Fairer Preferences Elicit Improved Human-Aligned Large Language Model Judgments

Han Zhou[1] **Xingchen Wan**[2*] **Yinhong Liu**[1] **Nigel Collier**[1]
Ivan Vulić[1] **Anna Korhonen**[1]
[1]Language Technology Lab, University of Cambridge
[2]Machine Learning Research Group, University of Oxford
{hz416, yl535, nhc30, iv250, alk23}@cam.ac.uk

Abstract

Large language models (LLMs) have shown promising abilities as cost-effective and reference-free evaluators for assessing language generation quality. In particular, pairwise LLM evaluators, which compare two generated texts and determine the preferred one, have been employed in a wide range of applications. However, LLMs exhibit preference biases and worrying sensitivity to prompt designs. In this work, we first reveal that the predictive preference of LLMs can be highly brittle and skewed, even with semantically equivalent instructions. We find that *fairer* predictive preferences from LLMs consistently lead to judgments that are better aligned with humans. Motivated by this phenomenon, we propose an automatic **Z**ero-shot **E**valuation-oriented **P**rompt **O**ptimization framework, ZEPO, which aims to produce fairer preference decisions and improve the alignment of LLM evaluators with human judgments. To this end, we propose a zero-shot learning objective based on the preference decision *fairness*. ZEPO demonstrates substantial performance improvements over state-of-the-art LLM evaluators, *without* requiring labeled data, on representative meta-evaluation benchmarks. Our findings underscore the critical correlation between preference fairness and human alignment, positioning ZEPO as an efficient prompt optimizer for bridging the gap between LLM evaluators and human judgments.

1 Introduction

Large language models (LLMs) (Brown et al., 2020; OpenAI, 2023; Anil et al., 2023a,b) have become the standard machinery for evaluating the quality of natural language generation over various aspects, such as coherence, fluency, and truthfulness, in a reference-free manner (Chen et al., 2023b; Zeng et al., 2024; Zheng et al., 2024b).

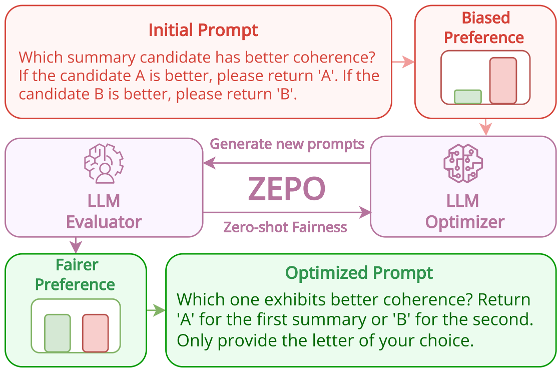

Figure 1: Illustration of the ZEPO pipeline. Given a manual prompt, the distribution of LLM preferences can be *biased* towards a certain class. ZEPO optimizes the prompt on a zero-shot fairness learning objective until the *balance* is achieved in the distribution.

Owing to the remarkable in-context learning capabilities of LLMs (Brown et al., 2020), prompting techniques further enable versatile use of LLM evaluators with user-defined evaluation criteria, where pairwise-preference-based evaluators have so far demonstrated superior human alignment to direct scoring (Liusie et al., 2024; Liu et al., 2024b).

However, LLMs have been known to exhibit preference bias (Wang et al., 2023), a priori propensity to predict certain classes over others unfairly, and display strong sensitivity to the actual prompts describing evaluation criteria (Zhou et al., 2023a; Sclar et al., 2024). The preference bias is argued to be largely due to various factors that result in a label distribution shift, such as position bias (Zheng et al., 2024b), verbosity bias (Saito et al., 2023), and contextual bias (Zhou et al., 2024a), where LLMs unfairly favor later and longer answers, or even follow repetitive answers in their demonstrations. We are thus motivated to explore the impact of preference biases on human alignment in the novel context of LLM evaluators. We start by conducting a systematic study examining the sensitivity of LLM evaluators to the provided instructions. By paraphrasing

*Now at Google. Code is available at https://github.com/cambridgeltl/zepo.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 1241–1252
November 12-16, 2024

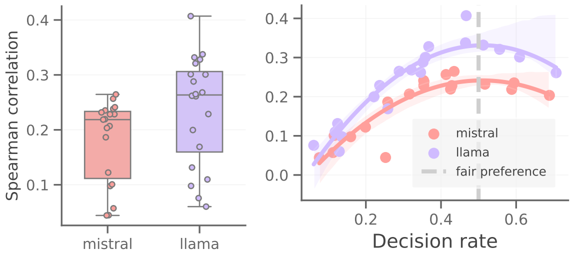

Figure 2: *LLM evaluators show strong sensitivity to instructions and fairer preference leads to better human-aligned LLM judgments.* Sensitivity and evaluation performance studies on preference fairness.

from a set of instructions, we find that the pairwise preference of LLMs largely varies even with semantically equivalent instructions, and different instructions exert different degrees of preference biases. Noticeably, we observe that fairer preferences consistently lead to better human-aligned judgments. Motivated by this empirical finding, we then propose an automatic ***Z**ero-shot **E**valuation-oriented **P**rompt **O**ptimization* (ZEPO) framework for steering LLM evaluators towards better agreements with humans; see Fig. 1. We design a new zero-shot fairness objective function by measuring the absolute difference between a uniform prior distribution and the model preference distribution. ZEPO, without any labeled data, shows substantial performance gains over state-of-the-art LLM evaluators with manually designed instructions on meta-evaluation benchmarks.

In sum, we provide the following contributions. **1)** We present a systematic analysis that reveals the strong sensitivity of LLM evaluators to instructions. Importantly, we find that *fairer preferences elicit better human-aligned LLM judgments*. **2)** We introduce a **Z**ero-shot **E**valuation-oriented **P**rompt **O**ptimization framework (ZEPO) for automatically optimizing LLM evaluators toward better human agreements without any labeled data. **3)** We demonstrate that ZEPO efficiently discovers the fairest instruction for LLM evaluators, delivering substantial gains in evaluation over representative tasks.

2 Related Work

LLMs as Evaluators. LLMs have been widely used to evaluate natural language generation tasks (Zhong et al., 2022; Chiang and Lee, 2023), enabling automatic and reference-free evaluations (Liu et al., 2023; Fu et al., 2023; Chen et al., 2023b; Dong et al., 2024). Recent studies show that LLM evaluators can serve as effective pairwise text rankers (Qin et al., 2023), where pairwise comparisons lead to better human-aligned judgments than Likert-score evaluations (Liusie et al., 2024; Liu et al., 2024b). Yet, there is still a prominent gap between LLM evaluators and human agreement (Shen et al., 2023). LLM evaluators are yet sensitive to exemplars (Wang et al., 2023) and exhibit unfair predictions due to position bias, verbosity bias, and self-preferences (Zheng et al., 2024b; Pezeshkpour and Hruschka, 2023; Panickssery et al., 2024; Liu et al., 2024a). Calibration methods have been proposed to alleviate biases (Li et al., 2023b,a; Zhou et al., 2024a), but are yet insufficient for addressing all aforementioned biases. In this work, we show that instructions exert large impacts on LLM evaluators, and searching for instructions with fairer preferences is a necessary and critical component in LLM-based evaluators.

Automatic Prompt Optimization. Unlike soft prompt tuning that requires 'white box' access to model parameters (Lester et al., 2021; Zhou et al., 2024b), hard prompt tuning directly searches for discrete prompts that are portable and 'black box' (Deng et al., 2022; Zhou et al., 2023a). Recent prompt optimization work further leverages LLMs as optimizers to generate more human interpretable prompts (Zhou et al., 2023b; Yang et al., 2024). Much effort has been devoted to more advanced search algorithms (Pryzant et al., 2023; Guo et al., 2024; Khattab et al., 2024; Wan et al., 2024; Liu et al., 2024c) but they heavily rely on labeled data. Instead, *zero-shot* prompt optimization is a rather underexplored research area, and previous work is mostly limited to entropy-based exemplar selection (Lu et al., 2022) or relies on model-synthesized data (Chen et al., 2023a). We explore the extreme, zero-shot learning setup and leverage LLM's self-predictive distribution to optimize toward fairer preferences. As we will show, our fairness objective shows the best correlation and outweighs other zero-shot metrics for LLM evaluators in Fig. 3.

3 Fairer Preferences Elicit Improved Human-Aligned Judgments

Prompt Sensitivity and Bias. We start by analyzing the sensitivity of LLM evaluators to variations in instructions. Formally, given some source text and corresponding response candidates as an input query x_i, we have the predicted label y_i as the model preference. Evaluation instruction I is formulated with the input query x_i

in a prompt template to form a complete context $C(x_i, I) = \texttt{Template}(x_i, I)$ for evaluation. LLM evaluators then make predictions by $y_i = \arg\max_{y\in\mathcal{Y}} p(y|C_i)$, where the verbalizer $\mathcal{Y}$ defines the set of preferences (i.e., A or B for pairwise preferences). To inspect prompt sensitivity, we leverage GPT-3.5 (OpenAI, 2023) to generate a set of semantically equivalent instructions $\mathcal{I} = \{I_1, ..., I_M\}$ by paraphrasing from an initial instruction I_1. In Fig. 2, we observe a severe fluctuation in human agreement scores by prompting Llama-3 8B (Touvron et al., 2023) model with $C_{I_m\in\mathcal{I}}(x, I_m)$. This reflects a high prompt sensitivity and poor robustness of standard LLM evaluators. The observation aligns with previous research in position biases (Zhao et al., 2021), and LLMs are sensitive to orders and formats of provided exemplars (Lu et al., 2022; Sclar et al., 2024).

Preference Fairness and Human Alignment. Following the previous finding, we hypothesize that the prompt sensitivity is mainly due to the preference bias incurred by spurious correlations from the instructions $\mathcal{I}$. We proceed to visualize the human agreement regarding preference distribution p_I by different instructions I across the entire query set $\{x_1, ..., x_N\}$, measured by $p_{I,A} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left(p(y_i = A|x_i, I) > p(y_i = B|x_i, I)\right)$, where $\mathbb{I}(\cdot)$ is an indicator function that counts the number of predictions that candidate *A* is preferred to *B* in pairwise evaluations. In Fig. 2, we show that the patterns are nearly perfectly fitted to a quadratic regression function, where the highest human agreement point is close to $p_I = 0.5$, and instructions with more skewed decision distributions always degrade the evaluation alignment. Therefore, p_I is a good indicator that connects decision fairness with human judgments, and instructions with *fairer* decision preferences can lead to *better* human-aligned LLM judgments.

4 ZEPO: Zero-Shot Prompt Optimization with Fairer Preferences

Zero-Shot Fairness Learning. Motivated by these findings, we now propose to automatically optimize the evaluation prompts for LLM evaluators toward fairer preferences, thereby achieving better human alignments. Importantly, the source preference distribution for an unbiased pairwise evaluator should naturally be uniform $p_S = 1/|\mathcal{Y}|$ (by the law of large numbers) given a sufficient number of randomly sampled pairwise candidates. Consequently,

Algorithm 1 ZEPO.

1: **Input**: Initial instruction prompt I; LLM optimizer $\mathcal{O}$; LLM evaluator $\mathcal{E}$; unlabeled data $\mathcal{D}$; number of classes J; number of epochs E; population size S.
2: **Output**: Optimized Instruction prompt I^*
3: Initialize the instruction $I^* \leftarrow I$.
4: **for** e in E **do**
5: Obtain new instruction candidates from the LLM optimizer $\mathcal{O}$: $\mathcal{I} \leftarrow \mathcal{O}(I^*)$, where $|\mathcal{I}| = S$.
6: **for** $I \in \mathcal{I}$ **do**
7: LLM evaluator $\mathcal{E}$ generates a preference distribution over $\mathcal{D}$ (i.e., the decision rate for each class y_i), $p_{I,y_i} = \mathcal{E}(I)$, measured by the equation in Sec. 3.
8: Compute the zero-shot fairness for each instruction candidate: $\texttt{fair}_{\mathcal{D}}(I) = -\frac{1}{J}\sum_{j=1}^{J} |\frac{1}{J} - p_{I,y_j}|$.
9: **end for**
10: Update the best instruction: $I^* \leftarrow \arg\max_{I\in\mathcal{I}} \texttt{fair}_{\mathcal{D}}(I)$.
11: **end for**
12: Return the optimized instruction I^*.

we propose a zero-shot fairness learning objective function as $\texttt{fair}_{x_i\sim\mathcal{D}}(I) = -\frac{1}{J}\sum_{j=1}^{J} |p_S - p_{I,y_j}|$ in an unsupervised set of data $\mathcal{D}$ by measuring the absolute difference between the source prior and preference distribution.

Automatic Prompt Optimization. In contrast with previous prompt optimization methods that heavily rely on labeled data, we propose ZEPO, an automatic **Z**ero-shot **E**valuation-oriented **P**rompt **O**ptimization framework. It is a more natural setup for reference-free LLM evaluations where human scores are usually unavailable in advance. ZEPO optimizes the evaluation prompts by maximizing the zero-shot fairness metric, such that $I^* = \arg\max_{I\in\mathcal{I}} \texttt{fair}_{x_i\sim\mathcal{D}}(I)$. We integrate an LLM paraphraser with a greedy search algorithm to update the instruction I iteratively, where the detailed ZEPO algorithm is shown in Algorithm 1. We refer to Appendix §A for more details on implementing ZEPO. It is worth noting that debiasing and calibration (Zheng et al., 2024a; Zhou et al., 2024a) methods can also control LLM evaluators for fairer preferences. We show in Figure 4 that ZEPO is a meta-method orthogonal to existing debiasing approaches and leads to further improvements. In addition, we report the initial (seed) prompt and ZEPO-optimized prompt with corresponding fairness scores in Table 5 and 6.

5 Experiments and Results

Datasets and Models. Following Zhong et al. (2022) and Fu et al. (2023), we evaluate ZEPO on representative meta-evaluation benchmarks, including two summarization tasks: News Room

Models	News Room				SummEval				Avg.
	COH	REL	INF	FLU	COH	FLU	CON	REL	
Other Metrics									
BertScore	0.15	0.16	0.13	0.17	0.28	0.19	0.11	0.31	0.19
GPTScore	0.31	0.35	0.26	0.31	0.28	0.31	0.38	0.22	0.30
Mistral 7B									
Scoring	0.32	0.39	0.20	0.26	0.23	0.19	0.37	0.19	0.27
G-Eval	0.36	0.36	0.24	0.39	0.25	**0.20**	**0.39**	0.25	0.31
Pairwise	0.33	**0.40**	0.19	0.19	0.06	0.01	0.07	0.16	0.18
ZEPO	**0.47***+14%*	0.38*-2%*	**0.44***+25%*	**0.48***+29%*	**0.29***+23%*	0.13*+12%*	0.32*+25%*	**0.30***+14%*	**0.35***+17%*
Llama-3 8B									
Scoring	0.42	0.41	0.30	0.29	0.35	0.23	**0.32**	**0.46**	0.35
G-Eval	0.38	0.34	0.26	0.26	0.34	0.22	0.29	0.42	0.33
Pairwise	0.49	0.51	0.46	0.45	0.24	0.12	0.30	0.21	0.35
ZEPO	**0.57***+8%*	**0.54***+3%*	**0.55***+9%*	**0.56***+11%*	**0.40***+16%*	**0.25***+13%*	0.30*+0%*	0.39*+18%*	**0.45***+10%*

Table 1: Spearman correlations on Mistral 7B and Llama-3 8B. We evaluate preference-based evaluators and direct-scoring evaluators in terms of Coherence (COH), Relevancy (REL), Informativeness (INF), Fluency (FLU), and Consistency (CON). We highlight the % improvement/degradation of ZEPO over "Pairwise" in +green/-red.

(Grusky et al., 2018) and SummEval (Fabbri et al., 2021), and one dialog task: TopicalChat (Mehri and Eskenazi, 2020) (see Appendix §A for further details). We examine ZEPO with state-of-the-art open-source LLMs, Mistral 7B (Jiang et al., 2023) and Llama-3 8B (Touvron et al., 2023).

Baselines. We provide baseline scores for reference-free evaluators in the zero-shot setup, including BERTScore (Zhang et al., 2020), GPTScore (Fu et al., 2023), and G-Eval (Liu et al., 2023). ZEPO is applicable to state-of-the-art pairwise ranking evaluators, and we report experimental results from Pairwise (Liu et al., 2024b) as the main baseline and provide direct scoring evaluation results named Scoring and G-Eval for reference.

Main Results. We present ZEPO on representative meta-evaluation benchmarks in Table 1. Notably, ZEPO yields substantial gains in alignment with human judgments over almost all aspects on the Pairwise baseline: 17% and 10% on average on Mistral 7B and Llama-3 8B, respectively. It shows that manually designed evaluation criteria and instructions (without prompt optimization) can expose strong preference bias with LLM evaluators. By conducting ZEPO on Pairwise in a zero-shot setup, the performance of pairwise evaluators can be largely recovered, outperforming fine-calibrated direct scoring and the G-Eval baselines. Furthermore, we notice that weaker models, e.g. Mistral 7B, can exhibit more catastrophic evaluations, suffering from preference biases (e.g., on COH and CON aspects in SummEval), whereas Llama-3 8B generates relatively more robust evaluations. In both cases, ZEPO constantly mitigates the preference bias and better aligns LLM evaluators. Overall, the results indicate that ZEPO is a label-free and efficient prompt optimizer for effectively aligning LLM evaluators with human judgments.

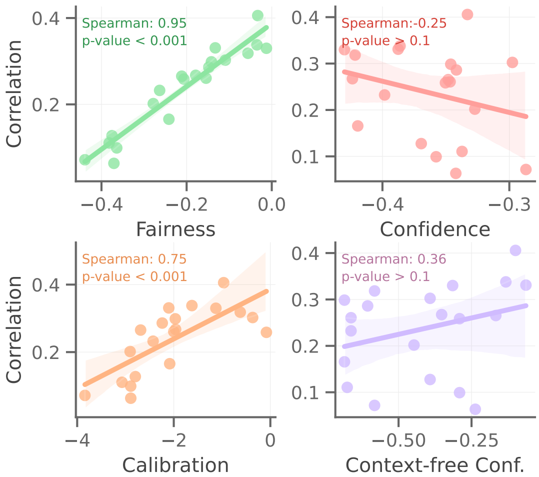

Figure 3: *Fairness shows the strongest correlation with LLM evaluation performance.* Correlation studies of zero-shot learning objectives and LLM evaluation performance. The growth of the x-axis indicates better/stronger fairness, confidence (conf.), and calibration.

Zero-shot Learning Objectives. We provide an in-depth analysis of the effectiveness of our proposed Fairness metric in comparison to other zero-shot objective functions as visualized in Fig. 3. We include model confidence, a commonly used zero-shot metric in exemplar selection (Lu et al., 2022; Wan et al., 2023a,b), measured as the negative of entropy. Calibration-based approaches have been effective in mitigating position biases (Zhao et al.,

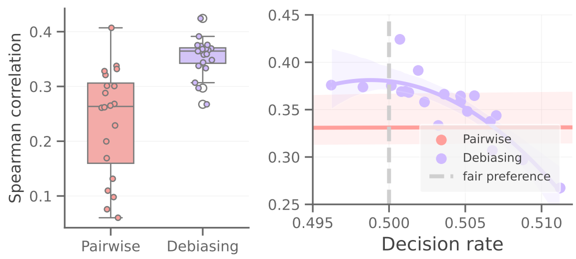

Figure 4: ZEPO *is orthogonal to debiasing approaches and brings further improved LLM judgments.* Sensitivity and evaluation performance studies on preference fairness before and after applying permutation debiasing on the COH aspect in SummEval from Llama-3 8B.

2021; Wang et al., 2023). We adopt a zero-shot calibration metric from Batch Calibration (Zhou et al., 2024a) and context-free confidence as another metric from Fair-Prompting (Ma et al., 2023), where overconfidence is argued to result in unfairness. First, `Fairness` shows the largest Spearman correlation with LLM evaluation performance, guaranteeing its effectiveness with ZEPO. Following fairness, `Calibration` is more weakly correlated, whereas `Confidence` metrics fail to serve as good objectives for ZEPO, with poorer correlations.

Complementarity with Debiasing. We further extend our study of ZEPO, focusing on its orthogonality/complementarity with debiasing approaches. We implement the *permutation debiasing* method which averages the probability for different orders/positions of the same candidates, also termed Balanced Position Calibration (Wang et al., 2023). Fig. 4 shows that the `Debias` method first improves the lower bar of the evaluation performance of LLMs. Secondly, when we inspect the preference distribution after applying `Debias`, we observe a fairer preference distribution where the decision rates become much closer to 0.5. However, LLM evaluators are still sensitive to semantically equivalent instructions even after debiasing, where the judgment alignment varies substantially from 0.26 to 0.43. In addition, we observe a similar quadratic curve in the second plot, indicating that our previous findings still hold: *fairer preferences lead to improved human-aligned LLM judgments*.

Following this observation, we conduct additional experiments on ZEPO *with* and *without* permutation debiasing. Table 2 shows that further gains can be achieved by integrating debiasing methods with prompt optimization. Therefore, we conclude that ZEPO is a meta-method on zero-

Methods	News Room COH	REL	INF	FLU	Avg.
Pairwise	0.49	0.51	0.46	0.45	0.48
ZEPO	0.57	0.54	0.55	0.56	0.56
Pairwise + Debias	0.60	**0.61**	0.64	**0.58**	0.61
ZEPO + Debias	**0.64** *+4%*	**0.61** *+0%*	**0.72** *+8%*	0.57*-1%*	**0.64***+3%*

Table 2: Spearman correlations on News Room with Llama-3 8B before and after applying permutation debiasing. We highlight the % improvement/degradation of ZEPO over "`Pairwise`" after debiasing in +green/-red.

shot prompt optimization while being orthogonal to other debiasing and calibration methods. In light of this work, we expect to build toward improved human-aligned LLM evaluators with a combination of prompt optimization, calibration, and advanced debiasing methods.

6 Conclusion

We first analyzed the relationship between preference fairness and human alignment; it revealed that LLM evaluators produce highly skewed preference distributions even with semantically equivalent instructions. We further showed that fairer preferences can yield improved human-aligned LLM judgments. Based on this insight, we proposed a zero-shot prompt optimization framework with a fairness-aware zero-shot proxy. It substantially improves alignments of pairwise LLM evaluators with humans, without any labeled data, and serves as a meta-method orthogonal to debiasing approaches.

Limitations

First, ZEPO is a zero-shot method that learns the zero-shot fairness metric from unlabeled data. It still requires a sufficient number of random *unlabeled* samples for pairwise evaluations to obtain a good estimation of preference distribution for fairness. We argue that such a data requirement is mild, as in the evaluation setup, the bottleneck lies in human-annotated labels, not unlabeled inputs. Second, ZEPO is primarily designed for preference-based evaluators, and we have widely examined the effectiveness of ZEPO in pairwise evaluations. Though pairwise evaluation appears to be the current leading standard, it is possible that future advances in LLM evaluators can achieve more efficient evaluation-by-ranking in multi-choice question formats with more than two classes, which have not been included in our current study. However, in principle, the proposed zero-shot fairness objective is a general learning metric scalable to any number of classes based on its uniform prior.

Lastly, ZEPO only integrates a basic LLM optimizer in exploring instruction candidates at a paragraph level with a greedy search algorithm. However, ZEPO is a meta-framework also orthogonal to LLM optimizers with more advanced search algorithms, and this synergy warrants further investigation in future work. ZEPO serves as a first step towards LLM evaluation with fairer preferences and is easy to extend with more exploitation-driven LLM optimizers in alternative search spaces.

Acknowledgements

The work has been supported by the UK Research and Innovation (UKRI) Frontier Research Grant EP/Y031350/1 (the UK government's funding guarantee for ERC Advanced Grants) awarded to Anna Korhonen at the University of Cambridge. The work has also been supported in part by a Royal Society University Research Fellowship (no 221137; 2022-) awarded to Ivan Vulić, and by the UK EPSRC grant EP/T02450X/1.

References

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. 2023a. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023b. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Wei-Lin Chen, Cheng-Kuang Wu, Yun-Nung Chen, and Hsin-Hsi Chen. 2023a. Self-ICL: Zero-shot in-context learning with self-generated demonstrations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15651–15662, Singapore. Association for Computational Linguistics.

Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023b. Exploring the use of large language models for reference-free text quality evaluation: An empirical study. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 361–374, Nusa Dua, Bali. Association for Computational Linguistics.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric Xing, and Zhiting Hu. 2022. RLPrompt: Optimizing discrete text prompts with reinforcement learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3369–3391, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yijiang River Dong, Tiancheng Hu, and Nigel Collier. 2024. Can llm be a personalized judge? *arXiv preprint arXiv:2406.11657*.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

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

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Qingyan Guo, Rui Wang, Junliang Guo, Bei Li, Kaitao Song, Xu Tan, Guoqing Liu, Jiang Bian, and Yujiu Yang. 2024. Connecting large language models with evolutionary algorithms yields powerful prompt optimizers. In *The Twelfth International Conference on Learning Representations*.

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. *arXiv preprint arXiv:2310.06825*.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan A, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2024. DSPy: Compiling declarative language model calls into state-of-the-art pipelines. In *The Twelfth International Conference on Learning Representations*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chengzu Li, Han Zhou, Goran Glavaš, Anna Korhonen, and Ivan Vulić. 2023a. On task performance and model calibration with supervised and self-ensembled in-context learning. *arXiv preprint arXiv:2312.13772*.

Zongjie Li, Chaozheng Wang, Pingchuan Ma, Daoyuan Wu, Shuai Wang, Cuiyun Gao, and Yang Liu. 2023b. Split and merge: Aligning position biases in large language model based evaluators. *arXiv preprint arXiv:2310.01432*.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

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.

Yinhong Liu, Zhijiang Guo, Tianya Liang, Ehsan Shareghi, Ivan Vulić, and Nigel Collier. 2024a. Measuring, evaluating and improving logical consistency in large language models.

Yinhong Liu, Han Zhou, Zhijiang Guo, Ehsan Shareghi, Ivan Vulic, Anna Korhonen, and Nigel Collier. 2024b. Aligning with human judgement: The role of pairwise preference in large language model evaluators. *arXiv preprint arXiv:2403.16950*.

Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. 2024c. Calibrating LLM-based evaluator. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2638–2656, Torino, Italia. ELRA and ICCL.

Adian Liusie, Potsawee Manakul, and Mark Gales. 2024. LLM comparative assessment: Zero-shot NLG evaluation through pairwise comparisons using large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 139–151, St. Julian's, Malta. Association for Computational Linguistics.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Huan Ma, Changqing Zhang, Yatao Bian, Lemao Liu, Zhirui Zhang, Peilin Zhao, Shu Zhang, Huazhu Fu, Qinghua Hu, and Bingzhe Wu. 2023. Fairness-guided few-shot prompting for large language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Shikib Mehri and Maxine Eskenazi. 2020. USR: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, Online. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Arjun Panickssery, Samuel R Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. *arXiv preprint arXiv:2404.13076*.

Pouya Pezeshkpour and Estevam Hruschka. 2023. Large language models sensitivity to the order of options in multiple-choice questions. *arXiv preprint arXiv:2308.11483*.

Reid Pryzant, Dan Iter, Jerry Li, Yin Lee, Chenguang Zhu, and Michael Zeng. 2023. Automatic prompt optimization with "gradient descent" and beam search. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7957–7968, Singapore. Association for Computational Linguistics.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, et al. 2023.

Large language models are effective text rankers with pairwise ranking prompting. *arXiv preprint arXiv:2306.17563*.

Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. 2023. Verbosity bias in preference labeling by large language models. *arXiv preprint arXiv:2310.10076*.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*.

Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. 2023. Large language models are not yet human-level evaluators for abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4215–4233, Singapore. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Xingchen Wan, Ruoxi Sun, Hanjun Dai, Sercan Arik, and Tomas Pfister. 2023a. Better zero-shot reasoning with self-adaptive prompting. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3493–3514, Toronto, Canada. Association for Computational Linguistics.

Xingchen Wan, Ruoxi Sun, Hootan Nakhost, and Sercan O Arik. 2024. Teach better or show smarter? on instructions and exemplars in automatic prompt optimization. *arXiv preprint arXiv:2406.15708*.

Xingchen Wan, Ruoxi Sun, Hootan Nakhost, Hanjun Dai, Julian Eisenschlos, Sercan Arik, and Tomas Pfister. 2023b. Universal self-adaptive prompting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7437–7462, Singapore. Association for Computational Linguistics.

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

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. 2024. Large language models as optimizers. In *The Twelfth International Conference on Learning Representations*.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. 2024. Evaluating large language models at evaluating instruction following. In *The Twelfth International Conference on Learning Representations*.

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

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 12697–12706.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024a. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024b. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Han Zhou, Xingchen Wan, Lev Proleev, Diana Mincu, Jilin Chen, Katherine A Heller, and Subhrajit Roy. 2024a. Batch calibration: Rethinking calibration for in-context learning and prompt engineering. In *The Twelfth International Conference on Learning Representations*.

Han Zhou, Xingchen Wan, Ivan Vulić, and Anna Korhonen. 2023a. Survival of the most influential prompts: Efficient black-box prompt search via clustering and pruning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13064–13077, Singapore. Association for Computational Linguistics.

Han Zhou, Xingchen Wan, Ivan Vulić, and Anna Korhonen. 2024b. AutoPEFT: Automatic Configuration Search for Parameter-Efficient Fine-Tuning. *Transactions of the Association for Computational Linguistics*, 12:525–542.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023b. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations*.

Models	TopicalChat			Avg.
	NAT	ENG	OVE	
Mistral 7B				
Pairwise	0.13	0.18	0.22	0.18
ZEPO	**0.14** +1%	**0.25** +7%	**0.28** +6%	**0.23**+5%
Llama-3 8B				
Pairwise	0.02	0.08	0.14	0.05
ZEPO	**0.16** +14%	**0.26** +18%	**0.46** +32%	**0.30**+25%

Table 3: Spearman correlations on TopicalChat with Mistral and Llama-3. We evaluate in terms of Naturalness (NAT), Engagement (ENG), and Overall quality (OVE). We highlight the % improvement/degradation of ZEPO over "Pairwise" in +green/-red.

A Implementation Details

ZEPO. In this section, we include implementation details to enable the reproducibility of our work. Regarding the template and prompt across all the experiments reported, we use the prompt template from Table 4. ZEPO evaluation results are conducted on top of the state-of-the-art pairwise evaluator, PairS (Liu et al., 2024b), which leverages pairwise comparisons between randomly sampled pairs and aggregates them into a ranked sequence with a sorting-based search algorithm. We use GPT-3.5-turbo as the LLM optimizer with a temperature of 0.9, which is instructed to generate diverse and creative paraphrasing of the initial instruction. Following that, we implement Mistral-7B-Instruct-v0.1 and Meta-Llama-3-8B-Instruct as our main LLM evaluators. In practice, we set 5 epochs with a population size S of 5 that sufficiently converges to the fairest instruction. For $|\mathcal{D}|$, we use 2,400 pairwise sampling (10 data points) per instruction for SummEval, 840 (20 data points) for News Room, and 1,200 (60 data points) for TopicalChat based on the number of candidates per data point. ZEPO serves as a first step towards fairer LLM evaluations, and we defer investigations on ZEPO with tighter, more sampling-efficient constraints to future work.

Zero-Shot Learning Objectives. Entropy is a commonly used zero-shot metric: $-\sum_j p_j \log p_j$. In Fig. 3, we use entropy as a confidence measurement for LLM evaluators and treat $\texttt{Confidence} = \sum_j p_j \log p_j$ in the negative of entropy averaged across $\mathcal{D}$. However, in the context of LLM evaluations, overconfidence may further misalign LLM evaluators with human judgments. Context-free confidence is computed with the same formulation

Prompt Templates for `Pairwise` and ZEPO in summarization.

```
Source text: [SOURCE_TEXT]

Summary A: [SUMMARY_1]

Summary B: [SUMMARY_2]

Question: [INSTRUCTION]
Answer: [OUTPUT]
```

Prompt templates for `Pairwise` and ZEPO in dialog.

```
Dialog history: [DIALOG_HISTORY]

Response Candidate A: [RESPONSE_1]

Response Candidate B: [RESPONSE_2]

Question: [INSTRUCTION]
Answer: [OUTPUT]
```

Prompt templates for LLM Optimizer to generate new instruction candidates.

```
Paraphrase the following instruction
for a pairwise comparison task.
Do not change the keyword "[ASPECT]".
Be diverse and creative in paraphrasing.
Return the instruction only.

Input: [INSTRUCTION]

Output: [NEW_INSTRUCTION]
```

Table 4: Prompt template for pairwise comparisons and the LLM optimizer to generate paraphrased instructions.

above but with a content-free input $C_I([\text{N/A}], I)$ adopted from the contextual calibration (Zhao et al., 2021). Context-free confidence is introduced in Fair-Prompting (Ma et al., 2023), where the main idea is to select exemplars with the lowest confidence with respect to a content-free input, such that the prediction for classes is more balanced with the prompt template alone. In addition, we adopted a zero-shot calibration metric from Batch Calibration (Zhou et al., 2024a): $\texttt{Calibration} = -|\frac{1}{N}\sum(\log p_A - \log p_B)|$, which measures the absolute distance in the marginalized logits between two classes.

It indicates a uniform prior in the logit space, and a better-calibrated model can generate fairer predictions in terms of their scores. In contrast with calibration, our fairness metric is based on a uniform prior in the preference (decision) distribution and demonstrates the strongest correlation with LLM evaluation performance.

Pointwise Baselines. We implement two pointwise evaluator baselines: direct `Scoring` and `G-Eval`. For both cases, the LLM evaluators are tasked with rating a specific aspect of the output candidate using an integer score on the Likert scale (Likert, 1932). In the `Scoring` approach, the evaluators assign a single score with the highest predictive probability to each output candidate. For the `G-Eval` baseline, the final score is calculated by taking the weighted average of the scores across all five score tokens. We use the same prompt templates and evaluation criteria from previous work (Liu et al., 2024c), which have been calibrated and deliver robust evaluations. As indicated in the main paper, ZEPO shows improved evaluation results in general over the aforementioned calibrated baselines.

Aspect	Instruction Prompt	Fairness
COH	Initial Prompt: Evaluate and compare the coherence of the two summary candidates for the given source text. Consider coherence aspects such as clarity and logical flow. A summary is coherent if it accurately captures the key information from the article, and presents them in a clear manner. Which summary candidate has better coherence? If the candidate A is better, please return 'A'. If the candidate B is better, please return 'B'. You must return the choice only. ZEPO-Optimized Prompt: Assess and contrast the coherence of the two summaries using the provided text. Take into account clarity and logical progression. A coherent summary efficiently conveys the main details from the text in a clear and organized manner. Which summary demonstrates stronger coherence? Select 'A' for option A or 'B' for option B. Indicate your chosen option.	Initial: -0.288 Optimized: **-0.007**
FLU	Initial Prompt: Evaluate and compare the fluency of the two summary candidates for the given source text. Which summary candidate has better fluency? If the candidate A is better, please return 'A'. If the candidate B is better, please return 'B'. You must return the choice only. ZEPO-Optimized Prompt: Evaluate the smoothness of each summary choice using the given text. Decide which summary showcases better fluency. Choose 'A' for candidate A or 'B' for candidate B. Please only submit your chosen option.	Initial: -0.417 Optimized: **-0.018**
CON	Initial Prompt: Evaluate and compare the consistency of the two summary candidates for the given source text. A summary is consistent with the article if it faithfully reflects the main points, facts, and tone of the article. A summary is inconsistent if it introduces any errors, contradictions, or distortions of the original article. Which summary candidate has better consistency? If the candidate A is better, please return 'A'. If the candidate B is better, please return 'B'. You must return the choice only. ZEPO-Optimized Prompt: Evaluate the consistency of two different ways of summarizing the given text. Find the summary that best captures the main ideas, details, and tone of the original text. Note any mistakes or differences in the summaries. Choose either 'A' for option A or 'B' for option B as the superior choice. Share your selected option.	Initial: -0.295 Optimized: **-0.012**

Table 5: Initial prompt and the ZEPO-found prompt. We report the fairness metric before and after optimization.

Aspect	Instruction Prompt	Fairness
REL	Initial Prompt: Evaluate and compare the relevance of the two summary candidates for the given source text. A summary is relevant if it captures the main points from the article, without leaving out any crucial details or adding any unnecessary or inaccurate ones. A summary is more relevant if it uses the same or similar terms and expressions as the article. A summary is less relevant if it omits some of the key facts from the article, or if it introduces irrelevant information that is not supported by the article. Which summary candidate has better relevance? If the candidate A is better, please return 'A'. If the candidate B is better, please return 'B'. You must return the choice only. ZEPO-Optimized Prompt: Assess the relevance of the two summaries presented for the text and pick the one that closely matches the main points of the article using similar language. Select 'A' for candidate A or 'B' for candidate B. Display your selection.	Initial: -0.3625 Optimized: **-0.0003**
INF	Initial Prompt: Evaluate and compare the informativeness of the two summary candidates for the given source text. Evaluate how each summary converts their input text to natural language text, without omitting, adding, or distorting any facts. Which summary candidate has better informativeness? If the candidate A is better, please return 'A'. If the candidate B is better, please return 'B'. You must return the choice only. ZEPO-Optimized Prompt: Assess and contrast the informativeness of two summaries based on the provided source material. Examine how accurately each summary reflects the original content. Determine which summary is more informative by selecting either 'A' or 'B'. Only indicate your choice.	Initial: -0.217 Optimized: **-0.001**

Table 6: Initial prompt and the ZEPO-found prompt. We report the fairness metric before and after optimization.