Leveraging Domain Knowledge for Efficient Reward Modeling 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 aligning Language Models (LMs) with human values/goals. The key to the strategy is learning a reward model (φ), which can reflect the latent reward model of humans. While this strategy has proven effective, the training methodology requires a lot of human preference annotation (usually in the order of tens of thousands) to train φ. Such a large-scale annotation is justifiable when it's a one-time effort, and the reward model is universally applicable. However, human goals are subjective and depend on the task, requiring task-specific preference annotations, which can be impractical to fulfill. To address this challenge, we propose a novel approach to infuse domain knowledge into φ, which reduces the amount of preference annotation required ($21\times$), omits Alignment Tax, and provides some interpretability. We validate our approach in E-Commerce Opinion Summarization, with a significant reduction in dataset size (to just 940 samples) while advancing the SOTA (~ 4 point ROUGE-L improvement, 68% of times preferred by humans over SOTA). Our contributions include a novel Reward Modeling technique and two new datasets: PROMPTOPINSUMM (supervised data for Opinion Summarization) and OPINPREF (a gold-standard human preference dataset). The proposed methodology opens up avenues for efficient RLHF, making it more adaptable to applications with varying human values.

1 Introduction

Reinforcement Learning from Human Feedback (RLHF) (Ziegler et al., 2019; Ouyang et al., 2022) is a prominent approach in aligning Language Models (LMs) with human values. Human values are represented by a function (φ), which ultimately acts as the reward in the RLHF training. 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. Preference Modeling 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: $\mathcal{D} = \{(X, Y_w, Y_l) \mid Y_w \succ Y_l\}$[1].

In contemporary works (Ziegler et al., 2019; Bai et al., 2022a; Ouyang et al., 2022; Rafailov et al., 2023), the reward functions are Large LMs (LLMs; pretrained Transformers) themselves. The text data, (X, Y_w) and (X, Y_l) are directly fed to φ, for training. Such a formulation necessitates large-scale human preference data to train the LLM (millions/billions of parameters). Typically the size of $\mathcal{D}$ varies from $20K$ (Nakano et al., 2021; Bai et al., 2022a) to $> 200K$ (Ethayarajh et al., 2022). Such a large-scale annotation is justifiable when it's a one-time effort, and the trained φ is universally applicable, irrespective of the nature of the

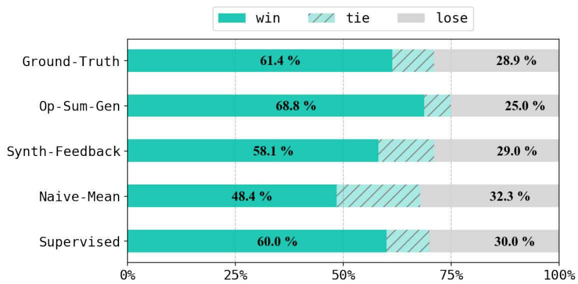

Figure 1: Human Eval: Pairwise win-tie-loss percentage of INDUCTIVE-BIAS model (our proposed model) vs. ground truth summary and summary from other models, for AMAZON benchmark. We see that our proposed approach (*infusing domain knowledge into φ to reap benefits of RLHF with modest human preference data*) helps INDUCTIVE-BIAS model achieve summaries which are always preferred (Section 5.2).

[1] $Y_w \succ Y_l$, in this entire paper, signifies that the output Y_w is preferred over the output Y_l; w: *win*, l: *loss*.

downstream task. However, human values are subjective (Jiang et al., 2022; Sorensen et al., 2023). For instance, *hallucination would be desired in Creative Writing, but not in Question-Answering*. This means that **depending on the downstream task, the reward function φ must have varying characteristics**. Collecting human preferences for all such tasks is impractical.

Motivated to resolve this need, we propose a novel reward modeling methodology, significantly reducing preference data requirements. We draw on the insight that φ is dependent on the downstream task and, hence, can utilize its task/domain[2] knowledge. Specifically, *φ lies in a low-dimensional manifold, whose dimensions can be deduced using domain knowledge*. Such an inductive bias *reduces the number of samples*[3] needed to train 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*. Specifically, we say that φ_τ (reward model for a domain τ) can be modelled by some numeric features $v_1, v_2, \cdots, v_n$. These n features fully characterize[4] the outputs from the LLM on some input. Thus, instead of training φ_τ on the text data $(\{(X, Y_w, Y_l) \mid Y_w \succ Y_l\})$, we use the n features. Such a formulation for φ **brings interpretability**—which features influence human preference the most (Section 6), and is **free from Alignment Tax** (degradation of language capabilities of an LLM post reward modeling; Bai et al. (2022a)) as we do not use an LLM to model φ.

We experimentally prove our hypothesis 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. In addition to advancing SOTA, we also analyze how our approach helps the model achieve alignment with human values for Opinion Summarization (Section 6).

Our contributions are:

1. A novel Reward Modeling 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[5] the dataset size by $> \mathbf{21}\times$. Our approach advances SOTA: at least $\sim$ **4**-point ROUGE-L improvement (Tables 1, 4 and 5; Section 5.2), and humans prefer our models' outputs **68**% over SOTA (Figure 1; Section 5.2).

2. Two new datasets: PROMPTOPINSUMM and OPINPREF. PROMPTOPINSUMM includes reviews and summaries for 25763 products (229521 summaries), for training and validation. OPINPREF is a gold-standard human preference dataset (with 940 instances) in the domain of Opinion Summarization.

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 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 an effective solution—by incorporating Reward Models, which reflect latent reward models within 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 (LLMs) as preference data generators. While this is a scalable approach to steering LMs, there is no guarantee that the preference dataset generated by LLMs reflects human goals. In our work, we propose a different solution, which promises to use human preference data but provides a way to reduce the required size

[2] We use task and domain interchangeably in the paper.

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

[4] Example of such characterization: Features like *fluency*, *coherence*, etc. can characterize text generated by an LLM.

[5] As compared to the smallest publicly available preference data. The smallest publicly available preference data is not in the domain of Opinion Summarization.

drastically. 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 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 LLM (`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 for Generic Text Summarization (Wang et al., 2023; Taori et al., 2023; Peng et al., 2023). 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, where $N-1$ reviews are used as input, and the left out review is used as a pseudo-summary (Section 2). Although these self-supervision datasets have helped further Opinion Summarization research, the approach has several shortcomings: the summaries always present a one-person rather than the consensus view, the summaries are reviews and might not have good coverage of aspects and opinions, etc. We move away from self-supervision to overcome these shortcomings and propose a new dataset. In the rest of this Section, we describe (*a*) PROMPTOPINSUMM: a new dataset to train Opinion Summarizers, (*b*) the benchmarks we used for evaluation, and (*c*) OPINPREF: gold-standard preference dataset for Opinion Summarization.

3.1 PROMPTOPINSUMM Dataset

We prompt the instruction-tuned `Mistral-7B` model (Jiang et al., 2023) to generate an opinion summary given product reviews. 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 due to cost. Appendix G includes examples and qualitative analysis. We use the Amazon dataset (He and McAuley, 2016), which has reviews for $\sim 180k$ products. We randomly sample reviews for 20763 products for `train` set and 5000 products for `validation` set. Specifically, we prompt the model to generate opinion summaries of 3 different qualities: Good (codenamed GOOD-SUM), Slightly Bad (codenamed SBAD-SUM), and Very Bad (codenamed VBAD-SUM). We generate multiple opinion summaries (3 at most) per quality. We provide reasoning for generating multiple summaries of different qualities in the extended discussion of our approach (Appendix B). We generate 184620 summaries for `train` set and 44901 summaries for `validation` set (see Appendix G).

3.2 Benchmarks for Evaluation

We use 9 Opinion Summarization benchmarks for evaluation. 3 of these benchmarks are the Amazon test set (Bražinskas et al. (2020), codenamed AMAZON), the Oposum+ test set (Amplayo et al. (2021), codenamed OPOSUM+) and the Flipkart test set (Siledar et al. (2023b), codenamed FLIPKART). AMAZON has reviews for 32 products from 4 domains, OPOSUM+ has reviews for 60 products from 6 domains and FLIPKART has reviews for 147 products from 3 domains.

Although these 3 benchmarks have been used widely, they have several shortcomings. For instance, AMAZON was developed by asking annotators to write a summary in first-person point-of-view. This causes problems such as summaries seeming personal rather than consensus opinions (which can include mixed sentiment), incomplete coverage of aspects and opinions, etc. Thus, using such *pseudo-summaries* for reference-based evaluations (ROUGE, BERTSCORE) on such a benchmark is not a correct portrayal of the models' performances. We highlight the shortcomings in detail in Appendix D. Siledar et al. (2024) recently provided 6 new benchmarks (AMAZON-R, AMAZON-RDQ, OPOSUM-R, OPOSUM-RDQ, FLIPKART-R, FLIPKART-RDQ) which are revamped versions (by getting rid of the shortcomings) of the aforementioned 3 benchmarks. We primarily rely on these 6 for our conclusions. Appendix D includes more details on domains and summary statistics.

3.3 OPINPREF Dataset

We create OPINPREF by asking humans to rank opinion summaries for given reviews. We utilize domain experts (annotator details in Appendix I) to perform the annotation. We believe that aligning with the internal reward model of domain experts would lead to better opinion summaries. We provide the domain expert with product reviews 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. We observe a Fleiss' Kappa (κ) score of 62.67% (substantial aggrement; aggrement is substantial when $60\% \leq \kappa < 80\%$). Appendix H includes statistics on the dataset.

4 Efficient Reward Modeling

We highlighted in Section 1 how the reward model (φ) can depend on the downstream task. Such dependence necessitates task/domain-specific human preference datasets, which are costly and time-consuming to create. This creates an obstacle in employing RLHF in task/domain-specific setups, thus hindering the steering of LLMs towards task/domain-specific human values.

We solve this challenge by leveraging domain knowledge. **The key insight is that we can use the domain knowledge to impart some inductive biases into the mathematical modeling of φ.** This would significantly reduce the amount of data required for training φ. Specifically, we say that φ_τ (reward model for a domain τ) can be modelled by some numeric features $v_1, v_2, \cdots, v_n$. These n features fully characterize[6] the outputs from the LLM on some input. Thus, instead of training φ_τ on the text data ($\{(X, Y_w, Y_l) \mid Y_w \succ Y_l\}$), we use the n features. Such a formulation for φ also brings interpretability and frees φ from Alignment Tax.

In Section 4.1, we detail our technique for the task/domain of E-Commerce Opinion Summarization—the task of summarizing user reviews for a product. Typically, user reviews discuss several aspects of a product and opinions/sentiments towards these aspects. An opinion summary must reflect all the aspects discussed by the input reviews and the opinions expressed towards these aspects. We discuss how we leverage such desirable properties to model φ.

4.1 Inducing Domain Knowledge

We identify desirable properties in an opinion summary with the help of domain experts[7]. We held multiple discussions to finalize the set of desirable properties. We show that **these properties are correlated to humans' judgement of summary** in Appendix A (Table 3). Based on these properties, we model φ_{op} (reward model for opinion summarization) as: $\varphi_{op} = f(v)$, where $v \in \{$`aspect-coverage`, `opinion-faithfulness`, `opinion-coverage`, `conciseness`, `relevance`, `hallucination`, `language-correctness`$\}$. The features `aspect-coverage`, `opinion-faithfulness` and `opinion-coverage` check if the generated opinion summary covers all mentioned aspects and opinions faithfully. The features `conciseness`, `relevance` , and `hallucination` check if the generated summary is concise, relevant to the input reviews, and is free from hallucination. The feature `language-correctness` checks if the generated text follows the language rules. We provide more details in Appendix A. These features, together, characterize the goodness of an opinion summary. We instruct `Mistral-7B` (Appendix A) to generate values for these features for an opinion summary, given reviews. We denote this transformation (from reviews and summary to 7 features) using Φ.

We train φ_{op} using OPINPREF, which is 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

[6] Example of such characterization: Features like *fluency*, *coherence*, etc. can characterize text generated by an LLM.

[7] Domain experts are from an E-Commerce platform.

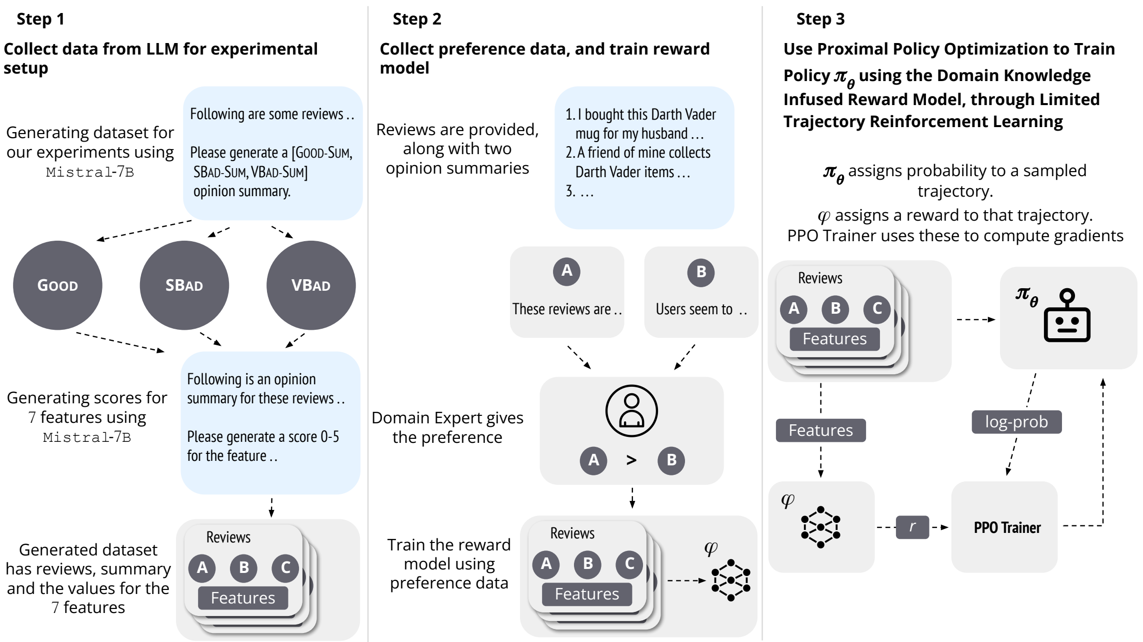

Figure 2: Overview of our approach. Step-1: We generate a new dataset for training Opinion Summarizers: PromptOpinSumm, by prompting `Mistral-7B` model. Again, we use `Mistral-7B` to compute values for the 7 features discussed in Section 4.1. Step-2: We ask humans (domain experts) for their preference, given reviews and two opinion summaries (A, B). We use the preference data and the features 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.

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; only φ_{op} is trainable, Φ is not).

After such an efficient reward modeling, we use φ_{op} for regular RLHF training (Appendix B) to get an Opinion Summarizer aligned with human goals. We illustrate the whole flow in Figure 2.

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

5 Experiments

We test our technique against the State-of-the-Art (SOTA) models, and strong Reinforcement Learning (RL) and RLHF baselines (our design and contemporary works). We list the questions we attempt to answer (through the experiments) in Section 5.1. We conduct automatic, human, and GPT-4 evaluations to verify our claim. 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).

5.1 Models & Objectives

We train the following models:

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 feature values 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 trained on the implicit preference Good-Sum $\succ$ SBad-Sum $\succ$ VBad-Sum. Kim et al. (2023) show that RLSF is an effective surrogate for RLHF when no human preference data is available. We train this reward model using Equation 1 too.

Inductive-Bias: This RLHF model is trained following our hypothesis (*infusing domain knowledge into φ to reap benefits of RLHF with modest human preference data*). We train φ_{op} using OpinPref dataset.

With these models, we ask the following questions in our experiments:

Scene-I: How effective is our technique (*infusing domain knowledge into φ to reap benefits of RLHF with modest human preference data*) over and above the usage of a good training dataset?

A comparative evaluation of SUPERVISED and INDUCTIVE-BIAS would answer this.

SCENE-II: How effective is our technique over and above vanilla RL? A comparative evaluation of NAIVEMEAN and INDUCTIVE-BIAS would answer this.

SCENE-III: How effective is our technique over contemporary RLHF techniques, which work without preference data? A comparative evaluation of SYNTH-FEEDBACK and INDUCTIVE-BIAS would answer this.

SCENE-IV: How effective is our technique, agnostic of the preference data? This question is raised to answer whether the gains are solely due to the good quality of OPINPREF, or the approach. A comparative evaluation between DPO (Rafailov et al. (2023), which uses OPINPREF in a supervised fashion) and INDUCTIVE-BIAS would answer this.

In addition to the above questions, we also check how our models fare against the SOTA (OP-SUM-GEN: Siledar et al. (2023a), MEDOS: Siledar et al. (2024), etc.). We *do not* use vanilla RLHF (Ziegler et al., 2019; Bai et al., 2022a) as a baseline, as it requires huge human preference data. Given that the goal of the paper is not to propose a new RLHF technique, but rather to propose a way to use RLHF with modest human preference annotations, omitting vanilla RLHF as a baseline does not affect our conclusions in any way.

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 approach. We include implementation details in Appendix E.

5.2 Evaluation Results

We test our approach on **9 benchmarks** (Section 3.2). In the main manuscript, we report automatic evaluation results on Amazon-based benchmarks (Table 1), human evaluation on the AMAZON benchmark, and GPT-4 evaluations on AMAZON, FLIPKART and OPOSUM+ benchmarks. Liu et al. (2023) show that GPT-4 evaluations correlate well with human evaluations for summarization; hence, in the interest of time and monetary expense, we resort to GPT-4 evaluation. We include automatic evaluation results for the rest of the benchmarks (Tables 4 and 5), and BERTSCORE based evaluations (Table 6) for all the benchmarks in Appendix C. We also include model generations for a randomly sampled product in Appendix C (Table 14). Due to the shortcomings highlighted in Section 3.2 and Appendix D, we complement our automatic evaluations of AMAZON, FLIPKART and OPOSUM+ with human and GPT-4 evaluations.

***Automatic Evaluation*.** From Table 1, we see that our proposed models are always better than the SOTA for AMAZON-R and AMAZON-RDQ. Supervised Fine Tuning on PROMPTOPINSUMM (SUPERVISED model) helps achieve significantly better ROUGE scores. This highlights the efficacy of our proposed PROMPTOPINSUMM dataset. From the automatic evaluations on AMAZON-R and AMAZON-RDQ, we see the following things:

Answer to SCENE-I: We see that INDUCTIVE-BIAS achieves gains over SUPERVISED. This answers the question in SCENE-I: Our technique is effective over and above using a good dataset.

Answer to SCENE-II: We see that INDUCTIVE-BIAS achieves gains over NAIVEMEAN. This answers the question in SCENE-II: Our technique is effective over vanilla RL.

Answer to SCENE-III: We see that INDUCTIVE-BIAS achieves gains over SYNTH-FEEDBACK. This answers the question in SCENE-III: Our technique is effective over the SOTA RLHF technique, which works without human preference data.

Answer to SCENE-IV: We see that INDUCTIVE-BIAS achieves gains over DPO. This verifies that gains of INDUCTIVE-BIAS can be safely attributed to the approach (not just the quality of OPINPREF).

***Human/*GPT-4 *Evaluation*.** We conduct human evaluation (Figure 1) for the AMAZON benchmark, using 3 domain experts (details in Appendix I). We observe a Fleiss' Kappa (κ) score of 56.25% (moderate agreement; agreement is moderate when $40\% \leq \kappa < 60\%$). We ask the experts to rank the summaries (anonymized and shuffled) given the reviews. 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 (SOTA) models and ground truth summaries. We include ground truth summaries in the evaluation to verify our claims about the quality of the benchmarks. From Figure 1, we see that INDUCTIVE-BIAS wins significantly over the competitors, further proving the

	Model-Code	AMAZON			AMAZON-R			AMAZON-RDQ		
		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	–	–	–	–	–	–
	CopyCat (Bražinskas et al., 2020)	31.97	5.81	20.16	20.09	1.79	12.94	20.54	1.94	13.85
	PlanSum (Amplayo and Lapata, 2020)	32.87	6.12	19.05	20.49	1.76	12.44	19.09	1.58	12.02
	MultimodalSum (Im et al., 2021)	34.19	7.05	20.81	21.43	1.58	13.20	20.39	2.08	12.83
	OP-SUM-GEN (Siledar et al., 2023a)	**35.46**	7.30	**21.50**	–	–	–	–	–	–
	MEDOS (Siledar et al., 2024)	34.63	**7.48**	20.97	23.92	2.27	14.69	25.44	4.16	16.45
Ours'	DPO	23.96	4.54	14.27	26.37	4.25	15.03	25.13	3.84	14.86
	SUPERVISED	28.99	4.90	16.91	32.52	5.96	18.07	30.46	5.49	17.63
	NAIVEMEAN	28.08	4.81	16.77	**34.0**	6.30	18.81	30.97	5.25	18.36
	SYNTH-FEEDBACK	29.39	4.68	17.35	33.62	6.06	18.61	30.65	5.23	18.11
	INDUCTIVE-BIAS	28.41	4.65	16.90	33.95	**6.40**	**19.23**	**31.89**	**5.78**	**18.84**

Table 1: Reference-based Evaluation Results (R-1: ROUGE-1, R-2: ROUGE-2, R-L: ROUGE-*L*) for the AMAZON, AMAZON-R and AMAZON-RDQ benchmarks. We see the following things: (*a*) Our proposed dataset (PROMPTOPINSUMM) leads to *marked increased over the SOTA* (by ~ 4 R-L points), (*b*) INDUCTIVE-BIASproves to be the *winner in all the four scenarios*: SCENE-I, SCENE-II, SCENE-III and SCENE-IV (Section 5.1), *proving the efficacy of our technique*. We also see that for the AMAZON benchmark, our models lag behind. However, *this is expected*, as we highlight in Section 3.2.

efficacy of our technique.

We run GPT-4 evaluations for AMAZON, FLIPKART and OPOSUM+ benchmarks (Figures 3, 4, 5). We run GPT-4 evaluations for AMAZON, as the agreement in human evaluation was moderate. We arrive at the same conclusions as human evaluation. We prompt GPT-4 to rank the summaries (anonymized and shuffled) given the reviews. As before, we compute the fraction of wins, ties, and losses. Again, we see that INDUCTIVE-BIAS remains a clear winner.

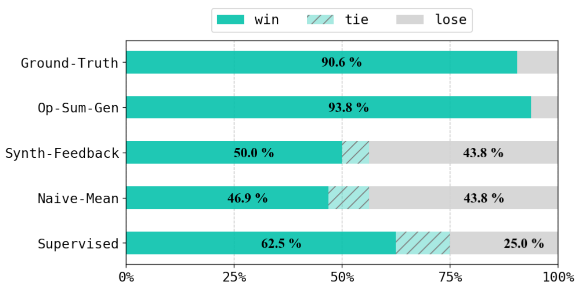

Figure 3: GPT-4 Eval: Pairwise win-tie-loss percentage of INDUCTIVE-BIAS model vs. competitors, for AMAZON benchmark.

6 Analysis

We perform a two-fold analysis: (*a*) First, we see the domain knowledge features influence for φ_{op}, (*b*) Second, we see how the ground truth summary and summary from trained models fare on the domain knowledge features. This two-fold analysis helps us understand: (*a*) which features influence

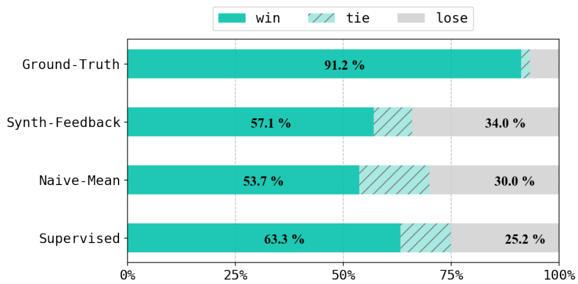

Figure 4: GPT-4 Eval: Pairwise win-tie-loss percentage of INDUCTIVE-BIAS model vs. competitors, for FLIPKART benchmark. Note that for the FLIPKART benchmark, we do not have results from OP-SUM-GEN, as Siledar et al. (2023a) only provide aspect-specific summaries.

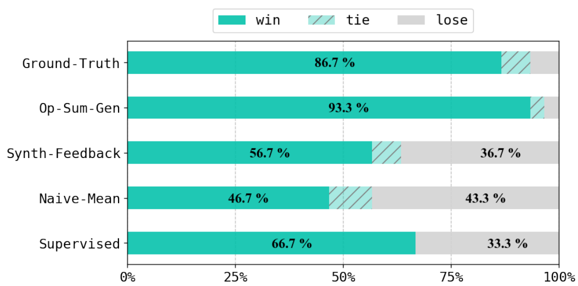

Figure 5: GPT-4 Eval: Pairwise win-tie-loss percentage of INDUCTIVE-BIAS model vs. competitors, for OPOSUM+ benchmark.

the latent reward model within humans[8] the most, and (*b*) how the ground truth summary and summary from trained models fare on these influential

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

features. Performing well 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[9] 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 2).

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

Figure 6 highlights the features' relative influence. We see that `hallucination` is most influential. This aligns with what our human preference annotators report—hallucination in summary is the primary cause of rejection. We see that hallucinations are mostly within the opinions in the summary. This is also reflected in Figure 6: `opinion-faithfulness` has significant influence. We also see that annotators prefer summaries with more specifics, i.e. they include more aspects: `aspect-coverage` has significant influence.

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 the AMAZON benchmark in the main manuscript, we include the rest in Appendix J. Table 2 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.

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): *A domain-knowledge infused reward model (φ_{op}) can help achieve alignment with latent reward model of humans for a task, with modest human preference annotations.*

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

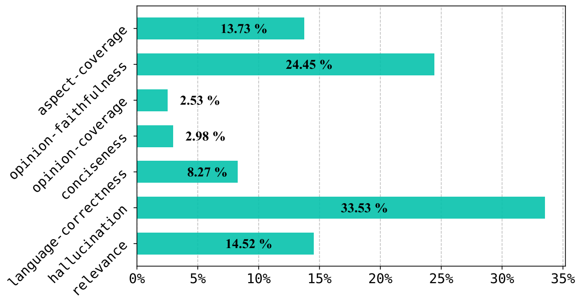

Figure 6: Relative Influence of all features in φ_{op}. All the influences sum to 1.

Models	AC ↑	OPF ↑	RE ↑	HL ↑
IB	**3.60**	**3.93**	**4.06**	**4.07**
SF	3.43	3.73	4.04	3.94
NM	3.57	3.91	4.04	3.09

Table 2: Scores on domain knowledge-based features (AC: `aspect-coverage`, RE: `relevance`, OPF: `opinion-faithfulness`, HL: `hallucination`) on the AMAZON benchmark for top-3 models (IB: INDUCTIVE-BIAS, NM: NAIVEMEAN, SF: SYNTHFEEDBACK). Note that for `hallucination`, Φ gives a higher score for less hallucination in the text.

7 Summary, Conclusion and Future Work

In this work, we propose a novel Reward Modeling 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 Modeling is a viable solution to reduce human preference annotations for downstream tasks. In the future, we will 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, to the best of our knowledge, 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 see empirically that Φ works well for Opinion Summarization, to extract the scores for the 7 features. However, there is no guarantee that such out-of-the-box performance would be reflected in another domain. Some fine-tuning might be necessary.

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.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. 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.

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.

Tejpalsingh Siledar, Rupasai Rangaraju, Sankara Sri Raghava Ravindra Muddu, Suman Banerjee, Amey Patil, Sudhanshu Shekhar Singh, Muthusamy Chelliah, Nikesh Garera, Swaprava Nath, and Pushpak Bhattacharyya. 2024. Product description and qa assisted self-supervised opinion summarization.

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 Modeling

We use 7 domain specific features for the reward model φ_{op}. We identify these features after extensive discussions with the domain experts. For each feature we prompt `Mistral-7B` to generate a score within 0 and 5. We give elongated instructions, including rough rubriks, and rules to generate the scores. This is a reason why we use an instruction-tuned model. 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.

We also analyze how these features correlate with humans' judgement of goodness of opinion summaries. We do this by looking at the scores for these features for preferred and dis-preferred summaries in the OPINPREF dataset. In Table 3, we see that the preferred summaries clearly have a higher score on all the features, than the dis-preferred ones. This shows that the scores correlate well with humans' judgement of goodness.

Feature	Pref.	Dis-pref.
`aspect-coverage` (↑)	3.69	2.84
`opinion-faithfulness` (↑)	4.02	3.05
`opinion-coverage` (↑)	3.92	3.22
`conciseness` (↑)	4.05	3.44
`relevance` (↑)	4.10	3.10
`hallucination` (↑)	3.99	2.79
`language-correctness` (↑)	4.50	3.32

Table 3: Scores for the domain knowledge based features. We see that for all the features, the human preferred (Pref.) summaries have higher scores than the ones rejected by humans (Dis-pref.). This shows that these features correlate well with humans' judgement of goodness of an opinion summary.

B 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 3). 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.

$$\mathcal{L}_{PPO} = -\mathbb{E}_{(R,s,\Phi(s))}\Bigg[\varphi_{op}(\Phi(R, s)) - \beta \log\left(\frac{\pi_\theta^{RL}(s|R)}{\pi^{SFT}(s|R)}\right)\Bigg] \quad (3)$$

Model-Code		Flipkart			Flipkart-R			Flipkart-RDQ		
		R-1 ↑	R-2 ↑	R-L ↑	R-1 ↑	R-2 ↑	R-L ↑	R-1 ↑	R-2 ↑	R-L ↑
	Medos (Siledar et al., 2024)	25.97	**5.29**	16.05	26.29	4.03	16.59	22.92	4.30	16.35
Ours'	DPO	28.85	4.10	15.55	34.23	7.86	18.62	29.96	5.25	17.28
	Supervised	27.38	4.09	15.37	39.32	10.52	22.56	32.25	6.88	19.04
	NaiveMean	28.34	4.38	**16.20**	**40.56**	10.68	22.74	32.57	6.67	19.39
	Synth-Feedback	26.37	4.18	15.48	38.77	10.99	22.97	31.04	6.98	18.59
	Inductive-Bias	**27.42**	4.21	15.71	39.10	**11.03**	**23.30**	**33.08**	**7.30**	**19.46**

Table 4: Reference-based Evaluation Results (R-1: Rouge-1, R-2: Rouge-2, R-L: Rouge-L) for the Flipkart, Flipkart-R and Flipkart-Rdq benchmarks. We see the following things: (*a*) Our proposed dataset (PromptOpinSumm) leads to *marked increased over the SOTA* (Medos; by $\sim$ 6 R-L points), (*b*) Inductive-BiasProves to be the *winner in all the four scenarios*: Scene-I, Scene-II, Scene-III and Scene-IV (Section 5.1), *proving the efficacy of our technique*. We also see that for Flipkart benchmark, despite the shortcomings, our models perform similar to the SOTA.

Model-Code		Oposum+			Oposum-R			Oposum-RDQ		
		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)	26.25	4.62	16.49	–	–	–	–	–	–
	CopyCat (Bražinskas et al., 2020)	27.98	5.79	17.07	22.41	2.30	13.94	22.38	2.03	14.06
	PlanSum (Amplayo and Lapata, 2020)	30.26	5.29	17.48	22.37	2.05	13.32	22.64	2.25	13.71
	MultimodalSum (Im et al., 2021)	33.08	7.46	19.75	23.35	2.98	14.53	23.73	2.80	14.70
	Op-Sum-Gen (Siledar et al., 2023a)	36.44	8.50	22.03	25.65	3.56	15.83	24.66	3.25	15.54
	Medos (Siledar et al., 2024)	**36.57**	8.79	**21.35**	26.82	3.67	15.92	26.32	3.34	16.10
Ours'	DPO	27.64	7.34	16.50	33.69	6.62	18.55	30.95	5.89	17.60
	Supervised	30.57	8.02	16.90	38.32	9.10	20.35	35.69	8.17	19.28
	NaiveMean	31.47	8.0	16.99	40.16	9.84	21.74	35.90	8.33	20.13
	Synth-Feedback	31.66	**8.86**	17.91	41.32	**10.40**	**22.23**	**37.85**	8.94	20.71
	Inductive-Bias	31.15	8.15	17.46	**41.58**	10.32	22.02	37.56	**9.21**	**20.88**

Table 5: Reference-based Evaluation Results (R-1: Rouge-1, R-2: Rouge-2, R-L: Rouge-L) for the Oposum+, Oposum-R and Oposum-Rdq benchmarks. We see the following things: (*a*) Our proposed dataset (PromptOpinSumm) leads to *marked increased over the SOTA* (Medos; by $\sim$ 6 R-L points), (*b*) Inductive-BiasProves to be the *winner in almost all of the four scenarios*: Scene-I, Scene-II, Scene-III and Scene-IV (Section 5.1), *proving the efficacy of our technique*. We also see that for Oposum+ benchmark, our models lag behind. However, *this is expected*, as we highlight in Section 3.2.

C Additional Automatic Evaluation Results

In addition to the Amazon-based benchmarks (Table 1), we also report results for Flipkart and Oposum+ based benchmarks (Tables 4 and 5). As before, we see that Inductive-Bias is almost always the winner. As before, we draw similar conclusions for Scene-I, Scene-II and Scene-III: Inductive-Bias wins, further strengthening the conclusion that our methodology is effective. We also see that, inspite of the shortcomings of the Flipkart benchmark, our models perform similar to the SOTA.

We also include BertScore evaluations for all the 9 benchmarks in Table 6. We see similar trends as Rouge Evaluation: our models are significantly better than the SOTA in majority of the benchmarks.

For a qualitative understanding, we include generations from several models on a randomly picked sample from the Amazon benchmark in Table 14.

D Details on the Benchmark Datasets

In this section we discuss details about the benchmarks, such as the domain of the products, summary statistics and finally highlight some shortcomings in the Amazon, Oposum+ and Flipkart datasets. Amazon has reviews for 32 products from 4 domains: "*electronics*", "*home & kitchen*", "*personal care*", and "*clothing, shoes & jewellery*". Oposum+ has reviews for 60 products from 6 domains: "*laptop bags*", "*bluetooth headsets*", "*boots*", "*keyboards*", "*television*", and "*vacuums*". Flipkart has reviews for 147 products from 3 domains: "*laptops*", "*mobiles*", and "*tablets*". Table 7 includes summary statistics for

Model Code	AMAZON	AMAZON-R	AMAZON-RDQ	OPOSUM+	OPOSUM-R	OPOSUM-RDQ	FLIPKART	FLIPKART-R	FLIPKART-RDQ
OP-SUM-GEN (Siledar et al., 2023a)	**88.78**	86.94	86.76	**86.63**	86.96	86.95	–	–	–
DPO	86.45	86.60†	86.37†	84.39	87.35*	86.90	83.75	86.61	85.40
SUPERVISED	87.79	88.23*	87.76*	85.13	88.59*	88.02*	84.21	88.11	86.40
NAIVEMEAN	87.95	88.29*	87.81*	85.25	88.96*	88.39*	84.32	88.29	86.52
SYNTH-FEEDBACK	87.81	88.28*	87.74*	85.22	89.08*	88.45*	84.27	88.28	86.49
INDUCTIVE-BIAS	87.98	**88.41***	**88.16***	85.33	**89.09***	**88.46***	**84.33**	**88.34**	**86.61**

Table 6: BERTSCORE evaluation results on the 9 benchmark datasets. We observe a similar trend as ROUGE evaluations: SOTA is better than our models for the AMAZON and OPOSUM+ benchmarks, which is expected (Section 3.2). For the rest of the datasets, we see that our models are significantly better. We do not include SOTA results for Flipkart-based benchmarks, as OP-SUM-GEN only provide aspect-specific summaries for the same. * denotes gain is statistically significant compared to SOTA with significance level 1%, † denotes gain is statistically significant compared to SOTA with significance level 5%.

the 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 7: Statistics of the benchmark datasets. OPOSUM+ represents the statistics of all OPOSUM+ based benchmarks (OPOSUM+, OPOSUM-R and OPOSUM-RDQ). Similar is the case for AMAZON and FLIPKART.

Finally, now we highlight the shortcomings of the benchmark datasets in the rest of the discussion.
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.*

E 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 8. We train all of our models using 2× A100 GPUs (80GB)

Hyperparameter	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 8: 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.

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 9 includes details on the hyperparameters.

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

Table 9: 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.

F Generated Summary Lengths

We analyze the generation lengths of the models, and the ground truth summary. Table 10 lists the summary lengths. We see that theDPO model generates very verbose summary. Additionally, we also see that the INDUCTIVE-BIAS model generates very concise summaries.

Model	AMAZON	OPOSUM+	FLIPKART
Ground-Truth	60.65	85.86	129.91
NAIVEMEAN	91.09	114.67	75.48
SYNTH-FEEDBACK	80.31	115.37	71.11
OP-SUM-GEN	55.84	62.93	-
INDUCTIVE-BIAS	81.62	88.63	73.57
SUPERVISED	81.31	117.03	74.56
DPO	138.50	141.50	131.40

Table 10: Generation Length Statistics: number of words in summaries. We use NLTK to tokenize the text.

G Details on PROMPTOPINSUMM

Here we provide more details on the generated PROMPTOPINSUMM dataset. Table 13 includes summary statistics of the generated dataset. We include an example from PROMPTOPINSUMM dataset (Table 11). 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 (∼ 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.

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 11: Example summaries from PROMPTOPINSUMM.

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.

H Statistics of the OPINPREF dataset

We look at the summary statistics for the OPINPREF dataset. Table 12. We see that, interestingly, annotators prefer longer summaries—this is because these summaries contain more specifics and details from the reviews.

Characteristic	Value
# words in reviews	641.21
# reviews	13.08
# words in summaries	73.16
# words in preferred summaries	85.41
# words in unpreferred summaries	66.91

Table 12: Statistics of the OPINPREF dataset. We use NLTK to tokenize the text.

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 13: Statistics of PROMPTOPINSUMM dataset. We use NLTK to tokenize the text.

I 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.

J All Evaluation Results

We include all of the evaluation results in this section. In Tables 15, 16, 17 and 17 we include pairwise comparison results, in a `win/tie/loss` format. We also include results on evaluation on how the models perform on the domain features in Tables 19, 20 and 21.

Reviews: i really like these boots. they can be a beast to get on, like any boot fit to last; once on, they are incredibly comfortable. I have had them for 4 years and they still look great - worn in but not worn out... What great boots! They do take some breaking in - like all real leather does. Run maybe a tad small in the general spectrum, but hey, remember the break in period. Look fabulous, will last me for the next twenty years.... They bridge the fashion vs. function quandary beautifully. Every woman needs a pair of these Frye boots! Fit is true to normal size. I wear a standard 9. These fit in the shaft for any ladies who have bigger calves as I do. You have to fight to get them on first few times and they wear stiff until fully broken in. Worth the money and break in time. Boots excellent quality, beautiful skin, a modern model. But! they are very narrow! This is the second pair of shoes is a brand and my 8.5 size fits me perfectly, even he little more than usual. But these are very narrow! go to a cobbler, let him try to stretch. Be careful! bought these for the wife to have something to kick around in and keep up with the kids, but she doesn't like them. Her heel lifts with every step, and they're stiff and heavy. Not an everyday casual shoe. Don't get me wrong, they're really good boots, but they used to make them better. My last pair from a few years ago were much higher quality. Thicker sole, better full grain leather, now they only make the black in 'grease', which stains all my socks black. Keeping them, but still disappointed in Frye. i bought it just for the heavy snow cuz my UGG got wet through.even though it is a little expensive, but the boots are nice and it was shipped to me for only two days.!!!!!! i love it They were comfortable the moment I put them on. The sizes are exact. The color the way they are made are what I expect from Frye which I've been wearing since the 70's. With a 20year break from not wearing Frye it was a pleasant surprise the quality has stood the test of time.
SUPERVISED: These Frye boots are a popular choice for women due to their comfort and durability. However, they can be a bit challenging to put on and require a break-in period. Some users have reported that they run small in size and are narrow, requiring a cobbler's assistance for a proper fit. The leather quality is generally good, but some users have noted that the leather may not last as long as they had hoped.
NAIVEMEAN: Users generally appreciate the comfort and durability of the Frye boots, with some mentioning the need for a break-in period. Sizes run true to normal, but some users find them to be a bit narrow. The boots are praised for their stylish appearance and ability to bridge the fashion and function quandary. However, some users have noted that they take some time to break in and are not suitable for everyday casual wear.
SYNTH-FEEDBACK: These Frye boots are praised for their comfort and durability, with some users noting the need for a break-in period. However, some users have noted that they run small in size and require stretching, while others find them narrow and stiff. The leather is described as beautiful and of good quality, with a modern model. Some users have had them for several years and have not worn them out.
INDUCTIVE-BIAS: These Frye boots are praised for their comfort and durability, with users expressing satisfaction with their fit and appearance. However, some users have noted that they require a break-in period and can be a bit narrow. The leather quality is praised for its durability and modern model, with some users expressing disappointment with the lack of improved quality in recent years.

Table 14: Example generation (randomly sampled) for some input reviews from all the models. Olive implies faithful/correct generation, while red indicates hallucinated text, or repetition. We see that only INDUCTIVE-BIAS is free from red text. The model closest in performance to INDUCTIVE-BIAS, the NAIVEMEAN model, misses out on two aspects: `leather-quality` and `quality-degradation`. INDUCTIVE-BIAS covers both, while being concise. We do not include DPO model in this comparison, as it was too verbose.

·	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 15: Pairwise Win/Tie/Loss Results for all models in Human Evaluation for AMAZON benchmark. We format the data as: `win/tie/loss`, `win` specifies how many time the *row* won over the *column*.

·	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 16: Pairwise Win/Tie/Loss Results for all models in GPT-4 Evaluation for AMAZON benchmark. We format the data as: `win/tie/loss`, `win` specifies how many time the *row* won over the *column*.

·	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 17: Pairwise Win/Tie/Loss Results for all models in GPT-4 Evaluation for FLIPKART benchmark. We format the data as: `win/tie/loss`, `win` specifies how many time the *row* won over the *column*.

·	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 18: 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 19: Intrisic Evaluation results on the AMAZON benchmark for all the models. Legend: AC: `aspect-coverage`, OPF: `opinion-faithfulness`, OPC: `opinion-coverage`, CC: `conciseness`, RE: `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	$\mathbf{4.33 \pm 0.45}$	$\mathbf{4.61 \pm 0.29}$
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 20: Intrisic Evaluation results on the FLIPKART benchmark for all the models. Legend: AC: `aspect-coverage`, OPF: `opinion-faithfulness`, OPC: `opinion-coverage`, CC: `conciseness`, RE: `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	$\mathbf{3.54 \pm 0.22}$	$\mathbf{3.50 \pm 0.06}$	$\mathbf{3.57 \pm 0.06}$	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 21: Intrisic Evaluation results on the OPOSUM+ benchmark for all the models. Legend: AC: `aspect-coverage`, OPF: `opinion-faithfulness`, OPC: `opinion-coverage`, CC: `conciseness`, RE: `relevance`, HL: `hallucination`, LC: `language-correctness`.