Leveraging Domain Knowledge for Efficient Reward Modelling in RLHF: A Case-Study in E-Commerce Opinion Summarization

Anonymous ACL submission

Abstract

Reinforcement Learning from Human Feedback (RLHF) has become a dominating strategy in steering Language Models (LMs) towards human values/goals. The key to the strategy is employing a reward model (φ) which can reflect a latent reward model with humans. While this strategy has proven to be effective, the training methodology requires a lot of human preference annotation (usually of the order of tens of thousands) to train φ. Such large-scale preference annotations can be achievable if the reward model can be ubiquitously used. However, human values/goals are subjective and depend on the nature of the task. This poses a challenge in collecting diverse preferences for downstream applications. To address this, we propose a novel methodology to infuse domain knowledge into φ, which reduces the size of preference annotation required. We validate our approach in E-Commerce Opinion Summarization, with a significant reduction in dataset size (just 940 samples) while advancing the state-of-the-art. Our contributions include a novel Reward Modelling technique, a new dataset (PROMPTOPINSUMM) for Opinion Summarization, and a human preference dataset (OPINPREF). The proposed methodology opens avenues for efficient RLHF, making it more adaptable to diverse applications with varying human values.

1 Introduction

Reinforcement Learning from Human Feedback (RLHF) (Ziegler et al., 2019; Ouyang et al., 2022) has emerged as a dominant paradigm in steering Language Models (LMs) towards human values. In the context of RLHF, human values are represented by a function (φ). For an output Y $(= y_1, y_2, \cdots, y_n)$ to some input X $(= x_1, x_2, \cdots, x_m)$, φ performs the mapping $(X, Y) \rightarrow r$. The reward function φ is latent to humans and manifests in human preferences.

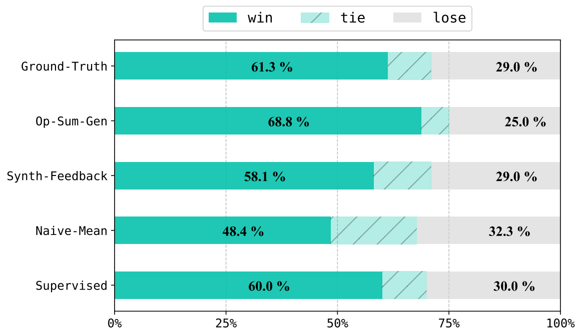


Figure 1: Human Eval: Pairwise win-tie-loss percentage of INDUCTIVE-BIAS model (our proposed model) vs. ground truth summary and summary from other models, for Amazon benchmark. We see that our proposed technique (domain knowledge infused reward model φ) helps INDUCTIVE-BIAS model achieve summaries which are always preferred more than the competitors.

Preference Modelling techniques, such as Bradley-Terry model (Bradley and Terry, 1952), Plackett-Luce models (Plackett, 1975; Luce, 2012) are used to learn φ from preference data, of the form: $\mathcal{D} = \{(X, Y_w, Y_l) \mid Y_w \succ Y_l\}$.

In contemporary works (Ziegler et al., 2019; Bai et al., 2022a; Ouyang et al., 2022; Rafailov et al., 2023), the reward functions are parameterized as LMs themselves. Thus the raw text, (X, Y_w) and (X, Y_l) are directly fed to φ. Such a formulation necessitates large-scale human preference data, to train the large LM (typically millions/billions of parameters). Typically the size of $\mathcal{D}$ varies from $20K$ (Ziegler et al., 2019; Nakano et al., 2021; Bai et al., 2022a) to $> 200K$ (Ethayarajh et al., 2022). Such a large-scale annotation might be achievable if the trained φ can be used ubiquitously, irrespective of the nature of downstream application. However, human values are subjective in nature (Jiang et al., 2022; Sorensen et al., 2023). For instance, hallucination would be desired in the task of Creative Writing, but undesired in Question-Answering. This means that depending on the context/downstream application, the reward function φ can have varying characteristics. Collecting human preferences for all such downstream applications would be impractical.

Motivated to resolve this need, we propose a novel methodology for reward model training. We draw on the insight that φ is dependent on the downstream application, and hence, can utilize its domain knowledge. Specifically, *φ lies in a low-dimensional manifold, whose dimensions can be deduced using domain knowledge*. Such an inductive bias *reduces the amount of samples*[1] needed to learn the parameters of φ, while *increasing bias*[2] of φ. Concretely our **hypothesis** is: *An inductive bias infused φ can help achieve alignment with human values for a task, with modest human preference annotations*. We experimentally prove this in the domain of E-Commerce Opinion Summarization (Bražinskas et al., 2020; Amplayo et al., 2021; Siledar et al., 2023b) – the task of summarizing user reviews for a product. First, we show that it helps advance the state-of-the-art in Opinion Summarization (Section 5). Second, we analyse how our approach helps the LM achieve alignment with human values for Opinion Summarization (Section 6).

Our contributions[3] are:

1. A novel Reward Modelling technique for RLHF, which leverages Domain Knowledge, to achieve alignment with human values, while significantly reducing human preference annotation. In the domain of Opinion Summarization, we achieve alignment while reducing the dataset size by $> 21\times$ (as compared to the smallest publicly available preference data[4]). Our approach advances the state of the art: humans prefer our models' outputs $> 68\%$ over the SOTA.

2. A new dataset, called PROMPTOPINSUMM, including reviews and summaries for 25763 products (229521 summaries), to train and validate Opinion Summarizer models.

3. A new human preference dataset, called OPINPREF, in the domain of Opinion Summarization, consisting of 940 instances.

2 Related Works

Steering Language Models (LMs) towards human goals: Steering LMs towards human goals/values refers to the task of training LMs to generate text which is more aligned with human values, such as '*text should not have harmful content*', '*it should be polite*', etc. Such a task necessitates a human presence in the training loop of the these LMs. In recent times, Reinforcement Learning from Human Feedback (RLHF) (Ziegler et al., 2019; Askell et al., 2021; Bai et al., 2022a; Ouyang et al., 2022; Liu et al., 2022) has emerged as a viable solution – by incorporating Reward Models, which reflect latent reward models wthin humans, into the training pipeline. These reward models are trained on human preference datasets (Ziegler et al., 2019; Nakano et al., 2021; Ethayarajh et al., 2022), which are typically of the order of tens of thousands, in size. Dependence on high-quality large-sized preference data is an obstacle for RLHF.

Recently, Reinforcement Learning from AI Feedback (RLAIF) (Bai et al., 2022b; Kim et al., 2023; Lee et al., 2023) has emerged as an alternative. It attempts to reduce the dependence on human preference datasets, by using Large LMs as preference data generators. While this is a scalable approach to steering LMs, there is no guarantee that the preference dataset generated by Large LMs actually reflects human goals – it is still a very open question. In our work, we propose a different solution – which promises to use human preference data, but provides a way to drastically reduce the required size. To the best of our knowledge, we are the first to attempt this.

Opinion Summarization: Opinion Summarization (Hu and Liu, 2004; Bražinskas et al., 2020; Amplayo et al., 2021; Siledar et al., 2023b) is the task of summarizing user reviews. Specifically, we look at E-Commerce Opinion Summarization, where user reviews are on products. These reviews contain aspects of the product, and users' sentiments/opinions towards those aspects. Previous works (Bražinskas et al., 2020; Siledar et al., 2023a) in E-Commerce Opinion Summarization

[1]An example: For a function, $f : (x_1, x_2, x_3, \cdots, x_m) \to y$, assuming that f is a linear combination of x_i (Linear Regression) reduces the training data requirement. Whereas, assuming no functional form (Feed-Forward Neural Network), would lead to more data requirement.

[2]Bias of a Machine Learning model is the expected difference between its output and the true value. Say $f^* : X \to y$ is a function, and $\hat{f}$ is an estimate of f^*, from some training data, then bias is represented as: $\mathbb{E}[\hat{f}(X) - f^*(X)]$.

[3]We would release the datasets and code publicly in the camera ready paper

[4]The smallest publicly available preference data is not in the domain of Opinion Summarization.

have used *Self-Supervised* training methodology. In this context, self-supervision refers to picking one of the N available reviews as a summary, commonly called *pseudo-summary*, and training the model on the remaining $N-1$ reviews to generate the pseudo-summary. The theme of solutions (Chu and Liu, 2018; Bražinskas et al., 2020; Siledar et al., 2023b,a) have mostly centered around Supervised Learning. The core problem has always been getting good synthetic datasets for training. More recently, Prompting (Bhaskar et al., 2023) has been explored to solve the task. Bhaskar et al. (2023) move away from making a better synthetic dataset generation pipeline, and test GPT-3.5 for Opinion Summarization.

We do not propose a new synthetic dataset generation methodology. Rather, we generate training data using an open-source Large LM (`Mistral-7B`), to test our hypothesis. To the best of our knowledge, we are the first to propose such a dataset for training Opinion Summarizers. Such an approach has been explored in earlier works (Wang et al., 2023; Taori et al., 2023; Peng et al., 2023) for general domain text generation. Taori et al. (2023) fine-tune `LLaMA-7B` (Touvron et al., 2023a) using Instruction-Tuning dataset generated using GPT-3. Peng et al. (2023) fine-tune `LLaMA-7B` using a dataset generated by GPT-4.

3 Dataset

Previous works (Bražinskas et al., 2020; Siledar et al., 2023a) in Opinion Summarization have used *Self-Supervised* training methodology. In the context of Opinion Summarization, self-supervision refers to picking one of the N available reviews as a summary, commonly called *pseudo-summary*, and training the model on the remaining $N-1$ reviews to generate the pseudo-summary. Among early works in this direction, Bražinskas et al. (2020) randomly selected a review as a pseudo-summary to construct a training sample. While, more recently, Siledar et al. (2023a) propose a sophisticated synthetic data creation pipeline for self-supervised opinion summarization.

Although these self-supervision datasets have helped further State-of-the-Art in Opinion Summarization, the trained models have several pitfalls. A few shortcomings are: (*a*) the generated opinion summaries are in first-person, (*b*) the review chosen as pseudo-summary might not cover all aspects and opinions, etc. To overcome these shortcomings, we move away from the self-supervised training regime in our work, and propose a new dataset. In the rest of this Section, we describe (*a*) PROMPTOPINSUMM – (*b*) a new dataset to train Opinion Summarization models, the benchmarks we used to test our proposed technique, and (*c*) OPINPREF – human preference annotated dataset for Opinion Summarization.

3.1 PROMPTOPINSUMM Dataset

We prompt the `Mistral-7B` model (Jiang et al., 2023) to generate opinion summary given reviews of a product. We also tried other open-source LLMs available at the time of the work, such as `LLaMA2-7B`, `LLaMA2-13B` (Touvron et al., 2023b), `Vicuna-7B`, `Vicuna-13B` (Chiang et al., 2023), `Zephyr-7B` (Tunstall et al., 2023). However, we found that `Mistral-7B` leads to better summaries. We limit ourselves to open-source models mainly to limit the expenditure. Appendix E includes an examples along with qualitative analysis. We use the Amazon dataset (He and McAuley, 2016), which has reviews for $\sim 180k$ products. We sample reviews for 20763 products for `train` set and 5000 products for `validation` set to prompt the `Mistral-7B` model. Specifically, we prompt the model to generate opinion summaries of 3 different qualities – Good (codenamed GOOD-SUM in the rest of the paper), Slightly Bad (codenamed SBAD-SUM in the rest of the paper) and Very Bad (codenamed VBAD-SUM in the rest of the paper). Again, per quality, we generate multiple opinion summaries (3 at most). We provide a reasoning of generating multiple summaries of different qualities in the discussion of our technique (Section 4.2). We generate a total of 184620 summaries for `train` set and 44901 summaries for `validation` set. Table 1 includes summary statistics of the generated dataset.

Split	Characteristic	μ	σ
`train`	# reviews	13.24	10.07
	# summaries	8.90	0.34
	# words in review	49.0	10.78
	# words in summary	78.28	34.45
`validation`	# reviews	10.53	6.80
	# summaries	8.98	0.16
	# words in review	48.65	10.63
	# words in summary	74.26	34.27

Table 1: Statistics of PROMPTOPINSUMM dataset.

3.2 Datasets for Evaluation

We test all our models on 3 popular benchmarks: Amazon (Bražinskas et al., 2020), Oposum+ (Amplayo et al., 2021), and Flipkart (Siledar et al., 2023b). Amazon dataset has reviews for 32 products from 4 domains: "*electronics*", "*home & kitchen*", "*personal care*", and "*clothing, shoes & jewellery*". Oposum+ dataset has reviews for 60 products from 6 domains: "*laptop bags*", "*bluetooth headsets*", "*boots*", "*keyboards*", "*television*", and "*vacuums*". Flipkart dataset has reviews for 147 products from 3 domains: "*laptops*", "*mobiles*", and "*tablets*". Table 2 includes relevant statistics of these benchmarks.

Characteristic	Oposum+	Amazon	Flipkart
# domains	6	4	3
# products	60	32	147
# reviews per product	10	8	10
# summaries per product	3	3	1

Table 2: Statistics of the benchmark datasets.

While we acknowledge the widespread use of these benchmarks, we notice several shortcomings. We highlight them in Appendix B. Due to these shortcomings, in our analysis, we do not rely much on overlap-based evaluations, such as ROUGE. Rather we rely on human evaluations.

3.3 OPINPREF Dataset

We create OPINPREF by asking humans to rate opinion summaries for given reviews. We utilize domain experts (annotator details in Appendix G) to perform the annotation. We believe that aligning to the internal reward model of domain experts would lead to better opinion summaries. We provide the domain expert with reviews of a product and two opinion summaries (products are sampled from the PROMPTOPINSUMM dataset). The domain expert notifies which of the two summaries they prefer. We use this to construct a dataset of the form: $\mathcal{D}_h = \{(R, s_w, s_l) \mid s_w \succ s_l\}$, where R is the set of reviews and s_w and s_l are opinion summaries. We construct a dataset of 940 samples. Appendix F includes statistics on the dataset (such as length of reviews and summaries).

4 Technique

We test our hypothesis in the domain of Opinion Summarization in E-Commerce – the task of summarizing user reviews for a product. Typically, user reviews discuss several aspects of a product, opinions/sentiments towards these aspects. An opinion summary has to reflect all the aspects that the input reviews discuss, along with the opinions/sentiments expressed towards these aspects. The section is structured as follows: in Section 4.1, we discuss our reward modelling technique, human annotation, etc., and in Section 4.2, we discuss the RLHF training pipeline using the novel reward model.

4.1 Domain Knowledge Induced Reward Modelling

We leverage insights on desirable properties in an opinion summary from domain experts[5]. Based on these insights, we characterize our reward model, φ_{op}, as follows: $\varphi_{op} = f(v)$, where $v \in$ {`aspect-coverage`, `opinion-faithfulness`, `opinion-coverage`, `conciseness`, `relevance`, `hallucination`, `language-correctness`}. The features `aspect-coverage`, `opinion-faithfulness` and `opinion-coverage` are directly relevant to Opinion Summarization. They check if the generated opinion summary covers all mentioned aspects and opinions. The features `conciseness`, `relevance` and `hallucination` pertain to Generic Text Summarization (Nallapati et al., 2016). They check if the generated summary is concise, relevant to the input reviews, and is free from hallucination. The features `language-correctness` pertains to the grammar of the language, checking if the generated text follows the rules of the language. These features, together, characterize the goodness of an opinion summary. We use `Mistral-7B` (see Appendix A for details) to obtain values for these features for an opinion summary, given reviews. We denote this transformation using Φ.

We train φ_{op} using OPINPREF, which is of the form: $\mathcal{D}_h = \{(R, s_w, s_l) | s_w \succ s_l\}$, where R is the set of reviews and s_w and s_l are opinion summaries. We parameterize φ_{op} using a Feed-Forward Neural Network, and train it using the *Elo*-loss (Ouyang et al., 2022; Glaese et al., 2022) (Equation 1; $\Phi(R, s_i)$ uses `Mistral-7B` to compute the 7 features). Such a formulation for φ_{op} brings interpretability – which features influence human preference the most, and is free from Alignment Tax (Bai et al., 2022a) – degradation of LM on benchmarks after reward model training.

[5] We consulted domain experts from an E-Commerce platform.

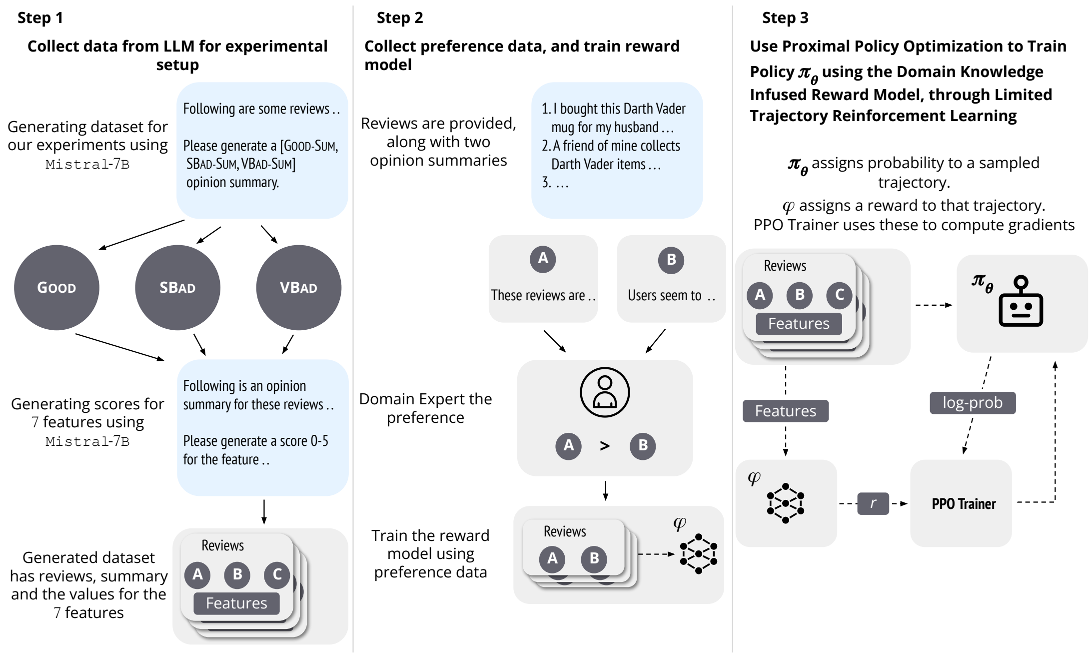


Figure 2: Overview of our approach. Step-1: We generate a new dataset for training Opinion Summarizers: PromptOpinSumm, by prompting `Mistral-7B` model. Again, we use `Mistral-7B` to compute values for the 7 features discussed in Section 4.1. Step-2: We ask humans (domain experts) to given their preference, given reviews and two opinion summaries (A, B). We use the preference data to train the reward model, φ_{op}. Step-3: We sample instances from PromptOpinSumm dataset; φ_{op} assigns a score to the sampled summaries, the policy, π_θ, assigns *log probabilities* to these summaries. Proximal Policy Optimization uses these to update π_θ.

$$\mathcal{L}_{pr} = -\mathbb{E}_{(R,s_w,s_l)\sim\mathcal{D}_h}\Big[\log\sigma\big(\varphi_{op}(\Phi(R,s_l)) - \varphi_{op}(\Phi(R,s_w))\big)\Big] \quad (1)$$

4.2 RLHF Training Pipeline

Using the trained reward model, we follow a similar training pipeline as Bai et al. (2022a); Ouyang et al. (2022), with a modification: *Limited Trajectory Reinforcement Learning*. Computing the transformation Φ for each generation online (during training) is expensive, especially with limited compute resources. To circumvent this, we limit the trajectories that are explored by our policy, π_θ. Specifically, we limit it to the Good-Sum, SBad-Sum and VBad-Sum trajectories in the PromptOpinSumm dataset. Having varying levels of quality in PromptOpinSumm is of use here – it lets the model still explore trajectories of several quality. Thus, we have an offline experience buffer, with Φ precomputed, for π_θ learn from.

We use Proximal Policy Optimization (PPO) (Schulman et al., 2017) to train our model (Equation 2). For each training step, we sample $(R, s, \Phi(R, s))$ tuples from PromptOpinSumm. We use the trained φ_{op} to compute the reward for s $(= \varphi_{op}(\Phi(R, s)))$. PPO uses this to update the log probability assigned by π_θ. We parameterize π_θ using a Transformer model, which takes reviews as input, and generates an opinion summary.

$$\mathcal{L}_{PPO} = -\mathbb{E}_{(R,s,\Phi(s))}\Big[\varphi_{op}(\Phi(R,s)) - \beta\log\Big(\frac{\pi_\theta^{RL}(s|R)}{\pi^{SFT}(s|R)}\Big)\Big] \quad (2)$$

5 Experiments and Results

Empirically, we test our technique exhaustively against the current State-of-the-Art model, and strong RL and RLHF baselines (our own design and from contemporary literature). We find that in overlap-based evaluation (Rouge-1, Rouge-2, Rouge-L), our proposed model (and other RL baselines) falls short of the SOTA. However, this is expected due to several shortcomings (Section 3.2) in the benchmark datasets. We conduct human and Gpt-4 to reliably verify efficacy of our technique. We find that our proposed technique excels significantly. In the rest of the section, we describe our models (Section 5.1) and evaluation results (Section 5.2).

	Model-Code	Amazon			Flipkart			Oposum+		
		R-1 ↑	R-2 ↑	R-L ↑	R-1 ↑	R-2 ↑	R-L ↑	R-1 ↑	R-2 ↑	R-L ↑
Prior Works	MeanSum (Chu and Liu, 2018)	29.20	4.70	18.15	–	–	–	26.25	4.62	16.49
	CopyCat (Bražinskas et al., 2020)	31.97	5.81	20.16	–	–	–	27.98	5.79	17.07
	PlanSum (Amplayo and Lapata, 2020)	32.87	6.12	19.05	–	–	–	30.26	5.29	17.48
	MultimodalSum (Im et al., 2021)	34.19	7.05	20.81	–	–	–	33.08	7.46	19.75
	OP-SUM-GEN Siledar et al. (2023a)	35.46	7.30	21.50	–	–	–	36.57	8.79	21.35
Ours'	SUPERVISED	28.22	4.91	16.79	28.10	4.18	15.37	30.15	7.26	16.74
	NAIVEMEAN	28.09	4.89	16.75	26.42	4.14	15.0	30.62	7.62	17.11
	SYNTH-FEEDBACK	24.82	4.30	15.48	27.40	3.64	15.25	29.10	6.19	16.20
	INDUCTIVE-BIAS	29.16	4.74	17.07	27.14	4.0	15.0	30.75	7.45	16.82

Table 3: Overlap-based Evaluation Results (R-1: ROUGE-1, R-2: ROUGE-2, R-L: ROUGE-L). Our models are not near the state-of-the-art in these metrics. We expect such a trend, and we present the reasons in Section 3.2. We include these results for completeness. Note that for Flipkart benchmark, we do not have results from OP-SUM-GEN, as Siledar et al. (2023a) only provide aspect specific summarization model for the Flipkart benchmark.

5.1 Models

We train the following models on PROMPTOPINSUMM:

SUPERVISED: This is a supervised model trained using Maximum Likelihood Estimation.

NAIVEMEAN: This is a Reinforcement Learning model, where the reward is computed by averaging the metric scores obtained using Φ.

SYNTH-FEEDBACK: This is a Reinforcement Learning from Synthetic Feedback (RLSF) (Kim et al., 2023) model. For this, we use a reward model which is trained on the implicit preference GOOD-SUM $\succ$ SBAD-SUM $\succ$ VBAD-SUM within the PROMPTOPINSUMM dataset. Kim et al. (2023) show that RLSF is a viable surrogate for RLHF. We train this reward model using Equation 1 too.

INDUCTIVE-BIAS: This is a RLHF model, trained following our hypothesis. We train φ_{op} using OPINPREF dataset.

In addition to the models above, we also use the current SOTA (Siledar et al., 2023a) for human evaluations. Siledar et al. (2023a) propose opinion summarizer models for general opinion summarization and aspect specific opinion summarization. As all of our models are for general opinion summarization, we use the same from Siledar et al. (2023a). We refer to this model by the name OP-SUM-GEN in the rest of the paper. Note that we do not use Direct Preference Optimization (DPO) (Rafailov et al., 2023) and vanilla RLHF (Ziegler et al., 2019; Bai et al., 2022a) as baselines. This is because both these techniques require huge human preference data. Additionally, the goal of the paper is not to propose a new RLHF technique, rather to propose a way to achieve alignment with domain-specific human goals with modest human preference annotation, using RLHF.

We use BART-Large (Lewis et al., 2020) for all of our models. The choice of the model is governed by two factors: (*a*) It provides a similar environment (model size) for comparison with SOTA, (*b*) We find that LLMs (`Mistral-7B`, `LLaMA2-7B`, `Zephyr-7B`, etc.) are already quite good at opinion summarization, thus any performance benefits (over SOTA) in those models cannot be reliably attributed to our novel reward modelling technique. We perform a hyperparameter sweep on `batch size`, `learning rate`, and `learning rate warmup` while training our models. We choose the model which performs best on the `validation` set of PROMPTOPINSUMM to generate summaries for the benchmark datasets. We include more details on implementation in Appendix C.

5.2 Evaluation Results

On the models above, we perform overlap-based evaluations (ROUGE). We report the scores in Table 3. In overlap-based evaluations, we find that our models do not perform better than existing models. We expect this behaviour (Section 3.2), which is why we do not draw conclusions from this evaluation. We report the metrics for completeness.

We also perform GPT-4 and human evaluations. For human evaluations, we use domain experts (annotator details in Appendix G) for the evaluations; we observe a Fleiss' Kappa score of 56.25% (moderate agreement). In both these settings, we ask the evaluator (GPT-4 or domain expert) to rank the generated summaries given the reviews. We anonymize all the summaries and shuffle the orders for each evaluation instance. Given the rankings, we compute the fraction of pairwise wins, ties and losses among all the models. We compare summaries from SUPERVISED, NAIVEMEAN, SYNTH-FEEDBACK, INDUCTIVE-BIAS, OP-SUM-GEN models and ground truth summaries. We include ground truth summaries in the evaluation to verify our claims about the quality of the benchmarks. Figure 1 shows the performance of INDUCTIVE-BIAS model over the others for human evaluations for the Amazon benchmark. We see that our proposed technique (domain knowledge based reward model φ_{op}) helps INDUCTIVE-BIAS model achieve summaries which are always preferred more than the competitors. We can see a similar trend for GPT-4 evaluation too, for all the benchmarks (Figures 3, 4 and 5). Both these evaluations empirically show that RLHF, using modest preference annotation, through domain knowledge infusion, can lead to performance gains.

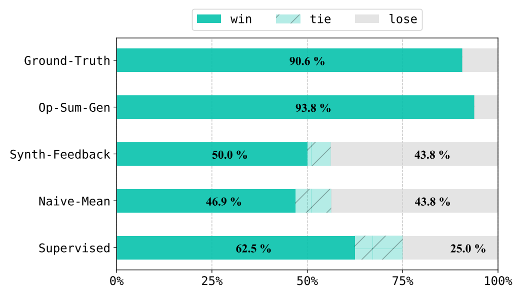


Figure 3: GPT-4 Eval: Pairwise win-tie-loss percentage of INDUCTIVE-BIAS model vs. competitors, for Amazon benchmark.

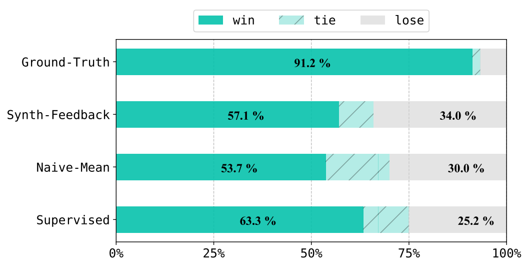


Figure 4: GPT-4 Eval: Pairwise win-tie-loss percentage of INDUCTIVE-BIAS model vs. competitors, for the Flipkart benchmark. Note that for Flipkart benchmark, we do not have results from OP-SUM-GEN, as Siledar et al. (2023a) only provide aspect specific summarization model for Flipkart benchmark.

6 Analysis

We perform a two-fold analysis: (*a*) First, we see the domain knowledge feature influences for φ_{op}, (*b*) Second, we see how the ground truth summary

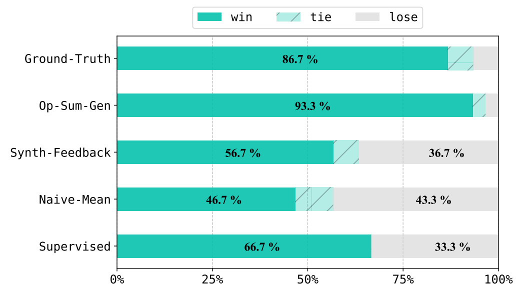


Figure 5: GPT-4 Eval: Pairwise win-tie-loss percentage of INDUCTIVE-BIAS model vs. competitors, for Oposum+ benchmark.

and summary from trained models fare on the domain knowledge features. This two-fold analysis helps us understand: (*a*) which features influence the latent reward model within humans[6] the most, and (*b*) how the ground truth summary and summary from trained models fare on these influential features. Performing good on influential features would mean the summary aligns well with the latent reward model within humans.

6.1 Analysis of φ_{op}

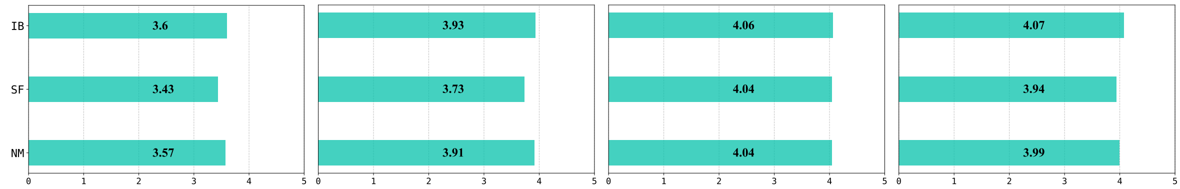
φ_{op} model has been trained on a set of features specified by domain experts. We analyze the relative influence of each feature on the final score assigned by φ_{op}. Doing this helps us understand an approximate importance[7] of each of these features. We do this by varying each feature by δ $(= 0.1)$ while keeping the other features constant, over multiple possible values of all features (Equation 3).

$$\Delta_i = \frac{1}{2\delta} \sum_{\mathbf{x}} \big(f(x_1, \cdots, x_i + \delta, \cdots, x_n) - f(x_1, \cdots, x_i - \delta, \cdots, x_n) \big) \quad (3)$$

Figure 7 highlights the relative influence of all the features. We see that `hallucination` is most influential. This aligns with what our human preference annotators report – hallucination in a summary is the primary cause of rejection. We see that, in majority cases, hallucinations are within the opinions in the summary. We see that this is also reflected in Figure 7 – `opinion-faithfulness` has significant influence. We also see that annotators prefer summaries which have more specifics about a product, i.e. they include more aspects – `aspect-coverage` has significant influence.

[6] Note that the trained φ_{op} represents latent human reward model.

[7] We call this approximate importance as influence of a features on the output is not necessarily its importance.



(a) `aspect-coverage` **(b)** `opinion-faithfulness` **(c)** `relevance` **(d)** `hallucination`

Figure 6: Scores on domain knowledge based features on the Amazon benchmark for top-3 models (`IB`: INDUCTIVE-BIAS, `NM`: NAIVEMEAN, `SF`: SYNTH-FEEDBACK). We see that INDUCTIVE-BIAS model manages to stay ahead of the competitors, notably for `hallucination`. This increases our confidence on the INDUCTIVE-BIAS model, as summaries generated by this model align more with the domain features.

6.2 Analysis of Summaries

We analyze the top-3 performing models (in human and GPT-4 evaluations) for the following features: `opinion-coverage`, `opinion-faithfulness`, `hallucination` and `relevance`. We show the analysis only for Amazon benchmark in the main body of the paper, we include rest in Appendix H. Figure 6 shows the performance on these features. We see that INDUCTIVE-BIAS model fares much better than the competitors on `hallucination` (the most influential metric). For `relevance`, `aspect-coverage` and `opinion-faithfulness`, our model is fairly better than the other models.

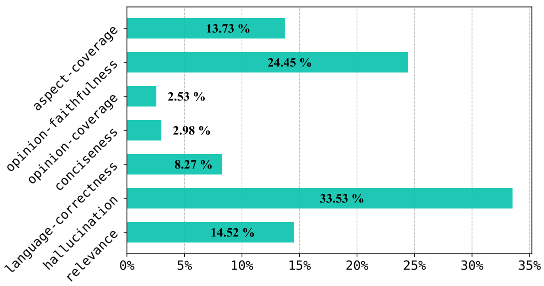


Figure 7: Relative Influence of all features in φ_{op}. All the influences sum to 1.

This shows that our technique helps INDUCTIVE-BIAS model perform well on features that influence the latent reward model within humans for opinion summarization. This means that our technique helps INDUCTIVE-BIAS model achieve a significant alignment with the latent reward model. This conclusion verifies our hypothesis (in the domain of opinion summarization) – *An inductive bias infused reward model* (φ_{op}) *can help achieve alignment with latent reward model of humans for a task, with modest human preference annotations*.

7 Summary, Conclusion and Future Work

In this work, we propose a novel Reward Modelling technique via Domain Knowledge Infusion. We verify our approach for E-Commerce Opinion Summarization, where we achieve state-of-the-art, while significantly reducing the amount of human preference annotations required (just 940 samples). In addition to advancing SOTA and reducing preference annotations, our technique provides another two-fold benefits: (*a*) No Alignment Tax, and (*b*) Interpretability. Due to the interpretable nature, we find that our model does achieve alignment with human goals for Opinion Summarization through analysis. From the results and analysis, we conclude that Domain Knowledge Infusion into Reward Modelling is a viable solution to reduce human preference annotations for downstream tasks. In future, we would verify this for other domains.

8 Ethical Considerations

We contribute two datasets in our work: PROMPTOPINSUMM, OPINPREF. These datasets are generated using an open-source model `Mistral-7B` (Jiang et al., 2023). We would release the datasets to further research in Opinion Summarization. For the OPINPREF, we have seen that it does not contain any harmful content, such as social biases, stereotypes, etc. However, we have seen that it contains products of explicit nature (sexual products). For the PROMPTOPINSUMM dataset, to the best of our knowledge, there is no presence of harmful content, such as social biases, stereotypes etc. We urge the research community to use the datasets with caution and check for potential harmfulness, based on their use-cases.

9 Limitations

A limitation of our work is we have tested our approach for one domain – Opinion Summarization.

However, we do not believe that this weakens our argument, as we have exhaustively shown that our approach not only advances SOTA, but also interpretably achieves alignment with humans. Future work in other domains would help in verifying this claim for other domains. Another limitation is: we train our RLHF models using the *limited trajectory* trick. However, this limitation is imposed by the available compute resources. With larger compute resources, this study can be extended to extensive exploration during RLHF training.

References

Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. 2021. Aspect-controllable opinion summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6578–6593, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Reinald Kim Amplayo and Mirella Lapata. 2020. Unsupervised opinion summarization with noising and denoising. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1934–1945, Online. Association for Computational Linguistics.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. 2021. A general language assistant as a laboratory for alignment.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, T. J. Henighan, Nicholas Joseph, Saurav Kadavath, John Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Christopher Olah, Benjamin Mann, and Jared Kaplan. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *ArXiv*, abs/2204.05862.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022b. Constitutional ai: Harmlessness from ai feedback.

Adithya Bhaskar, Alex Fabbri, and Greg Durrett. 2023. Prompted opinion summarization with GPT-3.5. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9282–9300, Toronto, Canada. Association for Computational Linguistics.

Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020. Unsupervised opinion summarization as copycat-review generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5151–5169, Online. Association for Computational Linguistics.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Eric Chu and Peter J. Liu. 2018. Meansum: A neural model for unsupervised multi-document abstractive summarization. In *International Conference on Machine Learning*.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5988–6008. PMLR.

Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Soňa Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. 2022. Improving alignment of dialogue agents via targeted human judgements.

Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of* *the 25th International Conference on World Wide Web*, WWW '16, page 507–517, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, page 168–177, New York, NY, USA. Association for Computing Machinery.

Jinbae Im, Moonki Kim, Hoyeop Lee, Hyunsouk Cho, and Sehee Chung. 2021. Self-supervised multimodal opinion summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 388–403, Online. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Liwei Jiang, Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny Liang, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jon Borchardt, Saadia Gabriel, Yulia Tsvetkov, Oren Etzioni, Maarten Sap, Regina Rini, and Yejin Choi. 2022. Can machines learn morality? the delphi experiment.

Sungdong Kim, Sanghwan Bae, Jamin Shin, Soyoung Kang, Donghyun Kwak, Kang Yoo, and Minjoon Seo. 2023. Aligning large language models through synthetic feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13677–13700, Singapore. Association for Computational Linguistics.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback.

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.

Ruibo Liu, Ge Zhang, Xinyu Feng, and Soroush Vosoughi. 2022. Aligning generative language models with human values. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 241–252, Seattle, United States. Association for Computational Linguistics.

R.D. Luce. 2012. *Individual Choice Behavior: A Theoretical Analysis*. Dover Books on Mathematics. Dover Publications.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Ouyang Long, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. Webgpt: Browser-assisted question-answering with human feedback. *ArXiv*, abs/2112.09332.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

R. L. Plackett. 1975. The analysis of permutations. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 24(2):193–202.

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.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms.

Tejpalsingh Siledar, Suman Banerjee, Amey Patil, Sudhanshu Singh, Muthusamy Chelliah, Nikesh Garera, and Pushpak Bhattacharyya. 2023a. Synthesize, if you do not have: Effective synthetic dataset creation strategies for self-supervised opinion summarization in E-commerce. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13480–13491, Singapore. Association for Computational Linguistics.

Tejpalsingh Siledar, Jigar Makwana, and Pushpak Bhattacharyya. 2023b. Aspect-sentiment-based opinion summarization using multiple information sources. In *Proceedings of the 6th Joint International Conference on Data Science & Management of Data (10th ACM IKDD CODS and 28th COMAD), Mumbai, India, January 4-7, 2023*, pages 55–61. ACM.

Taylor Sorensen, Liwei Jiang, Jena Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, Maarten Sap, John Tasioulas, and Yejin Choi. 2023. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem 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, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

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.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Daniel M. Ziegler, Nisan Stiennon, Jeff Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *ArXiv*, abs/1909.08593.

A Features for Reward Modelling

We use 7 domain specific features for the reward model φ_{op}. For each feature we prompt `Mistral-7B` to generate a score within 0 and 5. For each feature, 0 means the model is doing bad on the feature, and 5 means the model is doing good on the feature. We define all the features below:

`aspect-coverage`: This feature considers the aspect coverage within an opinion summary. The feature assumes a value 5 if all the aspects of the product, mentioned in the reviews, are mentioned in the summary. If none of the aspects are picked, the feature assumes a value 0.

`opinion-faithfulness`: This feature considers whether the mentioned opinions/sentiments in the summary are correct, that is, they are picked correctly from the reviews. For example, if an user mentions that they are *happy* with the battery of a phone, and the summary mentions that users are *unhappy* with the battery, the summary will not be considered faithful to opinion in the review. The feature assumes a value 5 if all the opinions are faithfully reflected. If no opinion is faithfully reflected, the value would be 0.

`opinion-coverage`: This feature considers whether all the opinions in the input reviews are picked by the opinion summary. The feature assumes a value 5 if all the opinions are picked up. If none of the opinions are picked up, the feature assumes a value 0.

`relevance`: This feature checks if the summary is relevant to the input reviews (that is the product). The feature assumes a value 5 if summary is completely relevant. If it is completely irrelevant, the feature assumes a value 0.

`conciseness`: This feature considers the conciseness and completeness of the opinion summary. The feature assumes a value 5 if the summary is concise and complete – not one phrase/sentence can be dropped off. It assumes a value 0 if the summary is totally incomplete, or very verbose.

`hallucination`: This feature considers the factuality of the opinion summary. The feature assumes a value 5 if the summary is totally factual, with respect to the input reviews. If there are a lot of hallucinations, the feature assumes a value 0.

`language-correctness`: This feature checks the correctness of language/text in the opinion summary. The feature assumes a value 5 if the summary is grammatically fully correct. It assumes a value 0 if the summary is very poor linguistically.

For conciseness, we do not include the prompts in the paper, we would release them as separate artifacts, with the datasets, in the camera ready version.

B Example Summaries in Existing Benchmarks

We highlight the shortcomings of the benchmark datasets here.

Amazon: Bražinskas et al. (2020) designed the test-set in such a way that the summary has to read like a review, for instance, summary would contain '*I think the quality has come down over the years.*', instead of '*Users think that quality has come down over years*'. Due to this writing style, the summaries read like reviews and are often in first person – high overlap would not necessarily mean a better summary, it would rather mean a better review.

Flipkart: Siledar et al. (2023b) generate this dataset by listing out the aspect-wise pros and cons presented within the reviews. We form an opinion summary by concatenating these pros and cons. Due to this, the summaries have frequent incoherent sentences.

Oposum+: Amplayo et al. (2021) create this benchmark by extracting sentences from the input reviews. Hence, this dataset has similar drawbacks as the Amazon benchmark.

Amazon

Nice boots but run a bit narrow. They look great but I think the quality has come down over the years. Still comfortable but I wish they broke in easier. I recommend these for any lady who is patient and looking for comfort.

Oposum+

great product for the cost . very easy to use and compatible with all of my phones ! it holds a charge great , is light enough and fits perfectly in my ear . the sound quality is great , the style is very cool and the unit feels top quality . it would drop and reconnect every 10 seconds nobody could hear me i could n't get it to unpair from the phone , there 's apparently no noise-cancellation in these . the battery life is ... bizarre . cheap , plastic-y , and poor sound quality .

Flipkart

Summary

Pros

Design: The full-metal Infinix INBook X1 Core i3 has a top notch and premium design.

35.56 cm (14 inch) 1920 x 1080 Pixel Full HD IPS Display: 100% sRGB with 300nits brightness ensures an excellent display.

Battery: Long-lasting battery. Gives around 8 hours of backup on normal usage.

Performance: The combination of Intel Core processor chip, high RAM size and sufficient storage capacity gives this laptop a high-speed performance experience.

Price: "Totally worth it in this price range.

Cons

Charging: Some current leakage during charging. Sometimes the laptop won't charge.

Trackpad: Not upto the mark.

***Verdict**: This laptop comes with a i3 10th gen dual core processor which is suitable for normal tasks like web browsing, online classes and watching movies. Not recommended as a gaming laptop.*

***Additional Information**: Can handle video editing and expandable SSD.*

C Implementation Details

We use BART-Large (Lewis et al., 2020) as our policy (π_θ) in all of the models. We do this to have a fair comparison with the state-of-the-art in Opinion Summarization. We use AdamW Optimizer to train the models, with a weight decay of 0.05. We use a `cosine` learning rate scheduler. We run a hyperparameter sweep on `batch size`, `learning rate`, and `learning rate warmup`. We include the possible values for the sweep in Table 4. We train all of our models using 2× A100 GPUs (80GB)

For the reward model, φ_{op}, we use a Feed Forward Network for the Policy Model. We use AdamW Optimizer to train the models, with a weight decay of 0.05. As before, we run a hyperparameter sweep on `batch size`, `learning rate`, and `learning rate warmup`. Table 5 includes details on the hyperparameters.

Hyperparam	Values
`batch size`	[64, 128, 256]
`learning rate`	$\sim \mathcal{U}(5e^{-6}, 5e^{-5})$
`learning rate warmup`	$\sim \mathcal{U}(0.2, 0.4)$

Table 4: Possible Values for Hyperparameters. For `learning rate warmup`, we sample the fraction of total steps the learning should be warmed up. For example, if the `learning rate warmup` is 0.2, it means that the learning rate will have a linear warmup for 20% of the total training steps.

Hyperparam	Values
`batch size`	[32, 64, 128]
`learning rate`	$\sim \mathcal{U}(5e^{-3}, 1e^{-1})$

Table 5: Possible Values for Hyperparameters for the Reward Model. For `learning rate warmup`, we sample the fraction of total steps the learning should be warmed up. For example, if the `learning rate warmup` is 0.2, it means that the learning rate will have a linear warmup for 20% of the total training steps.

D Generated Summary Lengths

We analyze the generation lengths of the models, and the ground truth summary. Table 6 lists the summary lengths.

Model	Amazon	Oposum	Flipkart
Ground-Truth	60.65	85.86	129.91
NAIVEMEAN	79.93	99.76	71.59
SYNTH-FEEDBACK	87.59	115.00	84.18
OP-SUM-GEN	55.84	62.93	-
INDUCTIVE-BIAS	73.53	68.63	68.63
SUPERVISED	81.84	126.00	77.80

Table 6: Generation Length Statistics: number of words in summaries. We use NLTK to split into tokens.

E Details of the Generated Dataset

We include an example from PROMPTOPINSUMM dataset (Table 7). We show one sample from GOOD-SUM, SBAD-SUM and VBAD-SUM quality each. We do not include the reviews for conciseness. However, we incluce the salient aspects of the reviews. The reviews talk about the following things:

1. Great price, Nice looking / Good color, Good utility / Good protection, Quick shipping, Nice fitting, Good accessibility of the laptop while the cover is on, Good finish quality.
2. Not a good fit for older macbooks, Broken / Unusable for the original ($\sim$ 2009 - 2010) white macbooks, cannot return return.

GOOD-SUM manages to discuss all of these things – it was able to detect that the cover does not fit specific models (highlighted in blue). It also detects that a few users like the "personal touch addition" factor of the cover.

VBAD-SUM is also a good representation of a bad summary – it totally ignores the "positive" aspects of the product and presents the "negative" aspects only. It fails at Aspect Coverage, Opinion Faithfulness and Opinion Coverage.

SBAD-SUM maintains almost a similar quality as the Good one. However, it fails to draw out certain aspects, such as "pizzaz", "personal touch addition", etc.

F Statistics of the OPINPREF dataset

We look at the summary statistics for the OPINPREF dataset. Table 8. We see that, interestingly, annotators prefer longer summaries – this is because these summaries contain more specifics and details from the reviews.

G Annotator Details

We include two disjoint sets of annotators in our work – first for creation of OPINPREF (3 annotators), second for human evaluation (3 annotators). For both annotations, we use domain experts. The domain experts are NLP researchers (age group: $24-30$) who have worked in Opinion Summarization for a long time, with publication experience (in *A*/*A** conferences). The domain experts for human evaluation also have a similar profile. The annotators have been paid generously, based on the standard annotation rates in the geographical location.

H All Evaluation Results

We include all evaluation results in this section. In Tables 9, 10, 11 and 11 we include pairwise comparison results, in a `win/tie/loss` format. We also include results on evaluation on the features in Tables 13, 14 and 15.

Good: The users have shared their positive experiences with the protective covers for Macbooks. They appreciate the ease of application, vibrant colors, quick shipping, and great quality. Some have mentioned the covers fit their Macbooks perfectly, while others have shared their disappointment when they realized it did not fit their specific model. A few users have noted the covers add a personal touch and pizzaz to their devices. However, some users have reported issues with the covers not fitting their Macbooks or falling off, leaving them feeling disappointed and frustrated.
Slightly bad: Some users have reported positive experiences with the Macbook cover, praising its protective qualities, easy application, and vibrant colors. Others, however, have encountered issues with it not fitting properly on their devices or being returned due to size mismatches. Overall, the product has been described as cute, great quality, and worth the investment, though some buyers have experienced disappointment with its size compatibility and easy detachment.
Very Bad: This is a terrible product for people with old Macbooks, it doesn't fit at all! And don't even bother looking at the description carefully before buying.

Table 7: Example summaries of generated dataset

Characteristic	**Value**
Avg. no of Words in reviews	641.21
Avg. no of reviews	13.08
Avg. no of Words in summaries	73.16
Avg no of Words in preferred summaries	85.41
Avg no Words in dispreferred summaries	66.91

Table 8: Statistics of the OPINPREF dataset.

·	Supervised	NaiveMean	Synth-Feedback	Inductive-Bias	Op-Sum-Gen
NaiveMean	0.50/0.06/0.38				
Synth-Feedback	0.44/0.12/0.44	0.40/0.09/0.5			
Inductive-Bias	**0.56**/0.09/0.28	**0.46**/0.18/0.31	0.56/0.12/0.28		
Op-Sum-Gen	0.31/0.28/0.38	0.25/0.12/0.56	0.25/0.21/0.5	0.25/0.06/**0.68**	
Ground-Truth	0.46/0.06/0.48	0.31/0.18/0.44	0.40/0.15/0.40	0.28/0.09/**0.59**	0.5/0.09/0.38

Table 9: Pairwise Win/Tie/Loss Results for all models in Human Evaluation for Amazon Benchmark. We format the data as: win/tie/loss.

·	Supervised	NaiveMean	Synth-Feedback	Inductive-Bias	Op-Sum-Gen
NaiveMean	0.63/0.12/0.25				
Synth-Feedback	0.59/0.12/0.28	0.5/0.06/0.44			
Inductive-Bias	**0.62**/0.12/0.25	**0.46**/0.09/0.44	**0.5**/0.06/0.44		
Op-Sum-Gen	0.06/0.03/0.9	0.09/0.0/0.90	0.12/0.09/0.78	0.06/0.0/**0.93**	
ground-truth	0.12/0.06/0.81	0.09/0.06/0.84	0.16/0.06/0.78	0.09/0.0/**0.90**	0.68/0.09/0.22

Table 10: Pairwise Win/Tie/Loss Results for all models in GPT-4 Evaluation for Amazon Benchmark. We format the data as: win/tie/loss.

·	Supervised	NaiveMean	Synth-Feedback	Inductive-Bias
NaiveMean	0.57/0.12/0.30			
Synth-Feedback	0.57/0.06/0.36	0.52/0.12/0.36		
Inductive-Bias	**0.63**/0.12/0.25	**0.54**/0.16/0.30	**0.57**/0.08/0.34	
Ground-Truth	0.10/0.06/0.84	0.06/0.01/0.92	0.07/0.01/0.91	0.06/0.02/**0.91**

Table 11: Pairwise Win/Tie/Loss Results for all models in GPT-4 Evaluation for Flipkart Benchmark. We format the data as: win/tie/loss.

·	Supervised	NaiveMean	Synth-Feedback	Inductive-Bias	Op-Sum-Gen
NaiveMean	0.56/0.03/0.4				
Synth-Feedback	0.5/0.16/0.34	0.46/0.1/0.44			
Inductive-Bias	**0.66**/0.0/0.33	**0.46**/0.1/0.44	**0.56**/0.06/0.36		
Op-Sum-Gen	0.1/0.06/0.83	0.06/0.03/0.9	0.03/0.03/0.93	0.03/0.03/**0.93**	
Ground-Truth	0.13/0.13/0.73	0.1/0.033/0.8666	0.06/0.06/0.86	0.06/0.06/**0.86**	0.7/0.1/0.2

Table 12: Pairwise Win/Tie/Loss Results for all models in GPT-4 Evaluation for Oposum+ Benchmark. We format the data as: win/tie/loss.

·	AC	OPF	OPC	CC	RL	HL	LC
Supervised	3.43 ± 0.20	3.71 ± 0.37	3.67 ± 0.26	3.79 ± 0.31	4.04 ± 0.37	3.89 ± 0.39	4.55 ± 0.35
NaiveMean	3.56 ± 0.22	3.91 ± 0.50	3.76 ± 0.38	3.89 ± 0.36	4.04 ± 0.48	3.99 ± 0.48	4.60 ± 0.27
Synth-Feedback	3.55 ± 0.40	3.87 ± 0.71	3.71 ± 0.43	3.94 ± 0.50	4.04 ± 0.61	3.94 ± 0.68	4.38 ± 0.92
Inductive-Bias	$\mathbf{3.60 \pm 0.17}$	$\mathbf{3.95 \pm 0.40}$	$\mathbf{3.85 \pm 0.25}$	3.99 ± 0.35	4.06 ± 0.34	$\mathbf{4.07 \pm 0.43}$	$\mathbf{4.65 \pm 0.32}$
Op-Sum-Gen	3.34 ± 0.68	3.92 ± 0.79	3.70 ± 0.54	$\mathbf{4.0 \pm 0.50}$	$\mathbf{4.08 \pm 0.72}$	3.87 ± 1.08	4.05 ± 1.31
Ground-Truth	3.55 ± 0.50	3.93 ± 0.46	3.56 ± 0.31	4.08 ± 0.32	4.04 ± 0.46	3.81 ± 0.86	4.40 ± 0.45

Table 13: Intrisic Evaluation results on the Amazon Benchmark for all the models. Legend: AC: aspect-coverage, OPF: opinion-faithfulness, OPC: opinion-coverage, CC: conciseness, RL: relevance, HL: hallucination, LC: language-correctness.

·	AC	OPF	OPC	CC	RL	HL	LC
SUPERVISED	3.61 ± 0.22	4.10 ± 0.39	3.84 ± 0.33	4.04 ± 0.28	4.21 ± 0.31	4.19 ± 0.42	4.53 ± 0.27
NAIVEMEAN	3.56 ± 0.21	4.13 ± 0.41	3.84 ± 0.34	4.0 ± 0.32	$\mathbf{4.31 \pm 0.36}$	4.26 ± 0.34	4.54 ± 0.39
SYNTH-FEEDBACK	3.56 ± 0.25	4.09 ± 0.40	3.79 ± 0.32	$\mathbf{4.02 \pm 0.30}$	4.19 ± 0.34	4.19 ± 0.36	4.53 ± 0.29
INDUCTIVE-BIAS	$\mathbf{3.63 \pm 0.20}$	$\mathbf{4.22 \pm 0.39}$	$\mathbf{3.85 \pm 0.30}$	4.01 ± 0.28	4.26 ± 0.29	$\mathbf{4.33 \pm 0.45}$	$\mathbf{4.61 \pm 0.29}$
OP-SUM-GEN	x	x	x	x	x	x	x
Ground-Truth	3.59 ± 0.15	3.88 ± 0.53	3.68 ± 0.27	4.02 ± 0.28	3.87 ± 0.59	3.67 ± 0.78	4.35 ± 0.44

Table 14: Intrisic Evaluation results on the Flipkart Benchmark for all the models. Legend: AC: `aspect-coverage`, OPF: `opinion-faithfulness`, OPC: `opinion-coverage`, CC: `conciseness`, RL: `relevance`, HL: `hallucination`, LC: `language-correctness`.

·	AC	OPF	OPC	CC	RL	HL	LC
SUPERVISED	3.47 ± 0.14	3.38 ± 0.26	3.49 ± 0.06	$\mathbf{3.64 \pm 0.19}$	$\mathbf{3.81 \pm 0.26}$	3.22 ± 0.56	3.96 ± 0.32
NAIVEMEAN	3.49 ± 0.05	3.48 ± 0.06	3.5 ± 0.0	3.56 ± 0.13	3.66 ± 0.22	3.52 ± 0.33	4.1 ± 0.33
SYNTH-FEEDBACK	3.50 ± 0.03	3.41 ± 0.26	3.5 ± 0.0	3.63 ± 0.24	3.62 ± 0.20	3.32 ± 0.63	$\mathbf{4.03 \pm 0.38}$
INDUCTIVE-BIAS	$\mathbf{3.54 \pm 0.22}$	$\mathbf{3.50 \pm 0.06}$	$\mathbf{3.57 \pm .06}$	3.62 ± 0.19	3.65 ± 0.23	$\mathbf{3.68 \pm 0.36}$	4.0 ± 0.29
OP-SUM-GEN	3.39 ± 0.3	3.46 ± 0.45	3.49 ± 0.28	3.61 ± 0.40	3.58 ± 0.82	3.43 ± 0.92	3.79 ± 1.18
Ground-Truth	3.42 ± 0.22	3.475 ± 0.28	3.5 ± 0.0	3.57 ± 0.16	3.49 ± 0.28	3.21 ± 0.48	3.56 ± 0.23

Table 15: Intrisic Evaluation results on the Oposum+ Benchmark for all the models. Legend: AC: `aspect-coverage`, OPF: `opinion-faithfulness`, OPC: `opinion-coverage`, CC: `conciseness`, RL: `relevance`, HL: `hallucination`, LC: `language-correctness`.