

Neural Text Generation Conditioned on Title-Verdict Pair on Subreddit r/AITA Data

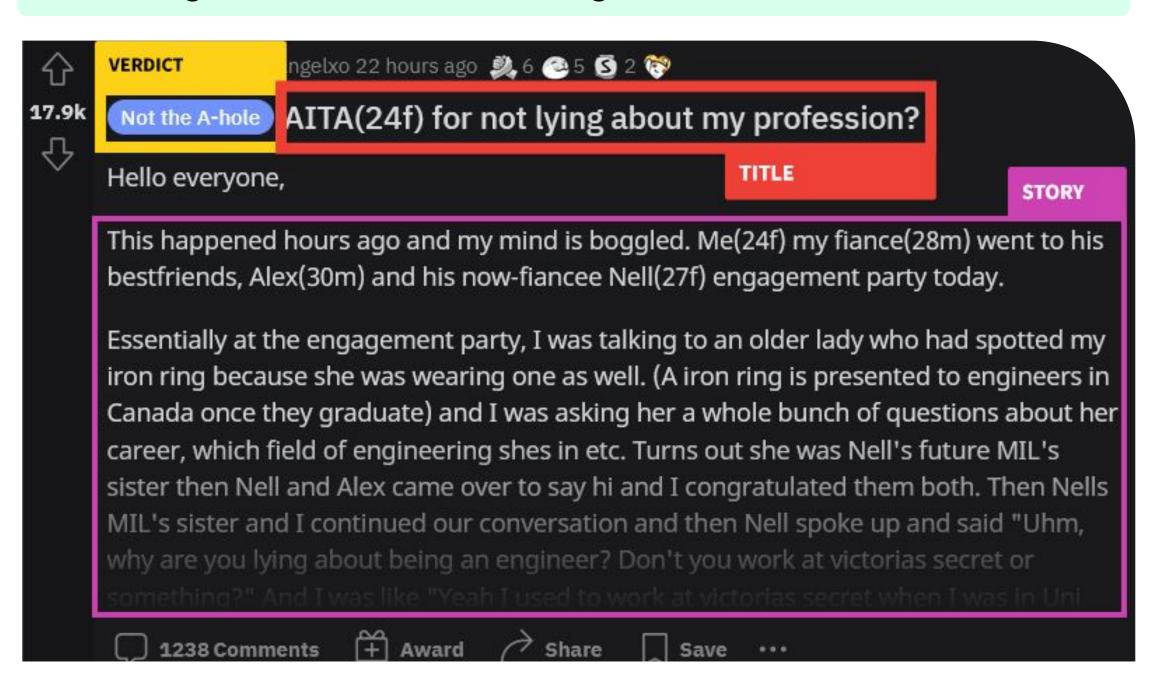
Team: Chomsky's Chimps

{esivaram, kkarthik, ssundarr, swaralia, baranwal}@usc.edu



INTRODUCTION

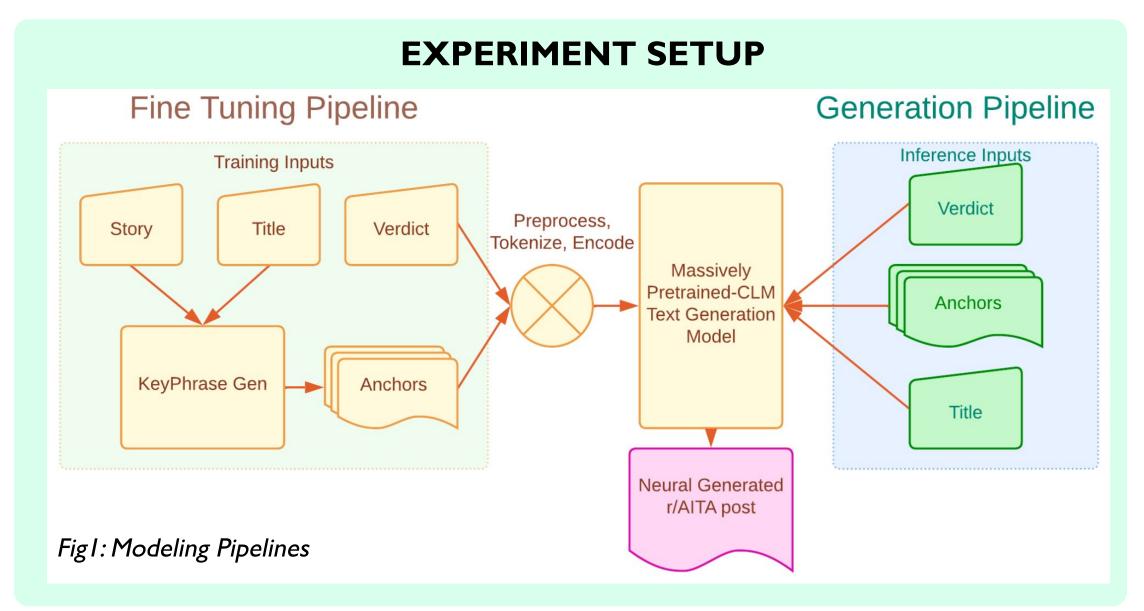
- I. The subreddit **r/AITA** is an online community where users express morally gray personal situations from their lives as short narratives, while other users pass a verdict on whether or not the narrator was at fault in the described situation.
- 2. We attempt to improve automatic storytelling in a constrained setting i.e. generating a story conditioned on a **<TITLE,VERDICT>** pair.
- 3. We propose a novel anchor-based approach for gaining control and introducing event coherence over text generation.



DATASET

- I. The raw dataset consists of over 120k posts made on the subreddit r/AITA between January I, 2014 and January I, 2022 which were scraped and structured using **Redshift API**.
- 2. The data set was further cleaned to remove links, garbled texts and outliers (documents over >596 tokens and under <105 tokens).
- 3. Only posts with YTA* and NTA* verdicts were retained
- 4. Dataset Size: Training: 3755 | Inference: 963
- 5. For extracting keyphrase Anchors, we used **RAKE** (which combines word frequency and graph-based metrics to weigh the importance) and **KeyBERT** (Neural) techniques.

*YTA = You're the A-hole, *NTA = Not the A-hole



- I) We trained models in three different fine tuning Pipelines to find out the best written stories:
 - a) Model: GPT2 Inputs: Title and Verdict
 - b) **Model:** GPT2 **Inputs:** Title, Verdict and Anchors
 - c) **Model:** GPT-Neo **Inputs:** Title and Verdict
 - Baseline: Prompt input to GPT2 with no fine-tuning on reddit data
- 3) Best Model Selection: The model parameters with the lowest loss during validation were stored as the representative weights for each model.

EVALUATION METRICS

Prompt Ranking Accuracy: Measures the magnitude of dependence of the generated text on the conditioning variables. Suitable for an open-ended generative task like ours.

BLEU: Automatic evaluation metrics based on n-gram overlap are ideally not adaptable for creative tasks like story generation which do not have solid ground truths

RESULTS

| Model | Training Inputs | Perplexity | BLEU | Prompt Ranking |
|-----------------------------------|------------------------------------|------------|-------|----------------|
| GPT2 Fine-tuned | Title + Verdict | 17.67 | 0.304 | 0.09 |
| GPT2 Fine-tuned + Rake anchors | Title + Verdict + Rake keywords | 11.45 | 0.365 | 0.10 |
| GPT2 Fine-tuned + KeyBERT | Title + Verdict + KeyBERT N-grams | 12.17 | 0.380 | 0.12 |
| GPT - Neo | Title + Verdict | 5.40 | 0.194 | - |

Table 1: Experiment Results

KEY FINDINGS

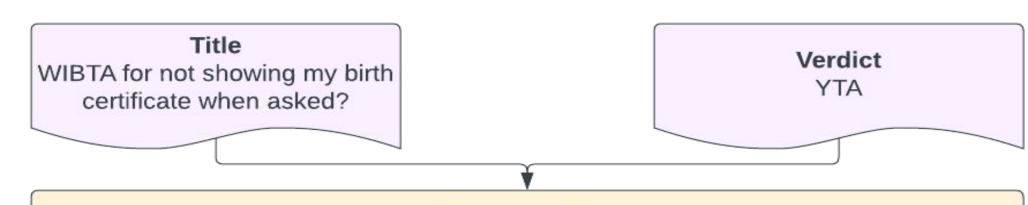
We compared the quality of the stories generated by the above models **GPT2 Fine-tuned**: Struggles with short context horizon and inter-story repetition. The perplexity score also reflects its inferiority **GPT2 Fine-tuned + RAKE anchors**: Rake keywords act as good anchors for the pre-trained model to maintain context. The scores reflect its superiority.

GPT2 Fine-tuned + KeyBERT: The n grams generated by KeyBERT are contextually important keyphrases and hence lead to comparable performance with the RAKE version.

Observations:

- I. As the maximum length of a story grows, the amount of **inter-story** repetition grows proportionally.
- 2. Despite fine-tuning and additional context from anchors, GPT2 based models do not outperform bigger pretrained models (such as GPT-3, GPT-Neo with Billions of parameters).

SAMPLE RESULT



Generated Text

As a Japanese living in the USA, my next door neighbors are constantly asking me to go back to the Japan over the summer. They also like to make fun of me for not having a Japanese passport. They brag to my family about how they learned English as a child but I our native language is still English. Even with my English I can't communicate with them very well, and my family often me they are too busy7me to listen to my side of the story.

One day I was on the train home from work, and the neighbors were talking with my neighbor and was pointing her phone at me7and talking about how jealous she is that I'm not a Japanese. 8She even screams out loud about how she's always touchy with me and how I should come and visit her, even though she knows I'm not a Japanese.

7lt's almost like they hate me because I don't have a Japanese passport.

I feel like their behavior is ridiculous, but at the same time, it is disrespectful to ask me to go back to Japanese Main society when my employer has me meet with them.

WIBTA if I told them I have nothing to do with Japan and even though I came here for tax reasons, I won't be showing them my birth certificate while they make unnecessary comments about it?

Fig3: Sample Text Generation

FUTURE WORK

- Better Common Sense Grounding: To achieve better common sense awareness in the generated story.
- Human Evaluation Metrics: Human evaluation metrics like triplet pairing task, where groups of 3 stories are presented to humans judges and are asked to pick the correct pairing.
- **Better Anchors**: To try more sophisticated key-phrase extraction / topic modeling.