Tiny Love with Large Models: Applying LLMs to Generate Short Love Stories

CHRISTOPHER T. KANG and DHRUV SINHA*, The University of Chicago, USA

Large Language Models (LLMs) are lauded for their ability to consume massive volumes of text. This ability is pertinent when consuming corpora containing long-form text, e.g., books, scientific papers, or when generating long-form text. However, specific genres — like poetry — are defined by their brevity. In this work, we train GPT2 Neo via prompt engineering and fine-tuning on the New York Times' *Tiny Love Stories*. Human evaluation suggests that prompt engineered stories are indistinguishable from human-written stories. We explore what this implies for the capabilities of LLMs in concise, emotion-rich regimes.

ACM Reference Format:
Christopher T. Kang and Dhruv Sinha. 2022. Tiny Love with Large Models: Applying LLMs to Generate Short Love Stories. In *DL Sys '22: Poster Session on Deep Learning Systems, December 7, 2022, Chicago, IL.* ACM, New York, NY, USA, 9 pages. https://doi.org/6712684.6712684

1 INTRODUCTION

"What is love?" The age-old question seems to have answers far and wide: books, plays, podcasts, songs, poems, and a plethora of other material has been produced to attempt to answer this question. However, while many answers lean to the verbose, others lean to the concise, like the New York Times' Tiny Love Stories. These stories, comprised of a title and body of fewer than 100 words, are emotionally touching stories about heartbreak, loss, joy, belonging, and love.

Unfortunately, parsing, interpreting, and generating these stories seems like a fundamentally challenging task for LLMs. These models are trained on long corpora of text, like Wikipedia or news articles, which almost always exceed 100 words. Furthermore, the topic material of love seems inherently challenging to effectively instill within LLMs, especially given the diversity of types of love, complex emotions, and nuances relating to love. Together, these factors suggest that LLMs will struggle to effectively generate stories akin to the NYT's Tiny Love Stories.

In this project, we attempt to tackle short story generation relating to love. In particular, we produce a dataset of 750+ Tiny Love Stories published by the NYT, test two methods to train GPT variants to generate tiny love stories, and finally conduct human studies to evaluate the indistinguishability of generated text.

The remainder of the paper is organized as follows. In Section 2, we describe the methodology used to build an LLM that can generate a tiny love story given a title. In Section 3, we describe the process of scraping data from the New York Times and perform a qualitative analysis of the entries. In Section 4, we analyze the quality of text generated via prompt engineering and fine-tuning. In Section 5, we overview our human evaluation approach and the results of the trials. In Section 6, we summarize our results and analyze what it suggests about the future of concise text generation.

2 TASK AND METHODOLOGY

In our text generation task, we provide the model with a real NYT Tiny Love Stories title and seek to generate a realistic love story within 100 words. To achieve this generation, we proceed with the project in three phases:

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2022 Association for Computing Machinery.
Manuscript submitted to ACM

(1) **Gather data**: We scrape the title and text of Tiny Love Stories, then perform a qualitative analysis of the obtained stories.
(2) **Training**: We test two training strategies, prompt engineering and fine-tuning, and analyze their performance when generating text.
(3) **Evaluation**: We compare the generated and actual text, then conduct human studies to test the indistinguishability of LLM-generated text.

3 DATA

3.1 Acquisition

While the NYT has an API, it only provides links to the actual articles. Luckily, because NYT login information is stored in cookies, we can use typical web scraping techniques to obtain the stories. In particular, we take the following steps to obtain full stories:

(1) **Link generation**: Using API requests, obtain a list of URLs to Tiny Love Stories posts.
(2) **Story scraping**: Using Selenium and BeautifulSoup, we first insert the NYT login cookies, then load a URL from the prior step. Finally, we scrape the relevant text and store in a JSON file.

Additionally, though each story also includes an image, we choose to focus on language models (instead of multi-modal models). Thus, we do not import the media.

The scraping code yielded 751 stories (title/body pairs). We chose to use 450 stories for fine-tuning, 151 for validation, and 150 for testing.

3.2 Qualitative analysis

We analyze the properties of the collected stories. To begin, there are some properties enforced by the NYT prior to publication:

(1) **Length**: stories are explicitly restricted to be less than 100 words.
(2) **Grammar**: stories have proper English grammar, consistent with typical editorial standards.
(3) **Theme**: stories are unified by a common theme of "love."

However, there is an immense diversity of qualities in different dimensions:

(1) **Topic**: while many of the stories focus on romantic love, Tiny Love also includes stories about love between parent and child, platonic friendships, and other relationships.
(2) **Content and tone**: the authors come from diverse backgrounds, varying greatly in age (writers include college students to seniors), gender, sexuality, etc. This influences both the content of the stories and their sentiment, which ranges from celebratory to regretful, retrospective, grateful, etc.
(3) **Dialogue**: stories range from totally expository to containing 25+ words of dialogue. This greatly influences text generation, which may also need to be able to generate dialogue inline with the short story.

To illustrate these differences, we provide two examples of stories below:

> *Running, Singing, Tattooed* from 'Sex Cures Everything'
> A workaholic who had been single for a long time, I got breast cancer. Survived, and realized I only live once, and should do things I really like. Started running, got tattoos and took up singing lessons. Then I

> met Tracy, who loves running, tattoos and singing. We're now the running and singing tattooed couple.
> — *Erika Kato*
>
> *Let's Not Keep In Touch* from 'You Never Call Me Anymore'
> "You never call me anymore," I said to my mother. Four years ago, when I was in my first semester of college and she was newly divorced after 25 years of marriage, we talked every day. Our conversations about the politics of dorm living, or a bath for our family dog, could stretch for hours. Now our calls are less frequent, but I find joy in knowing that they're also less needed. I have made friends who will be hard to leave after graduation, and my mother has met someone who loves her almost as much as I do.
> – *Liv Coron*

Contrasting these stories immediately reveals key differences:

(1) **Topic**: *Running, Singing, Tattooed* focuses on romantic love, while *Let's Not Keep In Touch* focuses on parental love.
(2) **Content and tone**: *Running, Singing, Tattooed* is written with pride, contentment, and joy. In contrast, *Let's Not Keep In Touch* is more mellow, exuding retrospection, bittersweet tone, and satisfaction.
(3) **Dialogue**: *Running, Singing, Tattooed* has no inline dialogue, while *Let's Not Keep In Touch* does.

This diversity in story attributes immediately poses significant challenges to generation at scale. For example, while individual stories may be emulatable, encompassing the broad diversity of topics, tones, and presence of dialogue is challenging to embed within the language model. Furthermore, we would ideally seek that these parameters would be tunable/selectable. I.e., we would like to be able to generate stories with a specific type of love, emotional affect, presence of dialogue, and cultural influences. We leave these challenges to future work, though identify potential avenues to achieving this diversity.

4 TRAINING

Because our task is text generation, we choose to use the GPT family of models. In particular, the GPT Neo models are well-known and have been optimized. We consider two of the newer models, GPT2-Neo 2.7B [2] and GPT2-NeoX 20B [1]. Using these two models, we test two strategies to generate content:

(1) **Prompt engineering**, or providing examples within the prompt to the LLM, is tested using GPT2-Neo 20B.
(2) **Fine tuning**, or providing example texts and modifying the weights to align, is testing using GPT2-Neo 2.7B.

We ultimately identify that the 20B parameter model is far more successful and use it for evaluation; however, we do not rule out the possibility of fine tuning a 20B parameter model or larger. In fact, it is possible that, provided the physical hardware to train the 20B model, the fine tuned model would exceed the prompt engineered model.

4.1 Prompt engineering

Our prompt engineering approach is intuitive " we provide the following prompt structure to the model:

> Title: [[*Example title*]]
> Story: [[*Example story*]]
>
> Title: [[*Test title*]]
> Story:

During generation, we randomly select an example title/story pair from the training dataset. We then requested the model generate between 100-200 tokens and truncated any excess text.

While other applications would find this prompt approach challenging due to the context limit, the text length restriction imposed by the dataset makes prompt engineering feasible. It is challenging to think of examples where 100 word stories would exceed 1024 tokens, even at a character-level tokenization.

When using the 20B model, our generation consistently generated syntactically correct stories. Furthermore, the length of stories generated often mimicked the input story's length, thus implicitly approximating the NYT length requirement.

Unfortunately, the prompt generation approach implicitly imposes hidden biases in the generated text. During experimentation, we noticed that the content of the example story could be represented in the generated text. For example, consider the following example story:

> *After Secrets, Acceptance* from 'When I Hate My Husband'
> I learned that my father was a spy from a total stranger. "This is a C.I.A. base," the guard said, handing me a form to sign. At 20, I had long suspected this. Still, I was angry that my dad hadn't told me himself. Choosing to break my family's history of secrecy, I came out as gay, which my father rejected. I was done with him. But in his 70s, something shifted. He invited me and my wife for a visit. He never said, "I accept you," but I could tell he did. Just like that, I wasn't angry anymore. – *Leslie Absher*

When used as the prompt, generated stories frequently included themes of being gay or being in the closet and often had melancholic themes. This suggests that the prompt approach should be structured to better control the influence of the example story.

4.2 Fine-tuning

To fine-tune the model, we ultimately used the HuggingFace Accelerate module for PyTorch[1] This module provides native support for multi-GPU and multi-node training; we first experimented with multi-GPU training using a single compute node with four A100s.

We provided the GPT2-Neo 2.7B model with our 450 training examples. However, this nearly exceeded the memory capacity, typically consuming almost all of the 160GB of VRAM available. Training the 20B model would thus be infeasible with a single-node setup. Furthermore, while multi-node setups are theoretically possible with HuggingFace Accelerate, it was unclear how to establish the necessary networking infrastructure to facilitate inter-node communication.[2] Thus, we constrained our analysis to the 2.7B model.

Unfortunately, the outcome of fine-tuning was often incomprehensible – the text was typically syntactically incorrect and amounted to repetitions of pronouns like 'I.' This could be from a learning rate that is too high, but we believe that the model size itself was an inherent limitation. This hypothesis emerges because prompt generation on the 2.7B model already struggled to generate syntactically and thematically correct stories.

Given the limitations with fine-tuning – namely with resource cost to train, lackluster results, and poor baseline results of the model – we decided to focus on appropriate prompt engineering.

[1]This only came after many hours of failed experimentation with HuggingFace Trainer and the multi-GPU extensions. Thank you Peng for the attempts to debug. Not a fun time.
[2]We did attempt to follow the steps provided by Argonne on the multi-node setup to train GPT2-NeoX 20B, but ultimately ran out of time and found the instructions too sparse to follow accurately.

5 EVALUATION

5.1 Truncated Content Generation (TCG)

The limited length of generated text is a key constraint imposed by the NYT. However, LLMs do not natively enable word-level length limitations, instead having an intrinsic token-level limit. (Tokenizers frequently operate at a more granular level than words, like character levels). Though text generated through prompt engineering was frequently of appropriate length, it's unclear whether the LLM had the explicit constraint built into it's generation. We were particularly concerned with the case where an individual could solely use the the length of the paragraph to determine whether the story was written by a human or generated by our model.

To overcome this, we employ the Truncated Content Generation (TCG) approach. In TCG, we clip the last n words of both human written and model-generated stories, then proceed with the typical evaluation. The intuition is that this will make generation fairer, as both human and robot text have a length requirement that is imposed with some uncertainty. Thus, in all the stories, the last sentence is typically incomplete (or omitted) and the total word length is between 90-100 for all stories.

Generated stories are *indistinguishable* on the TCG model if humans cannot determine whether a given passage was written by human or robot, i.e., the accuracy of human classification is 50%.

5.2 Experimental design

To evaluate whether stories written by humans and those generated by models are indistinguishable, we created and distributed Google Forms surveys. In the forms, there were six stories, three written by humans and three generated by the model (though this information was not provided to participants). For each title/text pair, evaluators were asked whether the text was generated by a human or the LLM.[3] We then conducted an A/B test, producing five different types of forms and sending them to our friends. Each of these forms had a distinct set of six stories. Provided our experiments could scale, using A/B tests allows us to better analyze how qualitative features of the passages affect distinguishability.

Our experimental design intentionally prioritized simplicity — this is only a first step in analyzing the quality of LLM-generated responses. We focused solely on distinguishability, rather than asking whether passages were emotionally moving or pertinent to the title (though, it is conceivable that evaluators could use these criteria in their assessments of authorship).

5.3 Results

We received 13 responses over three forms. Again, we contextualize our results by noting that the text generated by the LLM is highly idiosyncratic and may not be representative of it's full capabilities. Thus, we would recommend scaling up our evaluations before producing concrete conclusions.

(1) Form 2: In this form, we received 4 responses with an average score of 2.75.
(2) Form 3: In this form, we received only one response with a score of 5.
(3) Form 4: In this form, we received 8 responses with an average score of 2.88.

If we collate the responses from all the forms we received responses on, the average score was 50% for 13 responses. This means that an evaluator, on average, correctly answered 3 out of 6 questions.

[3] As an example, feel free to take a survey on Form 2 or Form 3.)

Fig. 1. Summary of responses for Form 2

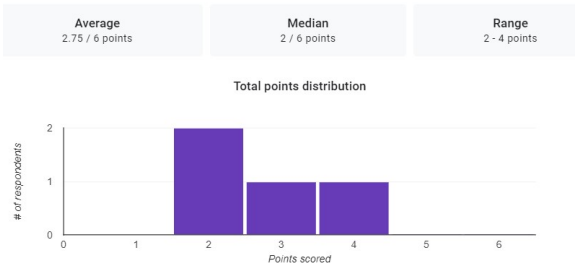


Fig. 2. Summary of responses for Form 3

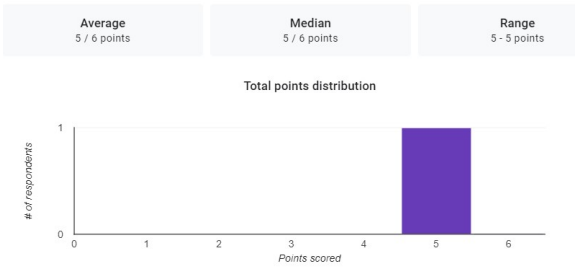
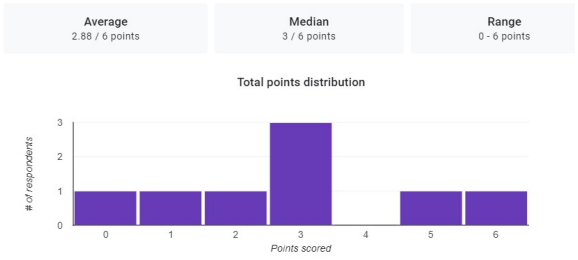


Fig. 3. Summary of responses for Form 4



If we only consider the stories written by humans, 53.84% of them were correctly predicted. This means that on all the human-written stories in the survey, only around 54% of them were correctly guessed as human-written. Model-generated stories followed a similar trend- around 46.15% of them were correctly predicted. While our sample size is small, this supports the hypothesis that stories generated by humans and LLMs are indistinguishable.

6 DISCUSSION AND FUTURE WORK

6.1 Indistinguishability

Our primary aim, indistinguishability, is supported when analyzing our quantitative results (50% accuracy). When studying which stories were typically judged as human-generated, we noticed three core attributes:

(1) **Subtle emotional content**: Tiny Love Stories are fundamentally emotional — they aim to invoke an emotional response in the reader. Evaluators preferred stories with an emotional appeal.
(2) **Information persistence**: While content was important, we also observed the importance of inter-sentence syntax. Because the 20B model largely was correct syntactically for individual sentences, evaluators instead relied on the overall composition of passages.
(3) **Relevance to provided title**: Unsurprisingly, the alignment of the story with respect to the title was a key piece of information, even in spite of TCG evaluation method.

For example:

> *Who's There?*
> My 88-year-old mother looks at the screen, squinting. "Who's there, you say?" My brother explains that the boxes frame her other children. Three thousand miles away, in Chicago, I wait for my 97-year-old father to sit down. Zoom, coronavirus, Lima, Chicago, Florida and dementia collide on our screens. I usually travel to Peru every other month to care for my parents. Without international flights, I feel as lost as my mother. I shoo away questions that start with "What if…?" Today, when my mother asks, "Who's there?" I say, "Tu hija." ("Your daughter.")

All evaluators correctly guessed that this story was human-generated. In this story, we hypothesize the use of the phrase 'Tu Hija' led reviewers to judge this as human-written. This could be because 'Tu Hija' is a Spanish phrase which is consistent with the context that her parents live in Peru, leading to a sense that information is consistently and gradually revealed. As a second example:

> *Music From Myanmar*
> The music is like nothing he has experienced before, but that is the kind of guy Taw Oo is. Originally from Myanmar, he spent nearly two decades as a rock star in Chicago before he found himself back in Myanmar, in a country ravaged by human rights abuses and war. Now he runs a popular pop-up concert in Yangon, his homeland, where he performs songs full of hope for a peaceful future, even though no one there can hear them. There's a reason he keeps the

Though this story was generated by the LLM, most evaluators guessed that it was written by a human. We hypothesize that this is explainable by the presence of a general emotional appeal and continuity between sentences (e.g., Taw Oo seems like a reasonable name from Myanmar; Myanmar has faced human rights abuses).

Causality claim? While these results are impressive, the size of responses do not yet suggest that stories generated by the language models are indistinguishable (However, it is definitely hinting in that direction). To solidify the indistinguishability claim, we want to collect several responses (at least 100) and see if the average score is statistically equal to or less than 3. Not only that, we want to understand to what extent did evaluators correctly guess (in separate analyses) model-generated and human-written stories. Ideally, on average, each evaluator should get a score of 1.5/3 or less on human-written as well as model-generated stories.

6.2 Generalizability

We believe there are major evaluative challenges to consider when extending LLMs to other emotion-driven text generation domains. Existing work [3] already suggests that our evaluation task may be inappropriate; robot-generated text is virtually indistinguishable from human-written text. Our analysis builds upon this in saying that even primitive LLMs can fool untrained readers simply by providing an emotional appeal and ensuring broader syntactic continuity.

To address this gap, we consider how evaluation tasks could be modified to more rigorously assess LLMs within this generation regime and provide a potential development roadmap.

An immediate response to challenges in human evaluation is to respond with *robot* evaluation, i.e. to use another LLM or ML technique to discriminate between generated responses. This would be inappropriate in most generation regimes (e.g., these discriminator LLMs likely could not assess the factual validity of generated longform text), but would be especially inappropriate for this emotion-driven task. In particular, a primary goal of the Tiny Love Stories is to evoke an emotional response, whether that is joy, melancholy, contentment, loss, etc. Until robots can feel, humans should be the primary evaluators of this content.

We then ask: how should humans in the loop evaluate the generated content? We suggest potential criteria below:

(1) **Indistinguishability**: The TCG task we provided could be modified; instead, both human- and robot-written passages could be provided. The evaluator then would need to decide *between* the two passages which is human. This task is likely far harder for the robot, which faces many of the aforementioned challenges with respect to length requirements, emotional affect, and relevance to title.
(2) **Emotional impact**: As stated earlier, one goal of the Tiny Love Stories is emotional effect. Asking evaluators to judge the emotional impact of the story could be a more compelling metric for whether robot-generated stories can achieve human-level emotional performance. However, this reward domain is likely far sparser than syntactic correctness, meaning that LLMs will need to attain mastery over more complex skills like paragraph flow and word choice.

It is unclear whether indistinguishability in the modified TCG task is easier than creating emotional impact; thus, future LLM experiments could consider evaluation using both tasks. Furthermore, we also suggest that future work analyzes how the emotional tone and level of dialogue within generated text can be parametrized (i.e. is it possible to programmatically set the level of sadness of a piece?).

6.3 Scaling up fine-tuning

The inferiority of fine-tuning for generation suggests that the difference in the quality of foundational models is insurmountable, i.e., a fine-tuned model may often be outperformed by much larger models which are simply prompted.

Thus, the next step for our project is the fine-tuning of a 20B model. To do this, we would need to extend our single-node, multi-GPU training to multi-node, multi-GPU training. This is achievable on the Polaris cluster (the tutorial is provided here), but the overall setup complexity and cost are unclear. As stated earlier, our 2.7B model required all four A100s to train; assuming linear scaling, the 20B model would easily require 10 nodes, with training time likely also expanding rapidly when accounting for inter-node communication costs.

Even if the fine-tuned model has superior performance to the prompted model, it is unclear whether this motivates further focus into fine-tuning. Namely, the fixed costs of fine-tuning are high — the infrastructure cost and setup complexity are driving costs — implying that most users will likely prefer to use foundation models which can be prompted, requiring almost no training or setup cost. Instead, the cost when running larger models is amortized over the additional inference cost. Thus, we are swayed to believe that further analysis of prompted generation for foundational models is motivated.

6.4 Qualitatively analyzing how prompt example stories affect foundational generation

Experiments with prompted generation suggests that the example story provided affects the generated story's tone and content. Provided that there is a major shift to prompted generation over fine-tuned small models, we recommend exploring the relationship between the example story and generated story.

One approach would be to perform topic modeling on both the generated and prompt stories. For example, HCI researchers are adept at coding interviews and written materials with general themes and tones — porting techniques from this sub-field may also reveal approaches to coding human- and robot-generated responses. From our preliminary view, we expect to find some intersection between the themes of the generated story and the stories given in the prompt.

Another variable to consider is the title provided, which itself may have an emotional effect (which may or may not conflict with the example story). To test the interplay between the title and example story, we could generate a series of stories for each title, then explore whether there are similar themes in content or tone for stories with the same/different titles and same/different example stories.

7 CONCLUSION

As humans and LLMs interact more frequently — especially in short-form systems like text or conversation — the ability to effectively navigate human emotions will prove a crucial skill. Our project contributes a base of work to begin further study. This work has resulted in the creation of a 750+ story dataset of NYT Tiny Love Stories and benchmark approach to the generation of short-form love stories consistent with the NYT. We demonstrate how the prompted GPT Neo 20B model appears to produce stories that are indistinguishable from human-generated stories in the TCG task. We contextualize the factors that may lead to this indistinguishability, including qualitative aspects of the text and broader challenges in the human evaluation of LLM-generated text. Finally, we introduce avenues for further study, including completing a roadmap to overcome challenges in evaluation, fine-tuning larger models on our provided dataset, and qualitatively analyzing the impact of prompt text on output.

REFERENCES

[1] Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. GPT-NeoX-20B: An Open-Source Autoregressive Language Model. https://doi.org/10.48550/ARXIV.2204.06745

[2] Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. *GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow.* https://doi.org/10.5281/zenodo.5297715 If you use this software, please cite it using these metadata..

[3] Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A Smith. 2021. All that's' human'is not gold: Evaluating human evaluation of generated text. *arXiv preprint arXiv:2107.00061* (2021).