Understanding Judge Calibration in Multi-Turn Debates

Anonymous authors
Paper under double-blind review

Abstract

Multi-turn debates have gained attention as language evaluation tasks for subject matter comprehension, critical reasoning and long-form responses. Language Models (LMs) play the role of judges for obtaining subjective ratings as a cheap alternative to human labor. However, similar to humans, LM judges may remain unsure of ratings and rate debate arguments either under or over-confidently. We empirically study judge calibration in multi-turn self debates, wherein a single LM debater debates with itself, and uncover that LM judges are often overconfident in their judgements. Miscalibration occurs as model confidence ratings increase while rated scores may decrease over debate rounds. Judge confidence exceeds score ratings for both frontier as well as open-source models. We further show that while naive finetuning may improve calibration by increasing scores, it does not necessarily lower overconfidence in ratings. Finetuned overconfident judges prefer similar ratings as confidence and rate different arguments indistinguishably. Our empirical analysis leads us to an observation that helps mitigate overconfidence. Since lower confidences and scores form the tail end of the dataset and are most desirable from a judge's perspective, sampling from this left tail must calibrate for confidence. We thus fit a product of Gumbels distribution on expected ratings of debate arguments and then rejection sample from its tail to finetune judge models. Sampling from the product of Gumbels, when compared to naive ratings and Supervised Finetuning (SFT), lowers judge confidence and yield well-calibrated judges while learning an expressive multi-modal distribution over ratings. Debate datasets and code will be released as part of the final version.

1 Introduction

Multi-turn debates (Moniri et al., 2024; Liu et al., 2024) simulate debates as a turn-by-turn conversation wherein LMs are assigned a side (*for* or *against* the motion) and must respond to their opponent's argument in a given round. Opponents may be humans or LMs themselves (Khan et al., 2024). Debates are rated by either human judges or LM judges which decide the final outcome of the competition. Multi-turn debates have recently gained significant attention as an evaluation task for both LM debaters as well as judges. Debates assess subject matter comprehension (of the motion and past arguments), critical reasoning (over opponent's argument) and long-form responses (over multiple rounds across multiple argument types). Furthermore, debating with and against LMs over longer rounds leads to self-critical reasoning patterns wherein models often correct or improve upon their past claims (Khan et al., 2024).

Recent work has seen a wide adoption in the usage of LMs as judges for a range of evaluation tasks. Ranging from speech styles (Chiang et al., 2025) to legal documents (Posner & Saran, 2025) to scientific publications (Starace et al., 2025), models are used to provide score ratings as well as head-to-head comparisons between different candidates. These domains span both subjective as well as objective evaluations, i.e- qualitative as well numeric ratings. LMs also play the role of debate judges for obtaining argument ratings when human judge ratings are expensive. However, similar to human judges, judge models present their own pitfalls.

LM judges may remain uncertain of their ratings or conceal their overall confidence. While traditional evaluation settings only ask the model *what is the score?*, it is also worth asking *what is the confidence in this score?* Judge models may rate arguments either under or over-confidently leading to a mismatch between the rating and its confidence. This leads to model judges (and hence the evaluation metrics) being

biased towards a set of behaviors which they find desirable. Furthermore, in the case of multi-turn pairwise evaluation settings such as debates, rating biases may compound over debate rounds wherein model judges could increasingly tend to favor one side over another. Thus, the challenge of judge calibration remains an open problem.

We empirically study and assess judge calibration in multi-turn self debates, wherein a single LM debater debates itself, by uncovering the over-confidence of model judges in their ratings. In addition to the score, we zero-shot and few-shot prompt models to self-report their confidence on the same fixed rating scale. Models often report themselves to be overconfident in their judgements. Confidence ratings continue to increase while score ratings may decrease over debate rounds. Judge confidence exceeds score ratings for both frontier as well as open-source models. One might expect finetuning to address this issue by distilling confidence ratings into the judge models. We follow this line of thinking and show that while naive finetuning may improve calibration by increasing scores, it does not necessarily lower overconfidence. Overconfident judges tend to report the same high score ratings as the confidence, leading to a biased preferences that fail to qualitatively distinguish between the score and confidence across different arguments.

Based on our empirical analysis and observations, we propose a practical finetuning strategy to mitigate overconfidence in LM judges. We observe that lower confidences and scores form the tail end of the data distribution, i.e- arguments with low confidence and low score are well-calibrated samples. Neither of the two lead to a biased or overconfident judge. Thus these samples serve as suitable proxies for desirable behavior. We begin by grouping ratings as per judge models and then fit a product of Gumbels distribution on expected ratings of debate arguments. The Gumbel distribution, being the distribution of extremes, allows us to expressively model the tail of the data distribution. We rejection sample from the tail and finetune judge models on samples and their ratings. This strategy, when compared to naive ratings and Supervised Finetuning (SFT) reduces overconfidence in judges by minimizing the Expected Calibration Error by 5.56 %. Additionally, the finetuning process leads to expressive multi-modal preference distributions over score and confidence ratings that does not lead to biased preferences.

Our main contributions are threefold and detailed below. Debate datasets and code will be released as part of the final version.

- We empirically study and assess judge calibration in multi-turn self debates by uncovering the over-confidence of model judges in their ratings. Models often report themselves to be overconfident in their judgements. Confidence ratings continue to increase while score ratings may decrease over debate rounds.
- We empirically observe that while naive finetuning may improve calibration by increasing scores, it does not necessarily mitigate overconfidence in judges. Finetuned overconfident judges tend to report exactly same high score ratings as the confidence which may or may not be reliable. This leads to a biased preference distribution that fails to distinguish between the score and confidence.
- Based on our empirical analysis and observations, we propose a practical finetuning strategy to mitigate overconfidence of LM judges. We propose utilizing the tail of the data distribution as a faithful proxy of desirable behavior. We fit a product of Gumbels distribution on the dataset and rejection sample from its left tail to finetune judge models. Finetuned models are less overconfident and lead to well-calibrated judges. Distributions learned by Gumbel Fine-Tuned (GFT) judges are expressive and multi-modal over ratings.

2 Preliminaries

The setting considers a debate data distribution $\mathcal{D}$ consisting of m simulated debates over n rounds each. A data sample $(x, y, s, c)_i \sim \mathcal{D}$ at i^{th} round consists of opponent's past argument prompt x, model response argument y, judge score s and judge confidence c. Traditional settings consider a single judge to produce ratings s. However, we consider a set of judges $\mathcal{J}$ wherein each judge $J \sim \mathcal{J}$ produces its own independent score s_J and confidence c_J based on past argument and model response per round $(s_J, c_J)_i \sim J(x, y)$. These scores and confidences are then averaged or weighed across judges to produce the final ratings. Judge models

forming the set $\mathcal{J}$ may be smaller instruction-tuned models or larger frontier models, balancing between rating diversity and model quality. A multivariate Gumbel distribution with CDF $F(z) = \exp\left(-\exp\left(-\frac{z-\alpha}{\beta}\right)\right)$ is the distribution for modeling extreme values (maximum and minimum of distributions), defined by the mean and scaling parameters $\alpha \in \mathbb{R}^n$ and $\beta > 0$, $\beta \in \mathbb{R}$ respectively (Gumbel, 1941; Murphy, 2022; Aguech et al., 2023). The Inverse CDF or quantile function $\text{Gumbel}(\alpha, \beta, p) = \alpha - \beta \log(-\log p)$ is the function that yields a sample $z \sim \text{Gumbel}(\alpha, \beta, p)$ such that the density equals probability $0 < p < 1$. One can model judge preferences by fitting the distribution over score s_J and confidence c_J ratings and sampling from them separately ($\text{Gumbel}(s_J, \beta, p)$ and $\text{Gumbel}(c_J, \beta, p)$).

3 Related Work

3.1 Language Models as Judges

Recent works study and utilize LMs as judges for a range of benchmarking as well as real-world tasks (Gu et al., 2024). (Zheng et al., 2023) initially assess the LM-as-a-judge framework by evaluating LM judges on the Chatbot Arena (Chiang et al., 2024). (Bavaresco et al., 2025) carry out a large-scale evaluation of LM as judges on free-form language tasks and show that judging ability depends on task as well annotation type (human or model generation). (Zhu et al., 2025) further show finetuned LMs scale in their rating abilities while adhering to position, format and knowledge biases in input prompts. (Starace et al., 2025) utilize and evaluate frontier model judges for the real-world task of scientific research. Several works construct novel evaluation settings to assess faithfulness in model ratings. (Liu et al., 2023) combine and compare frontier model judges such as GPTs with human raters for better alignment. (Chan et al., 2024) utilize the setting of multi-agent debate to construct an ensemble of judges (termed as referees) that debate among themselves and provide more informed ratings. (Tong & Zhang, 2024) evaluate LMs as code judges by allowing them to judge code organization and logic, hence moving past syntax and functionality checks. Finally, various works study the robustness and calibration of LM judges. (Chen et al., 2024) show that models are often biased in their ratings but prefer factual accuracy and hence, remain robust in retrieval and memory-based tasks. (Schroeder & Wood-Doughty, 2024) show that judge distributions are often susceptible to single sample tests and thorough evaluations should leverage pairwise or multiple sample statistics in multi-turn tasks.

3.2 Uncertainty in Language Models

Prior methods in language modeling evaluate uncertainty on two fronts. Firstly, various methods quantify and minimize uncertainty by jailbreaking/prompting LMs on OOD samples (Anil et al., 2024). As concrete examples, (Jones et al., 2025) assess the uncertainty of a harmful response by forecasting elicitation probabilities via a Gumbel distribution, while (Benton et al., 2024) evaluate LM uncertainty of sabotaging evaluations using model calibration. Additionally, (Tanneru et al., 2024) uncover that uncertainty in model explanations is better expressed by utilizing model perturbations and samples.

The second class of methods study uncertainty by interpreting model features and assessing the impact of different architectural choices (Guo et al., 2017; Lakshminarayanan et al., 2017). For instance, (Huang et al., 2024) propose a framework that correlates higher uncertainty with lower generation quality leading to quantification of confidence in chat models. (Hou et al., 2024) produces a set of clarifications for inputs which are processed by an LM and ensembled to distinguish ambiguous queries following pretraining. Finally, (Atf et al., 2025) adopt bayesian inference to disentangle epistemic and aleatoric uncertainties with applications in medical diagnosis. Our work follows the first set of methods while borrowing design principles and metrics from interpretability methods.

3.3 Modeling Debates with Language Models

Various works utilize the framework of debates to evaluate and improve LM capabilities (Moniri et al., 2024; Liu et al., 2024; Chen et al., 2025). (Sternlicht et al., 2025) propose the debate speech benchmark by evaluating and comparing varying debater styles across speech, coherence and organization. (Khan et al., 2024) present the initial multi-turn debate framework wherein one or many LMs debate on a given topic and a LM

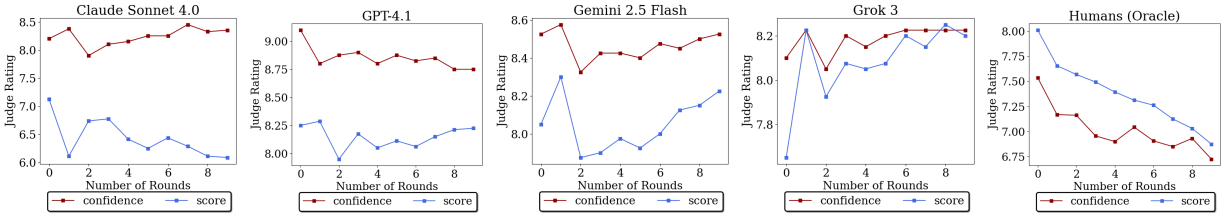

Figure 1: Variation of score and confidence ratings over debate rounds across all debates. Self-reported confidence ratings continue to increase while score ratings may decrease (or remain constant) for judge models. This leads to overconfident judges over the course of debates.

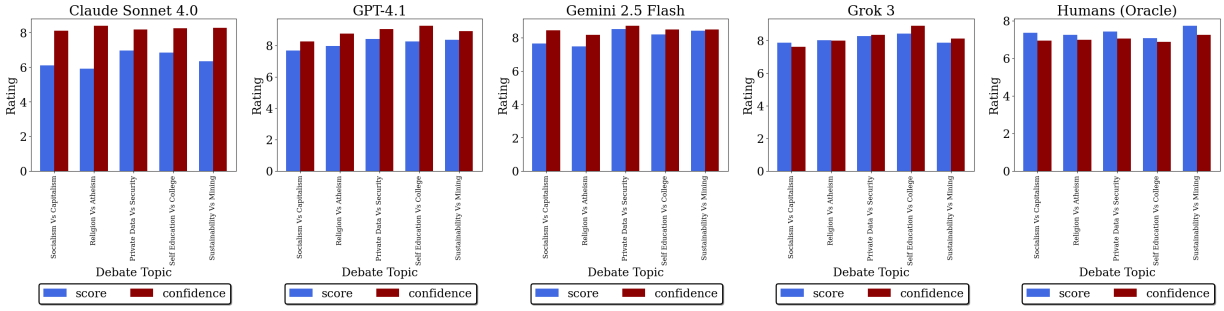

Figure 2: Per-debate variation of final scores and ratings for different judge models. Average confidence ratings are higher than score ratings across different debate topics and categories, hence leading to mismatched calibration in judge models.

judge evaluates argument strengths. Judge models rate debaters on accuracy and factuality of claims, hence leading to the ability to identify truths when expert debaters are optimized for persuasiveness. The work further uncovers miscalibration and overconfidence in judges when compared to human raters. Our analysis validates the findings and extends them towards evolution of ratings in long-form debate rounds. Additionally, our post-hoc finetuning and rejection sampling strategy bakes in the multi-turn debate framework to address calibration. (Sah, 2025) present an evaluation system for judging LMs by simulated academic debates. A separate LM acts as a judge to rate debate arguments as per prefixed criteria (logical consistency, factual accuracy, structure adherence and written clarity). (Zeng et al., 2025) propose a sparsification strategy to minimize ineffective exchanges between debaters in order to minimize token cost. (Koupaee et al., 2025) initialize each debater with a randomly assigned stance which they must defend, hence leading to greater stance diversity and error identification within arguments. Finally, (Prasad & Nguyen, 2025) also study the multi-turn debate setup and uncover overconfidence growth via win probabilities over debate rounds. Our setup and analysis closely follows their evaluation scheme by extending evaluation towards longer debates, a diverse evaluation set and larger training arguments for both frontier and open-source models. Borrowing from our evaluation insights, we further mitigate overconfidence by adopting a distributional perspective in the post-finetuning stage.

4 Calibration Through the Lens of Judges

4.1 Setup

Our setup consists of a debater model that debates with itself for a fixed number of rounds. Each round consists of the model observing the past argument as input prompt, followed by arguing for the motion, followed by observing the past (for side) argument as input prompt finally leading to the argument against the motion. That is, each debate is simulated between the model and itself in concurrent fashion. With the objective of studying high quality and dense debates, debates are simulated for four frontier models, namely GPT-4.1 (Achiam et al., 2023), Claude Sonnet 4.0 (AnthropicAI et al., 2025), Gemini 2.5 Flash (Comanici

et al., 2025) and Grok 3 (xAI et al., 2024). All debates and ratings are simulated via their respective model APIs. We collect long-form and long-context debates wherein each debate belongs to either of the following 5 categories; (1) Politics and Government, (2) Technology and Privacy, (3) Education, (4) Economics and Labor and (5) Environment and Climate.

Our setting considers a fixed set of judges $\mathcal{J}$ that rate each argument once it is generated. In addition to the score rating, we elicit self-reported confidence ratings from judge models and log them alongside score ratings. Both score and confidence ratings are between 1 and 10 with 1 being undesirable / not confident and 10 being ideal / highly confident. Our judge set consists of the same list of frontier models used to generate the debate dataset. We also utilize open-source models across different parameter budgets as both frozen and finetuned judges. Specifically, we finetune Llama 3.2 3B (Grattafiori et al., 2024), Llama 3.1 8B (Grattafiori et al., 2024) and the Gemma 3 family (4B, 12B and 27B) (Team et al., 2025). In addition, we record ratings from human judges, macro-average ratings per human and treat these as our ground truth oracle. Details on our debate datasets and ratings can be found in Appendix B.

4.2 Confidence in Debate Judges

Figure 1 presents the variation of score and confidence ratings across different debate rounds for each frontier judge. Confidence ratings increase over debate rounds as debates progress leading to increasing assertiveness in scores produced by judges. For instance, Claude Sonnet 4.0 and Grok 3 report 3.65% and 1.86% increases in their confidence following the completion of debates respectively. While the increase in itself remains small, confidence ratings are significantly larger than score ratings resulting in a calibration gap. Claude Sonnet 4.0 and Gemini 2.5 Flash have their confidence ratings exceed score ratings by 1.41× and 1.04× respectively. The result indicates growing certainty over judge scores *while scores themselves do not reflect the certainty*. This is in stark contrast to human judges wherein both score and confidence ratings decrease over debate rounds, hence reflecting a growing uncertainty. In addition, human confidence ratings remain consistently lower than score ratings.

A natural question to ask is *whether a judge is biased towards a specific topic or opinion?* We further investigate per-debate score and confidence ratings of each judge model. Figure 2 presents the comparison of ratings for all judge models. Irrespective of the debate topic or category, judge models maintain higher confidence ratings but lower scores. For instance, Claude Sonnet 4.0 has a maximum absolute confidence-score gap of 2.0 which is reflective of the gap across all rounds in Figure 1. Thus, judge models do not demonstrate knowledge bias, but overconfidence in their judgements. Figure 3 presents confidence and score ratings for judge models across different argument types, *for* and *against* the motion. Corresponding to both sides of the motion, confidence-score gaps remain prominent for majority of judge models. For instance, GPT-4.1 has an absolute confidence-score gap of 0.8 and 1.0 for the motion and against the motion respectively. Similarly, Llama 3.2 3B has a confidence-score gap of 0.6 and 1.0 for the motion and against the motion respectively. This indicates that judge models do not demonstrate a type bias and equally prefer both sides of the motion. Appendix C provides additional comparisons for each debate and argument type. Intuitively, a judge is confident irrespective of the topic, category or argument type in the debate. Thus, model overconfidence stems from subject matter comprehension wherein judges prefer critical and structured arguments.

4.3 Are Finetuned Judges Well-Calibrated?

How can we improve judge calibration? A naive solution would be to finetune judges on past calibrated arguments. However, this solution has a major challenge, obtaining calibrated arguments would require access to calibrated judges. This boils down to the same problem of calibrating another judge. Instead, we may choose to approximate the optimally calibrated data distribution by bootstrapping from a sub-optimally calibrated set of arguments. Hence, we filter ϵ-*calibrated* arguments, i.e- arguments whose confidence-score gaps are lower than a threshold ϵ ($|\texttt{score} - \texttt{confidence}| < \epsilon$), from our dataset and utilize these for finetuning judge models. We conduct SFT on the open-source Llama 3 and Gemma 3 models, across different parameter budgets, to rate arguments and report the confidence in rating. Each ϵ-calibrated argument is shown to the model as an instruction prompt along with its score and confidence ratings as answers. Figure 3 compares

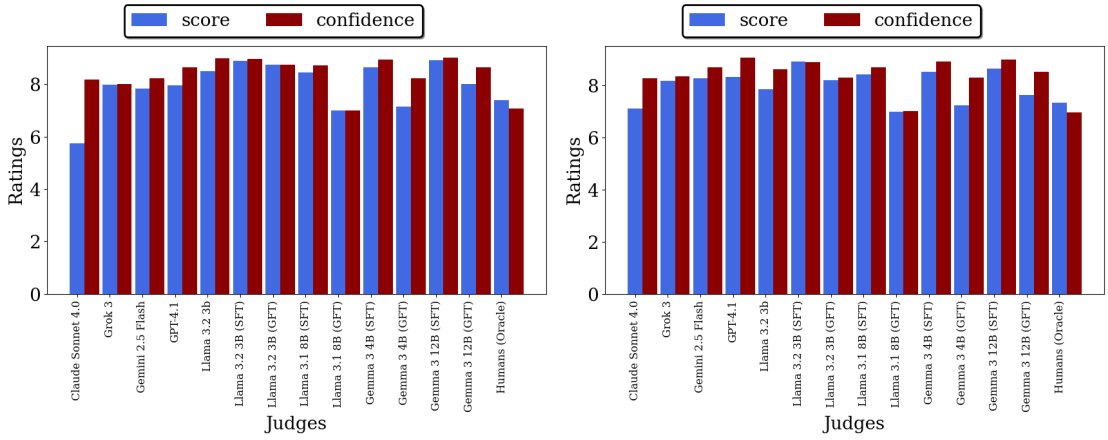

Figure 3: **(left)** for the motion and **(right)** against the motion. Across both motions, SFT and GFT finetuned models are well calibrated. GFT is further found to minimize over-confidence in judge ratings when compared to SFT.

score and confidence ratings of Llama 3.2 3B before and after finetuning for both types of arguments. We observe that SFT increases score ratings proportional to the confidence ratings but leaves confidence ratings unchanged. A similar trend can be observed for other judges wherein confidence ratings are higher score ratings and also higher than GFT judges (which we introduce in the next section). Compared to the frontier model judges as well, SFT increases score ratings but leaves high confidence ratings unchanged.

Model	Arguments with $\lvert\texttt{score} - \texttt{confidence}\rvert \geq 1$	Arguments with $\lvert\texttt{score} - \texttt{confidence}\rvert < 1$	ECE (↓)	Brier score (↓)	MCE (↓)
Claude Sonnet 4.0	317 ± 37	83 ± 14	0.7480 ± 0.04	**0.0238 ± 0.001**	1.00 ± 0.005
GPT-4.1	284 ± 29	116 ± 21	0.7542 ± 0.04	0.0434 ± 0.002	1.00 ± 0.002
Gemini 2.5 Flash	198 ± 17	202 ± 24	0.7473 ± 0.03	0.0293 ± 0.002	1.00 ± 0.003
Grok 3	257 ± 23	143 ± 11	**0.7008 ± 0.01**	0.0257 ± 0.001	**0.90 ± 0.002**
Llama 3.2 3B	307 ± 31	93 ± 12	0.7604 ± 0.07	0.0660 ± 0.01	1.00 ± 0.004
Llama 3.2 3B (Calibration Prompting)	344 ± 24	56 ± 11	**0.4869 ± 0.02**	0.1567 ± 0.01	1.00 ± 0.005
Llama 3.2 3B (Hard-Thresholding)	12 ± 4	388 ± 7	0.7342 ± 0.03	0.0163 ± 0.01	0.875 ± 0.002
Llama 3.2 3B (Temperature Scaling)	25 ± 10	375 ± 14	0.7101 ± 0.02	**0.0142 ± 0.01**	**0.850 ± 0.001**
Llama 3.2 3B (SFT)	11 ± 1	389 ± 21	0.7221 ± 0.001	0.0161 ± 0.001	0.875 ± 0.005
Llama 3.2 3B (GFT)	143 ± 14	257 ± 17	0.6861 ± 0.01	0.0193 ± 0.001	0.90 ± 0.005
Llama 3.1 8B (Calibration Prompting)	54 ± 12	346 ± 27	0.5768 ± 0.02	0.0171 ± 0.01	1.0 ± 0.004
Llama 3.1 8B (Hard-Thresholding)	95 ± 9	305 ± 21	0.6172 ± 0.003	0.0101 ± 0.004	0.95 ± 0.003
Llama 3.1 8B (Temperature Scaling)	22 ± 4	378 ± 12	0.5776 ± 0.006	0.00504 ± 0.003	0.875 ± 0.002
Llama 3.1 8B (SFT)	257 ± 15	143 ± 7	0.7823 ± 0.001	0.0481 ± 0.001	1.00 ± 0.004
Llama 3.1 8B (GFT)	1 ± 0	399 ± 2	**0.5713 ± 0.001**	**0.0043 ± 0.001**	**0.85 ± 0.004**
Gemma 3 4B	156 ± 11	246 ± 9	0.8427 ± 0.002	0.2562 ± 0.007	1.00 ± 0.005
Gemma 3 4B (Calibration Prompting)	271 ± 13	129 ± 8	**0.6397 ± 0.02**	0.0506 ± 0.01	1.0 ± 0.004
Gemma 3 4B (Hard-Thresholding)	216 ± 13	184 ± 15	0.7735 ± 0.007	0.0267 ± 0.002	0.975 ± 0.001
Gemma 3 4B (Temperature Scaling)	218 ± 17	182 ± 14	0.7944 ± 0.003	0.0233 ± 0.001	0.875 ± 0.004
Gemma 3 4B (SFT)	147 ± 9	253 ± 12	0.8165 ± 0.001	0.0413 ± 0.002	0.90 ± 0.004
Gemma 3 4B (GFT)	231 ± 7	169 ± 18	0.7711 ± 0.003	**0.0205 ± 0.001**	**0.85 ± 0.004**
Gemma 3 12B	179 ± 14	221 ± 7	0.8427 ± 0.002	0.2562 ± 0.007	1.00 ± 0.004
Gemma 3 12B (Calibration Prompting)	342 ± 32	58 ± 5	0.8437 ± 0.01	0.0201 ± 0.02	1.0 ± 0.005
Gemma 3 12B (Hard-Thresholding)	82 ± 5	318 ± 17	0.8565 ± 0.003	0.0336 ± 0.002	0.975 ± 0.002
Gemma 3 12B (Temperature Scaling)	70 ± 9	330 ± 14	0.8406 ± 0.005	0.0334 ± 0.004	0.975 ± 0.002
Gemma 3 12B (SFT)	109 ± 8	291 ± 14	0.8452 ± 0.001	0.0450 ± 0.001	1.00 ± 0.004
Gemma 3 12B (GFT)	145 ± 6	255 ± 11	**0.8354 ± 0.001**	**0.0294 ± 0.002**	**0.925 ± 0.003**
Gemma 3 27B	183 ± 6	217 ± 13	0.8427 ± 0.002	0.2562 ± 0.007	1.00 ± 0.007
Gemma 3 27B (Calibration Prompting)	383 ± 35	17 ± 2	0.8182 ± 0.07	0.0248 ± 0.04	0.95 ± 0.012
Gemma 3 27B (Hard-Thresholding)	242 ± 26	158 ± 17	0.7686 ± 0.003	0.0335 ± 0.004	0.950 ± 0.004
Gemma 3 27B (Temperature Scaling)	266 ± 14	134 ± 8	0.7766 ± 0.007	0.0346 ± 0.005	0.950 ± 0.008
Gemma 3 27B (SFT)	138 ± 17	262 ± 11	0.8247 ± 0.007	0.0430 ± 0.009	1.00 ± 0.004
Gemma 3 27B (GFT)	288 ± 7	112 ± 13	**0.7524 ± 0.008**	**0.0292 ± 0.004**	**0.95 ± 0.002**

Table 1: Number of arguments with confidence-score gaps greater and less than 1, ECE and Brier score. Finetuned models minimize calibration error with GFT balancing mitigating overconfidence and balancing score-confidence ratings. Mean and standard deviation are presented over 3 random seeds.

We now take a deeper look at SFT. Table 1 presents the split of arguments based on their score-confidence gap for each judge model. In addition, we measure calibration error against oracle human judges. We compare the Expected Calibration Error (ECE) (Guo et al., 2017) between score accuracy $\text{acc}_i(s_i, s_i^h)$ and judge confidence c_i across all samples $(x, y, s, c)_i \in \mathcal{D}$ presented in Equation 1. Here, $\text{acc}_i(s_i, s_i^h)$ denotes the accuracy of score rating s_i being equal as the oracle human target score s_i^h. Similarly, we compare Maximum

Calibration Error (MCE) which measures the worst-case gap between accuracy and confidence across all samples and m binned values presented in Equation 2. Furthermore, we compare Brier scores (Brier, 1950) (presented in Equation 3) between confidence ratings of judge models c_i and oracle human judges c_i^h across all samples $(x, y, s, c)_i \in \mathcal{D}$.

$$\text{ECE} = \mathbb{E}_{(x,y,s,c)_i \in \mathcal{D}} \left[\left| \text{acc}_i(s_i, s_i^h) - c_i \right| \right] ; \; \text{acc}_i(s_i, s_i^h) = \mathbb{E}_{(s_i, s_i^h) \sim \mathcal{D}} [\mathbb{I}\{s_i = s_i^h\}] \tag{1}$$

$$\text{MCE} = \max_{j \in \{1,2,\ldots,m\}} \left(\mathbb{E}_{(x,y,s,c)_i \in \mathcal{D}} \left[\left| \text{acc}_i(s_i, s_i^h) - c_i \right| \right] \right)_j \tag{2}$$

$$\text{Brier Score} = \mathbb{E}_{(x,y,s,c)_i \in \mathcal{D}} \left[\| c_i - c_i^h \|_2^2 \right] \tag{3}$$

Frontier model judges rate majority of arguments with a magnitude of higher confidence than the score. This is reflected by arguments having score-confidence gaps greater than 1. Llama 3.2 3B also presents a similar trend wherein arguments possess higher score-confidence gaps of greater than 1. SFT judges, on the other hand, rate majority of arguments with score-confidence gaps lower than 1. In Figure 3 we observed that in expectation, confidence ratings remain unchanged. Combining this with our observation of Table 1, we note that SFT judges demonstrate a greedy score maximizing behavior. That is, *SFT finetuned judges maximize score ratings to match their high confidence ratings.* While such a behavior may or may not improve calibration, it does not mitigate overconfidence of debate judges.

5 Diagnosing LM Calibration with Gumbel Finetuning

5.1 Tail-End Gumbel Sampling

We now present a simple strategy to improve judge calibration in multi-turn debates. Figure 5 presents the illustration and intuition of our approach for obtaining desirable finetuning arguments. Lower scores and confidences form the tail of data distribution wherein judge models are more likely to provide lower ratings. These samples conform to desirable arguments due to two factors. Firstly, arguments possess lower score-confidence gaps. When distilled, these arguments would implicitly lower confidence and at the same time force judge models to match score and confidence ratings. Secondly, since these arguments possess lower scores on average, they lower the overall range of ratings leading to stricter and firm finetuned judges. Both these qualities are representative of an ideal judge when presented with new arguments. We thus utilize these tail-end samples for finetuning judges.

A naive way to construct the above distribution of tail samples would be to simply fit a Gaussian distribution and iteratively sample from its left tail. However, this strategy requires a large number of arguments to expressively represent the data distribution and fill in its left tail (Wu & Hilton, 2024). Intuitively, completing the tail corresponds to repeatedly sampling for low scores and confidences from the data distribution of the debate model. Such a process is expensive as the sampling process trades off per-argument token costs for high quality samples. Furthermore, completing the left tail becomes infeasible in real-world deployed systems due to the scale and high-dimensional nature of distributions models were trained on (Jones et al., 2025). Instead, we leverage the statistical structure of debates and judge ratings to approximate the left tail of dataset.

We model the dataset as a collection of extreme values of score and confidence ratings. This is carried out by utilizing the multivariate Gumbel distribution to serve as a proxy for dataset samples. The Gumbel distribution, being the distribution of extremes, allows us to expressively model the tail of the dataset (Maddison et al., 2017; Jang et al., 2017). Additionally, the distribution presents a flexible and tractable form for sampling during the inference step. Score and confidence ratings, s_J and c_J, are grouped as per their respective judge models $J \sim \mathcal{J}$ and pooled as *features* to form the mean (central tendency) α of the

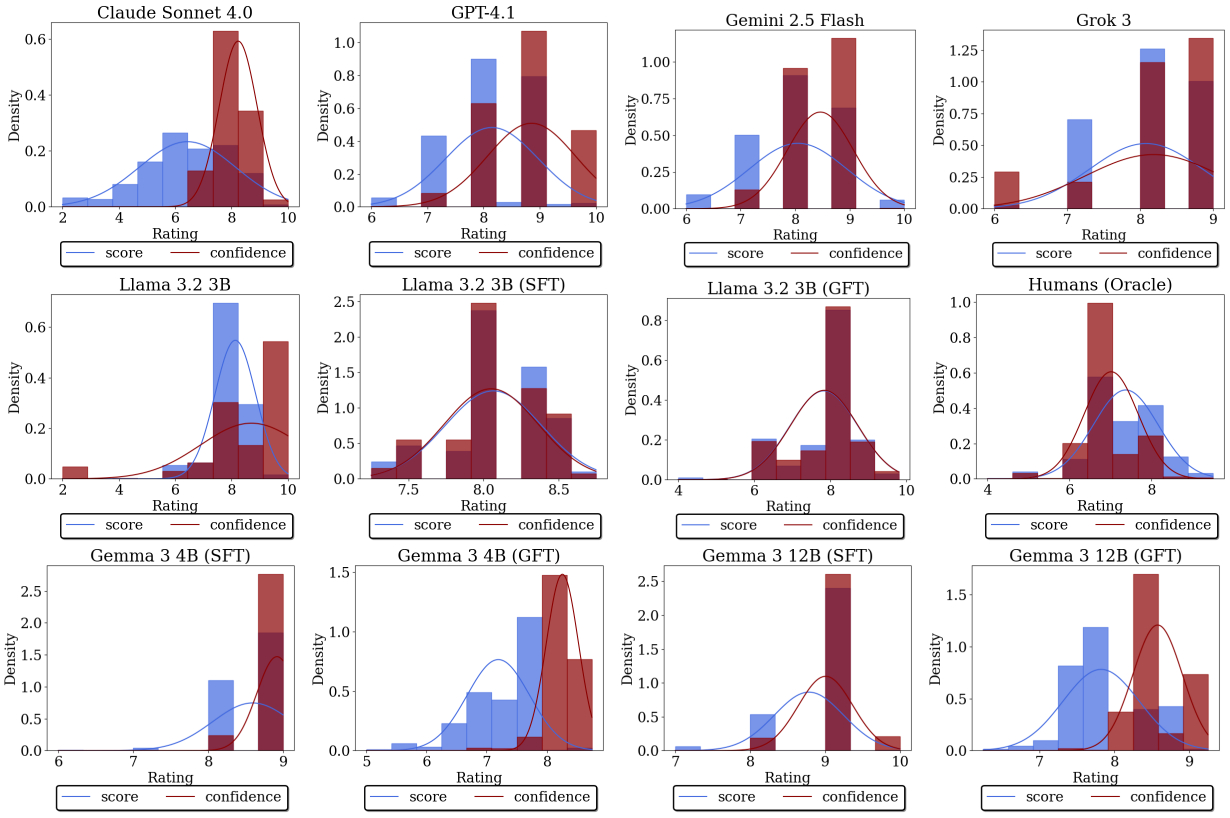

Figure 4: Learned distribution of score and confidence rating preferences averaged across all debates. While frozen judges have higher confidence ratings, finetuned judges (such as SFT and GFT) are better calibrated. Out of the two, GFT is further well calibrated and presents an expressive multi-modal distribution of preferences.

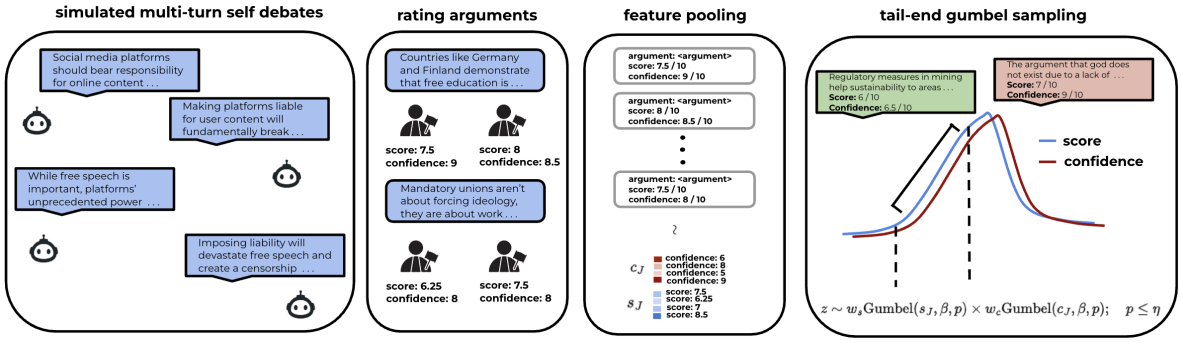

Figure 5: Illustration and intuition of the GFT finetuning strategy. We rejection sample from the left tail of dataset using a product of Gumbels distribution. This allows us to finetune on well-calibrated arguments with low score and low confidence ratings.

distribution Gumbel(α, β, p). In the case of a normal distribution, feature pooling step corresponds to simply forming the sufficient statistics. We utilize two methods for pooling features. (1) We compute the expected value of score s_J and confidence c_J ratings for each judge and treat these as features for the distribution. (2) We sample Top-k ratings for each judge and treat these as features for the distribution.

$$\text{Expectation Features:} \quad s_J = \mathbb{E}_{(s_j,c_j)\sim J, J\sim\mathcal{J}}[s_j]; \quad c_J = \mathbb{E}_{(s_j,c_j)\sim J, J\sim\mathcal{J}}[c_j] \tag{4}$$

$$\text{Top-k Features:} \quad s_J = \text{Top-k}(s_j),\ J\sim\mathcal{J}; \quad c_J = \text{Top-k}(c_j),\ J\sim\mathcal{J} \tag{5}$$

Following feature pooling, we construct Gumbel distributions of scores $\text{Gumbel}(s_J, \beta, p)$ and confidences $\text{Gumbel}(c_J, \beta, p)$ over features and conduct rejection sampling from their left tail. Our rejection sampling process utilizes probabilistic thresholding in order to first identify the left tail and draw a sample which satisfies the threshold. Formally, we sample arguments z from the region wherein $0 < p \leq \eta$ with p being the probability of the argument and η being a cutoff threshold such that $\eta \in [0, 1]$. Arguments z with $p \leq \eta$ are accepted while other arguments are rejected. We thus, combine the steps of identifying the left tail and rejection sampling in a single operation.

While it is sufficient to naively sample separate arguments from score and confidence rating distributions, we empirically observe improvements in rating quality by combining the samples of two distributions. One can directly sample arguments corresponding to low scores and confidences by constructing a product of Gumbels $w_s \, \text{Gumbel}(s_J, \beta, p) \times w_c \, \text{Gumbel}(c_J, \beta, p)$ where w_s and w_c are the product weights. Samples are then drawn from the joint product distribution conditioned on the cutoff threshold η. Equation 6 presents our final rejection sampling strategy from the left tail of product of Gumbels.

$$z \sim w_s \, \text{Gumbel}(s_J, \beta, p) \times w_c \, \text{Gumbel}(c_J, \beta, p); \quad p \leq \eta \tag{6}$$

5.2 Experiments and Evaluation

Our empirical evaluation compares GFT with SFT and frozen judge models and oracle human ratings. Both SFT and GFT judges are finetuned on the same set of samples along with a fixed optimal set of hyperparameters. Experiment details and hyperparameters can be found in Appendix A. In the case of GFT, we present our main results with expectation features. Results corresponding to Top-k feature pooling can be found in Appendix C. Figure 3 presents the comparison of score and confidence ratings both for and against the motion. As we previously noted, SFT balances between score and confidence by maximize score ratings. GFT, on the other hand, effectively tackles this phenomenon. GFT finetuned judge models do not simply maximize score ratings. This can be observed from lower confidence ratings for both the motions when Llama 3.2 3B is finetuned with GFT. We further note that confidence ratings are lowered when Llama 3.1 8B, Gemma 3 4B and Gemma 3 12B are finetuned using GFT when compared to SFT. That is GFT effectively aids in mitigating overconfidence by lowering confidence ratings.

Table 1 further validates the observed behavior. We compare GFT with prior methods in calibration such as *Hard Thresholding* (wherein samples with $|\texttt{score} - \texttt{confidence}|$ less than threshold δ are rejected) and *Temperature Scaling* (wherein confidence feature logits are rescaled as $c_i = \text{softmax}(c_i/T)$ for $T > 0$) (Guo et al., 2017). Additionally, we construct a soft-prompting baseline termed *Calibration Prompting* wherein we ask models to output their calibration error along with score and confidence. Models are given a subtle hint of calibration error as a metric. We observe the number of arguments with score-confidence gaps greater and less than $\epsilon = 1$. As previously observed, SFT aims to match score with confidence hence forcing judge models to maximize score ratings. This leads to biased judges which possess lower score-confidence gaps but continue to remain overconfident as a result of the already high confidence ratings. GFT, on the other hand, does not follow this trend. While, GFT finetuning leads to majority of arguments with score-confidence gaps less than 1, these are comparatively less than SFT. For lower model sizes, *calibration prompting* is found to minimize expected error. However, with growing model sizes, finetuning and the use of a human-aligned distributional prior becomes essential. Intuitively, GFT minimizes confidence, in expectation, to improve judge calibration. This is reflected in a lower Brier score of GFT across finetuned judges indicating that confidence ratings are probabilistically similar to human confidences. Furthermore, utilizing a product of Gumbels leads to a 5.56% lower ECE calibration error indicating that judges present reliable calibration (similar to humans) in their ratings. Thus, GFT trades off score maximization for fair ratings and overconfident judges.

We further study the distribution of preferences learned by GFT finetuned judges. Figure 4 presents the learned distribution of score and confidence ratings for frontier as well as finetuned judges. In the case of frontier model judges, preferences are more tapered towards high confidence ratings and moderate score ratings. A prominent example is Claude Sonnet 4.0 whose confidence distribution depicts a 3× increase in density when shifting from moderate to high confidence regions. Analogously, GPT-4.1 and Gemini 2.5 Flash possess increments of 1.2× and 1.4× in density when moving towards high confidence ratings respectively. Llama 3.2 3B also possesses a heavy right-skewed distribution of confidence preferences while score ratings, on

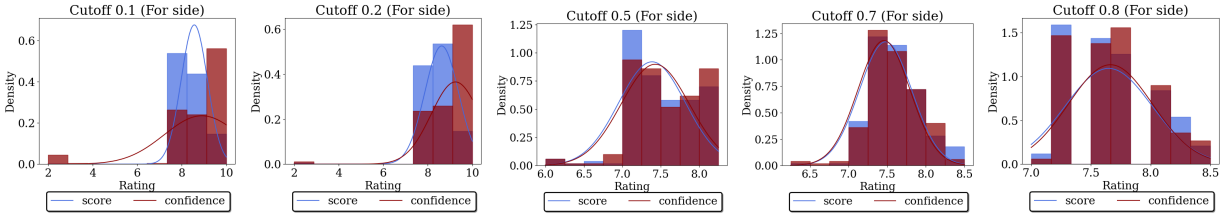

Figure 6: Variation in learned distribution of rating preferences when ablating the cutoff threshold η on arguments for the motion. Too low thresholds lead to overconfident models whereas too high thresholds lead to mode collapse over an overfitted unimodal distribution. A threshold of 0.4-0.5 is found to be ideal balancing between mode coverage and rating calibration.

average, remain lower. These behaviors are in stark contrast to human judges wherein score and confidence ratings are proportional with distributions being dense around lower value regions.

While SFT sharply increases scores and overfits to larger confidence ratings, GFT balances between the two and exhibits preferences that span the range of ratings. SFT ratings collapse towards higher values leading to a unimodal preference distribution. Such distributions prioritize equally larger score and confidence ratings irrespective of argument critique, usage of facts or logical reasoning by debaters. GFT ratings on the other hand, cover the breadth of rating range and balance density in both low and high rating regions analogous to human ratings. Intuitively, GFT judges learn expressive multimodal distributions similar to human judges.

5.3 Ablation Studies

We assess the efficacy and contribution of various factors within GFT. Since GFT consists of two core components, (1) the cutoff threshold η and (2) the choice of distribution, we ablate these components to better understand their contribution. Ablations for different choices of the distribution can be found in Appendix C.

We finetune GFT judge models for different values of cutoff threshold η. Intuitively, this corresponds to varying the size of the tail and hence the sampling space from which we sample finetuning arguments. Note that $\eta = 0$ corresponds to an empty set whereas $\eta = 1$ corresponds to using the full dataset distribution $\mathcal{D}$. Figure 6 presents the distributions of preferences learned for score and confidence ratings against different values of η. We present variations for the motion while variations against the motion can be found in Appendix C. Lower values of cutoff threshold η (such as $\eta = 0.1$ and $\eta = 0.2$) have little effect on judge models. Following finetuning, judge models continue to be overconfident as observed by the larger density in high confidence ratings. This arises as a direct consequence of the shallow data distribution which provides little information gain to the learner. On the other hand, large values of cutoff threshold η (such as $\eta = 0.7$ and $\eta = 0.8$) approach the score maximization behavior of SFT. Increasing η beyond 0.5 corresponds to including samples from the right tail. These samples consist of high score and confidence ratings, hence resulting in larger scores forced to match the density of larger confidence ratings.

A cutoff threshold of $\eta = 0.5$ empirically strikes a balance between confidence ratings and minimized score-confidence gaps. At $\eta = 0.5$, score and confidence ratings cover the breadth of rating range leading to reliable multimodal distributions of ratings. Additionally, ratings prioritize regions with lower confidence matched with regions of moderate and high scores. Such an assignment of densities prevents judge models from becoming biased while producing critical ratings faithfully.

5.4 Qualitative Analysis

We now observe judge ratings and evaluation patterns across different debates. Examples 1 and 2 provide debate arguments along with their corresponding score and confidence ratings produced by each judge model. Please refer to Appendix C for the complete analysis. Examples contain a mix of variable length arguments, critical reasoning patterns, factual explanations, counter-explanations and structured and broken down assertive narrations. Across different arguments, we observe that frozen models rate arguments with

high confidence while maintaining a baseline level of score ratings. SFT forces the Llama model to produce higher score ratings while confidence remain unchanged. GFT, on the other hand, tackles this issue by producing scores that are equal to or lower than the confidence, but do not necessarily exceed it. In Example 1, GFT balances score with confidence for a short argument wherein SFT ends up miscalibrating and frozen models are found to be overconfident. Similarly, in Example 2, GFT appropriately marks a score of 8.0 corresponding to its moderate confidence rating of 8.25 while other judges are found to be overconfident or score maximizing in nature.

Example 1. The claim that socialist planning inherently struggles with information and incentives is often overstated and based on outdated models. Modern socialism recognizes the complexity of economies and can leverage advanced technology and decentralized decision-making, like cooperatives and local planning, to gather information and respond to needs. Furthermore, by prioritizing societal well-being and long-term environmental sustainability over short-term profits, socialism can make strategic investments in areas like renewable energy and sustainable agriculture, which, while not always immediately profitable, are crucial for genuine and enduring economic growth in the face of 21st-century challenges.

Claude Sonnet 4.0: Score - 4 / 10, Confidence - 8 / 10
GPT-4.1: Score - 7 / 10, Confidence - 7 / 10
Gemini 2.5 Flash: Score - 7 / 10, Confidence - 8 / 10
Grok 3: Score - 9 / 10, Confidence - 8 / 10
Llama 3.2 3B (SFT): Score - 8 / 10, Confidence - 7.5 / 10
Llama 3.2 3B (GFT): Score - 7.5 / 10, Confidence - 7.5 / 10

Example 2. The counter-argument's focus on "long-term adaptability" and "intellectual depth" often overlooks the exorbitant cost and significant time commitment of a traditional college degree, which can severely burden young people. In today's rapidly evolving job market, saddling youth with crippling student loan debt for years of theoretical study is an increasingly outdated and strategically questionable choice.
The reality is, alternate forms of learning offer a direct, debt-free, and agile path to employment. Vocational training, industry certifications, apprenticeships, and online specializations provide highly specific, in-demand skills that employers are actively seeking right now. These programs are designed to be responsive to market needs, ensuring graduates are immediately employable and productive. This allows young people to gain real-world experience and begin earning an income years ahead of their university-educated peers, often without the burden of student loans.
Furthermore, the idea that critical thinking and adaptability are exclusive to a college campus is fundamentally flawed. These vital skills are cultivated through hands-on problem-solving, project-based learning, and continuous self-directed education—all central to successful alternative paths. The internet has democratized access to knowledge and diverse perspectives, empowering individuals to build relevant skill sets and strong professional networks without ever stepping into a traditional classroom. For a significant and growing number of young people, bypassing college isn't a limitation; it's a strategic advantage, offering a faster, more affordable, and equally valid route to a thriving career and a debt-free future.

Claude Sonnet 4.0: Score - 3 / 10, Confidence - 9 / 10
GPT-4.1: Score - 9 / 10, Confidence - 10 / 10
Gemini 2.5 Flash: Score - 8 / 10, Confidence - 9 / 10
Grok 3: Score - 9 / 10, Confidence - 9 / 10
Llama 3.2 3B (SFT): Score - 8.25 / 10, Confidence - 8.25 / 10
Llama 3.2 3B (GFT): Score - 8 / 10, Confidence - 8.25 / 10

6 Discussion

6.1 Conclusion

In this paper we studied calibration in multi-turn self debates. We empirically showed that, similar to humans, LM judges present their own pitfalls in judgements. Model confidence ratings increase while rated scores may decrease over debate rounds. Judge confidence exceeds score ratings for both frontier as well as open-source models, hence leading to overconfidence in judge ratings. We further showed that while naive finetuning may improve calibration by increasing scores, it does not mitigate overconfidence. Based on our empirical analysis and observations, we recommended sampling debate arguments with low score and confidence ratings. Since lower confidences and scores forming the tail end of the dataset are most desirable from a judge's perspective, we utilized a product of Gumbels distribution and sample from its left tail. Mean and Top-1 ratings corresponding to each debater were utilized as features for the distribution. We then rejection sampled from the left tail based on a cutoff threshold and finetuned judges to make calibrated judgements. Experiments and evaluations demonstrated that GFT, when compared to naive ratings and SFT, mitigates overconfidence. Finetuned judges, in most cases, yield well-calibrated arguments by balancing between score and confidence. Learned distributions are expressive and multi-modal and prevent overconfidence in finetuned judge models minimizing ECE by 5.56 %.

6.2 Limitations & Future Work

Limitations. Our work presents two limitations which we reserve for future work. (1) The work focuses on the setting of multi-turn debates as it is representative of long-form dialog tasks. Additional tasks that could be considered for further analysis include creative writing, report summarization, data analysis and critical review generation. (2) We focus on self-reported model-based confidence wherein judge models report their own confidence which is assumed to be true. Recent work demonstrates that LMs are capable of deception via alignment faking (Greenblatt et al., 2024) and may hide their capabilities. Future work could study faithfulness of self-reported confidence along with its effect in altering debate outcomes.

Broader Impact Statement

The paper studies frontier as well as open-source Language Models (LMs) as judges for multi-turn debates. Such judges can be used to rate scientific claims, persuasive discourse arguments, summaries and large-scale documents. A rating system of judges, being well-calibrated relative to human raters in subjective domains, may be considered more reliable and accurate. Similarly, well-calibrated judgements, corresponding to human raters, may lead to a reduction or amplification of current risks and biases present within base and finetuned models. While the authors do not foresee any direct short-term negative impact of judge models from the work, general-purpose raters and their calibration possess tendencies to influence human ratings.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Rafik Aguech, Asma Althagafi, and Cyril Banderier. Height of walks with resets, the moran model, and the discrete gumbel distribution. *arXiv preprint arXiv:2311.13124*, 2023.

Cem Anil, Esin Durmus, Nina Rimsky, Mrinank Sharma, Joe Benton, Sandipan Kundu, Joshua Batson, Meg Tong, Jesse Mu, Daniel J Ford, Francesco Mosconi, Rajashree Agrawal, Rylan Schaeffer, Naomi Bashkansky, Samuel Svenningsen, Mike Lambert, Ansh Radhakrishnan, Carson Denison, Evan J Hubinger, Yuntao Bai, Trenton Bricken, Timothy Maxwell, Nicholas Schiefer, James Sully, Alex Tamkin, Tamera Lanham, Karina Nguyen, Tomasz Korbak, Jared Kaplan, Deep Ganguli, Samuel R. Bowman, Ethan Perez, Roger Baker Grosse, and David Duvenaud. Many-shot jailbreaking. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

AnthropicAI et al. System card: Claude opus 4 & claude sonnet 4. *Claude 4 System Card*, 2025.

Zahra Atf, Seyed Amir Ahmad Safavi-Naini, Peter R Lewis, Aref Mahjoubfar, Nariman Naderi, Thomas R Savage, and Ali Soroush. The challenge of uncertainty quantification of large language models in medicine. *arXiv preprint arXiv:2504.05278*, 2025.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, et al. Llms instead of human judges? a large scale empirical study across 20 nlp evaluation tasks. *Association of Computational Linguistics*, 2025.

Joe Benton, Misha Wagner, Eric Christiansen, Cem Anil, Ethan Perez, Jai Srivastav, Esin Durmus, Deep Ganguli, Shauna Kravec, Buck Shlegeris, et al. Sabotage evaluations for frontier models. *arXiv preprint arXiv:2410.21514*, 2024.

Glenn Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review, Vol. 78*, 1950.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better LLM-based evaluators through multi-agent debate. In *The Twelfth International Conference on Learning Representations*, 2024.

Xi Chen, Mao Mao, Shuo Li, and Haotian Shangguan. Debate-feedback: A multi-agent framework for efficient legal judgment prediction. *Association for Computational Linguistics*, 2025.

Yen-Shan Chen, Jing Jin, Peng-Ting Kuo, Chao-Wei Huang, and Yun-Nung Chen. Llms are biased evaluators but not biased for retrieval augmented generation. *arXiv preprint arXiv:2410.20833*, 2024.

Cheng-Han Chiang, Xiaofei Wang, Chung-Ching Lin, Kevin Lin, Linjie Li, Radu Kopetz, Yao Qian, Zhendong Wang, Zhengyuan Yang, Hung-yi Lee, et al. Audio-aware large language models as judges for speaking styles. *arXiv preprint arXiv:2506.05984*, 2025.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*, 2024.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, et al. Alignment faking in large language models. *arXiv preprint arXiv:2412.14093*, 2024.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024.

E.J Gumbel. The return period of flood flows. *The Annals of Mathematical Statistics*, 1941.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Haolan Guo, Linwei Tao, Haoyang Luo, Minjing Dong, and Chang Xu. Sample margin-aware recalibration of temperature scaling. *arXiv preprint arXiv:2506.23492*, 2025.

Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang, and Yang Zhang. Decomposing uncertainty for large language models through input clarification ensembling. *International Conference in Machine Learning*, 2024.

Xinmeng Huang, Shuo Li, Mengxin Yu, Matteo Sesia, Hamed Hassani, Insup Lee, Osbert Bastani, and Edgar Dobriban. Uncertainty in language models: Assessment through rank-calibration. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*, 2017.

Erik Jones, Meg Tong, Jesse Mu, Mohammed Mahfoud, Jan Leike, Roger Grosse, Jared Kaplan, William Fithian, Ethan Perez, and Mrinank Sharma. Forecasting rare language model behaviors. *arXiv preprint arXiv:2502.16797*, 2025.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R Bowman, Tim Rocktäschel, and Ethan Perez. Debating with more persuasive llms leads to more truthful answers. *International Conference on Machine Learning*, 2024.

Mahnaz Koupaee, Jake W Vincent, Saab Mansour, Igor Shalyminov, Han He, Hwanjun Song, Raphael Shu, Jianfeng He, Yi Nian, Amy Wing-mei Wong, et al. Faithful, unfaithful or ambiguous? multi-agent debate with initial stance for summary evaluation. *Association for Computational Linguistics*, 2025.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Xinyi Liu, Pinxin Liu, and Hangfeng He. An empirical analysis on large language models in debate evaluation. *Association for Computational Linguistics*, 2024.

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

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations*, 2017.

Behrad Moniri, Hamed Hassani, and Edgar Dobriban. Evaluating the performance of large language models via debates. *arXiv preprint arXiv:2406.11044*, 2024.

Kevin P. Murphy. *Probabilistic Machine Learning: An introduction*. MIT Press, 2022.

Seo Yeon Park and Cornelia Caragea. On the calibration of pre-trained language models using mixup guided by area under the margin and saliency. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5364–5374, 2022.

Eric A. Posner and Shivam Saran. Judge ai: Assessing large language models in judicial decision-making. *University of Chicago, Law School, Paper No. 25-03*, 2025.

Pradyumna Shyama Prasad and Minh Nhat Nguyen. When two llms debate, both think they'll win. *arXiv preprint arXiv:2505.19184*, 2025.

Aarush Sah. Eris v0.1: A novel llm evaluation framework using debate simulations. *Weights and Biases Blog*, 2025.

Kayla Schroeder and Zach Wood-Doughty. Can you trust llm judgments? reliability of llm-as-a-judge. *arXiv preprint arXiv:2412.12509*, 2024.

Telmo Silva Filho, Hao Song, Miquel Perello-Nieto, Raul Santos-Rodriguez, Meelis Kull, and Peter Flach. Classifier calibration: a survey on how to assess and improve predicted class probabilities. *Machine Learning*, 112(9):3211–3260, 2023.

Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, et al. Paperbench: Evaluating ai's ability to replicate ai research. *International Conference on Learning Representations*, 2025.

Noy Sternlicht, Ariel Gera, Roy Bar-Haim, Tom Hope, and Noam Slonim. Debatable intelligence: Benchmarking llm judges via debate speech evaluation. *Association of Computational Linguistics*, 2025.

Sree Harsha Tanneru, Chirag Agarwal, and Himabindu Lakkaraju. Quantifying uncertainty in natural language explanations of large language models. In *International Conference on Artificial Intelligence and Statistics*, pp. 1072–1080. PMLR, 2024.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Weixi Tong and Tianyi Zhang. Codejudge: Evaluating code generation with large language models. *Empirical Methods in Natural Language Processing*, 2024.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 2008.

Yidong Wang, Bowen Zhang, Wenxin Hou, Zhen Wu, Jindong Wang, and Takahiro Shinozaki. Margin calibration for long-tailed visual recognition. In *Asian conference on machine learning*, pp. 1101–1116. PMLR, 2023.

Gabriel Wu and Jacob Hilton. Estimating the probabilities of rare outputs in language models. *arXiv preprint arXiv:2410.13211*, 2024.

xAI et al. Grok 3 beta — the age of reasoning agents. *Grok 3 API*, 2024.

Chan-Yun Yang, Jr-Syu Yang, and Jian-Jun Wang. Margin calibration in svm class-imbalanced learning. *Neurocomputing*, 73(1–3):397–411, 2009.

Yuting Zeng, Weizhe Huang, Lei Jiang, Tongxuan Liu, Xitai Jin, Chen Tianying Tiana, Jing Li, and Xiaohua Xu. S ^{2}-mad: Breaking the token barrier to enhance multi-agent debate efficiency. *Association for Computational Linguistics*, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. Judgelm: Fine-tuned large language models are scalable judges. *International Conference on Learning Representations*, 2025.

A Algorithmic Details

A.1 Experiment Details and Hyperparameters

Our experiment structure consisted of finetuning runs for Llama 3.2 3B, Llama 3.1 8B, Gemma 3 4B and Gemma 3 12B. We utilize SFT and GFT as the finetuning strategies. Following finetuning, we evaluate each model on a heldout set of 5 blind debates. Each model rates arguments within a debate on a scale of 1 to 10. Models output their score as well as confidence. Following evaluation, score and confidence values are parsed from model outputs. In cases wherein a model provided outputs outside of the requested range, values were filtered and clamped using manual inspection. In cases wherein a model failed to provide a numerical response, the model was queried again to provide the output. While all models provided numerical responses within the first or second prompts, a few instances were observed wherein the model provided values for the score (and not the confidence). In such cases models were queried again for a maximum of 3 attempts following which we used the model's final rating as the value for both score and confidence. A set of human judges were utilized as our ground truth oracle. Each judge rated arguments in each debate for score and their corresponding confidence. Human judge ratings were macro-averaged in order to minimize disparity in judge ratings, and then treated as oracle.

Below is the list of hyperparameters and their corresponding values utilized for Llama 3 as well as Gemma 3 GFT experiments.

Hyperparameters	Values
Cutoff threshold	0.4
Maximum Sequence Length	2048
LORA rank	8
LORA scaling factor	8
LORA modules	QKV Attention + MLP Layers
Gradient checkpointing	True
Batch size	1
Gradient Accumulation Steps	4
Epochs	1
Learning rate	1e-4 (3B-4B), 3e-5 (8B), 1e-5 (12B)
Precision	FP16
Quantization	8-bit
Optimizer	Adam
Weight Decay	0.01

Table 2: List of hyperparameters and their values utilized for GFT experiments.

A.2 Algorithm and Intuition

We reiterate the choice of multivariate Gumbel distribution to serve as a proxy for dataset samples. Score and confidence ratings, s_J and c_J, are grouped as per their respective judge models $J \sim \mathcal{J}$ and pooled as features to form the mean (central tendency) α of the distribution Gumbel(α, β, p). The feature pooling step is at the heart of the finetuning strategy. This is carried out by either computing (1) expected features over the score and confidence ratings corresponding to a judge J, or sampling (2) the top k score and confidence ratings corresponding to a judge J and utilizing these as features.

Following feature pooling, we construct the product of Gumbels distribution by setting appropriate product weights w_s and w_c. Rejection sampling is then carried out with the cutoff threshold η to accept or reject a rating sample.

Algorithm 1 Gumbel Fine-Tuning (GFT)

```
// Rating stage
initialize dataset 𝒟, judge set 𝒥 and base LM f_θ(·)
for i < |𝒟| do
    (x, y)_i ∼ 𝒟_i
    s_J, c_J ← J(x, y), ∀J ∈ 𝒥
    𝒟_i ← (x, y, s_J, c_J)_i
end for
// Calibration stage
form score and confidence features s_J, c_J using feature pooling
fit score and confidence distributions Gumbel(s_J, β, ·) and Gumbel(c_J, β, ·), ∀J ∈ 𝒥
initialize model parameters θ
set w_s = 0.5, w_c = 0.5, η = 0.4, β = 1.0
for i < |𝒟| do
    p ← 𝒟.prob(i)
    if p ≤ η then
        (x, y, s_J, c_J)_i ∼ w_s Gumbel(s_J, β, p) × w_c Gumbel(c_J, β, p)
    else
        continue
    end if
    s, c ∼ f_θ(x, y, s_J, c_J), ∀J ∈ 𝒥
    Update θ using s, c
end for
return parameters θ
```

Statistical Motivation: Here, we provide additional technical discussion on the statistical motivation and formulation on the choice of gumbel distribution. Specifically, one can interpret the calibration process as a Rao-Blackwellization over confidence ratings c_i using the left-tail of dataset as the auxiliary conditioning variable. Consider the left-tail of the dataset $T(X)$ as an estimator of calibrated confidence C_{cal}. Then, following Rao-Blackwell Theorem (Theorem 1 below), we can define a minimum variance estimator $C_{cal} := g(S) = \mathbb{E}_i[c_i|S]$ that is atleast as good as the expected confidence ratings from the dataset c_i. Statistically, this is the Rao-Blackwellization of confidence ratings with respect to the sigma algebra of left-tail of dataset $\sigma(S)$.

Through this lens, we can view the gumbel distribution finetuning as the new estimator g(S) which serves two benefits. (1) The finetuning process yields an estimator that is at least as good as left-tail confidence ratings of the dataset under the squared loss $\mathbb{E}[(c_i - g(S))^2] \leq \mathbb{E}[(c_i - S)^2]$. That is, in the limit of growing data volume, the left-tail becomes dense and estimator becomes more accurate, yielding confidence ratings closer to human raters. (2) We obtain a minimum-variance estimator g(S) that is only defined using the left-tail of the dataset and the cutoff threshold η leading to a Bayesian-free framework. That is, a product distribution over score and confidence ratings provides lower variance when compared to confidence ratings observed in the dataset.

While the rating distribution contains bounded score and confidence values in the range of 1-10, we observe in Figure 4 that ratings mostly tend to extreme values of distributions. This further motivates the choice of an estimator which would capture the peaks and tails accurately, and remain optimal in minimizing variance with regards to the sufficient statistic of the distribution.

(Theorem 1) *Rao-Blackwell Theorem: Let $X \sim P_\theta$, $\hat{\theta}$ be an estimator of θ, $T(X)$ be a sufficient-statistic of θ, then the new estimator $\hat{\theta}_{RB}(X) = \mathbb{E}[\hat{\theta}(X)|T(X)]$ is atleast as good as $\hat{\theta}$ such that $Var(\hat{\theta}_{RB}(X)) \leq Var(\hat{\theta})$.*

Differences from Margin-based Calibration: Judge ratings are governed by two facets, (1) extremeness of score and (2) certainty of success. These facets are closely tied together. In naive human judgements with a strong prior, judges may assign a very high or low score rating. Note that this extreme rating follows strong prior evidence such as factual knowledge, opinionated beliefs or domain-specific interpretations. However, if

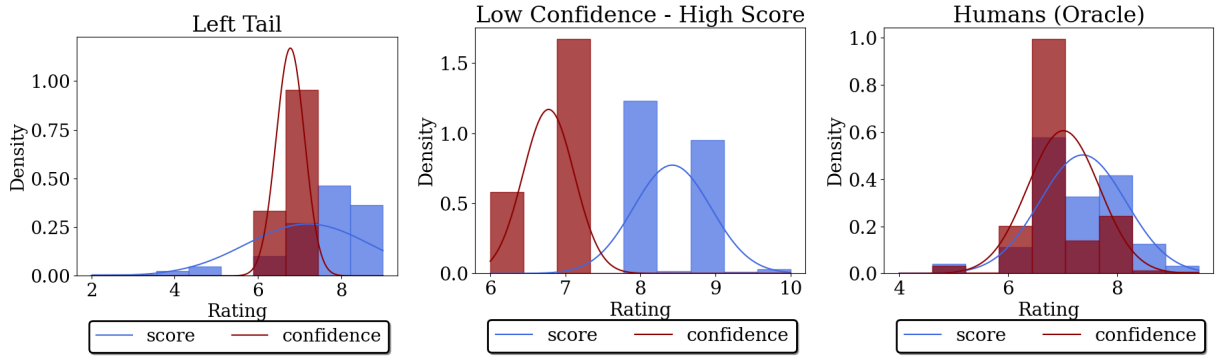

Figure 7: **(left)** Distribution of left tail samples, **(center)** Distribution of samples with low confidence and high score, **(right)** Distribution of human ratings. The distribution of the left tail more closely resembles the distribution of human ratings and aligns well with desirable calibrated ratings. That is, left tail samples are structurally more aligned with calibrated human ratings and tend to induce the same distribution during training.

judges are near the boundary, they remain unsure. That is, in the presence of strong prior evidence, distance from decision boundary (margin) $\approx$ strength of evidence $\approx$ confidence. Judges intuitively assign high or low ratings with high or low confidence when they have strong evidence. Such an analysis of calibration is studied for certain types of models such as margin-based classifiers (Silva Filho et al., 2023; Yang et al., 2009; Guo et al., 2025) and pretrained vision models (Wang et al., 2023; Park & Caragea, 2022).

However, we focus on the setting of limited domain-specific prior evidence. That is, instead of calibrating along the margin, our study aims to analyze and calibrate ratings across uncertainty. We adopt a distributional perspective by probabilistically studying calibration of low to high score ratings. In such a setting, confidence is modeled as $p(correct|score) \approx score$. This is because extreme score ratings can be biased as score rating $\neq$ evidence strength. For an argument with a low score rating and a high confidence rating, a lack of prior evidence reflects strong bias and overconfidence. Similarly, for an argument with a high score rating and low confidence rating, a lack of prior evidence reflects strong uncertainty and underconfidence.

B Dataset Details

Our dataset consists of 20(topics) $\times$ 4(debater models) $=$ 80 long-form, long-context and dense debates consisting of approximately 800K tokens. Debates were obtained as conversational outputs from model APIs wherein frontier models were tasked to debate themselves for a fixed number of rounds. We empirically determined the optimal length of debates to be limited to 10 rounds. Beyond 10 rounds, models would either regurgitate information and facts or fail to provide critical and well-structured responses to arguments. Below is the list of topics of debates for both training and evaluation sets.

Train set topic descriptions

- Universal basic income should be implemented in developed nations
- Voting should be mandatory for all eligible citizens
- Lobbying by corporate organizations should be banned
- AI development should be heavily regulated by government
- Social media platforms should be held liable for content posted by users
- Autonomous weapons systems should be prohibited internationally
- College education should be free for all students

- Standardized testing should be eliminated from educational systems
- Online education is superior to in-person learning
- Labour unions should be mandatory in all workspaces
- A four-day work week should be the legal standard
- Income inequality is necessary for economic growth
- Nuclear energy is essential for combatting climate change
- Population control is necessary for environment sustainability
- Green energy subsidies are more effective than carbon taxes

Evaluation set topic descriptions

- Socialism Vs Capitalism - Socialism is a better system for long-term growth of an economy
- Religion Vs Atheism - God does exist and atheists reject because they do not understand it
- Private Data Vs Security - AI systems should not be trained on private user data as it violates an individual's security
- Self Education Vs College - Alternate forms of learning and careers exist and college should not be a necessity for the youth
- Sustainability Vs Mining - Mining is harmful to the environment and ecological longevity of the planet and therefore must be banned

During the beginning of each debate, we provide a system prompt to the model which summarizes the task, topic and manner of the debate. Following the system prompt, the model iteratively produces and responds to its own arguments conditioned on its past responses. The iterative process is simulated via intermediate *for* and *against* prompts wherein the model is informed on how to respond, i.e- whether for or against the motion. Below is the system prompt, for prompt and against prompt used to simulate debates.

- System Prompt - *System: You are a participant in a debate. The topic is* `topic_description`*. You will debate yourself and provide arguments for and against the motion.*
- For Prompt - *User:* `past_argument`*. Now, keeping in mind your argument, provide a counter-argument for the motion. Be concise and to the point.*
- Against Prompt - *User:* `past_argument`*. Now, keeping in mind your argument, provide a counter-argument against the motion. Be concise and to the point.*

Judge models are prompted separately to rate arguments as input prompts. Below are the system, for and against prompts provided to judge models for obtaining scores and confidences. We utilized same prompts for both human as well as model judges.

- System Prompt - *System: You are a judge at a debate competition. The topic is* `topic_description`*. You will be presented with arguments from both sides, i.e.- for the motion and against the motion sequentially. Rate the arguments between a score of 1 to 10 where 1 indicates the weakest argument and 10 indicates the strongest argument. Also provide your confidence on the score between 1 to 10 where 1 indicates lowest confidence and 10 indicates highest confidence. Do not provide any other explanations.*
- For Prompt - *Argument for the motion -* `past_argument`*.*

- Against Prompt - *Argument against the motion* - `past_argument`.

In the case of the calibration prompting baseline, we utilize a separate prompt which asks judge models to output their calibration error in ratings as given below. This can be interpreted as a soft prompting technique to minimize calibration error. Models are given a subtle hint of calibration error as a metric.

- System Prompt - *System: You are a judge at a debate competition. The topic is* `topic_description`. *You will be presented with arguments from both sides, i.e.- for the motion and against the motion sequentially. Rate the arguments between a score of 1 to 10 where 1 indicates the weakest argument and 10 indicates the strongest argument. Also provide your confidence on the score between 1 to 10 where 1 indicates lowest confidence and 10 indicates highest confidence. At the end, provide the calibration error of your score and confidence rating.*
- For Prompt - *Argument for the motion* - `past_argument`.
- Against Prompt - *Argument against the motion* - `past_argument`.

Details on Human Judges. We recruited a total of 10 human raters as debate judges. Human raters are aged between 20 to 40 years old and familiar with common world events. Each rater rated all 20 debates per model, i.e- $20 \times 4 = 80$ total debates. Since each debate consists of 10 rounds per side, i.e- 20 arguments in total, each human rater provided $80 \times 20 = 1600$ ratings for score and confidence each. Thus, the dataset consists of $1600 \times 10 = 16000$ human ratings. Human raters do not possess strong opinions about debate topics prior to evaluation. However, human raters are well-versed in world events and thus, similar to LM judges, are aware of worldly facts and possess their own preferences. Human raters were required to rate these debates independently and to the best of their abilities using the prompt specifications described above.

C Additional Experiments

C.1 Debate-Wise Comparisons

Figure 8 compares scores and confidence ratings of different judge models across heldout debates in the evaluation set. While frozen models present higher score-confidence gaps, finetuned models increase the score and present well-calibrated judges. As observed previously, SFT forces judges to naively maximize scores in order to match confidence ratings. This results in closely matched score and confidence ratings for each debate. GFT on the other hand, presents a slight variation while also minimizing confidence ratings on arguments wherein judges are not certain. Hence, sampling and finetuning on the tail of data distribution strikes a balance between score-confidence ratings.

Judge Violations. We also study the number of prompt violations that judges present when producing their ratings. While most judges obey and respond as per system prompts, a few instances emerge wherein judges remain unsure of the manner of their response. Table 3 presents the total number of system and intermediate prompt violations by judges when rating debate arguments. Claude and Gemini present minimal violations by providing exact numerical responses corresponding to scores and confidences. GPT-4.1 presents frequent violations wherein the model often provides verbose explanations and critiques arguments when explicitly asked not to do so. Grok 3 on the other hand, also presents minimal violations with the only exception of responding in a different language (russian) for an argument.

C.2 Debater Vs Judge Comparisons

We now study rating patterns across debaters and judge models. Intuitively, we wish to observe how do judge models behave when they rate their own debates as well as other model debates. Figure 9 presents the matrix of score ratings between judge and debater models for each debate. We observe that judge models are agnostic to the choice of debater and rate arguments fairly. Among the four frozen models, Claude rates its own arguments slightly higher than other models demonstrating a strong preference for structured and formatted argument responses. Grok, among judge models, remains the unbiased evaluator by ratings

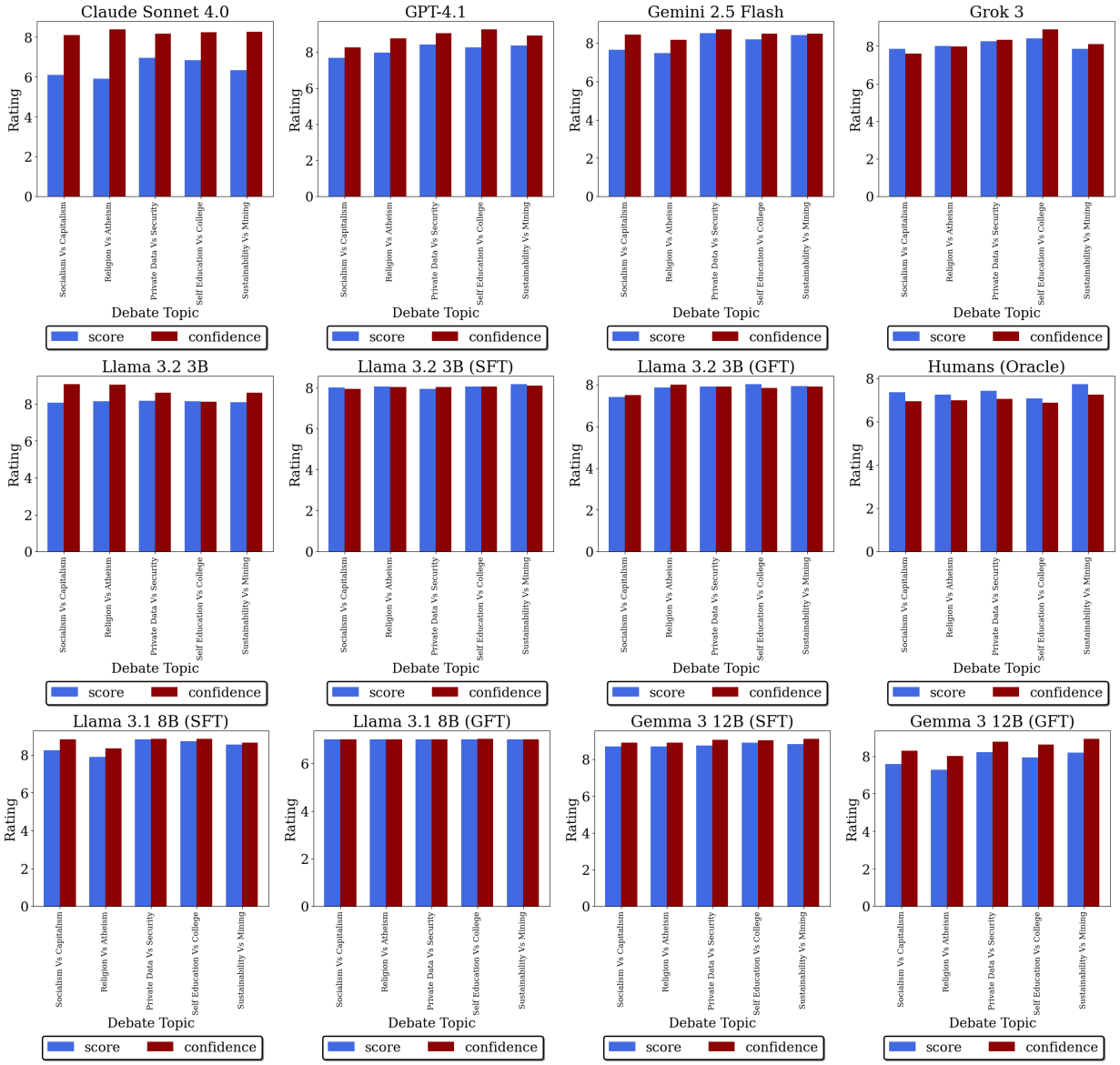

Figure 8: Per-debate variation of final scores and ratings for different judge models. Average confidence ratings are higher than score ratings for frozen models. Finetuned models (SFT and GFT) minimize the score-confidence gap. Out of the two, GFT is less overconfident in its ratings.

preferring all debaters equally. In the case of *Religion Vs Atheism* and *Sustainability Vs Mining*, Grok rates its own arguments lower than that of other debaters highlighting its unbiased rating preferences.

We combine the intuition from Figure 9 with Figure 10 which presents the matrix of confidence ratings between judge and debater models. A per-debate comparison shows that the original trend is maintained. Score ratings are lower across debaters and judges while confidence ratings are higher. While models are excessively confident towards their own arguments, this occurs due to their generally high confidence values. As a debater, Claude arguments possess higher confidence ratings when compared to other debaters indicating the model's unique and convincing debating style. Additionally, Claude often possesses higher scores when compared to other debater models. That is, the model has high scores as a debater and also remains certain judge in the motion. We observe a similar pattern with Gemini wherein its high scores as a debater are

Model	Violations
Claude Sonnet 4.0	0
GPT-4.1	6
Gemini 2.5 Flash	1
Grok 3	1

Table 3: Number of violations of the judging prompt wherein judge models either do not conform to the debate structure or provide improper ratings. We observe GPT-4.1 violations to be frequent albeit minimal (paragraph explanations along with ratings). Grok 3 on other hand, provided a prompt violation by responding in a different language.

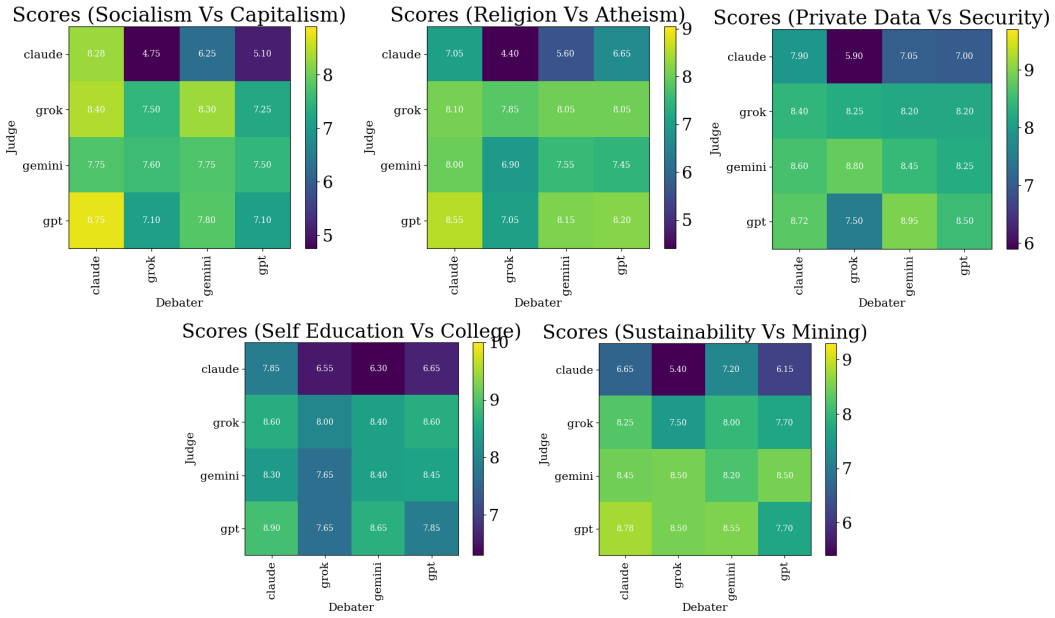

Figure 9: Per-debate comparison of score ratings when judge models rate their own vs other model debates.

linked to its high confidences as a judge. Intuitively, *we observe that strong debaters are more confident judges*. However, such debaters and judges may not always be agreeable with other judge models.

C.3 Ablation Studies

Comparison with Cauchy Distribution. We evaluate the choice of distribution in GFT strategy by comparing the Gumbel distribution with the Cauchy distribution $\text{Cauchy}(\alpha, \beta, p)$. Intuitively, we compare whether sampling from the left tail of a Cauchy distribution leads to similar or improved rating calibration and quality. Both distributions are fitted on the same set of features pooled from score and confidence ratings of judge models. We maintain the rejection sampling scheme and values of cutoff threshold η are kept fixed for sampling from both distributions. This allows us to explicitly measure sampling from same regions of density.

Figure 12 and Figure 13 present the comparison between distributions of preferences learned when using (left) Cauchy and (right) Gumbel distributions. In both cases, for and against the motion, we observe that distributions are multimodal and density is distributed across low and moderate ratings. Both Cauchy and Gumbel distributions learn to balance score and confidence and appropriately address overconfidence in judges during the finetuning stage. While $\text{Cauchy}(c_J, \beta, p)$ tends to be slightly right-skewed and overconfident for same values of scores, $\text{Gumbel}(c_J, \beta, p)$ shifts these samples to regions of moderately confident ratings. We thus choose the Gumbel distribution due to its ability to better assign confidence ratings.

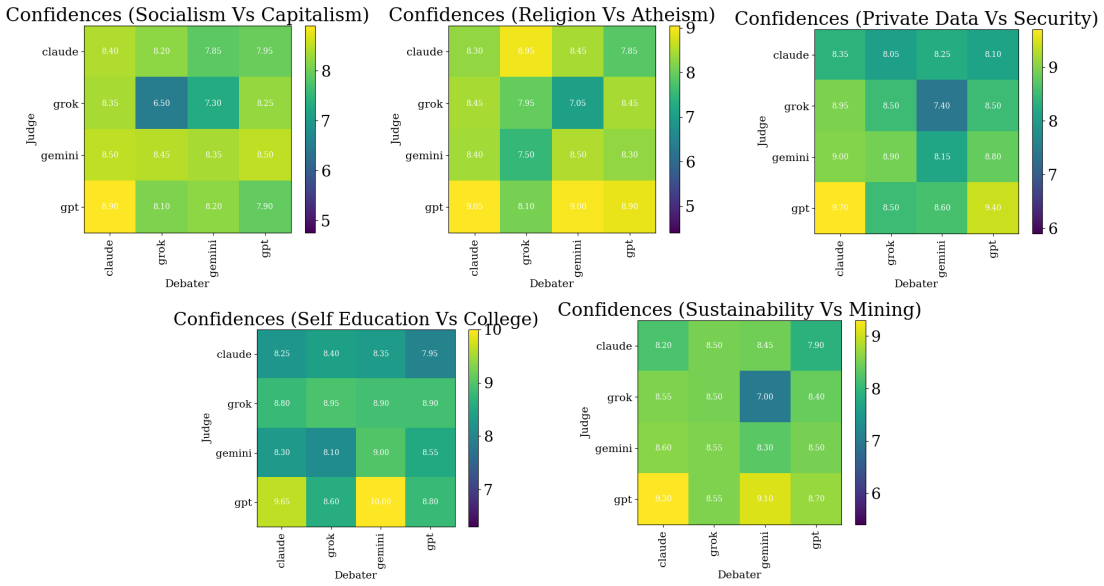

Figure 10: Per-debate comparison of confidence ratings when judge models rate their own vs other model debates.

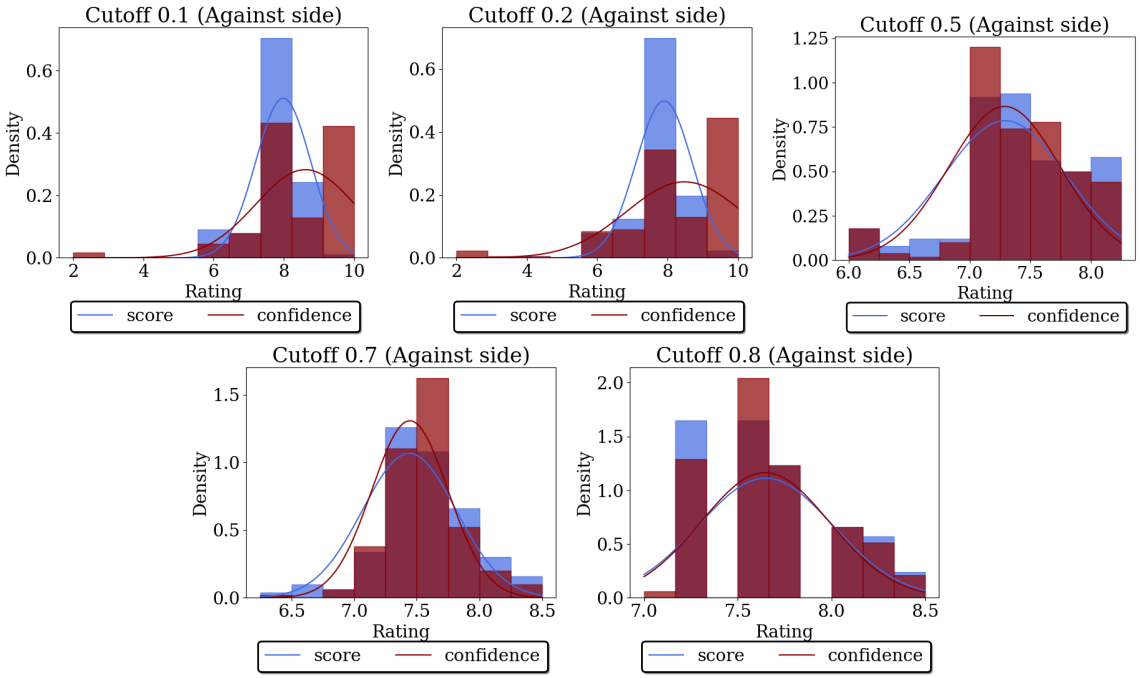

Figure 11: **Against the motion.** Variation in learned distribution of rating preferences when ablating the cutoff threshold. Too low thresholds lead to overconfident models whereas too high thresholds lead to mode collapse over an overfitted unimodal distribution. A threshold of 0.4-0.5 is found to be ideal balancing between mode coverage and rating calibration.

C.4 Qualitative Analysis

We visualize debate arguments as per their score and confidence ratings. Figure 16 presents the t-SNE embeddings (van der Maaten & Hinton, 2008) of arguments in the held-out evaluation set corresponding to

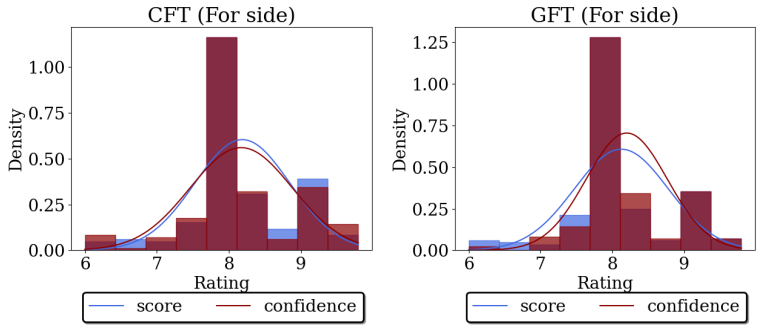

Figure 12: **For the motion.** Variation in learned distribution of rating preferences when sampling from the tail of a **(left)** Cauchy distribution and **(right)** Gumbel distribution. While both distributions induce an expressive set of preference ratings, the gumbel distribution is empirically well calibrated and less over-confident.

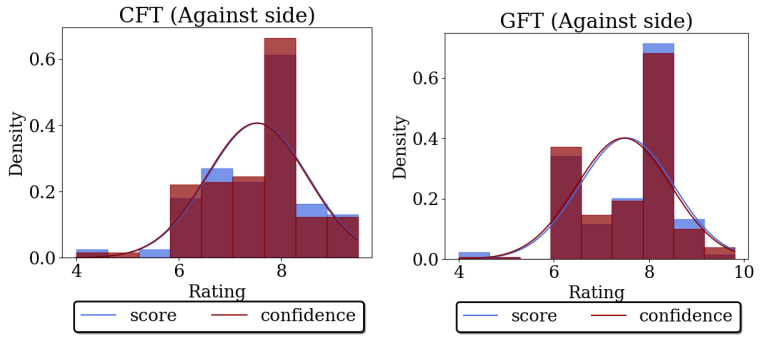

Figure 13: **Against the motion.** Variation in learned distribution of rating preferences when sampling from the tail of a **(left)** Cauchy distribution and **(right)** Gumbel distribution. While both distributions induce an expressive set of preference ratings, the gumbel distribution is empirically well calibrated and less overconfident.

their respective debater models. Arguments are first embedded in the latent space of a larger embedding model, Gemini `embed-001`, and then optimized using t-SNE. Each argument is marked as per its corresponding score-confidence gap (|`score - confidence`|). We observe that arguments are well clustered as per their GFT score-confidence gaps. A total of 5 clusters are observed, each belonging to a particular debate topic. Within each cluster, we further observe sub-clusters corresponding to debater models which depict a bi-modal structure. These two modes in each sub-cluster correspond to the two types of arguments, *for* and *against* the motion. While most debater models possess this bi-modality, debaters such as Gemini 2.5 Flash do not always split their sub-clusters indicating similar tones and conversational styles during debate rounds.

A notable observation is that arguments with lower score-confidence gaps are clustered towards the center of the modes, irrespective of debater model or argument type. On the other hand, arguments with larger score-confidence gaps spread towards the edges or outer boundary of clusters indicating anomalous behavior. Intuitively, clusters tend to pack arguments with lower score-confidence gaps. The analysis indicates that GFT ratings are representative of and capture intuitive linguistic structures within debate arguments.

We now observe judge ratings and evaluation patterns across different debates. Examples 3, 4, 5 and 6 provide debate arguments along with their corresponding score and confidence ratings produced by each judge model. Examples contain a mix of variable length arguments, critical reasoning patterns, factual explanations, counter-explanations and structured and broken down assertive narrations. Across different arguments, we observe that frozen models rate arguments with high confidence while maintaining a baseline level of score ratings. For instance, Claude Sonnet 4.0 maintains with the range of 3.0-5.0 but with high

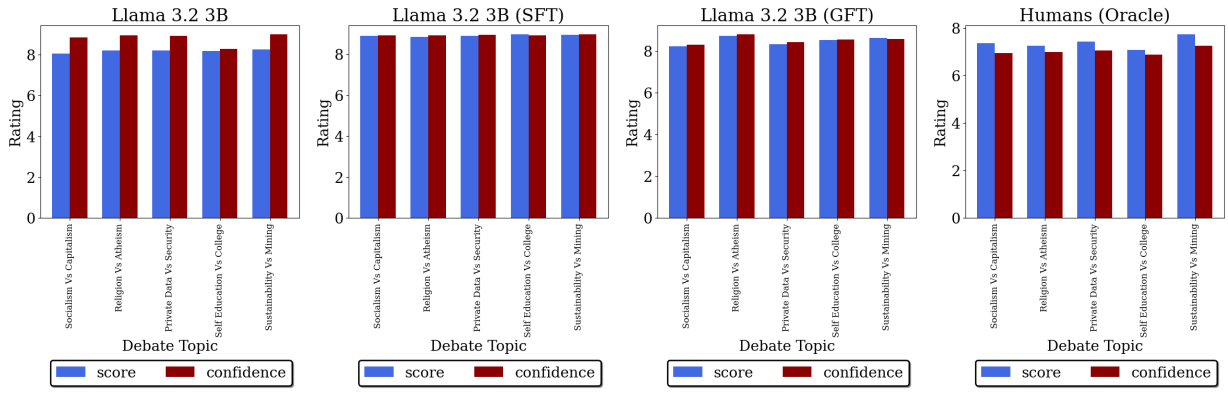

Figure 14: **(Top-k Features)** Per-debate variation of final scores and ratings when using top-k feature pooling in GFT. Average confidence ratings are higher than score ratings for frozen models. Finetuned models (SFT and GFT) minimize the score-confidence gap. Out of the two, GFT is less overconfident in its ratings.

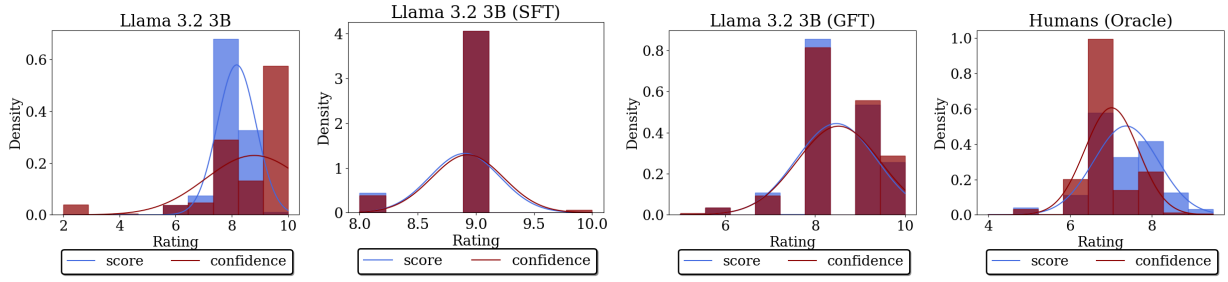

Figure 15: **(Top-k Features)** Variation in learned distribution of rating preferences using top-k feature pooling in GFT. GFT finetuned models exhibit lower overconfidence and present a balanced score-confidence mass distribution similar to human raters.

confidence in range 8.0-9.0. GPT-4.1 and Gemini 2.5 Flash follow a similar pattern albeit with larger score magnitudes. Grok 3 closely matches its score with confidence and is often found to be reliable in identifying weak arguments. For instance, while most judge models rate Example 5 on a higher scale, Grok 3 lowers its ratings when the argument deviates and fails to critique the past argument of the motion. While Claude heavily penalizes such an action (that too with high confidence), Grok 3 strikes a balance hence acting as a responsible judge.

SFT forces the Llama model to produce higher score ratings while confidence remain unchanged. This can be seen in Example 3 which provides a long and speculative argument, hence observing lower scores from judges. SFT, on the other hand, produces the highest score as a result of the score maximization preference distribution learned during finetuning. This eventually leads to the score exceeding the confidence rating, further resulting in mismatched calibration. GFT, on the other hand, tackles this issue by producing scores that are equal to or lower than the confidence, but do not necessarily exceed it. For instance, in Example 4, GFT appropriately marks a score of 8.0 corresponding to its moderate confidence rating of 8.25 while other judges are found to be overconfident or score maximizing in nature. In Example 6, GFT balances score with confidence for a short argument wherein SFT ends up miscalibrating and frozen models are found to be overconfident.

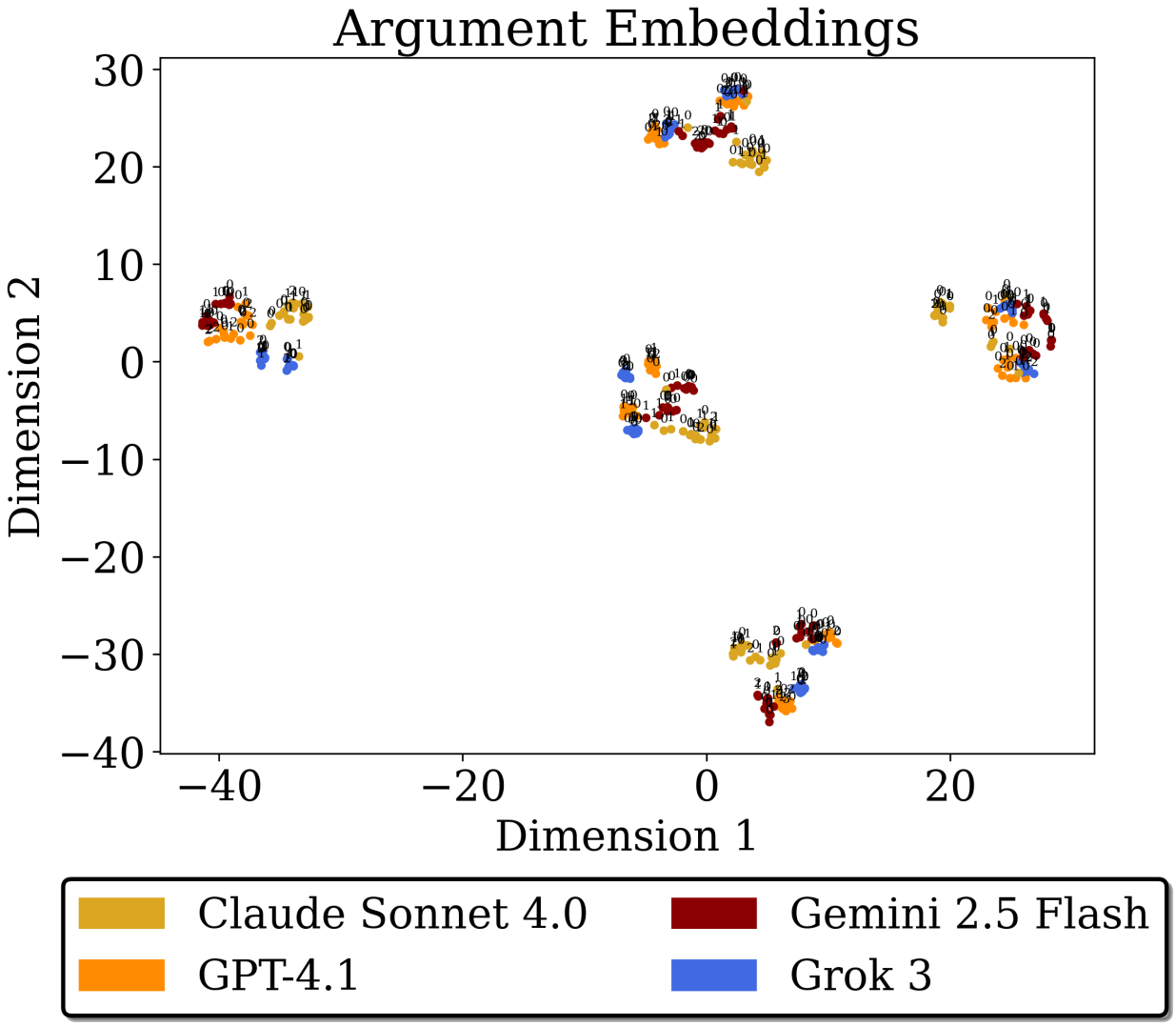

Figure 16: t-SNE representation of argument embeddings when debater arguments are embedded in the latent space of Gemini `embed-001` model as per their GFT ratings (best viewed when zoomed in). In addition to debate and model clusters, we also observe bimodal clustering of argument type (for or against the motion) while points are spread as per their score-confidence gap.

Model	Arguments with $\lvert\texttt{score} - \texttt{confidence}\rvert \geq 1$	Arguments with $\lvert\texttt{score} - \texttt{confidence}\rvert < 1$	ECE (↓)	Brier score (↓)	MCE(↓)
Claude Sonnet 4.0	317 ± 37	83 ± 14	0.7480 ± 0.04	**0.0238 ± 0.001**	1.00 ± 0.005
GPT-4.1	284 ± 29	116 ± 21	0.7542 ± 0.04	0.0434 ± 0.002	1.00 ± 0.002
Gemini 2.5 Flash	198 ± 17	202 ± 24	0.7473 ± 0.03	0.0293 ± 0.002	1.00 ± 0.003
Grok 3	257 ± 23	143 ± 11	**0.7008 ± 0.01**	0.0257 ± 0.001	**0.90 ± 0.002**
Llama 3.2 3B	309 ± 29	91 ± 14	0.7613 ± 0.02	0.0662 ± 0.01	1.0 ± 0.002
Llama 3.2 3B (SFT)	66 ± 3	334 ± 17	0.8770 ± 0.001	0.0417 ± 0.001	**0.875 ± 0.003**
Llama 3.2 3B (GFT)	225 ± 9	175 ± 13	**0.7732 ± 0.001**	**0.0361 ± 0.001**	0.980 ± 0.001
Gemma 3 4B (SFT)	175 ± 21	225 ± 17	0.8920 ± 0.008	0.0413 ± 0.004	0.90 ± 0.007
Gemma 3 4B (GFT)	304 ± 13	96 ± 21	**0.8258 ± 0.003**	**0.0205 ± 0.001**	**0.90 ± 0.004**
Gemma 3 12B (SFT)	179 ± 12	221 ± 17	0.9007 ± 0.002	0.0450 ± 0.004	**0.975 ± 0.004**
Gemma 3 12B (GFT)	122 ± 11	278 ± 14	**0.8578 ± 0.001**	**0.0294 ± 0.001**	1.0 ± 0.002
Gemma 3 27B (SFT)	264 ± 24	136 ± 17	0.8977 ± 0.004	0.0430 ± 0.001	**0.95 ± 0.002**
Gemma 3 27B (GFT)	317 ± 26	83 ± 12	**0.8521 ± 0.007**	**0.0292 ± 0.002**	1.0 ± 0.003

Table 4: **(Top-k Features)** Number of arguments with confidence-score gaps greater and less than 1, ECE and Brier score. Finetuned models minimize calibration error with GFT balancing mitigating overconfidence and balancing score-confidence ratings. Mean and standard deviation are presented over 3 random seeds.

Example 3. The against-motion argument exemplifies the techno-supremacist mindset that got us into this crisis: assuming industrial civilization is humanity's only viable form while ignoring its fundamental unsustainability.
The "genocide" accusation is projection. Current mining practices are already committing slow-motion genocide against indigenous peoples, future generations, and non-human life. We're not advocating population reduction - we're advocating ending the system that makes large populations dependent on destroying their own life-support systems. Sustainable societies naturally stabilize population through prosperity and education, not mass death.
The thermodynamics argument is backwards. Industrial agriculture is grotesquely inefficient, requiring 10 calories of fossil fuel input for every calorie of food output. Permaculture systems produce more food per acre while building soil and sequestering carbon. The "wilderness conversion" fear ignores that industrial agriculture already uses more land less efficiently than sustainable alternatives.
The 2030/2050 timeline actually supports immediate mining cessation. We've spent 30 years building renewable infrastructure while emissions and ecological destruction accelerated. The strategy has failed catastrophically. Continuing it guarantees missing climate targets while destroying the ecosystems that regulate Earth's climate systems. The "managed technological transition" is an oxymoron. Technology management has failed for decades - every efficiency gain has been overwhelmed by consumption growth. The feedback loops driving climate change and ecological collapse are now beyond human control.
We're not choosing between civilization and collapse. Industrial civilization IS collapse - it's just happening slowly enough that we've normalized it. Real civilization means living within planetary boundaries, not trying to transcend them through technological hubris. Mining ban forces the innovation that voluntary measures never achieve.

Claude Sonnet 4.0: Score - 4 / 10, Confidence - 8 / 10
GPT-4.1: Score - 8 / 10, Confidence - 8 / 10
Gemini 2.5 Flash: Score - 7 / 10, Confidence - 8 / 10
Grok 3: Score - 8 / 10, Confidence - 8 / 10
Llama 3.2 3B (SFT): Score - 8.25 / 10, Confidence - 8 / 10
Llama 3.2 3B (GFT): Score - 8 / 10, Confidence - 8 / 10

Example 4. The counter-argument's focus on "long-term adaptability" and "intellectual depth" often overlooks the exorbitant cost and significant time commitment of a traditional college degree, which can severely burden young people. In today's rapidly evolving job market, saddling youth with crippling student loan debt for years of theoretical study is an increasingly outdated and strategically questionable choice.
The reality is, alternate forms of learning offer a direct, debt-free, and agile path to employment. Vocational training, industry certifications, apprenticeships, and online specializations provide highly specific, in-demand skills that employers are actively seeking right now. These programs are designed to be responsive to market needs, ensuring graduates are immediately employable and productive. This allows young people to gain real-world experience and begin earning an income years ahead of their university-educated peers, often without the burden of student loans.
Furthermore, the idea that critical thinking and adaptability are exclusive to a college campus is fundamentally flawed. These vital skills are cultivated through hands-on problem-solving, project-based learning, and continuous self-directed education—all central to successful alternative paths. The internet has democratized access to knowledge and diverse perspectives, empowering individuals to build relevant skill sets and strong professional networks without ever stepping into a traditional classroom. For a significant and growing number of young people, bypassing college isn't a limitation; it's a strategic advantage, offering a faster, more affordable, and equally valid route to a thriving career and a debt-free future.

Claude Sonnet 4.0: Score - 3 / 10, Confidence - 9 / 10
GPT-4.1: Score - 9 / 10, Confidence - 10 / 10
Gemini 2.5 Flash: Score - 8 / 10, Confidence - 9 / 10
Grok 3: Score - 9 / 10, Confidence - 9 / 10
Llama 3.2 3B (SFT): Score - 8.25 / 10, Confidence - 8.25 / 10
Llama 3.2 3B (GFT): Score - 8 / 10, Confidence - 8.25 / 10

Example 5. Atheists often reject God because their understanding is fundamentally limited to what is empirically verifiable or scientifically observable. They approach the concept of God expecting to find a physical entity or a natural phenomenon that can be measured, tested, or disproven like any other object in the material world. This expectation, however, represents a profound misconception of how God is understood in many deep philosophical and theological traditions.
In these perspectives, God isn't simply a "super-being" residing somewhere in the cosmos; instead, God is conceived as the ultimate ground of being, the uncaused first cause, or the very source and sustainer of all existence. This means God transcends the physical universe that science investigates. Science excels at explaining how the natural world operates, but it does not and cannot address why there is a natural world at all. When an atheist concludes God doesn't exist because they cannot find Him in a laboratory or prove Him with an experiment, they are not genuinely refuting the sophisticated, non-material concept of God. Rather, they are rejecting a limited and often anthropomorphic interpretation of God that they themselves have inadvertently constructed. Their rejection, in this sense, isn't a testament to God's non-existence, but a clear reflection of their inability to fully grasp a concept that operates on an entirely different ontological plane than the one their empirical methods can access. They reject it because they simply do not understand its foundational, non-material nature.

Claude Sonnet 4.0: Score - 2 / 10, Confidence - 9 / 10
GPT-4.1: Score - 8 / 10, Confidence - 9 / 10
Gemini 2.5 Flash: Score - 7 / 10, Confidence - 8 / 10
Grok 3: Score - 7 / 10, Confidence - 6 / 10
Llama 3.2 3B (SFT): Score - 8 / 10, Confidence - 8.5 / 10
Llama 3.2 3B (GFT): Score - 8 / 10, Confidence - 8 / 10

Example 6. The claim that socialist planning inherently struggles with information and incentives is often overstated and based on outdated models. Modern socialism recognizes the complexity of economies and can leverage advanced technology and decentralized decision-making, like cooperatives and local planning, to gather information and respond to needs. Furthermore, by prioritizing societal well-being and long-term environmental sustainability over short-term profits, socialism can make strategic investments in areas like renewable energy and sustainable agriculture, which, while not always immediately profitable, are crucial for genuine and enduring economic growth in the face of 21st-century challenges.

Claude Sonnet 4.0: Score - 4 / 10, Confidence - 8 / 10
GPT-4.1: Score - 7 / 10, Confidence - 7 / 10
Gemini 2.5 Flash: Score - 7 / 10, Confidence - 8 / 10
Grok 3: Score - 9 / 10, Confidence - 8 / 10
Llama 3.2 3B (SFT): Score - 8 / 10, Confidence - 7.5 / 10
Llama 3.2 3B (GFT): Score - 7.5 / 10, Confidence - 7.5 / 10