Batch Calibration: Rethinking Calibration for In-Context Learning and Prompt Engineering

Presenter: Han Zhou

Paper | Blog

ICLR 2024

Google Research

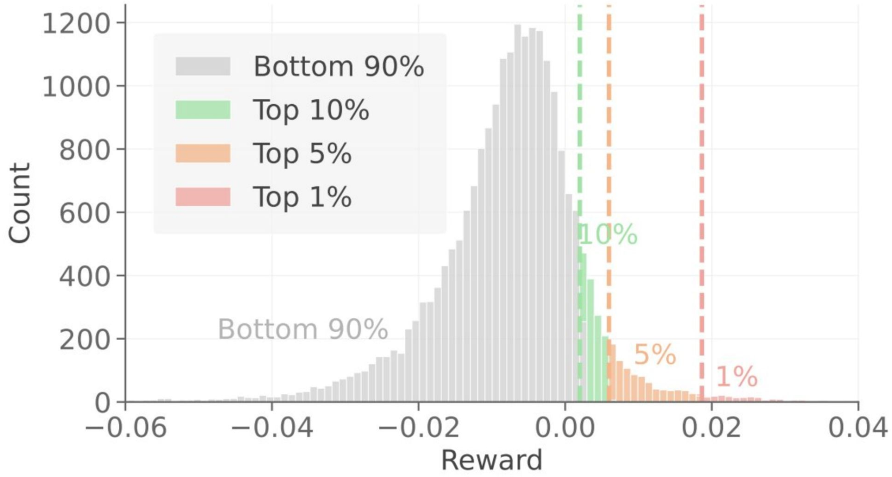
Agenda

01

Introduction

Prompt Sensitivity

- The distribution of influence over the vocabulary of tokens is heavily **non-uniform**.
- **The vast majority of tokens** actively harm LLM predictions.
- Only **a small fraction of tokens** improve performance.



Distribution of the incremental reward ΔR(v) evaluated on 16-shot RTE samples with Flan-T5 base. The top-{1,5,10}% tokens in terms of their incremental reward are highlighted in color.

Survival of the Most Influential Prompts: Efficient Black-Box Prompt Search via Clustering and Pruning.
EMNLP 2023

Google Research

In-Context Learning (ICL)

- The predictions of LLMs are sensitive and even *biased* to:
  - The prompt template
  - Label spaces
  - Choice of examples
  - Order of examples
  - …

- We refer this behavior to the *Contextual Bias*:
  - A-priori propensity of LLMs to predict certain classes over others unfairly given the context.
  - This is mainly due to the pretraining statistics and corpora.

- The biased prediction hinders the potential of LLM.
  - A phenomenal behavior is never predicting some classes (maybe vs. yes, no)

Calibration

- **Definition of calibration:**
  - Calibration is to correct the biased prediction.
  - It mitigates the contextual bias.

- **Existing calibration techniques:**
  - Contextual Calibration (CC) [1]
  - Domain-context Calibration (DC) [2]
  - Prototypical Calibration (PC) [3]

[1] Zhao TZ, Wallace E, Feng S, Klein D, Singh S. Calibrate Before Use: Improving Few-Shot Performance of Language Models. ICML 2021.
[2] Fei Y, Hou Y, Chen Z, Bosselut A. Mitigating Label Biases for In-context Learning. ACL 2023.
[3] Han, Z., Hao, Y., Dong, L., Sun, Y., & Wei, F. (2022). Prototypical calibration for few-shot learning of language models.ICLR 2023

02

Analysis

Overview

- Contextual Calibration (CC):
  - Calibrate via content free tokens: “Review: N/A, Sentiment: ”
- Domain-context Calibration (DC)
  - Random in-domain tokens: “Review: [random text], Sentiment: ”
- Prototypical Calibration (PC)
  - Learn clusters of each class in the prototypical space.

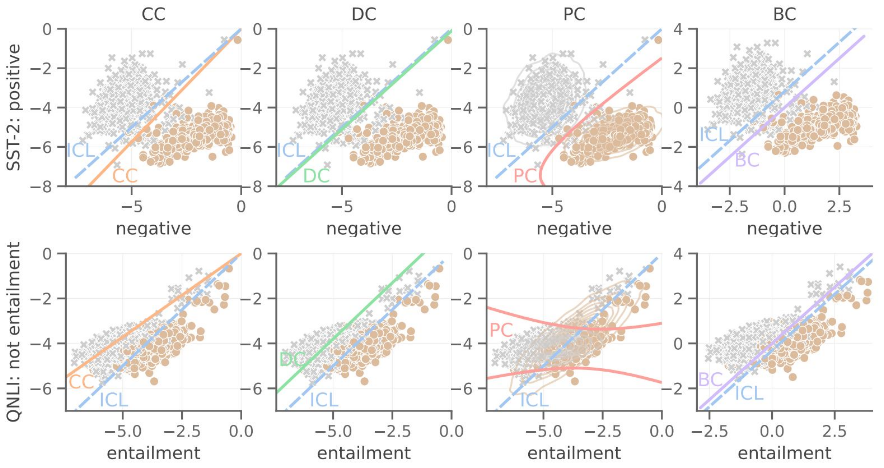
Method	Token	#Forward	Comp. Cost	Cali. Form	Learning Term	Decision Boundary $h(\mathbf{p})$	Multi-Sentence	Multi-Class
CC	N/A	$1+1$	Inverse	$\mathbf{W}\mathbf{p}+\mathbf{b}$	$\mathbf{W}=\mathrm{diag}(\hat{\mathbf{p}})^{-1}, \mathbf{b}=\mathbf{0}$	$p_0=\alpha p_1$	✗	✓
DC	Random	$20+1$	Add	$\mathbf{W}\mathbf{p}+\mathbf{b}$	$\mathbf{W}=\mathbf{I}, \mathbf{b}=-\frac{1}{T}\sum_t \mathbf{p}(y\|\text{text}_j, C)$	$p_0=p_1+\alpha$	✗	✓
PC	-	1	EM-GMM	-	$\sum_j \alpha_j P_G(\mathbf{p}\|\boldsymbol{\mu}_\mathbf{j}, \boldsymbol{\Sigma}_\mathbf{j})$	$P_G(\mathbf{p}\|\mu_0,\Sigma_0)=P_G(\mathbf{p}\|\mu_1,\Sigma_1)$	✓	✗
BC (Ours)	-	1	Add	$\mathbf{W}\mathbf{p}+\mathbf{b}$	$\mathbf{W}=\mathbf{I}, \mathbf{b}=-\mathbb{E}_x\left[\mathbf{p}(y\|x,C)\right]$	$p_0=p_1+\alpha$	✓	✓

Design Principles

RQ1: What Constitutes a Better Decision Boundary for Calibrations?

a. Non-linear decision boundary is susceptible to overfitting and instability.

b. Linear boundary is empirically more robust.



Visualization of the decision boundaries of uncalibrated ICL, and after applying existing calibration methods and the proposed BC in representative binary classification tasks of SST-2 (top row) and QNLI (bottom row) on 1-shot PaLM 2-S.

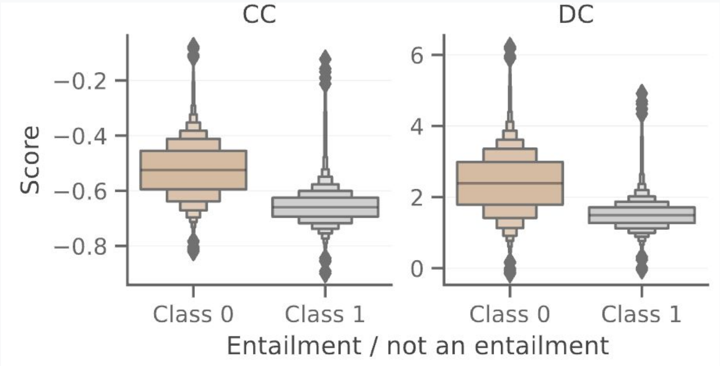
Google Research

Design Principles

RQ2: Is Content-free Input a Good Estimator of the Contextual Prior?

a. *Content-free inputs can be inappropriate.*

For example:
Question: N/A, Sentence: N/A, Answer: ->[entailment]



The distribution of ICL scores after applying CC and DC on QNLI. Due to an unfair content-free prior, the prediction by 1-shot PaLM-2 is biased towards entailment.

Google Research

03

Batch Calibration

Google Research

Batch Calibration

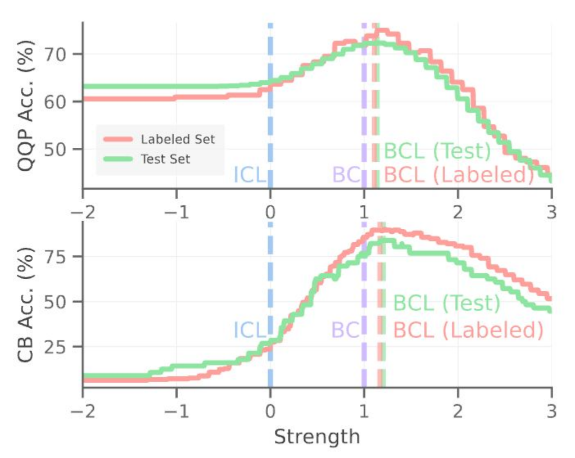
Zero-shot, inference-only:

$$\mathbf{p}(y = y_j|C) = \underset{x\sim P(x)}{\mathbb{E}}\left[\mathbf{p}(y = y_j|x, C)\right]$$
$$\approx \frac{1}{M}\sum_{i=1}^{M}\mathbf{p}(y = y_j|x^{(i)}, C)\,\forall\, y_j \in \mathcal{Y}$$

$$\hat{y}_i = \underset{y\in\mathcal{Y}}{\arg\max}\,\mathbf{p}_{\mathrm{BC}}(y|x_i, C) = \underset{y\in\mathcal{Y}}{\arg\max}\left[\mathbf{p}(y|x_i, C) - \hat{\mathbf{p}}(y|C)\right]$$

Adjustable Batch Calibration Layer (BCL):

$$\mathbf{p}_{\mathrm{BCL}}(y|x_i, C) = \mathbf{p}(y|x_i, C) - \gamma\hat{\mathbf{p}}(y|C)$$



BC benefits from labeled data: The performance of an adaptable batch calibration layer (BCL) compared to the zero-shot BC with a changing strength. The *strength* at 0 and 1 represent the uncalibrated ICL and BC, respectively. We highlight the optimal strength learned from a labeled set by a red vertical line and the best test strength by a green line.

Google Research

Batch Calibration

Illustration of Batch Calibration (BC). Batches of demonstrations with in-context examples and test samples are passed into the LLM. Due to implicit bias sources in the context, the score distribution from the LLM becomes highly biased. BC is a modular and adaptable layer option appended to the output of the LLM/VLM (vision language model). BC generates calibrated scores. Highlighted symbols indicate the distribution means (visualized for illustration only).

Google Research

Batch Calibration

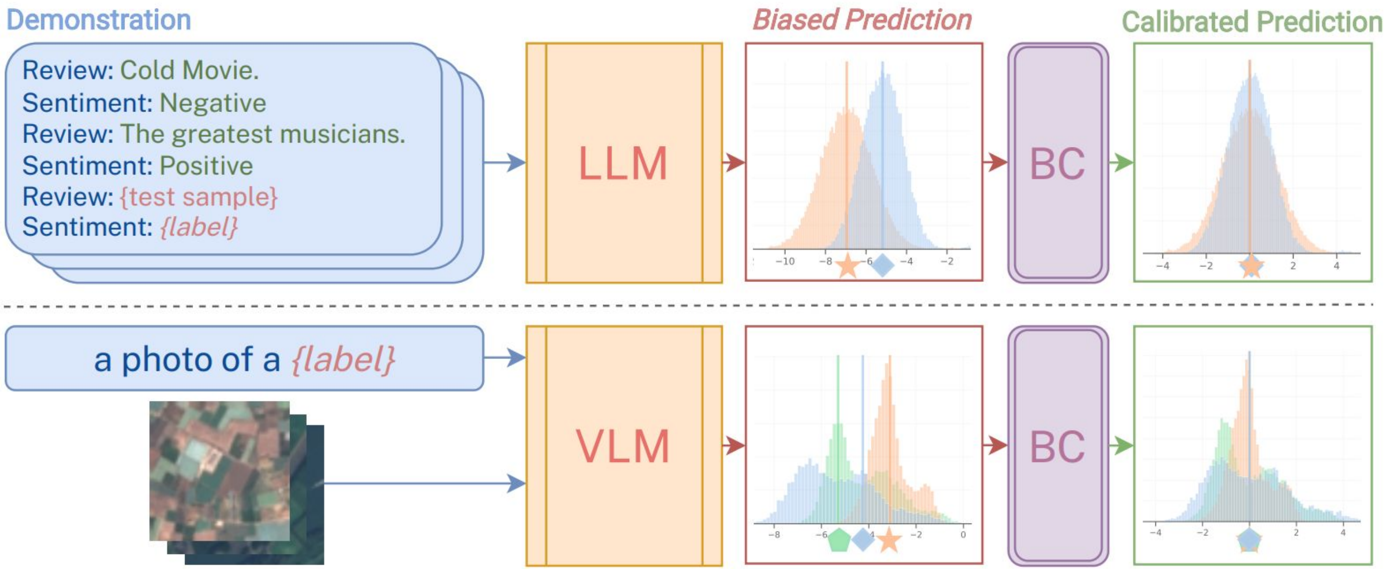


Illustration of Batch Calibration (BC). Batches of demonstrations with in-context examples and test samples are passed into the LLM. Due to implicit bias sources in the context, the score distribution from the LLM becomes highly biased. BC is a modular and adaptable layer option appended to the output of the LLM/VLM (vision language model). BC generates calibrated scores. Highlighted symbols indicate the distribution means (visualized for illustration only).

Google Research

04

Experiments

Datasets and Models

Datasets:

- GLUE, SuperGLUE, PaLM 2 evaluation data sets

Models: PaLM 2-S, PaLM 2-M, PaLM 2-L, CLIP ViT-B/16

Baselines:

- ICL
- CC [1]
- DC [2]
- PC [3]

[1] Zhao TZ, Wallace E, Feng S, Klein D, Singh S. Calibrate Before Use: Improving Few-Shot Performance of Language Models. ICML 2021.
[2] Fei Y, Hou Y, Chen Z, Bosselut A. Mitigating Label Biases for In-context Learning. ACL 2023.
[3] Han, Z., Hao, Y., Dong, L., Sun, Y., & Wei, F. (2022). Prototypical calibration for few-shot learning of language models.ICLR 2023

05

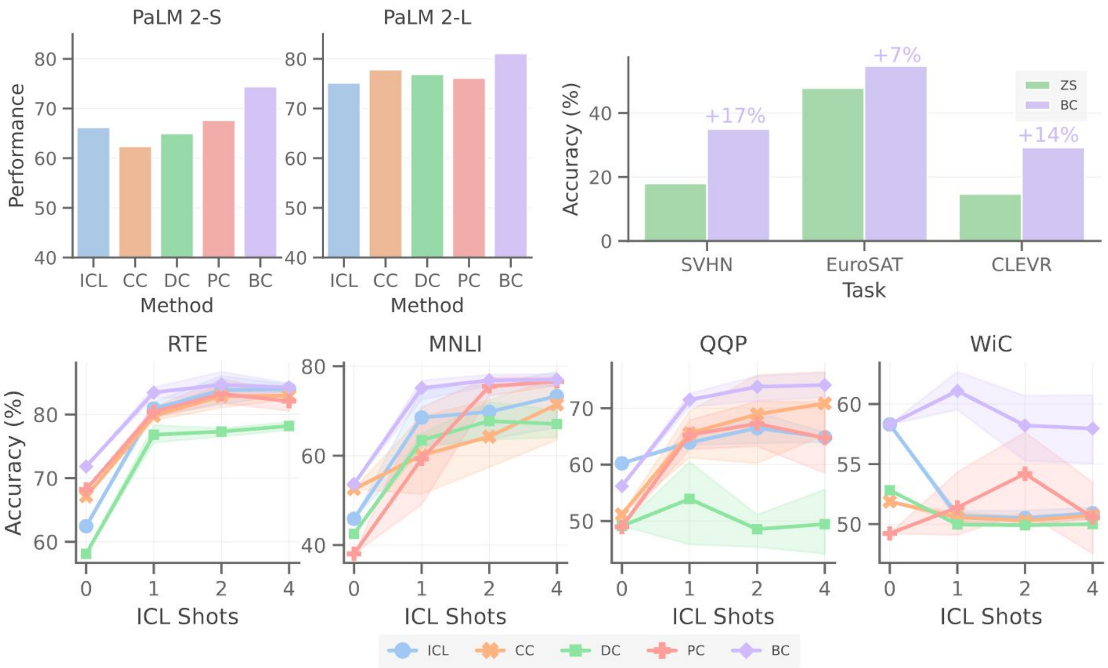
Results

Main results

Model	PaLM 2-S					PaLM 2-L				
Method	ICL	CC	DC	PC	BC	ICL	CC	DC	PC	BC
SST-2	93.62$_{0.62}$	**95.50**$_{0.25}$	94.29$_{0.32}$	**95.71**$_{0.10}$	95.44$_{0.15}$	93.16$_{5.18}$	**95.82**$_{0.62}$	94.91$_{2.01}$	95.64$_{0.47}$	**95.78**$_{0.55}$
MNLI	**68.52**$_{7.98}$	60.07$_{11.26}$	63.45$_{1.99}$	59.29$_{13.79}$	**75.12**$_{2.76}$	72.77$_{3.65}$	**79.45**$_{3.46}$	71.53$_{4.86}$	78.68$_{7.10}$	**81.34**$_{2.29}$
QNLI	**81.20**$_{1.90}$	56.86$_{3.29}$	65.62$_{3.53}$	69.82$_{17.73}$	**82.45**$_{1.82}$	64.68$_{3.53}$	**69.71**$_{4.89}$	68.97$_{3.27}$	61.01$_{15.26}$	**87.90**$_{1.24}$
MRPC	66.42$_{10.15}$	**70.44**$_{0.94}$	68.58$_{0.21}$	**71.86**$_{1.29}$	70.05$_{2.40}$	**73.19**$_{1.21}$	72.40$_{3.53}$	68.68$_{0.40}$	**75.39**$_{2.60}$	70.39$_{2.56}$
QQP	63.91$_{0.66}$	**65.55**$_{5.34}$	53.92$_{9.35}$	65.28$_{3.42}$	**71.48**$_{1.46}$	**82.57**$_{0.75}$	81.17$_{2.03}$	78.32$_{1.82}$	**81.42**$_{0.24}$	79.56$_{1.40}$
BoolQ	83.99$_{3.90}$	87.14$_{1.60}$	87.64$_{1.10}$	**88.70**$_{0.15}$	**87.83**$_{0.10}$	90.02$_{0.60}$	**90.15**$_{0.54}$	87.77$_{1.17}$	64.40$_{22.37}$	**90.10**$_{0.22}$
CB	45.71$_{10.61}$	29.64$_{7.85}$	65.71$_{3.20}$	**81.07**$_{9.42}$	**78.21**$_{3.19}$	**92.86**$_{2.19}$	85.72$_{7.78}$	**92.86**$_{2.82}$	89.29$_{7.25}$	**93.21**$_{1.49}$
COPA	**96.40**$_{2.30}$	95.80$_{2.05}$	**96.40**$_{2.88}$	**96.20**$_{2.05}$	**96.40**$_{2.07}$	**98.60**$_{1.14}$	97.20$_{1.10}$	97.40$_{0.89}$	**99.00**$_{0.71}$	97.00$_{1.00}$
RTE	**80.94**$_{1.29}$	79.78$_{0.92}$	76.82$_{1.72}$	80.43$_{1.07}$	**83.47**$_{1.10}$	75.09$_{2.11}$	80.00$_{2.48}$	79.21$_{1.95}$	**86.64**$_{2.62}$	**85.42**$_{2.48}$
WiC	50.69$_{0.59}$	50.56$_{0.50}$	49.97$_{0.13}$	**51.38**$_{3.56}$	**61.10**$_{2.07}$	51.35$_{1.90}$	55.58$_{6.38}$	54.67$_{6.02}$	**57.87**$_{11.08}$	**64.83**$_{8.59}$
ANLI-R1	**46.24**$_{4.21}$	42.54$_{3.20}$	40.26$_{3.66}$	40.28$_{6.46}$	**59.82**$_{0.51}$	63.06$_{2.63}$	71.92$_{3.71}$	**73.56**$_{3.88}$	72.30$_{8.05}$	**75.00**$_{3.03}$
ANLI-R2	40.44$_{0.90}$	38.36$_{0.82}$	38.44$_{3.46}$	**41.88**$_{4.50}$	**50.16**$_{0.82}$	58.40$_{1.19}$	65.36$_{3.75}$	**65.48**$_{1.91}$	64.98$_{2.94}$	**67.30**$_{2.34}$
ANLI-R3	42.53$_{0.99}$	38.78$_{1.04}$	**43.67**$_{5.25}$	37.50$_{0.81}$	**55.75**$_{1.66}$	61.35$_{3.14}$	**67.32**$_{0.98}$	66.23$_{0.72}$	63.03$_{6.03}$	**66.38**$_{0.74}$
Avg.	66.20	62.39	64.98	**67.65**	**74.41**	75.16	**77.83**	76.89	76.13	**81.09**

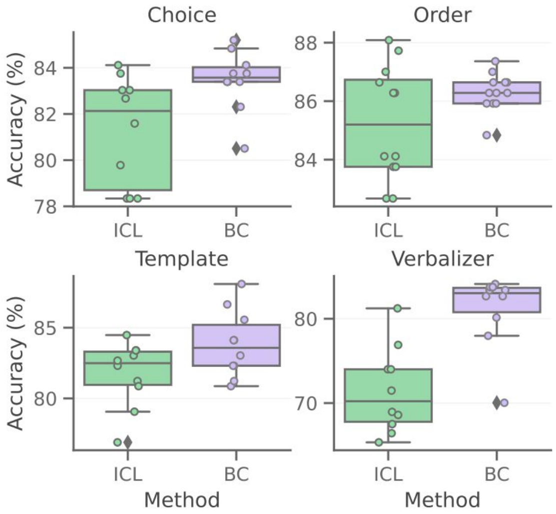
Accuracy (%) on natural language classification tasks with 1-shot PaLM 2-S and PaLM 2-L Models. We report the mean and standard deviation for all results for 5 different in-context examples. We reproduce all baselines. The **best** and **second-best** results are marked in bold fonts and ranked by color.

Additional results

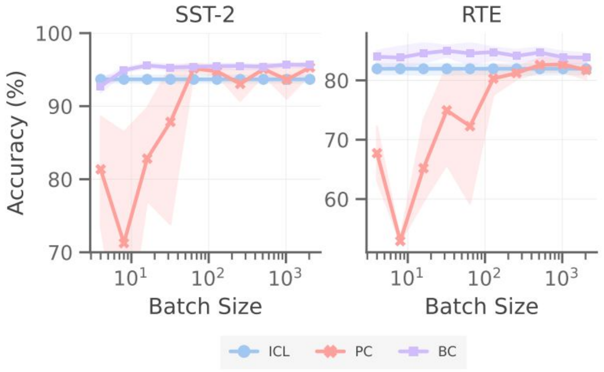


The ICL performance on various calibration techniques over the number of ICL shots on PaLM 2-S. Each shot indicates 1 example per class in the demonstration. Lines and shades denote the mean and standard deviation over 5 random seeds, respectively.

Ablation



BC makes prompt engineering easier: Performance of BC with respect to ICL choices, ICL orders, prompt templates, and verbalizers.



BC is data-efficient and insensitive to the batch size: Performance of BC across different sizes of an initial unlabeled set without using a running estimate of the contextual bias. We compare BC with the state-of-the-art PC baseline that also leverages unlabeled estimate set, and experiments are conducted on PaLM 2-S.

Google Research

06

Conclusion

Batch Calibration

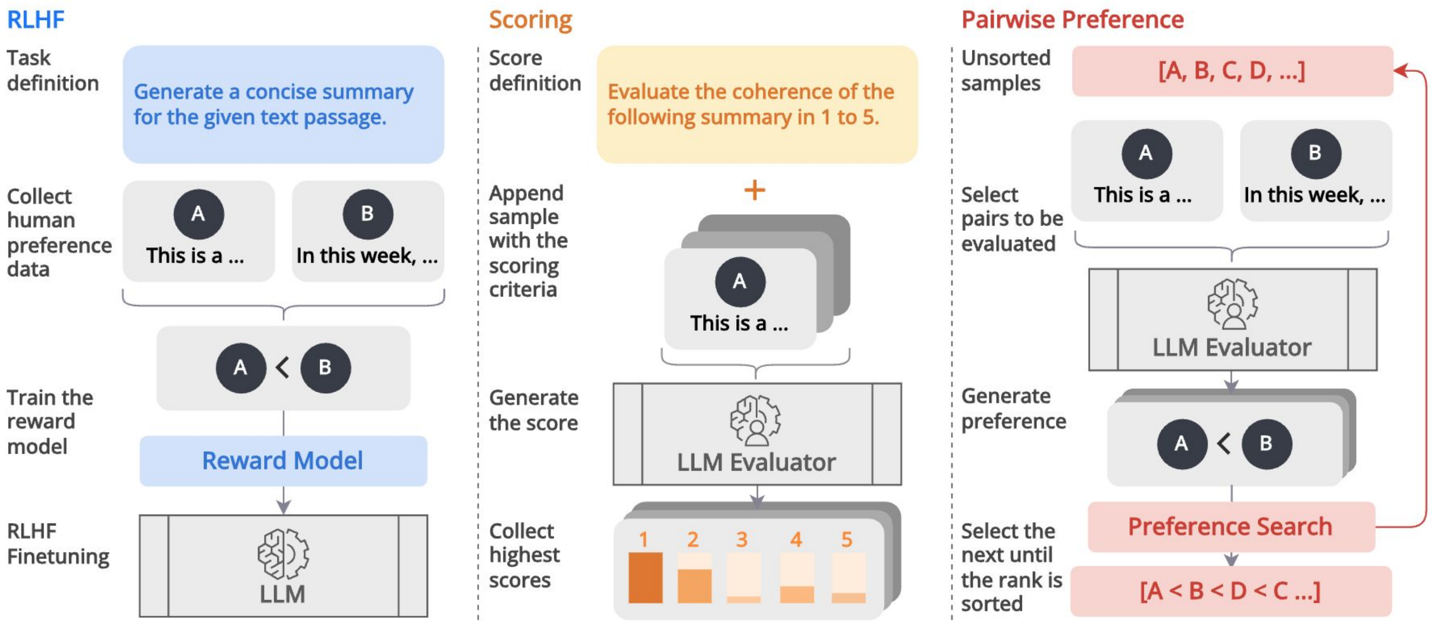
- **Zero-shot**: No ground-truth label used at any point in time.
- **Efficiency**: No additional cost in inference.
- **Black-box**: No internal access to model parameters & no gradients required.
- **Inference-only:** Applicable for all API-only services, e.g. Sax.
- **Easy-to-implement:** few lines of scripts.

Google Research

07

LLM Judgments

LLM-as-a-Judge



Aligning with Human Judgement: The Role of Pairwise Preference in Large Language Model Evaluators.
COLM 2024

The limitations of Calibration

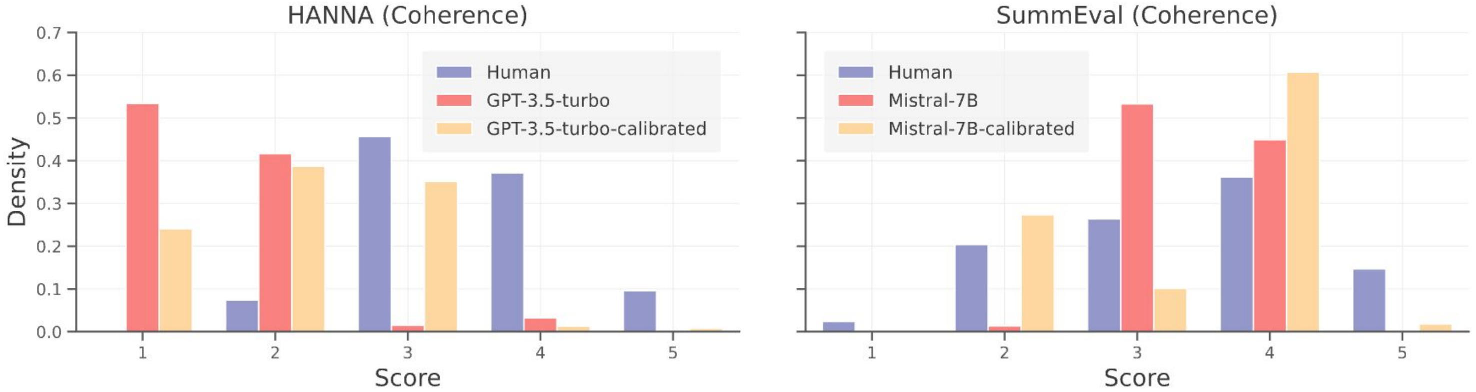


Figure 2: *LLM evaluations are misaligned with human judgements.* The score histograms on evaluating the coherence in HANNA (Chhun et al., 2022) and SummEval (Fabbri et al., 2021). We present the scores from gold human evaluations, LLMs, and LLMs after calibrations. The histograms can be interpreted as estimated score prior distributions via marginalization.

Aligning with Human Judgement: The Role of Pairwise Preference in Large Language Model Evaluators.
COLM 2024

Fairer Preference

- LLM evaluators are also very sensitive to evaluation instruction and criteria.
- However, if there is a fairer preference from LLMs, we notice a stronger correlation between LLM judgments and human preference.

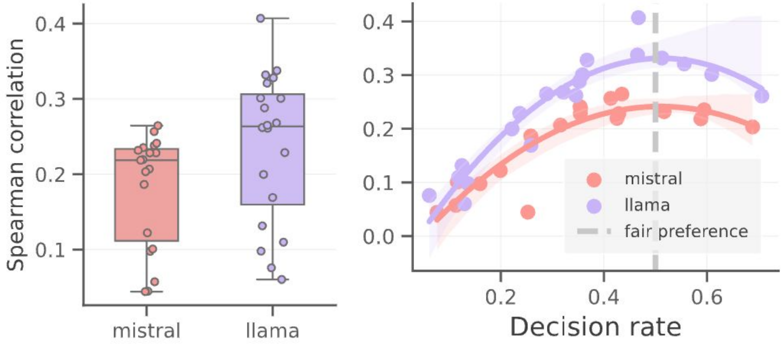
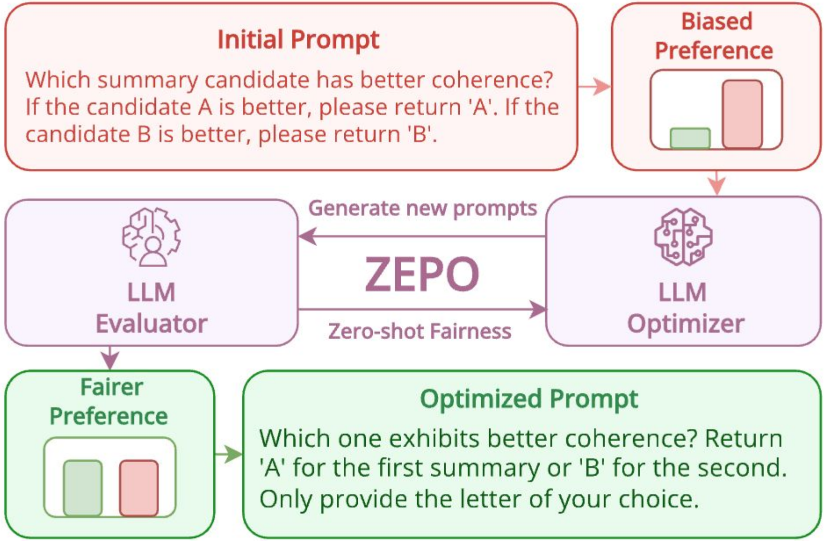


Figure 2: *LLM evaluators show strong sensitivity to instructions and fairer preference leads to better human-aligned LLM judgments.* Sensitivity and evaluation performance studies on preference fairness.

Fairer Preferences Elicit Improved Human-Aligned Large Language Model Judgments. Under Review

Zero-shot Prompt Optimization (ZEPO)

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.



Fairer Preferences Elicit Improved Human-Aligned Large Language Model Judgments. Under Review

Experiments

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.

Prompts

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

Thank You

Han Zhou
Student Researcher