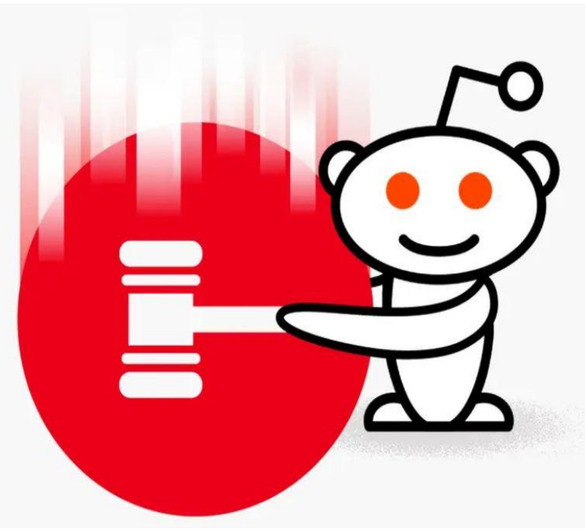
Neural Text Generation Conditioned on Title-Verdict Pair on Subreddit r/AITA Data

USC Viterbi School of Engineering

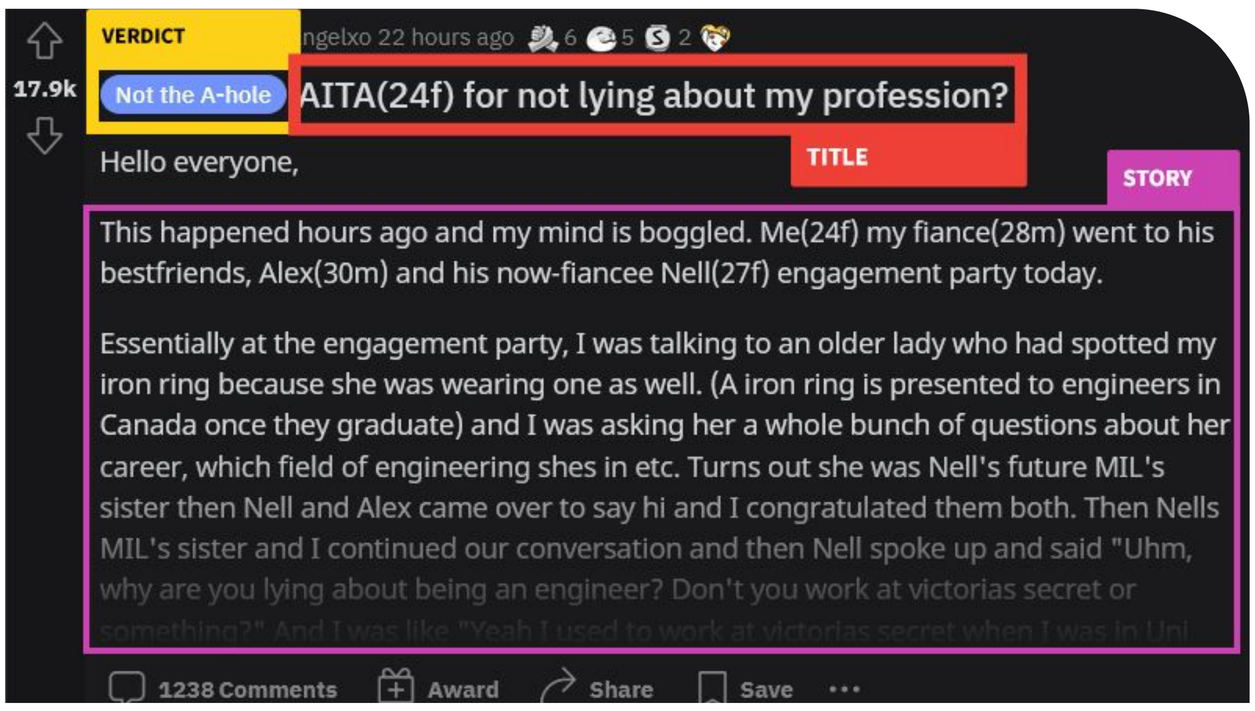
Team: Chomsky's Chimps

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



INTRODUCTION

1. 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

1. The raw dataset consists of over 120k posts made on the subreddit r/AITA between January 1, 2014 and January 1, 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: 37551 | 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

EXPERIMENT SETUP

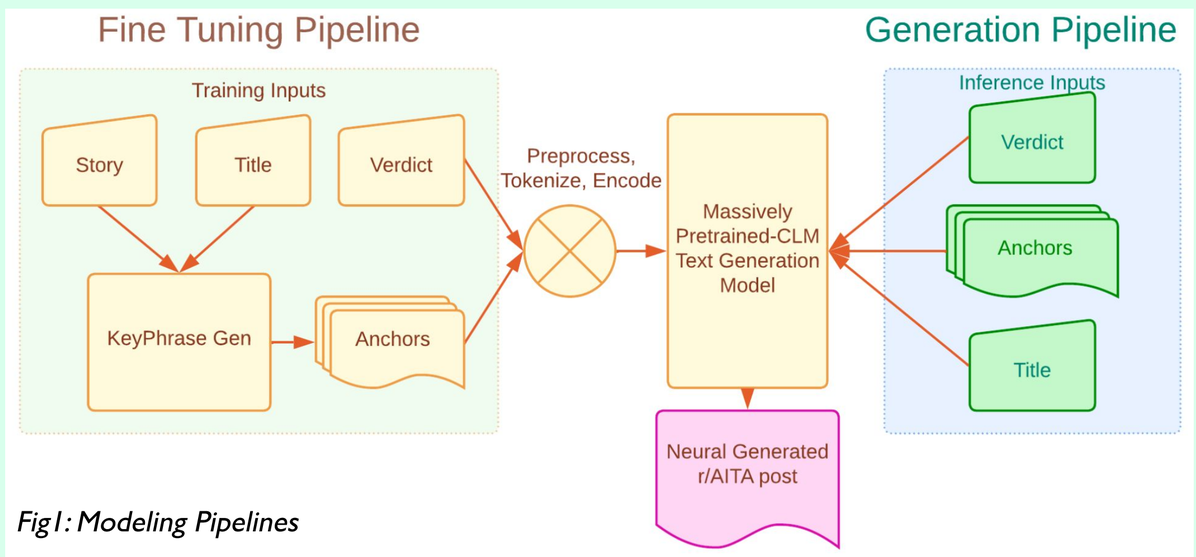


Fig1: Modeling Pipelines

1) 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
2) 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:

1. 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

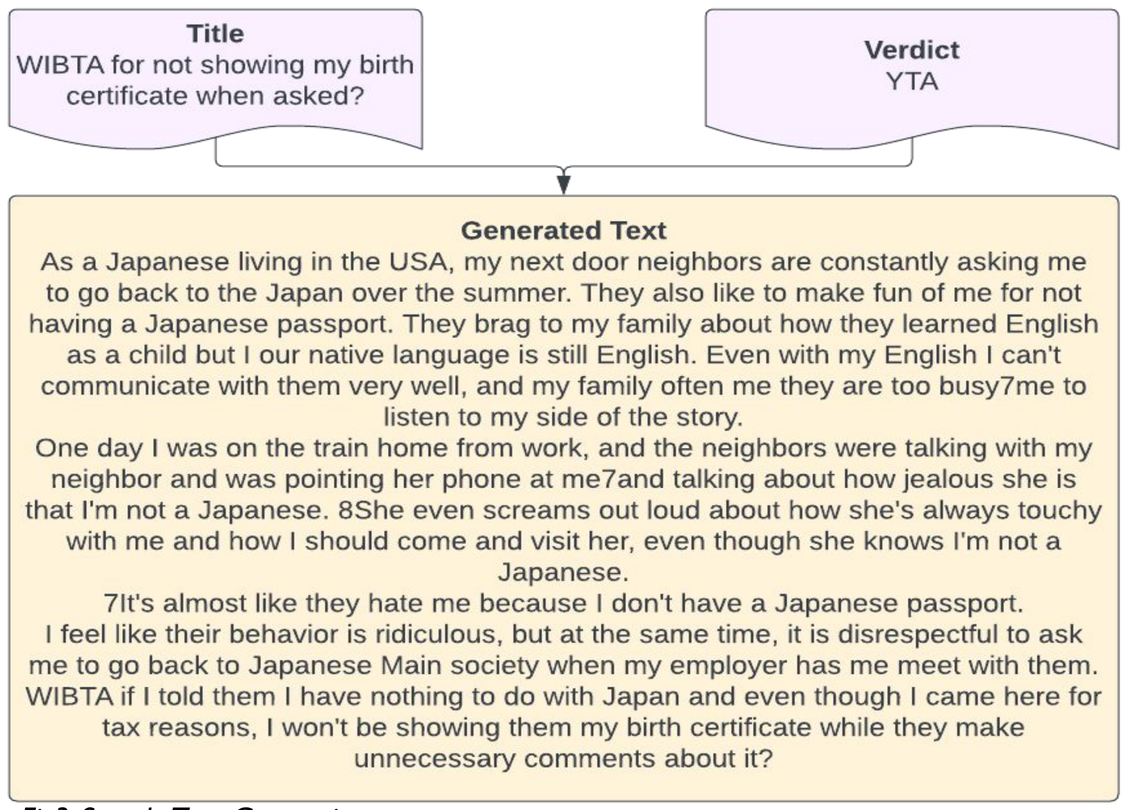


Fig3: Sample Text Generation

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.