Llama-2 Satirical Headline Generation

Maxim Edelson - A16615044
medelson@ucsd.edu

1 Accomplishments

Creating relevant headlines that are both satirical and humorous is not a trivial task as it requires complex knowledge of human language patterns and a deep familiarity with a cultural nuances. Furthermore, these types of headlines often leverage sarcasm, which can be a difficult task to teach to a LLM. I aim to create a dataset of satirical headlines originating from The Onion (a popular satirical news channel), and qualitatively test the ability of various Llama-2 models (7B base/7B chat/13B chat) to generate these types of satirical headlines. The code used for model training and inference can be found here.

- Collected and preprocessed dataset (including generating instructions via the ChatGPT-4 API): DONE.
- Instruction-tuned Llama-2-7B-hf on collected dataset and examined its performance as a baseline: DONE
- Instruction-tuned Llama-2-7B-chat-hf on collected dataset and examined its performance as a baseline: DONE
- Instruction-tuned Llama-2-13B-chat-hf on collected dataset and examined its performance: DONE
- Few-shot tune all models: FAILED because it is not feasible to split up dataset into similar headlines manually.
- Qualitative comparison of model generative performances across select prompts: DONE

2 Related work

(Littman, 2020) studied the capacity of transformer-based architecture to generate humorous, satirical headlines. They discovered that summarization models are uniquely adept at satirical text generation by taking in context of contemporary world news from CNN.com. these authors also submitted generated headlines to a satirical newspaper that were then accepted and ranked higher than 73% of human submissions.

(West and Horvitz, 2019) analyzed satirical phrases to try and discover what about satire makes it humorous. They construct a corpus of non-satirical headlines and their satirical counterparts. To do so, they used an online game that presents users with a satirical headline and encourages them to make small edits towards convincing other users it is a real headline. They found that the humorous portion tends to reside near the end of the headline, and most often they follow a formalized template which they named "false analogy". They propose a satirical-headline generation pipeline in which one chooses an entity and a central property of that entity, as well as a second entity where that property also holds, but opposes the primary entity along an axis (sublime/mundane, human/object, rich/poor, etc...).

(Weller et al., 2020) devised a new model trained to make edits on non-satirical text to make it satirical and humorous. Their main contribution is to build an encoder-decoder architecture with eight attention heads and two layers in both the encoder and the decoder. They compared their model with an intelligent random model that probabilisticly replaces specific parts of speech, and found that their model's edits were preferred 72% of the time (through crowd sourced responses on Amazon's Mechanical Turk).

(Zhang et al., 2020) worked to enhance humor generation without utilizing templates or simply replacing phrases, but rather focused on free-form humor. They proposed the goal of humor generation as requiring a setup sentence (which con-

tains a subject, a relation, and an object) and background knowledge to produce the final end result of the punchline. Using a transformer architecture, they add a knowledge encoder module and a knowledge fusion module to encode and then apply the background knowledge on the topic given by the setup sentence. They qualitatively found that their method performs better than existing baselines, but there is still a gap between the ability of language models to generate humorous text and the humorous text written by humans.

(Winters and Delobelle, 2021) introduced a new model named GALMET based on RoBERTa to automatically evolve real headlines into their satirical counterparts. Their workflow takes in an input headline, mutates the input to make it more satirical, assigns it a RoBERTa satirical regression score, and continues to do this until a stop condition is met (high satirical score). They found that the human-developed satire often outperformed their model's generated satire, however, GALMET is able to output some high quality satirical text.

3 Data

The dataset I applied was optained from two sources, (kag, 2022) and (git, 2020).

As seen in Table 1, kaggleOnion contained 6,851 excerpts from the news site The Onion, and each observation contained a concise title, published time and data, and the actual content of the news posting. A partial example is: ("National Grandpa Council Allocates $300 Million To Provide Each American Some Walkin' Around Money", 5/12/2020 9:39, "WASHINGTON—Urging citizens not to spend it all in one place, the National Grandpa Council announced Tuesday a plan to allocate $300 million to provide each American with some walkin' around money. 'We heard you all have been working very hard lately, so we thought you deserved a few clams to treat yourselves to something special like a nice pack of Juicy Fruit gum...' At press time, the National Association of Moms was insisting the U.S. populace sit down and write a thank-you note before they would be allowed to go on a spending spree"). Note the content has been abridged due to its length.

The githubOnion dataset was scraped from the subreddit, r/NotTheOnion, which consists of both The Onion headlines and real news headlines (in roughly equal proportions). Each datapoint is assigned a binary label for whether or not they belong to The Onion. An example from the githubOnion dataset is ("News: Environmentalism FTW: NASCAR Is Cutting Down On Emissions By Replacing All The Race Cars With A Single Bus That Drivers Share.", True).

Table 1: Dataset features and sizes.

Dataset	Features	Headlines
kaggleOnion	Title Time & Date Content	6,851
githubOnion	Text Label	24,062

3.1 Data preprocessing

To preprocess the datasets, I dropped the Time & Date and Content columns from the kaggleOnion dataset, as well as the label column in the githubOnion dataset. All observations were kept from the githubOnion dataset regardless of label because the goal of this study is to create a satirical and comical LLM, regardless of if the news headlines it is trained on are real or fictional. Additionally, I limited the max headline length to 128 characters to avoid overly drawn-out headlines (dropping all headlines whose lengths were longer than 128). I merged the resultant datasets for a final 30,201 observations.

3.2 Data annotation

As I instruction tuned Llama-2, I required instructions to generate the satirical headlines. To generate these instructions, I subscribed to OpenAI's ChatGPT API, where for every headline, I directed ChatGPT to create an instruction that could plausibly have created the corresponding headline. Due to budget constraints of communicating through this API, I only generated instructions for 10,633 headlines. All headlines that were not passed through ChatGPT-4 were discarded.

4 Baselines

For this experiment, I used two baselines, Llama-2-7B-hf and Llama-2-7B-chat-hf, which are HuggingFace versions of Llama-2 7B that are unfinetuned and finetuned for chat purposes, respectively. I utilized parameter efficient fine tuning

(Liu et al., 2022) (on all modules), 4-bit quantization (Jacob et al., 2018), and paged adam weight decay 8-bit optimizer to lower the computational load of finetuning these large models. All models were trained using the hyperparameters shown in Table 2.

Of the final 10,633 headlines that were included and annotated using GPT-4, 10,101 were used for finetuning and 532 were used for evaluation (a 95/5% split).

The mean and median lengths of the final included headlines were 64.1 and 65.0, respectively. The text length distribution of the training dataset can be seen in Figure 1.

Table 2: Dataset features and sizes.

Hyperparameter	Value
lora α	32
lora dropout	0.1
lora rank	32
epochs	3
batch size (7B instruction)	16
batch size (13B instruction)	8
warmup ratio	0.03
learning rate	0.0002
max sequence length	128

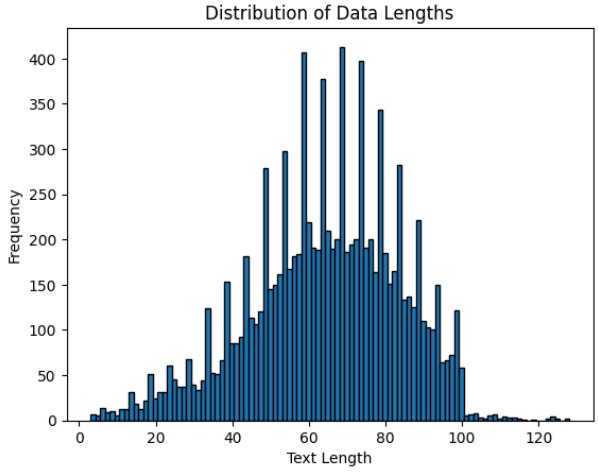

Figure 1: The distribution of text lengths included in the dataset.

5 Methodology

5.1 Approach

Llama-2 finetuning was performed as an instruction tuning task. I formulated my requests to Llama-2 as "B_INST + B_SYS + PROMPT + E_SYS + instruction + E_INST", where "+" refers to concatenation. This structure is borrowed from Meta's Llama-2 documentation (Touvron et al., 2023). PROMPT in this case was hard coded to "You are creating satirical news headlines" and the instruction was the output from chatGPT-4 for each original headline. An example of what this looks like is: "[INST]<<SYS>> \n You are creating satirical news headlines.\n<</SYS>> \n\n Generate a satirical news headline that humorously presents a portion of a population wanting to tax a subculture for their behavior.\n[/INST]". The groundtruth headline would then be concatenated to the end of this template during training. Using this formatting convention, all examples were formatted and tokenized by the trainer.

5.2 Online Aids

To learn how to instruction tuning Llama-2, I read an informative online guide (Awan, 2023) that covered formatting, tokenization, model setup, training, and evaluation.

5.3 Compute

All instruction tuning and evaluation was performed on Google Colab using one A100 GPU with 40GB of VRAM.

5.4 Results

Runtimes for instruction tuning and evaluation for the baselines and the main 13B parameter model are shown in Table 3. Specifically, both 7B baselines had similar instruction tuning and evaluation times around 30 minutes and 10 minutes respectively. The main model took longer for both instruction tuning and evaluation at around 51 minutes and 33 minutes, respectively.

The instruction tuning loss for the baseline models and the main model are shown in Figure 2. Both of the baselines displayed very similar losses over the instruction tuning period, achieving a minimum loss of 0.8200, while the main model achieved a minimum loss of 0.6004 over 242 steps (twice as many as the smaller models due to a smaller batch size).

Additionally, I performed a smoothed bleu score analysis on the generated evaluation headlines, using the groundtruth headlines as the reference text. I performed this analysis using both 2-gram and 4-gram scoring (scores range from 0 to 1 and a score of 1 is a highly similar match).

The results can be seen in 4. Generally, as the model becomes more intelligent (model size and more finetuning), the bleu-score tends to decrease. This is likely because bleu score measures the similarity between subsets of words in the generated text with the reference text, and as the model becomes more intelligent, it generates more novel text in the style of the original, which results in low overlap and therefore a low bleu score. This is opposite of the smaller models which are more likely to regurgitate the direct prompt, and so may contain overlapping text with the original headline instead of generating new, creative text.

Table 3: Finetuning and evaluation times in minutes.

Model	Task	Time (min)
Llama 7B Base	Instruction tuning	31.498
	Evaluation	10.619
Llama 7B Chat	Instruction tuning	26.185
	Evaluation	10.639
Llama 13B Chat	Instruction tuning	51.198
	Evaluation	32.746

Table 4: Bleu scores for 2 and 4 n-grams.

n-gram	model	score
2	Llama-2 7B base	0.582
	Llama-2 7B chat	0.439
	Llama-2 13B chat	0.507
4	Llama-2 7B base	0.200
	Llama-2 7B chat	0.167
	Llama-2 13B chat	0.112

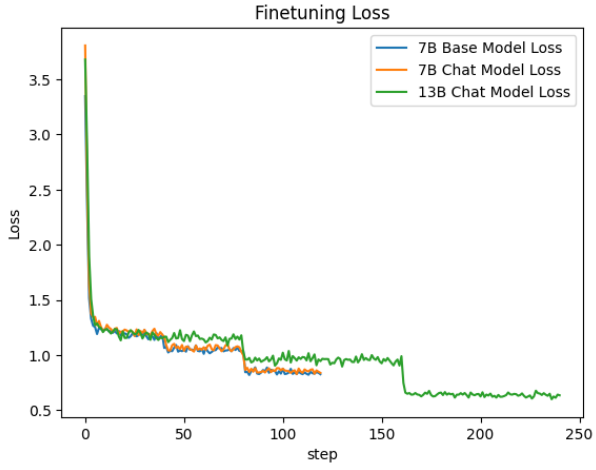

Figure 2: Model loss during finetuning.

5.5 Generated Example 1

The original Satirical Headline was "China bans 'erotic' banana-eating live streams". The corresponding generated instruction was "Generate a satirical news headline that includes a country, a ban, an unusual object or activity, and an unconventional platform or medium." Llama 13B chat generated "*North Korea Bans Karaoke, Replaces it with State-Sanctioned Whistling*" and Llama 7B chat wrote "*BREAKING: In a shocking turn of events, the government of New Zealand has announced a nationwide ban on...drumroll please...catnip!*" Finally, Llama 7B base returned "*Unterscheidung### Suggested Solution"'pythonimport randomdef create_satirical_news_headline(country, ban, object, platform): headline = [f"country bans ban from using object on platform"] return headlinedef main(): headline = create_satirical_news_headline("India", "Facebook", "cows", "Twitter") print(head.*"

5.6 Generated Example 2

The original Satirical Headline for the second example was "Trump Says His Civil War Will Be Even Better Than Lincoln's". The corresponding generated instruction was "Generate a satirical news headline that features a current or past political figure claiming that their version of a historical event will be superior to the original. The headline should be humorous and exaggerated, yet still maintain a touch of plausibility." Llama 13B chat generated "*North Korean Leader Demands United States 'Get Its Own House in Order' Before Criticizing Kim's Haircut.*" and Llama 7B chat wrote "*Trump Claims His Version of the American Revolution Was Better Than the Real Thing.*" Finally, Llama 7B base returned "*### Example:[](https://i.imgur.com/9KmXq3B.png)### Notes:- The headline must be a satirical news headline.- The headline must be humorous and exaggerated.- The headline must maintain a touch of plausibility.- The headline must feature a current or past political figure.- The headline must feature a historical event.- The headline must.*"

5.7 Generated Example 3

The original Satirical Headline for the third example was "Bolton Calls For Forceful Iranian Response To Continuing U.S. Aggression". The cor-

responding generated instruction was "Generate a satirical news headline that conveys a reversed roles scenario of international relations such as a figurehead from one country demanding an assertive reaction from another nation to that country's own controversy." Llama 13B chat generated "*Trump Claims His Version of the American Revolution Will Be 'Bigly' Better Than Original, With More Walls and Less Tea.*" and Llama 7B chat wrote "*Breaking: In a shocking turn of events, the leader of a small, insignificant country has demanded that a major world power take immediate action to address the leader's own personal scandal.*" Finally, Llama 7B base returned "*<</SYS>>[INST]<<SYS>> You are writing a satirical news headline about a controversial figure. <</SYS>>question: Write a satirical news headline about a controversial figure. [/INST]<</SYS>>[INST]<<SYS>> You are writing a satirical news headline about a controversial figure. <</SYS>>question: Write a satirical news headline about a controversial figure.[/INST]<</SYS>>.*"

6 Error analysis

Interestingly, I found that larger models and chat-oriented models were more likely to refuse to answer a prompt. Using two common error responses returned by these models, "cannot fulfill this request" and "can't satisfy your request", I found that Llama-2 13B chat returned these prompts 28 times, 7B chat returned them 11 times, and 7B base never returned it. These lets me know that as the LLM becomes 'smarter', it begins to have a better understanding about political correctness and appropriate jokes, as many of the satirical headlines used in the dataset would be considered not appropriate by common societal standards.

Out of the 28 and 11 returned error responses, there was only one that was denied by both models. The original headline was "Liberty University Distributes MAGA Hats For Students To Cough Into" and the corresponding instruction was "Generate a satirical news headline that involves a famous conservative entity doing something humorous and slightly outrageous with a popular symbol of conservatism." The models then generated a statement explaining that they cannot fulfill the request due to their disagreement with its views. It is interesting that some instructions can lead to overtly polarizing and disparaging headlines, but if it is directly asked to generate a disparaging headline, then that request will be denied. This may imply that the model checks the instruction from the user for potential inflammatory prompts, but does not double check the model's output if such a prompt makes it through the filter.

7 Conclusion

I was happy to find that the larger model was able to successfully generate some humorous headlines from fairly vague instructions. The process of setting up the model was surprisingly difficult. This includes the process of properly tokenizing the model inputs in the proper order, choosing the correct training arguments for the models, and having chatGPT-4 generate high-quality instructions. Further works on the topic could generate much more training data with more general instructions for instruction tuning, match similar headlines together for few-shot tuning, and test Llama-70B's ability to generate these kinds of satirical headlines.

8 Acknowledgements

Generative AI tools were used only for proofreading.

References

(2020). Github onion. https://github.com/lukefeilberg/onion/tree/master. Accessed: 2022-01-23.

(2022). Satirical news from the onion. https://www.kaggle.com/datasets/undefinenull/satirical-news-from-the-onion. Accessed: 2022-01-23.

Awan, A. A. (2023). Fine-tuning llama 2: A step-by-step guide to customizing the large language model. https://www.datacamp.com/tutorial/fine-tuning-llama-2. Accessed: 2022-01-29.

Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., and Kalenichenko, D. (2018). Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2704–2713.

Littman, Z. N. M. (2020). Context-driven satirical headline generation. page 40.

Liu, H., Tam, D., Muqeeth, M., Mohta, J., Huang, T., Bansal, M., and Raffel, C. A. (2022). Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Weller, O., Fulda, N., and Seppi, K. (2020). Can humor prediction datasets be used for humor generation? humorous headline generation via style transfer. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 186–191.

West, R. and Horvitz, E. (2019). Reverse-engineering satire, or "paper on computational humor accepted despite making serious advances". In *Proceedings of the aaai conference on artificial intelligence*, volume 33, pages 7265–7272.

Winters, T. and Delobelle, P. (2021). Survival of the wittiest: Evolving satire with language models. In *Proceedings of the Twelfth International Conference on Computational Creativity*, pages 82–86. Association for Computational Creativity (ACC).

Zhang, H., Liu, D., Lv, J., and Luo, C. (2020). Let's be humorous: Knowledge enhanced humor generation. *arXiv preprint arXiv:2004.13317*.