Small But Funny: A Feedback-Driven Approach to Humor Distillation

Anonymous ACL submission

Abstract

The emergence of Large Language Models (LLMs) has brought to light promising language generation capabilities, particularly in performing tasks like complex reasoning and creative writing. Consequently, distillation through imitation of teacher responses has emerged as a popular technique to transfer knowledge from LLMs to more accessible, Small Language Models (SLMs). While this works well for simpler tasks, there is a substantial performance gap on tasks requiring intricate language comprehension and creativity, such as humor generation. We hypothesize that this gap may stem from the fact that creative tasks might be hard to learn by imitation alone and explore whether an approach, involving supplementary guidance from the teacher could yield higher performance. To address this, we study the effect of assigning a dual role to the LLM – as a "teacher" generating data, as well as a "critic" evaluating the student's performance. Our experiments on humor generation reveal that the incorporation of feedback significantly narrows the performance gap between SLMs and their larger counterparts compared to merely relying on imitation. As a result, our research highlights the potential of using feedback as an additional dimension to data when transferring complex language abilities via distillation.

1 Introduction

NLP is on a trajectory towards creating increasingly large models (OpenAI, 2023; Touvron et al., 2023). LLMs achieve high performance across many tasks in both zero and few-shot settings. However, there are growing concerns about the computational efficiency and environmental sustainability of such approaches (Strubell et al., 2019).

Knowledge distillation (Hinton et al., 2015b) has thus gained a renewed interest, where the term has evolved to denote the process of distilling the

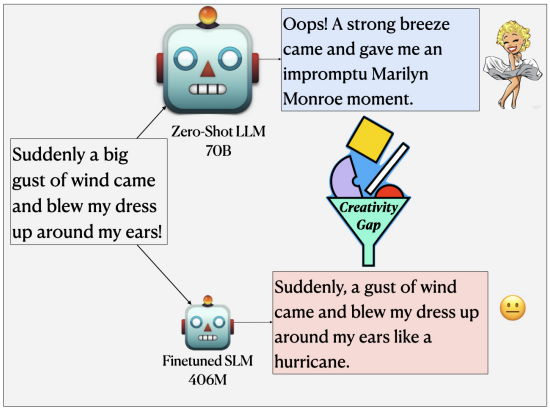

Figure 1: Performance gap between LLMs and SLMs: Generations from a teacher LLM (Llama2) and a student SLM (BART) finetuned on its outputs.

responses of LLMs to SLMs (West et al., 2022). Recent work has explored the distillation of commonsense knowledge (Bhagavatula et al., 2023), chain-of-thought reasoning (Li et al., 2023a), and summarization abilities (Liu and Chen, 2022; Jung et al., 2023) to smaller language models. However, the exploration of distilling creative abilities such as humor into SLMs remains an open and challenging area of research.

In this work, we explore the application of knowledge distillation in the context of the creative style-transfer task of *conditional humor generation*. Given a literal text, the goal is to generate a humorous meaning-preserving paraphrase, as shown in Figure 1. While LLMs do not match the subtleties of humor in human-written text (Hessel et al., 2023; Chakrabarty et al., 2023), they surpass their smaller counterparts (Radford et al., 2019; Chen et al., 2023; Hessel et al., 2023).

We argue that such creative tasks are challenging for SLMs to learn. First, due to the inherent constraints of SLMs, such as reduced model capacity, they are limited in their ability to explore diverse solution spaces and generate innovative outputs.

Second, although imitating a teacher model is a good starting point, it may result in superficial overfitting to the teacher's style rather than learning the task itself (Gudibande et al., 2023). Fig. 1 demonstrates that even after fine-tuning on the humorous responses from the LLM, the SLM outputs fall flat. Existing techniques that address this gap attempt to improve task understanding through distilling chains of thoughts (Mukherjee et al., 2023; Wang et al., 2023a). This approach is less applicable for creative generation tasks, which can't be scripted. For instance, it is possible to solve a math reasoning problem systematically, but it is difficult to come up with a recipe for a joke (Hessel et al., 2023; West et al., 2023).

To bridge this gap, and improve upon mimicry of the teacher, we propose a novel distillation framework involving both imitation and feedback. Following the typical imitation stage in which the SLM learns from the LLM's outputs, we use the LLM as a critic to provide feedback on the student's outputs, facilitating iterative improvement.

Evaluation of our student models on the proposed task based on EmpatheticDialogues (Rashkin et al., 2019) and Samsum (Gliwa et al., 2019) datasets confirms the advantage of learning from feedback; our student model, based on the small BART model, performs on par with LLMs that are orders of magnitude larger, such as Llama2-70B upto 65% of the time, and significantly outperform supervised fine-tuning by a margin of 18-20% . We assess the strengths and limitations of the critic in evaluating the SLM by comparing it against human judgments. We found that our critics can match human judgements with up to 76% accuracy, but can also suffer from biases due to length, position or other biases. We explore the effect of data size, frequency of critic intervention, and the effect of potential evaluation biases on narrowing the gap between the SLM and LLM.

Our work on distilling humor is a step towards more natural and engaging conversations (Ritchie et al., 2007), making SLMs more appealing for downstream applications where latency and computational efficiency need to be prioritized.[1]

[1]We will make the code available upon publication.

2 Related Work

2.1 Computational Humor Generation

Computational humor is an interdisciplinary field at the intersection of NLP and humor theory. Early efforts in computational humor revolved around rule-based systems and linguistically-motivated methods. Raskin (1979) introduced a semantic analysis of humor which laid the foundation for subsequent rule-based approaches to humor recognition (Mihalcea and Strapparava, 2005; Reyes et al., 2012; Chen and Soo, 2018; Weller and Seppi, 2019). Concerning the more challenging task of humor generation, early approaches were linguistically informed and focused on specific types of humor, *e.g.* puns or jokes (Ritchie, 2005; Petrović and Matthews, 2013).

With the advent of deep learning, the focus shifted towards data-driven and neural approaches. This ranges from using RNNs and GANs to create puns (Yu et al., 2018; Luo et al., 2019), to more recent transformer-based approaches (Garimella et al., 2020) that can complete or generate jokes. These approaches enabled the generation of more contextually relevant and natural humor. Since neural approaches are primarily data-driven, this concurrently led to the creation of large joke datasets (Weller et al., 2020) and benchmarks (Hossain et al., 2020).

Recent research has delved into the generation of figurative language such as sarcasm (Chakrabarty et al., 2022), puns (Mittal et al., 2022), as well as interactive chatbots with humor capabilities (Kulshreshtha et al., 2020). New multi-modal humor benchmarks (Hessel et al., 2023) have been developed to gauge the humor understanding and explanation abilities of LLMs. Weller et al. (2020) proposed the first model for humor style transfer building a transformer model that translates from regular to humorous, and leveraging humor prediction data from news headlines for humor generation.

2.2 LLMs as Evaluators

Automatic evaluation of Natural Language Generation (NLG) tasks is typically based on N-gram metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004) and embedding-based metrics such as BERTScore (Zhang et al., 2020) which require gold standard references. Recent research has explored a reference-free approach to assess NLG

tasks by leveraging the implicit knowledge and instruction following abilities of LLMs. Fu et al. (2023) proposed GPTScore, which prompts LLMs with instructions and aspect definitions (*e.g.* fluency or coherence). The score is computed by calculating the conditional probability of generating the target text. GEval (Liu et al., 2023a) is an alternative metric, that presents the instructions in a form-filling paradigm and uses the probabilities of output tokens from LLMs to normalize the scores (*e.g.* score between 1 and 5) resulting in finer-grained continuous scores. In contrast to GPTScore and GEval, LLM-Eval (Lin and Chen, 2023) uses a single prompt to evaluate multiple evaluation aspects, thus minimizing the calls to the LLM. Additionally, LLMs have also been used to assess the factuality of generated text (Zha et al., 2023). Mehri and Shwartz (2023) propose a learned evaluation metric via instruction tuning.

2.3 Imitation Learning from LLMs

The emergence of LLMs has caused a paradigm shift in NLP from traditional knowledge distillation (Hinton et al., 2015a) to an imitation-based data or response distillation. These approaches use LLMs as training data generators and train a smaller language model on this data (West et al., 2022). With subsequent work demonstrating that this may be a sparse form of distillation, leading to the student mimicking the style of the teacher, but not the reasoning abilities (Mukherjee et al., 2023; Gudibande et al., 2023), further extensions have been developed to distill a complete "Chain-of-Thought" (Wei et al., 2023) from the teacher model (Li et al., 2023a; Wang et al., 2023a; Shridhar et al., 2023; Magister et al., 2023; Hsieh et al., 2023), improving the performance of smaller language models.

2.4 LLM-based Alignment

A growing body of research leverages feedback from a language model to iteratively enhance its performance. This feedback may encompass written comments, numerical scoring, rankings, or explanations. Humpback (Li et al., 2023b) aims to construct a better instruction-tuning dataset through an iterative self-training algorithm. Self-Refine (Madaan et al., 2023) and REFINER (Paul et al., 2023) explore LLMs to engage in self-reflection, providing feedback and encouraging the use of this feedback to enhance their responses. I2D2 (Bhagavatula et al., 2023) demonstrates that small language models may improve if trained on their refined outputs generated using constrained decoding and filtered with a simple supervised critic model. Our work is inspired by BRIO (Liu et al., 2022, 2023b) for summarization, which reuses the generation model as the evaluation model to rank the candidates with contrastive learning. However, we avoid a ranking-based approach and use pairwise feedback due to the challenges in mitigating positional and length biases among multiple generations. Concurrent to our work, Zephyr (Tunstall et al., 2023) uses preference learning to align a student model to user intent by using preferences derived from different candidate models from a large teacher.

3 Method

Our proposed knowledge distillation framework is depicted in Fig. 2. Assuming that the teacher outperforms the student in the humor generation task, our objective is to bring the student's performance closer to that of the teacher. We attempt to achieve this in two phases. First, in the imitation phase (Sec 3.1), the student undergoes finetuning using a set of humorous outputs $O = O_1, O_2, ..., O_n$ generated by the teacher for a given input I. In the critique phase (Sec 3.2), the finetuned student generates candidate output pairs $P = P_1, P_2$, which are then evaluated by a critic model. The scores obtained from the critic model are employed to train the student through a feedback-incorporation method.

3.1 Imitation Phase

During the imitation phase, we construct a dataset consisting of < literal, humorous > pairs and use it to directly train the student. As depicted in Fig. 2, the process begins with a literal text input (I) and prompting the teacher to generate $N = 3$ humorous paraphrases (O) of the input in a zero-shot setting. The utilization of N outputs encourages diversity in student outputs, as it demonstrates multiple possible humorous paraphrases for a given input. The student is then finetuned on the constructed dataset by minimizing the cross entropy loss between the reference and predicted outputs.

3.2 Critique Phase

In the critique phase, we aim to improve the student's task comprehension by teaching the student to differentiate between effective and ineffective

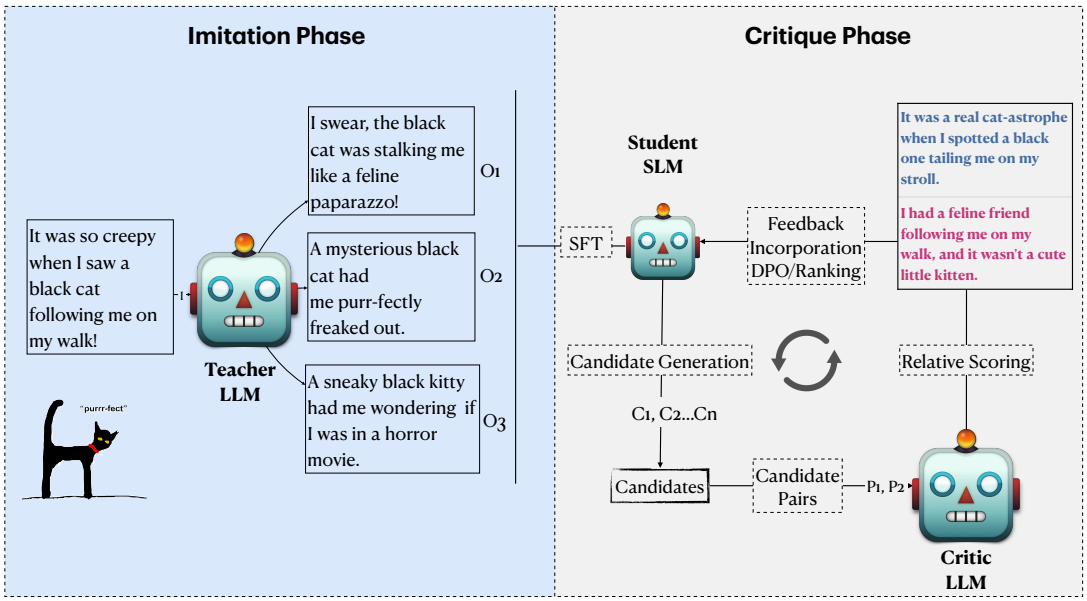

Figure 2: The proposed knowledge distillation framework: We perform task-specific distillation from a large, general language model, in two phases: an initial imitation phase, followed by a critical feedback phase which controls the quality of the generated humorous outputs from the student.

humor.

Critic The critic, an LLM (Large Language Model), evaluates the humor quality of outputs from the finetuned student. Drawing from prior work on computational humor (Valitutti, 2011) and LLM-based scoring (Liu et al., 2023a,b), we adopt a pairwise relative scoring approach to mitigate subjectivity as opposed to providing an absolute score on a humorous output. To obtain the preferred output from the pair, we use Multiple Choice Prompting (MCP) (Robinson and Wingate, 2023), which, presents both paraphrases simultaneously and asks the LLM to pick the better humorous output. For two humorous paraphrases, P1 and P2, of the same input text I, we present them as candidates labeled with symbols (e.g., "1" and "2"). The critic LLM predicts a token ("1" or "2"), with associated probabilities indicating preference, which we denote as the Win Tie Rate (WTR) of P1 against P2. [2]

Feedback Incorporation Subsequently, we leverage feedback from the pairwise scorer to refine the student model. Starting with the finetuned student (Fig. 2), we generate a set of $k = 6$ candidates using either diverse beam search or nucleus sampling. From this candidate pool, we select a pair of diverse humorous paraphrases, denoted as $P1$ and $P2$, where diversity is measured by a maximum pairwise n-gram-based edit distance score. The critic then scores these pairs, resulting in candidates categorized as either positive or negative based on their performance. Our objective here is to improve the finetuned student by discerning which output is preferred. To integrate this feedback, we explore the following feedback objectives.

1. *DPO* (Rafailov et al., 2023) provides an alternative to RLHF, aiming to align language models using human feedback without training an explicit reward model. When comparing two humorous paraphrases P_i and P_j, if P_i receives a higher quality score from the critic, DPO increases the likelihood of its completion over P_j. It employs the Bradley-Terry reward model, approximating reward modeling with a sigmoid loss function,

$$\mathcal{L}_{\text{DPO}}\left(\pi_\theta; \pi_{\text{ref}}\right) = - \mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta\left(y_w \mid x\right)}{\pi_{\text{ref}}\left(y_w \mid x\right)} - \beta\log\frac{\pi_\theta\left(y_l \mid x\right)}{\pi_{\text{ref}}\left(y_l \mid x\right)}\right.\right.$$

where, π_θ represents the ratio of chosen to rejected scores from the fine-tuned LM, while π_{ref} signifies the same ratio for an exact frozen copy of the model. The first term in the sigmoid denotes the shift in the preferred completion, while the second term indicates the shift in the dispreferred completion.

[2] The prompts used for WTR are shown in Appendix B

2. *BRIO* (Liu et al., 2022), a sequence-level contrastive (ranking) objective proposed for abstractive summarization. For a given input I, when faced with two candidate humorous paraphrases, P_i and P_j, and P_i attains a higher quality score from the critic, the student is guided to assign a probability to P_i that exceeds twice that of P_j.

$$\hat{\mathcal{L}}_{ctr}(\theta) = \sum_{S_i, S_j \in \mathcal{S}_c, i<j} \max\left(0, \bar{p}_s\left(S_j \mid D; \theta\right)\right.$$
$$\left. -\bar{p}_s\left(S_i \mid D; \theta\right) + \frac{1}{\lambda}\log 2(j-i)\right)$$

where λ corresponds to the average output length, and $\bar{p}_s$ corresponds to the length normalized log probabilities of paraphrases i and j. Following Liu et al. (2022), we combine the cross-entropy loss used in the imitation phase with the margin loss as a multi-task loss, to maintain the generation abilities of the student.

3. *BRIO-DPO*: We also experiment with combining both ranking and preference learning - for this variant, we obtain the contrastive pairs for the BRIO training from the teacher instead of the student, by letting the teacher rank two of its own responses. In this case, the BRIO loss is considered as supervision rather than feedback. The preference pairs for DPO still come from the student. To achieve this, we first train the student on BRIO, and use this model to initialize the DPO training.

We experiment with both one-shot and iterative feedback as shown in Figure 2, where the student may receive feedback from the teacher more than once.

4 Experimental Setup

Dataset. Similar to FLUTE (Chakrabarty et al., 2022), we pick literal input sentences from the EmpatheticDialogues dataset (Rashkin et al., 2019). Each conversation is obtained by pairing a speaker and a listener, where the speaker talks emotionally about personal matters, and the listener infers the underlying emotion and responds empathetically. We sample 12,000 sentences from the training set and generate $N = 3$ responses from the teacher to create 36,000 literal-humorous text pairs. We sampled 1,000 sentences from the validation and test sets for evaluating the student and 100 samples from the test set for human evaluation. For out-of-distribution (OOD) evaluation, we sample 500 literal inputs from Samsum (Gliwa et al., 2019).

Teacher Model. We use the 70B chat version of Llama 2 (Touvron et al., 2023) as the teacher model. We generate responses from the teacher using a temperature of 0.8 using nucleus sampling with top_p = 1.

Student Model. For student model, we focus on BART-large (Lewis et al., 2020). All the BART student responses are obtained using beam search with the number of beams set to 5 during generation.

Metrics. Traditional generation metrics such as ROUGE (Lin, 2004), may not work well for the creative task of generating humor. Inspired by prior work (Liu et al., 2023a; Fu et al., 2023), we use the LLM-based metric of **Win Tie Rate (WTR)** also described in Sec 3.2 for automatic evaluation. WTR aims to compare a pair of paraphrases and measures whether one paraphrase is equally (tie) or more (win) creative/humorous than the other, while preserving the meaning of the original input (consistency). In this work, we are more interested in humor than consistency - humor is essential, consistency can be subjective and optional based on the type of humor used (e.g. sarcasm). To make it more straightforward for human evaluation, we ask humans to provide individual WTR for the humor WTR separately from consistency WTR. We compute the automatic WTR of the student models using both the critic model based on Llama2-70B and GPT4 (OpenAI, 2023).

Length and Positional Biases The critic is prone to two major biases - Position Bias, where evaluation changes based on the encoding order of the responses and Length Bias, the tendency to favor longer responses. To mitigate length bias, during the feedback training phase, we incorporate only a subset of the candidate training pairs, filtered to maintain near-equal length between the two candidates (with a length ratio falling within the range of 0.8 to 1.2). This selection is intended to counteract the length bias resulting from the teacher's inclination towards longer outputs which has also been observed in prior work regarding AI (Saha et al., 2023a) as well as human feedback(Shen et al., 2023). To mitigate positional bias, we average all our win rates by letting the paraphrases take both position 1 and position 2.

Baselines. We finetune BART-large for 10 epochs on the teacher data using the Maximum Likelihood Estimation (MLE) objective (Sec. 3.1).

This baseline is denoted as BART-FT-WS (WS stands for Warm Start). We further refine this baseline to create students trained without and with feedback losses. For the first class of student models, we compare against,

- **BART-FT**: BART-FT-WS further finetuned using the same MLE objective.
- **BART-SD**: Indicating Self-Distillation (SD), BART-FT-WS is trained on only the positive candidates obtained from the critic(Sec. 3.2).

For the proposed feedback-based baselines,

- **BART-BRIO**: BART-FT-WS finetuned using the BRIO ranking objective for 10 epochs.
- **BART-DPO**: BART-FT-WS finetuned using the DPO objective for 10 epochs.
- **BART-BRIO-DPO**: BART-FT-WS finetuned using BRIO for 10 epochs, followed by DPO for an additional 10 epochs.

5 Experiments

Our overall goal is to investigate whether a creative generation ability such as humor can be taught through a combination of examples from the teacher, and feedback on the student responses. To answer this question, we focus our experiments on answering the following research questions:

① **RQ1**: How does the teacher perform as a critic?

② **RQ2**: How do students trained with and without feedback compare?

③ **RQ3**: How frequently should the teacher intervene?

④ **RQ4**: How does data size affect student performance?

⑤ **RQ5**: Does the length bias of the critic affect student responses?

5.1 RQ1: How does the teacher perform as a critic?

In order to validate the role of the teacher as a critic model, we conduct blind human evaluation by some of the authors ("annotators"). In Table 1, we analyze the alignment between human and LLM evaluation and compare two versions of the critic. Cloze prompting scores each paraphrase separately based on the perplexity of the template "The funny paraphrase of I is P_i" (for $i = 1, 2$). Conversely, Multiple Choice Prompting (MCP) presents the two paraphrases to the model and instructs it to choose the better one.

We ask the annotators to perform pairwise scoring of 100 blind pairs of humorous outputs. Annotators are asked to rank the paraphrases based on how humorous they are and how consistent they are with respect to the meaning of the input. We compare overall (WTR) and individual win tie rates for humor (WTR-H) and consistency (WTR-C) of the humorous outputs with the input using the critic model (Llama2-70B). We use the following metrics inspired by Saha et al. (2023b) to evaluate the two scoring methods:

AgH and AgC. We measure the LLM-Human Agreement (Ag) which is a score between [0, 1] indicating the percent of pairs that the annotator and the LLM agreed on. We compute agreement by independently matching each human judgment for each pair with the model judgment.

Positional Bias (PB). We measure the fraction of samples where the critic's scoring changes based on the order in which the paraphrases are presented to it.

Length Bias (LB). LB aims to measure the model tendency to favor longer responses when the human did not. We measure length bias as the fraction of samples where humans prefer the shorter response, but the critic model prefers the longer response.

We observe that overall, MCP performs significantly better in terms of human agreement when compared to the cloze style evaluation, with an agreement on humor up to 76% and consistency of up to 65% across both forward and backward positions. Concerning length bias, both methods exist length bias in about 20-25% of the cases where they tend to prefer longer outputs than the humans. Although sub-metrics based on Humor (WTR-H) and Consistency (WTR-C) look promising, we found that combining them using simple (MEAN, AND) or more complex methods like BSM (Saha et al., 2023b) results in amplifying positional biases to a huge extent for this task. Hence, we use the overall WTR as the automatic metric to evaluate model responses both during training and evaluation.

Arbiter	Type	Ag-H↑	Ag-C↑	PB↓	LB↓
Cloze	WTR-H	57	-	0	18
Cloze	WTR-C	-	46	0	21
Cloze	WTR	57	47	0	25
MCP	WTR-H	76	-	18	17
MCP	WTR-C	-	65	28	19
MCP	**WTR**	**76**	**59**	**15**	**20**

Table 1: Evaluation of Human Agreement with different Arbiters - Agreement with Humor wins (Ag-H), Agreement with Consistency wins (Ag-C) and Positional Bias (PB), Length Bias (LB), Scores are multiplied by 100.

Student	WTR_{GPT4}↑	WTR_{llama2}↑
BART-FT	30	28
BART-SD	35	36
BART-BRIO	48	53
BART-DPO	52	60
BART-BRIO-DPO	56	65

Table 2: LLM-based Evaluation of student models - Win Tie Rate(WTR) is measured against the teacher (Llama2-70B) by the critic llama2 (WR_{llama2}) or external critic GPT-4 (WR_{GPT4}) using the method described in Sec 3.2. WTR is scaled by 100.

Student	WTR_{llama2}↑
BART-FT	34
BART-BRIO	42
BART-DPO	47
BART-BRIO-DPO	49

Table 3: LLM-based Evaluation of student models on OOD test set (500 examples from (Gliwa et al., 2019)

Model	# Data	Frequency	WTR_{llama2}↑
BART-FT	36K	-	30
BART-FT	24K	-	26
BART-FT	12K	-	24
BART-BRIO	36K	1/10 epochs	53
BART-BRIO	24K	1/10 epochs	45
BART-BRIO	12K	1/10 epochs	43
BART-BRIO	12K	2/10 epochs	56
BART-BRIO	12K	10/10 epochs	66

Table 4: Effect of data size and frequency of feedback on win rate of student models against the teacher.

5.2 RQ2:How do students trained with and without feedback compare?

Table 2 shows the results of the automatic evaluation for the proposed student models using the Win Tie Rate (WTR) metric. To gauge the student's performance against the teacher that provided the training data, we compare the student outputs to the teacher's references. Hence, we measure the WTR of *the student vs. the teacher*.

When employing BRIO for ranking feedback, BART-BRIO surpasses the imitation-based students (BART-FT and BART-SD) with notable increases in performance—13-18% and 19-25% based on evaluations by both GPT-4 and Llama2-70B, respectively. This suggests that in approximately 50% of cases, the student achieves comparable or superior performance to the teacher. When incorporating the same critic feedback as a preference learning objective through DPO (BART-DPO), we observe a similar trend of performance enhancement in the student, with the DPO baseline outperforming BRIO by a modest margin of 4-6%. Subsequently, we explore a combined approach utilizing both BRIO and DPO objectives, resulting in a further improvement of 4-6% over using DPO or BRIO independently.

OOD Test Set We show the performance of the BART-based student models on the OOD test set in Table 3. We can observe a similar trend of student performance improving with both BRIO and DPO feedback objectives. However, the BRIO-DPO variant results in a very minor performance boost indicating that further investigation is needed in combining ranking and preference learning objectives.

5.3 RQ3: How frequently should the teacher intervene?

To test whether the student can benefit from an increased frequency of feedback, we compare different versions in which we seek feedback from the teacher iteratively after every K epoch which varies from 1, 5, and 10 (static). Table 4 shows that more frequent feedback has a positive impact on student performance, but the performance gains may saturate as feedback frequency is increased. There is also a trade-off between the communication cost with the teacher, which needs to be considered when providing more frequent feedback. This motivates future work in investigating this trade-off further, and approaches such as curriculum learning or active learning could be leveraged to choose which samples should receive feedback.

5.4 RQ4: How does data size affect student performance?

In this experiment, we assess the influence of the amount of data that is used to supervise the student. We examine the relationship between data

Input: I am a stay at home mom.	
BART-FT	**BART-DPO**
I'm a stay-at-home mom, and I'm loving every minute of it!	I'm a stay-at-home superhero, saving the world one diaper change at a time.

(a)

Input: i felt bad about sleeping in today.	
BART-FT	**BART-BRIO**
I was feeling a bit guilty about my morning snooze-fest.	I woke up feeling like a sloth who's been snoozed out for so long, I'm pretty sure I've heard the sound of snores echoing off the walls.

(b)

Input:My son failed a really important test.	
BART-FT	**BART-BRIO-DPO**
My son's test score was lower than a Kardashian's Instagram followers.	My son's test results were so bad, I'm pretty sure he's been secretly practicing his 'I'm-not-a-complete-failure' face for weeks.

(c)

Input:My wife and I are going to buy are very first brand new car this week	
BART-DPO	**BART-DPO-ITER2**
We're finally upgrading our ride from a rusty old clunker to a sleek, sexy beast of a car.	We're upgrading our transportation game from "legs" to "wheels" this week!

(d)

Table 5: Qualitative Examples showcasing variants trained with different types of feedback

size and performance by varying data size between 4,000, 8,000, and 12,000 literal sentences, in all cases providing 3 humorous paraphrases per example. Table 4 shows that simply increasing the data size may be insufficient to improving the performance. The performance is lower than that of the more sophisticated teaching approach, specifically incorporating feedback, which proved to be more beneficial and data-efficient. We can observe that, the model with feedback trained on 12K examples still performed better than a model trained without feedback with triple the number of examples.

5.5 RQ5: Does length bias of the critic affect student responses?

The length bias in the teacher is propagated to the student, especially when trained over longer periods and more so with the BRIO or ranking objective. Without any mitigation strategies, the length of the student outputs can almost double compared to the input text. As described in our Sec 3, we also filter pairs that are significantly longer than one another during training to prevent the student from learning to optimize for length. This is similar to the length correlation observed in human feedback (Singhal et al., 2023). Addressing this issue requires addressing length bias in Multiple Choice Prompting in LLMs, which is an evolving research area (Wang et al., 2023b; Shen et al., 2023).

6 Qualitative Analysis

Figure 5 highlights instances where each of the feedback objectives (BART-DPO, BART-BRIO, and BART-BRIO-DPO) outperforms the BART-FT baseline trained solely through imitation learning. In Figure 5a, BART-DPO produces a response that exhibits a higher degree of humor compared to BART-FT. In Figure 5b, both BART-FT and BART-BRIO generate humorous responses, with BRIO's output slightly edging in humor but at the expense of being longer. Figure 5c, BART-BRIO-DPO generates a more contextually relevant and amusing paraphrase in comparison to the BART-FT student. Lastly, Figure 5d demonstrates an instance where the BART-DPO variant, receiving feedback for two iterations instead of one, exhibits superior performance compared to its counterpart trained with only one iteration of feedback.

7 Conclusions

We present a novel framework for knowledge distillation in the context of humor generation, with the teacher LLM providing references and feedback on a smaller student model's performance. Our approach involves leveraging a critic to guide the student model toward generating more humorous outputs. The effectiveness of our method is demonstrated through evaluations conducted by both the LLM and humans. Additionally, we analyze the effect of various design choices on the performance, such as the frequency of feedback and training set size. Our analysis sheds light on the limitations of LLM-based critics, and serves as motivation for future research in mitigation of biases in LLM-based evaluation and AI feedback.

8 Limitations

Quantifying Humor Measuring humor, especially in computational contexts, is inherently subjective and may not be accurately captured using simple metrics.

Cultural references We observed that some of the generated humorous outputs were referring to celebrities and events in North America. As humor varies widely across individuals and cultures, the proposed models may not generalize well across diverse demographic and cultural groups.

Bias propagation Training on feedback data from the LLM may lead to the propagation of any existing biases in the LLM's training data. If the LLM's data lacks diversity, these biases could be intensified in the smaller model. Similarly, the biases in model evaluation can also be propagated to the student.

9 Ethical Considerations

Data. The datasets used to gather the literal inputs outlined in Sec 4 are publicly accessible. Some of these datasets include crowdsourced annotations about emotional events, which may contain offensive, biased or hateful content. We use the chat variant of the llama2 that is aligned on generating safe warnings and responses when offensive content may be present. Further, the Llama2 model generates such warnings for about 3-4% of our inputs, and we discared these inputs from the data distilled into the student.

Celebrity references. As shown in Figure 1, the teacher LLM makes references and analogies to celebrities primarily from North America. When used as analogies for negative connotations, the humor generated may be considered offensive to specific people.

References

Chandra Bhagavatula, Jena D. Hwang, Doug Downey, Ronan Le Bras, Ximing Lu, Lianhui Qin, Keisuke Sakaguchi, Swabha Swayamdipta, Peter West, and Yejin Choi. 2023. I2D2: Inductive knowledge distillation with NeuroLogic and self-imitation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9614–9630, Toronto, Canada. Association for Computational Linguistics.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2023. Art or artifice? large language models and the false promise of creativity. *ArXiv*, abs/2309.14556.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. FLUTE: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023. Minigpt-v2: Large language model as a unified interface for vision-language multi-task learning. *arXiv:2310.09478*.

Peng-Yu Chen and Von-Wun Soo. 2018. Humor recognition using deep learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 113–117, New Orleans, Louisiana. Association for Computational Linguistics.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *ArXiv*, abs/2302.04166.

Aparna Garimella, Carmen Banea, Nabil Hossain, and Rada Mihalcea. 2020. "judge me by my size (noun), do you?" YodaLib: A demographic-aware humor generation framework. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2814–2825, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. The false promise of imitating proprietary llms.

Jack Hessel, Ana Marasovic, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2023. Do androids laugh at electric sheep? humor "understanding" benchmarks from the new yorker caption contest. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 688–714, Toronto, Canada. Association for Computational Linguistics.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015a. Distilling the knowledge in a neural network.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015b. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531.

Nabil Hossain, John Krumm, Michael Gamon, and Henry Kautz. 2020. SemEval-2020 task 7: Assessing

humor in edited news headlines. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 746–758, Barcelona (online). International Committee for Computational Linguistics.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017, Toronto, Canada. Association for Computational Linguistics.

Jaehun Jung, Peter West, Liwei Jiang, Faeze Brahman, Ximing Lu, Jillian Fisher, Taylor Sorensen, and Yejin Choi. 2023. Impossible distillation: from low-quality model to high-quality dataset & model for summarization and paraphrasing.

Apoorv Kulshreshtha, Daniel De Freitas Adiwardana, David Richard So, Gaurav Nemade, Jamie Hall, Noah Fiedel, Quoc V. Le, Romal Thoppilan, Thang Luong, Yifeng Lu, and Zi Yang. 2020. Towards a human-like open-domain chatbot. *arXiv*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. 2023a. Symbolic chain-of-thought distillation: Small models can also "think" step-by-step. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2665–2679, Toronto, Canada. Association for Computational Linguistics.

Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Luke Zettlemoyer, Omer Levy, Jason Weston, and Mike Lewis. 2023b. Self-alignment with instruction backtranslation. *ArXiv*, abs/2308.06259.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yen-Ting Lin and Yun-Nung Chen. 2023. LLM-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pages 47–58, Toronto, Canada. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuo Wang, Ruochen Xu, and Chenguang Zhu. 2023a. G-eval: Nlg evaluation using gpt-4 with better human alignment. *ArXiv*, abs/2303.16634.

Yixin Liu, Alexander R. Fabbri, Pengfei Liu, Dragomir Radev, and Arman Cohan. 2023b. On learning to summarize with large language models as references.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022. BRIO: Bringing order to abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2890–2903, Dublin, Ireland. Association for Computational Linguistics.

Zhengyuan Liu and Nancy Chen. 2022. Learning from bootstrapping and stepwise reinforcement reward: A semi-supervised framework for text style transfer. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2633–2648, Seattle, United States. Association for Computational Linguistics.

Fuli Luo, Shunyao Li, Pengcheng Yang, Lei Li, Baobao Chang, Zhifang Sui, and Xu Sun. 2019. Pun-GAN: Generative adversarial network for pun generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3388–3393, Hong Kong, China. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Sean Welleck, Bodhisattwa Prasad Majumder, Shashank Gupta, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. *ArXiv*, abs/2303.17651.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2023. Teaching small language models to reason. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1773–1781, Toronto, Canada. Association for Computational Linguistics.

Shuhaib Mehri and Vered Shwartz. 2023. Automatic evaluation of generative models with instruction tuning.

Rada Mihalcea and Carlo Strapparava. 2005. Making computers laugh: Investigations in automatic humor recognition. In *Human Language Technology - The Baltic Perspectiv*.

Anirudh Mittal, Yufei Tian, and Nanyun Peng. 2022. AmbiPun: Generating humorous puns with ambiguous context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1053–1062, Seattle, United States. Association for Computational Linguistics.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4.

OpenAI. 2023. Gpt-4 technical report.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th*

Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Debjit Paul, Mete Ismayilzada, Maxime Peyrard, Beatriz Borges, Antoine Bosselut, Robert West, and Boi Faltings. 2023. Refiner: Reasoning feedback on intermediate representations.

Saša Petrović and David Matthews. 2013. Unsupervised joke generation from big data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 228–232, Sofia, Bulgaria. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Victor Raskin. 1979. Semantic mechanisms of humor. In *Annual Meeting of the Berkeley Linguistics Society*, volume 5, pages 325–335.

Antonio Reyes, Paolo Rosso, and D. Buscaldi. 2012. From humor recognition to irony detection: The figurative language of social media. *Data Knowl. Eng.*, 74:1–12.

Graeme Ritchie. 2005. Computational mechanisms for pun generation. In *Proceedings of the Tenth European Workshop on Natural Language Generation (ENLG-05)*, Aberdeen, Scotland. Association for Computational Linguistics.

Graeme Ritchie, Ruli Manurung, Helen Pain, Annalu Waller, Rolf Black, and Dave O'Mara. 2007. A practical application of computational humour. In *Proceedings of the 4th international joint conference on computational creativity*, pages 91–98. UK London.

Joshua Robinson and David Wingate. 2023. Leveraging large language models for multiple choice question answering. In *The Eleventh International Conference on Learning Representations*.

Swarnadeep Saha, Omer Levy, Asli Celikyilmaz, Mohit Bansal, Jason Weston, and Xian Li. 2023a. Branch-solve-merge improves large language model evaluation and generation. *ArXiv*, abs/2310.15123.

Swarnadeep Saha, Omer Levy, Asli Celikyilmaz, Mohit Bansal, Jason Weston, and Xian Li. 2023b. Branch-solve-merge improves large language model evaluation and generation.

Wei Shen, Rui Zheng, Wenyu Zhan, Jun Zhao, Shihan Dou, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. Loose lips sink ships: Mitigating length bias in reinforcement learning from human feedback. *ArXiv*, abs/2310.05199.

Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. 2023. Distilling reasoning capabilities into smaller language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7059–7073, Toronto, Canada. Association for Computational Linguistics.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. 2023. A long way to go: Investigating length correlations in rlhf.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment.

Alessandro Valitutti. 2011. How many jokes are really funny? towards a new approach to the evaluation of computational humour generators. In *Proceedings of International Workshop on Natural Language Processing and Cognitive Science (NLPCS 2011)*, pages 189–200, Denmark. Samfundslitteratur. International Workshop on Natural Language Processing and Cognitive Science (NLPCS 2011) ; Conference date: 20-08-2011 Through 21-08-2011.

Peifeng Wang, Zhengyang Wang, Zheng Li, Yifan Gao, Bing Yin, and Xiang Ren. 2023a. SCOTT: Self-consistent chain-of-thought distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5546–5558, Toronto, Canada. Association for Computational Linguistics.

Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023b. Large language models are not fair evaluators. *ArXiv*, abs/2305.17926.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Orion Weller, Nancy Fulda, and Kevin Seppi. 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, Online. Association for Computational Linguistics.

Orion Weller and Kevin Seppi. 2019. Humor detection: A transformer gets the last laugh. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3621–3625, Hong Kong, China. Association for Computational Linguistics.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic knowledge distillation: from general language models to commonsense models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4602–4625, Seattle, United States. Association for Computational Linguistics.

Peter West, Ximing Lu, Nouha Dziri, Faeze Brahman, Linjie Li, Jena D. Hwang, Liwei Jiang, Jillian Fisher, Abhilasha Ravichander, Khyathi Chandu, Benjamin Newman, Pang Wei Koh, Allyson Ettinger, and Yejin Choi. 2023. The generative ai paradox: "what it can create, it may not understand".

Zhiwei Yu, Jiwei Tan, and Xiaojun Wan. 2018. A neural approach to pun generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1660, Melbourne, Australia. Association for Computational Linguistics.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Prompts for Evaluation

In Figure 3, we show the prompt used to obtain the pairwise scores from the teacher Llama2-70B model.

```
You will be given an input text from a conversation. You will then
be given two paraphrase choices (1 or 2) for this input text.
Choose the better paraphrase, based on how humorous and
human-like it sounds, while staying true to the input text.
Input text: {I}
Paraphrases:
1. {P1}
2. {P2}
Format your answer in this way -
'Answer: Your choice of 1 or 2
Reason: A reason for your answer which is maximum 15 to 20
 words long.'
```

Figure 3: Pairwise evaluation using Multiple Choice Prompting

B Prompts for Generation

In Figure 3, we show the prompt used to obtain the humorous outputs from the teacher Llama2-70B model.

```
Generate 3 creative and humorous paraphrases
with the exact same meaning as the input text.
Format the paraphrases as a list numbered as 1., 2.,
3. etc. Answer with only the list of humorous
paraphrases and nothing else, no exceptions.

Input text: {I}
```

Figure 4: Humor Generation