses used to pre-train the models. Due to limited compute resources, we fine tuned our pre-trained models with only a 60,000-song training dataset. Fine tuning with the full 3 million-song dataset could lead to even greater improvements.



Increasing the size of the pre-trained model (T5: small vs. large; GPT-2: small vs. medium) produced decidedly more mixed results. For both GPT-2 and T5, the larger pre-trained models scored marginally *lower* than their smaller equivalents. The larger fine tuned GPT-2 models yielded similar results as their smaller fine tuned equivalents. The larger fine tuned T5 models performed marginally better than their smaller counter- parts.

Varying the generation hyperparameters also produced mixed results. Increasing the number of beams from 2 to 5 in T5 models generally led to slightly *lower* scores. Small improvements were only seen when the temperature on the T5 models was also increased from 1.0 to 1.5. Increasing the number of beams resulted in





little to no change in the scores for fine tuned GPT-2 models and slightly *lower* scores for pre-trained GPT-2 models. We hypothesize that varying the number of beams had very little impact because our task only generates the next song lyric, which typically consists of only 6 to 9 words.

Increasing the temperature from 1.0 to 1.5 had virtually no impact on the scores for the GPT-2 models, while the scores for the T5 models were slightly lower across the board.

Our top performing model variants were a fine-tuned T5 large checkpoint model with 2 beams and a fine-tuned GPT-2 medium checkpoint model with 5 beams. The results highlight the differences between the model architectures: T5 produced higher BLEU and ROUGE scores because its encoder-decoder architecture allowed it to better capture specific terms. GPT-2 had higher BLEURT scores because its generated lyrics exhibited greater coherence and better matched the tone of the reference lyrics.

model	size	type	beams	temp	bleu	bleurt	rouge1	rouge2	rougeL
T5	lg	baseline	2	1	0.0201	0.2531	0.1165	0.0363	0.1065
T5	**lg**	**fine-tuned**	**2**	**1**	**0.0666**	**0.2914**	**0.1718**	**0.0838**	**0.1653**
GPT2	med	baseline	2	1	0.0276	0.2312	0.0843	0.0324	0.0810
GPT2	**med**	**fine-tuned**	**5**	**1**	**0.0360**	**0.3331**	**0.1377**	**0.0567**	**0.1300**

The fine-tuned T5 large model's BLEU score of 0.067 outperforms the 0.051 BLEU score achieved by Ram, et all (2021) in their model for a similar lyric generation task. That team did, however, improve the BLEU score to 0.144 after incorporating additional generation constraints, such as number of syllables and rhyme.

Differences in Output due to Model Architecture

The T5 model, with its encoder-decoder, text-to-text architecture, holds a theoretical advantage over the GPT-2 model, which is based solely on a decoder architecture. This advantage suggests that T5 could offer more contextually accurate lyric generation.

Analyzing the song with ID = 2894076 (refer to Table 1), the input lyrics convey an artistic expression of dissatisfaction and a desire for self-improvement, culminating in the phrase "If I could only find." The generated lyric from the GPT-2 model, "a way to get out of this place," although coherent, lacks contextual relevance, as the original lyrics do not reference a specific place. Conversely, the T5 model's generated lyric, "I'd be a better person," is more contextually aligned with the input but might lack some coherence compared to the GPT-2 output.

We hypothesize that the disparities in generation could be influenced by the number of input song lyrics provided to the models. GPT-2 might perform better when given the initial few lines of a song, where there's limited context; while T5 could potentially excel when given more song lyrics that allow it to build a richer context to generate more accurate and coherent outputs.

Conclusion

Our experiments showed that modern pre-trained NLP models can serve as excellent starting points from which to build task-specific lyric generation models. With only a modest training set of 60,000 records, we were able to fine tune the pre-trained models to produce results that were on par with, if not better than, the results produced by models from prior work in this space.

Similar to what researchers have experimented with in the past, can additional parameters be incorporated during training and in the task prompt to improve results? For example, identifying the genre could yield generated lyrics that better reflect the unique themes and styles of songs of that genre. Identifying the part of the song (e.g. intro, chorus, interlude, etc.) could similarly yield lyrics that better match the unique cadence, rhyme, and structure of that song part.

Given sufficient compute and GPU resources, fine tuning with a much larger training data set with additional parameters could lead to substantially better results. Training NLP models using copyrighted materials, however, is fraught with ethical dilemmas and legal risks. Should songwriters be compensated for contributing their prior works to training an AI model? Should an AI-generated song receive copyright protections? Who owns the copyright to the AI-generated song? These are important issues that must be addressed before this new class of tools can be widely adopted.

Additional Resources

- Final Presentation Slide Deck
- Project GitHub Repository

References

Eric Malmi, Pyry Takala, Hannu Toivonen, Tapani Raiko, and Aristides Gionis. 2016. “DopeLearning: A computational approach to rap lyrics generation”. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 195–204.

Kento Watanabe, Yuichiroh Matsubayashi, Kentaro Inui, Tomoyasu Nakano, Satoru Fukayama, and Masataka Goto. 2017. “LyriSys: An interactive support system for writing lyrics based on topic transition”. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces (ACM IUI 2017)*, pages 559–563.

Naveen Ram, Tanay Gummadi, Rahul Bhethanabotla, Richard J. Savery, Gil Weinberg. 2021. “What? Collaborative Pop Lyric Generation Using Multitask Transfer Learning”. *Proceedings of the 9th International Conference on Human-Agent Interaction (2021)* 165-173.

Appendix

Table 1: Sample of Lyric Predictions

song_id	input	reference	GPT-2	T5
3101979	You should let me come thru, yeah Speeding down 40 west to you Yeah, hey Clouds in the city grey, whoa But the weekend's not too far away, hey You should let me come thru, yeah Speeding down 40 west to you Yeah, yeah, hey	Clouds in the city grey, whoa	Clouds in the city grey, whoa	Clouds in the city grey, whoa
2894076	Life is short Filled with stuff I don't know what for I ain't had enough I learned all I know By the age of nine But I could better myself If I could only find	Some new kind of kick	a way To get out of this place	I'd be a better person
3116564	Have I been keeping you sick? Here's me believing a bleeding heart's just a quick fix All you needed was me, not my solutions or views I wish I could take all this pain back I've put you through Guess it's too little too late At least I can assure you; I had the best intentions It's toxic, it's noxious, ironically honest Get out tonight, don't say farewell	Run for your life, go save yourself	Cause I've been keeping you sick Here's me believing a bleeding heart's just a	I'm gonna take you back to the place where you belong

Evaluation Metrics - T5

size	type	beams	temp	bleu	bleurt	rouge1	rouge2	rougeL
sm	baseline	2	1.0	0.0227	0.2645	0.1145	0.0353	0.1052
sm	finetuned	2	1.0	0.0533	0.2732	0.1512	0.0660	0.1468
lg	baseline	2	1.0	0.0201	0.2531	0.1165	0.0363	0.1065
lg	finetuned	2	1.0	0.0666	0.2914	0.1718	0.0838	0.1653
sm	baseline	5	1.0	0.0229	0.2583	0.1157	0.0363	0.1060
sm	finetuned	5	1.0	0.0571	0.2807	0.1530	0.0711	0.1476
lg	baseline	5	1.0	0.0183	0.2507	0.1145	0.0335	0.1046
lg	finetuned	5	1.0	0.0660	0.2908	0.1646	0.0818	0.1588
sm	baseline	2	1.5	0.0228	0.2515	0.1141	0.0361	0.1049
sm	finetuned	2	1.5	0.0347	0.2537	0.1341	0.0454	0.1281
lg	baseline	2	1.5	0.0172	0.2247	0.0980	0.0274	0.0910
lg	finetuned	2	1.5	0.0390	0.2674	0.1329	0.0531	0.1262
sm	baseline	5	1.5	0.0199	0.2594	0.1116	0.0326	0.1022
sm	finetuned	5	1.5	0.0546	0.2746	0.1482	0.0636	0.1428
lg	baseline	5	1.5	0.0189	0.2465	0.1151	0.0359	0.1064
lg	finetuned	5	1.5	0.0633	0.2931	0.1670	0.0791	0.1598

Comparative Metrics - T5

The tables below show the difference in evaluation metrics when model and generation hyperparameters are modified.

Fine Tuning: fine tuned vs. baseline models

model	bleu	bleurt	rouge1	rouge2	rougeL
sm-2-1.0	0.031	0.009	0.037	0.031	0.042
lg-2-1.0	0.047	0.038	0.055	0.047	0.059
sm-5-1.0	0.034	0.022	0.037	0.035	0.042
lg-5-1.0	0.048	0.040	0.050	0.048	0.054
sm-2-1.5	0.012	0.002	0.020	0.009	0.023
lg-2-1.5	0.022	0.043	0.035	0.026	0.035
sm-5-1.5	0.035	0.015	0.037	0.031	0.041
lg-5-1.5	0.044	0.047	0.052	0.043	0.053

Model Size: t5-large vs. t5-small models

model	bleu	bleurt	rouge1	rouge2	rougeL
base-2-1.0	-0.003	-0.011	0.002	0.001	0.001
fine-2-1.0	0.013	0.018	0.021	0.018	0.019
base-5-1.0	-0.005	-0.008	-0.001	-0.003	-0.001
fine-5-1.0	0.009	0.010	0.012	0.011	0.011
base-2-1.5	-0.006	-0.027	-0.016	-0.009	-0.014
fine-2-1.5	0.004	0.014	-0.001	0.008	-0.002
base-5-1.0	-0.001	-0.013	0.004	0.003	0.004
fine-5-1.0	0.009	0.018	0.019	0.016	0.017

Number of Beams: 5 vs. 2

model	bleu	bleurt	rouge1	rouge2	rougeL
sm-base-1.0	0.000	-0.006	0.001	0.001	0.001
sm-fine-1.0	0.004	0.007	0.002	0.005	0.001
lg-base-1.0	-0.002	-0.002	-0.002	-0.003	-0.002
lg-fine-1.0	-0.001	-0.001	-0.007	-0.002	-0.007
sm-base-1.5	-0.003	0.008	-0.003	-0.004	-0.003
sm-fine-1.5	0.020	0.021	0.014	0.018	0.015
lg-base-1.5	0.002	0.022	0.017	0.009	0.015
lg-fine-1.5	0.024	0.026	0.034	0.026	0.034

Temperature: 1.5 vs. 1.0

model	bleu	bleurt	rouge1	rouge2	rougeL
sm-base-2	0.000	-0.013	0.000	0.001	0.000
sm-fine-2	-0.019	-0.019	-0.017	-0.021	-0.019
lg-base-2	-0.003	-0.028	-0.019	-0.009	-0.015
lg-fine-2	-0.028	-0.024	-0.039	-0.031	-0.039
sm-base-5	-0.003	0.001	-0.004	-0.004	-0.004
lg-base-5	-0.002	-0.006	-0.005	-0.007	-0.005
sm-fine-5	0.001	-0.004	0.001	0.002	0.002
lg-fine-5	-0.003	0.002	0.002	-0.003	0.001

Evaluation Metrics - GPT-2

size	type	beams	temp	bleu	bleurt	rouge1	rouge2	rougeL
sm	baseline	2	1.0	0.0274	0.3254	0.1115	0.0402	0.1059
sm	finetuned	2	1.0	0.0303	0.3383	0.1297	0.0477	0.1230
med	baseline	2	1.0	0.0203	0.2999	0.1003	0.0341	0.0954
med	finetuned	2	1.0	0.0344	0.3377	0.1347	0.0537	0.1281
sm	baseline	5	1.0	0.0241	0.2567	0.0874	0.0307	0.0831
sm	finetuned	5	1.0	0.0344	0.3363	0.1322	0.0540	0.1255
med	baseline	5	1.0	0.0276	0.2312	0.0843	0.0324	0.0810
med	finetuned	5	1.0	0.0360	0.3331	0.1377	0.0567	0.1300
sm	baseline	2	1.5	0.0274	0.3255	0.1113	0.0404	0.1060
sm	finetuned	2	1.5	0.0304	0.3384	0.1296	0.0477	0.1229
med	baseline	2	1.5	0.0203	0.2992	0.1000	0.0342	0.0953
med	finetuned	2	1.5	0.0344	0.3377	0.1344	0.0538	0.1281
sm	baseline	5	1.5	0.0242	0.2567	0.0874	0.0306	0.0831
sm	finetuned	5	1.5	0.0344	0.3363	0.1322	0.0540	0.1255
med	baseline	5	1.5	0.0275	0.2312	0.0841	0.0328	0.0809
med	finetuned	5	1.5	0.0362	0.3331	0.1376	0.0562	0.1300

Comparative Metrics - GPT-2

The tables below show the difference in evaluation metrics when model and generation hyperparameters are modified.

Fine Tuning: fine tuned vs baseline models

model	bleu	bleurt	rouge1	rouge2	rougeL
sm-2-1.0	0.003	0.013	0.018	0.007	0.017
med-2-1.0	0.014	0.038	0.034	0.020	0.033
sm-5-1.0	0.010	0.080	0.045	0.023	0.042
med-5-1.0	0.008	0.102	0.053	0.024	0.049
sm-2-1.5	0.003	0.013	0.018	0.007	0.017
med-2-1.5	0.014	0.039	0.034	0.020	0.033
sm-5-1.5	0.010	0.080	0.045	0.023	0.042
med-5-1.5	0.009	0.102	0.054	0.023	0.049

Model Size: gpt-2 medium vs gpt-2 small models

model	bleu	bleurt	rouge1	rouge2	rougeL
base-2-1.0	-0.007	-0.025	-0.011	-0.006	-0.011
fine-2-1.0	0.004	-0.001	0.005	0.006	0.005
base-5-1.0	0.003	-0.026	-0.003	0.002	-0.002
fine-5-1.0	0.002	-0.003	0.005	0.003	0.005
base-2-1.5	-0.007	-0.026	-0.011	-0.006	-0.011
fine-2-1.5	0.004	-0.001	0.005	0.006	0.005
base-5-1.0	0.003	-0.026	-0.003	0.002	-0.002
fine-5-1.0	0.002	-0.003	0.005	0.002	0.005

Number of Beams: 5 vs. 2

model	bleu	bleurt	rouge1	rouge2	rougeL
sm-base-1.0	-0.003	-0.069	-0.024	-0.010	-0.023
sm-fine-1.0	0.004	-0.002	0.003	0.006	0.003
med-base-1.0	0.007	-0.069	-0.016	-0.002	-0.014
med-fine-1.0	0.002	-0.005	0.003	0.003	0.002
sm-base-1.5	-0.003	-0.069	-0.024	-0.010	-0.023
sm-fine-1.5	0.004	-0.002	0.003	0.006	0.003
med-base-1.5	0.007	-0.068	-0.016	-0.001	-0.014
med-fine-1.5	0.002	-0.005	0.003	0.002	0.002

Temperature: 1.5 vs 1.0

model	bleu	bleurt	rouge1	rouge2	rougeL
sm-base-2	0.000	0.000	0.000	0.000	0.000
sm-fine-2	0.000	0.000	0.000	0.000	0.000
med-base-2	0.000	-0.001	0.000	0.000	0.000
med-fine-2	0.000	0.000	0.000	0.000	0.000
sm-base-5	0.000	0.000	0.000	0.000	0.000
med-base-5	0.000	0.000	0.000	0.000	0.000
sm-fine-5	0.000	0.000	0.000	0.000	0.000
med-fine-5	0.000	0.000	0.000	-0.001	0.000