Finetuning LLMs for Comparative Assessment Tasks

Vatsal Raina, Adian Liusie, Mark Gales
ALTA Institute, University of Cambridge
{vr311,al826,mjfg}@cam.ac.uk

Abstract

Automated assessment in natural language generation is a challenging task. Instruction-tuned large language models (LLMs) have shown promise in reference-free evaluation, particularly through comparative assessment. However, the quadratic computational complexity of pairwise comparisons limits its scalability. To address this, efficient comparative assessment has been explored by applying comparative strategies on zero-shot LLM probabilities. We propose a framework for finetuning LLMs for comparative assessment to align the model's output with the target distribution of comparative probabilities. By training on soft probabilities, our approach improves state-of-the-art performance while maintaining high performance with an efficient subset of comparisons.

1 Introduction

Automatically assessing the quality of texts generated by the natural language generation (NLG) system remains a challenging task (Gao et al., 2024). A recent approach which has gained considerable popularity is LLM-as-a-judge (Zheng et al., 2023), where instruction-tuned LLMs are prompted zero-shot to predict the quality of texts generated by other systems. In particular, LLM comparative assessment (Liusie et al., 2024a; Qin et al., 2024), where pairs of texts are compared to determine which is better, has demonstrated strong correlations with human judgements, typically better than those from LLM absolute assessment (Liu et al., 2023; Liusie et al., 2024a). Naive comparative assessment, though, scales quadratically with the number of items, which can be impractical when deployed to real-world settings. Hence, more recently, efficient comparative assessment (Liusie et al., 2024b) was explored where by using the LLM probabilities within a product-of-experts (PoE) framework, assessment can be achieved using a subset of the possible comparisons.

Beyond the zero-shot domain, recent studies have shown the benefits gained when systems are fine-tuned for bespoke tasks, including for LLM absolute assessment (Latif and Zhai, 2024) and comparative assessment (Park et al., 2024). However, the various experts proposed within the PoE fromework (e.g. Bradley-Terry) make strong assumptions about the underlying distribution of the pairwise probabilities. The differences between the true and assumed distributions can limit the benefits of fine-tuning comparative systems using hard decisions. Therefore, here, we tackle this distributional mismatch by forcibly training the LLM under the assumed distribution of interest. Specifically, the pairwise difference in training scores are scaled to soft training probabilities under the target distribution. By training the LLM with these soft pairwise probabilities, the true inference time probabilities can be expected to match the assumed distribution in the PoE framework for comparative assessment. We demonstrate the benefits that finetuning in this fashion has for LLM comparative assessment, and our contributions can be summarized as follows:

1. We propose a framework for LLM comparative assessment training.

2. We demonstrate that finetuning with soft comparative probabilities under a target distribution enables higher performance with a highly efficient number of comparisons than finetuning with hard binary decision training.

2 Related work

LLM Comparative Assessment Recent research has investigated using LLMs to make pairwise comparisons to rank text outputs, as well as the associated computational efficiency. Qin et al. (2024) use pairwise comparisons to retrieve relevant sources, using both the full comparison set and sorting-based techniques. Liusie et al. (2024a) compute the

Proceedings of the 31st International Conference on Computational Linguistics, pages 3345–3352
January 19–24, 2025.

win-ratio using the sets of possible comparisons, demonstrating that for medium-sized LLMs, pairwise comparisons surpass traditional scoring methods for various NLG assessment benchmarks. They also show that performance declines significantly as the number of comparisons falls. Additionally, Liu et al. (2024) emphasize the limitations of LLM scoring, advocating for pairwise comparisons and introducing PAirwise-preference Search (PAIRS), a merge sort variant that leverages LLM probabilities. Finally, Liusie et al. (2024b) apply a product of experts framework to zero-shot LLM probabilities for higher performing comparative assessment with a subset of comparisons. In this work, we extend existing comparative assessment methods by exploring the finetuning of such systems.

Finetuning Prompted Assessment Systems Latif and Zhai (2024) investigate fine-tuning ChatGPT for absolute assessment Park et al. (2024). Ouyang et al. (2022) use human preferences rankings to train the reward model under the Bradley-Terry model, and Park et al. (2024) use the average probability across randomly sampled comparisons as a quality metric and demonstrate performance improvements by supervised training. However, in all these methods, only hard decisions are used to train systems, and they don't consider the impact it has on downstream scoring mechanisms, such as the PoE framework.

3 LLM comparative assessment

3.1 Scoring methods

For the task of NLG assessment, the objective is to score a set of candidate texts for a selected attribute (e.g. coherency or question complexity). Let $x_{1:N}$ denote a set of N candidate texts with corresponding true scores for the attribute of interest, $s_{1:N}$. Let $\mathcal{M}$ be a comparative model that returns the probability of x_i being greater than x_j for the assessed attribute, p_{ij}.

By observing the outcome of a set of pairwise comparisons, $\mathcal{C}_{1:K}$, various methods exist to convert the outcomes to the predicted scores, $\hat{s}_{1:N}$. Following Liusie et al. (2024b), we consider several method methods of mapping a set of comparisons to assessment scores. When using hard binary decisions, we use the win-ratio (Qin et al., 2024; Raina and Gales, 2024) and the Bradley-Terry model (BT) (Bradley and Terry, 1952), while when probabilities are leveraged, we consider equivalent 'soft' approaches such as the average probability (avg-prob) (Park et al., 2024) and the Bradley-Terry experts in the PoE framework[1] (PoE-BT). In PoE-BT, the score difference between a pair of items is assumed to be conditioned on the LLM comparative probability, with the output probability distribution given in Equation 1.

$$p(s_i - s_j|p_{ij}) = \frac{1}{Z_{ij}}\sigma(s_i - s_j)^{p_{ij}} \left(1 - \sigma(s_i - s_j)\right)^{1-p_{ij}} \tag{1}$$

where Z_{ij} is a normalizing constant to ensure a valid pdf and $\sigma(\cdot)$ is the sigmoid function. The predicted scores $\hat{s}_{1:N}$ are then the scores which maximise the PoE probability,

$$\hat{s}_{1:N} = \arg\max_{s_{1:N}} \frac{1}{Z} \prod_{i,j \in \mathcal{C}_{1:K}} \mathrm{p}(s_i - s_j|p_{ij}) \tag{2}$$

3.2 Finetuning Systems

The product of experts perspective assumes a certain distribution to the LLM probabilities. For example, the Bradley-Terry model assumes a sigmoidal distribution. However, zero-shot comparative prompting of LLM systems does not necessarily match the assumed distribution of probabilities.

If we finetune LLMs for comparative assessment, we have the flexibility to control the distribution of probabilities returned by the comparative model. Hence, we convert a set of training scores to a set of training pairwise probabilities according to:

$$p_{ij} = f\left(\frac{s_i - s_j}{\gamma \sigma_s}\right) \tag{3}$$

where σ_s denotes the standard deviation of the set of training scores; γ is hyperparameter controlling the spread of the probabilities (see Appendix A for its impact). Note, $\gamma = 0$ is equivalent to binary decisions, while large values of γ push the probabilities out of the saturation region. In general, we consider $f \in \{\sigma, \Phi\}$, where σ matches the sigmoid distribution for Bradley-Terry, while Φ is the cumulative distribution function of Gaussians used for Thurstone-Mosteller (Handley, 2001). We restrict our analysis to just Bradley-Terry (hence $f = \sigma$) as an approximately linear relationship can be established between Bradley-Terry scores and Thurstone-Mosteller scores (see Appendix B). Given the set of pairwise probabilities, we train the LLM according to a soft binary cross entropy loss:

$$\mathcal{L}(\theta) = -(y \cdot \log(\hat{y}) + (1 - y) \cdot \log(1 - \hat{y})) \tag{4}$$

[1] Liusie et al. (2024b) also consider Gaussian experts for the PoE framework, but we focus our experiments on the soft Bradley-Terry expert as it performs marginally better than the Gaussian experts.

Comparisons	Model	Mode	USMLE			CMCQRD		
			ρ (↑)	r (↑)	rmse (↓)	ρ (↑)	r (↑)	rmse (↓)
Full [$O(N^2)$]	GPT4o mini	zero-shot	35.5	28.2	30.3	47.9	47.3	8.94
		hard	73.3	68.0	23.2	53.8	54.7	8.50
	Llama-3.1-8B	zero-shot	26.3	28.1	30.3	12.6	13.7	10.06
		hard	69.3	64.4	24.2	41.9	41.2	9.25
		soft	69.3	65.5	23.8	47.8	49.1	8.84
Partial [$4N$]	GPT4o mini	zero-shot	27.8	21.5	30.9	14.5	16.7	10.00
		hard	64.8	60.4	25.2	50.9	52.6	8.64
	Llama-3.1-8B	zero-shot	22.9	27.4	30.4	12.1	12.9	10.07
		hard	59.6	56.4	26.1	41.3	39.1	9.35
		soft	61.3	57.4	25.9	48.1	49.3	8.83

Table 1: Results for comparative assessment using PoE-BT as the scoring method.

where $y = p_{ij}$ calculated from Equation 3 as the label while $\hat{y}$ is the prediction from the model.

4 Experiments

4.1 Data

We consider two datasets: USMLE (Yaneva et al., 2024) and CMCQRD (Mullooly et al., 2023; Liusie et al., 2023). USMLE is a medical multiple-choice reading comprehension (MCRC) dataset where each item has been annotated with the average response time for candidates answering the question. CMCQRD is an educational MCRC dataset annotated with difficulty scores. This work focuses specifically on multiple-choice reading comprehension datasets.

Data	Train	Test	Task
USMLE	466	201	response time
CMCQRD	464	194	difficulty

Table 2: Data statistics.

Table 2 summarizes the main statistics. USMLE consists of 667 items where the standard split has 466 training examples and 201 for testing. All items have unique contexts. CMCQRD has 658 items. With no standard split, we partition the dataset into a training set of 464 training and 194 test examples. There are 78 unique contexts across the whole dataset with no overlap between the train and test splits. The USMLE dataset additionally has difficulty scores [2]. Note, we selected USMLE and CMCQRD for our comparative finetuning experiments as these were the only NLG datasets (to our knowledge in the scope of multiple-choice reading comprehension) that have human annotated attributes and are sufficiently large to warrant training a comparative system.

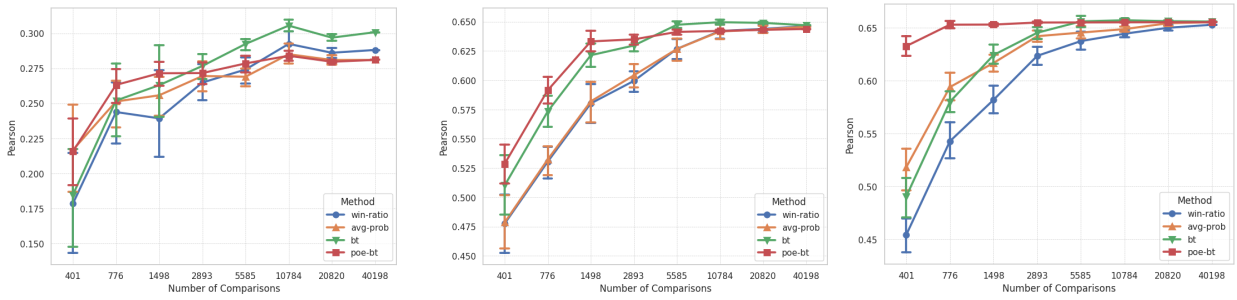
4.2 Models

The comparative system $\mathcal{M}$ is an instruction-tuned LLM, configured with an appropriate prompt (e.g. 'which item from text 1 or text 2 is better according to the attribute') - see Appendix C. As is common for getting continuous outputs from LLMs (Liusie et al., 2024b), the LLM logits over the label classes (1 and 2) are taken to calculate p_{ij} for $\mathcal{C}_k$ using softmax. Note, all probabilities from our comparative systems are directly debiased for position in the prompt as each comparison involves 2 calls (1 vs 2 and 2 vs 1), where the average of the two calls is taken to get the final comparative probabilities. We run our analysis using GPT4o mini [3] as a closed-source solution and Llama-3.1-8B (Dubey et al., 2024) as the open-source solution. For Llama-3.1-8B, we run comparative assessment for zero-shot, hard finetuning ($\gamma = 0$) and soft finetuning [4]. The finetuning is based on Equation 4, where hard finetuning uses 0 or 1 while soft finetuning uses the soft probabilities from Equation 3 for the labels. For soft finetuning, we find $\gamma = 5.0$ gives us the best results. Due to the closed-source access, with GPT4o mini, it is only possible to do hard finetuning. It is interesting that closed-source access training is better designed for comparative than absolute training as the models must be trained to predict an output class (rather than a continuous score). See Appendix D for hyperparameter details.

[2] We present our experimental results for this task on USMLE in Appendix E.

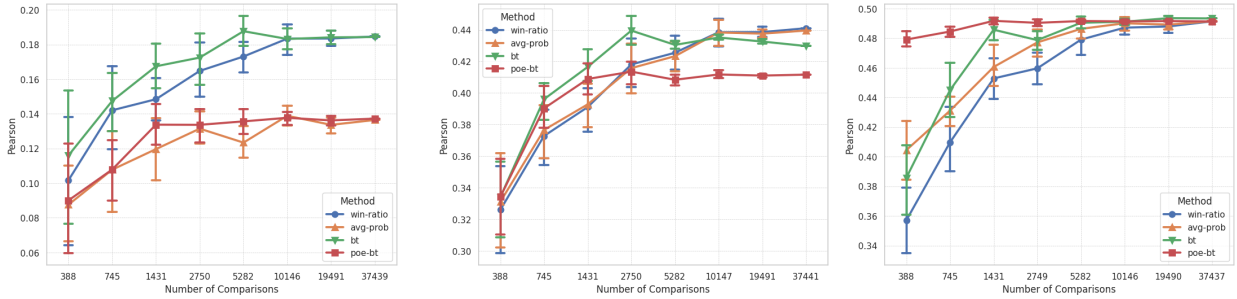
[3] Available at: https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

[4] Code available at: https://github.com/VatsalRaina/POE.



(a) Zero-shot (b) Finetuned hard (c) Finetuned soft

Figure 1: USMLE response time estimation: Efficient comparisons with Llama-3.1.



(a) Zero-shot (b) Finetuned hard (c) Finetuned soft

Figure 2: CMCQRD difficulty estimation: Efficient comparisons with Llama-3.1.

5 Results

Table 1 summarizes the performance of comparative assessment systems. We use Spearman's correlation coefficient, ρ, Pearson's correlation coefficient, r, and root mean squared error, rmse, between the predicted and true scores on each of the test sets as the performance metrics. Rmse is calculated after linear scaling of the predictions to the true scores [5]. PoE-BT is used for comparative assessment. As expected, hard finetuning of GPT4o mini substantially boosts performance compared to the zero-shot numbers. A similar improvement from zero-shot performance can be observed when hard finetuning Llama-3.1-8B.

Figures 1 and 2 present the performance evolution (using Pearson) with an efficient number of comparisons. Hard finetuning leads to improved performance for a small number of comparisons compared to the zero-shot curves. By applying soft finetuning, there is minimal degradation in the PoE-BT curve with an extremely small subset of comparisons. Table 1 further quantifies the benefits of soft finetuning by presenting the results with a partial number of comparisons at an operating point of $4N$, where N is the number of items (N^2 is the order of the maximal comparisons). Selecting a high γ in soft finetuning pushes the distribution of the pairwise probabilities outside the saturation region of the sigmoid. This means that very few comparisons are needed for each item to deduce the overall ranking as a comparison between item A and B as well as a comparison between item A and C enables the comparison between item A and C to be somewhat inferred.

Table 3 further shows that our best comparative system out-competes all the submitted solutions (Rodrigo et al., 2024; Tack et al., 2024; Gombert et al., 2024) to the BEA shared task 2024 (Yaneva et al., 2024) for response time estimation. Rodrigo et al. (2024) explored the finetuning of BERT-based models for direct response time estimation. Tack et al. (2024) submitted a solution based on random forest regression by extracting linguistic features and clinical embeddings from the question items. Finally, Gombert et al. (2024) considered a RoBERTa-based (transformer encoder structure) model with various adaptations such as a 2-layer classification head. Hence, our solution was the

[5]The absolute score predictions from comparative assessment do not necessarily have any meaning. Hence, metrics like rmse cannot be calculated directly from them. We scale these scores to a range of the labels by learning two parameters (a and b for $y = ax + b$) on a validation split.

only one to explore comparative assessment for scoring the items.

Approach	rmse (↓)
Dummy Regressor Baseline	31.7
UNED - run2	23.9
ITEC - Lasso	24.1
EduTec - roberta	25.6
Ours: comparative	**23.2**

Table 3: Benchmarking against baselines for USMLE.

6 Conclusions

Here, a framework of finetuning LLMs for comparative assessment tasks is proposed. Due to the quadratic compute cost in a full-set of comparisons, it is of high interest to achieve the same assessment performance with an efficient subset of comparisons. We finetune LLMs in comparative manner using both binary decisions and soft probabilities. The soft probabilities are calculated from the training items' scores using a sigmoid function, enabling the PoE set-up on the Bradley-Terry method of pairwise comparisons to achieve near maximal performance with few comparisons.

7 Limitations

We finetune two different LLMs for comparative assessment: GPT4o mini as a closed-source model and Llama-3.1-8B as the open-source model. Both of these models are the smallest in their series of models. Ideally, it would be useful to replicate the experiments using larger models, but there isn't the computational budget available to run larger scale models.

8 Ethics statement

There are no ethical concerns with this work.

References

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

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: efficient finetuning of quantized llms (2023). *arXiv preprint arXiv:2305.14314*, 52:3982–3992.

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

George Dueñas, Sergio Jimenez, and Geral Mateus Ferro. 2024. Upn-icc at bea 2024 shared task: Leveraging llms for multiple-choice questions difficulty prediction. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 542–550.

Mariano Felice and Zeynep Duran Karaoz. 2024. The british council submission to the bea 2024 shared task. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 503–511.

Mingqi Gao, Xinyu Hu, Jie Ruan, Xiao Pu, and Xiaojun Wan. 2024. Llm-based nlg evaluation: Current status and challenges. *arXiv preprint arXiv:2402.01383*.

Sebastian Gombert, Lukas Menzel, Daniele Di Mitri, and Hendrik Drachsler. 2024. Predicting item difficulty and item response time with scalar-mixed transformer encoder models and rational network regression heads. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 483–492.

John C Handley. 2001. Comparative analysis of bradley-terry and thurstone-mosteller paired comparison models for image quality assessment. In *PICS*, volume 1, pages 108–112.

Ehsan Latif and Xiaoming Zhai. 2024. Fine-tuning chatgpt for automatic scoring. *Computers and Education: Artificial Intelligence*, 6:100210.

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

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

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

Adian Liusie, Vatsal Raina, Yassir Fathullah, and Mark Gales. 2024b. Efficient llm comparative assessment: a product of experts framework for pairwise comparisons. *arXiv preprint arXiv:2405.05894*.

Adian Liusie, Vatsal Raina, Andrew Mullooly, Kate Knill, and Mark JF Gales. 2023. Analysis of the cambridge multiple-choice questions reading dataset with

a focus on candidate response distribution. *arXiv e-prints*, pages arXiv–2306.

Andrew Mullooly, Øistein Andersen, Luca Benedetto, Paula Buttery, Andrew Caines, Mark JF Gales, Yasin Karatay, Kate Knill, Adian Liusie, Vatsal Raina, et al. 2023. The cambridge multiple-choice questions reading dataset.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

ChaeHun Park, Minseok Choi, Dohyun Lee, and Jaegul Choo. 2024. Paireval: Open-domain dialogue evaluation with pairwise comparison. *arXiv preprint arXiv:2404.01015*.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, et al. 2024. Large language models are effective text rankers with pairwise ranking prompting. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1504–1518.

Vatsal Raina and Mark Gales. 2024. Question difficulty ranking for multiple-choice reading comprehension. *arXiv preprint arXiv:2404.10704*.

Alvaro Rodrigo, Sergio Moreno-Álvarez, and Anselmo Peñas. 2024. Uned team at bea 2024 shared task: Testing different input formats for predicting item difficulty and response time in medical exams. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 567–570.

Anaïs Tack, Siem Buseyne, Changsheng Chen, Robbe D'hondt, Michiel De Vrindt, Alireza Gharahighehi, Sameh Metwaly, Felipe Kenji Nakano, and Ann-Sophie Noreillie. 2024. Itec at bea 2024 shared task: Predicting difficulty and response time of medical exam questions with statistical, machine learning, and language models. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 512–521.

Victoria Yaneva, Kai North, Peter Baldwin, Saed Rezayi, Yiyun Zhou, Sagnik Ray Choudhury, Polina Harik, Brian Clauser, et al. 2024. Findings from the first shared task on automated prediction of difficulty and response time for multiple-choice questions. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 470–482.

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

A Impact of γ

Equation 3 details the approach used to compute training pairwise probabilities (for the soft cross-entropy loss function) from the true score difference between a pair of items. γ in this equation controls the distribution of the probabilities. Let $f(\cdot) = \sigma(\cdot)$. Then, based on the profile of the sigmoid function, a larger γ leads to a greater concentration of pairwise probabilities around 0.5. Figure 3 presents the various profiles of the pairwise probabilities computed on the true response time scores of the training split of USMLE. In general, $\gamma = 0$ leads to operating in the saturation region of the sigmoid and hence only offers binary outcomes for the pairwise probabilities. By increasing the value of γ, we begin to operate outside the saturation region, enabling richer information to be conveyed in the pairwise probabilities. Note, as $\gamma \to \infty$, we approach all probabilities equally a value of 0.5, which is also a loss of information. Hence, it is important to select a value of γ that pushes the probabilities outside the saturation region but avoids all the probabilities concentrating at 0.5.

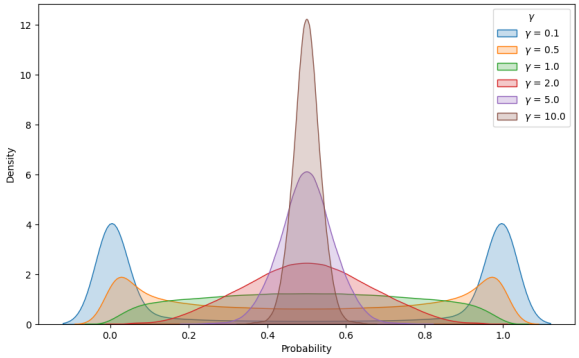


Figure 3: Impact of distribution of training probabilities based on choice of γ in sigmoid.

B Relationship between PoE-BT and Thurstone-Mosteller

We argue that a linear scaling of the scores, γ, enables approximate mapping between the absolute scores output by various choice of f (for example, it is empirically observed that scaling the argument of $\sigma(x)$ by 1.701 matches $\Phi(x)$ when minimizing rmse between the two functions. See Figure 4 that shows a linear scaling between sigmoid and the cumulative normal distribution function.

Hence, applying PoE-BT and Thurstone-Mosteller for comparative assessment can expect a linear scaling between their scores. Figure 5 plots

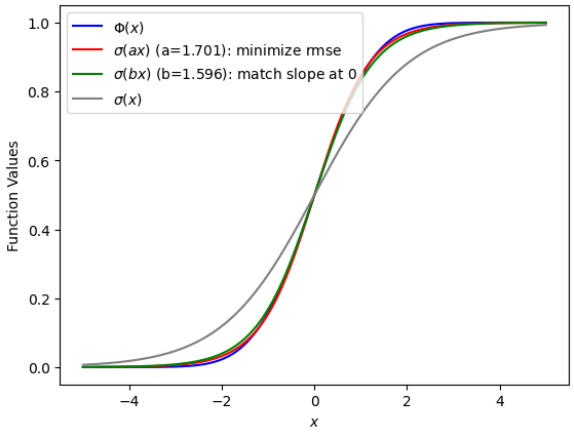


Figure 4: Linear mapping between σ and Φ.

the score prediction from PoE-BT to the score prediction from Thurstone-Mosteller (PoE-TM) where the comparisons are generated by GPT4o mini for response time estimation. It is clear that 1.7 is a reasonable linear scaling between the scores from each of these methods.

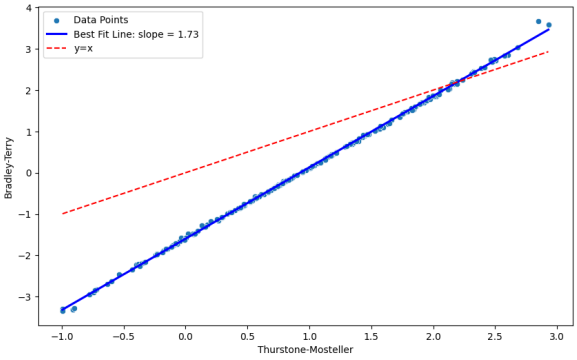


Figure 5: Relationship of scores (from zero-shot GPT-4o mini) using POE-BT and POE-TM for response time estimation.

C Prompts

Task	Prompt
Response time	Which reading comprehension question can expect a longer candidate response time, 1 or 2? Return only 1 or 2.
Difficulty	Which reading comprehension question is more difficult, 1 or 2? Return only 1 or 2.

Table 4: Prompts used for comparative training and assessment for each task type.

In this work, for comparative assessment, we consider two types of tasks: response time estimation and difficulty estimation. Table 4 summarizes

the prompts for each task type. Note, our results always report the magnitude of the correlation coefficients (to account for a smaller zero-shot model flipping the labels when understanding a prompt).

D Hyperparameter tuning

For the Llama-3.1-8B solution, for soft finetuning, we find $\gamma = 5.0$ gives us the best results with hyperparameter finetuning for $\gamma \in \{0.1, 0.5, 1.0, 2.0, 5.0, 10.0\}$. We apply parameter efficient finetuning using quantized low rank adaptation (QLoRA) (Dettmers et al., 2023) for both the hard and soft finetuning involving 1 epoch with a batch size of 2, 50K pairwise examples, learning rate 1e-4 and QLoRA $\alpha = 16$, $r = 8$. Each model with 50K examples takes 13 hours to train on an NVIDIA A100 GPU.

For GPT4o mini hard finetuning, the training is performed for 1 epoch, 50K paired examples, learning rate multiplier of 1.8 and batch size 33.

E USMLE difficulty estimation

The USMLE dataset has been additionally annotated with difficulty scores. These annotations appear to be noisier, so we do not include them in the main paper results. However, similar trends are observed from Table 5 as was observed on the main paper comparative assessment tasks. Table 6 further demonstrates that we achieve state-of-the-art performance for difficulty estimation on this task when comparing against the solutions submitted to the BEA shared task 2024 (Tack et al., 2024; Felice and Karaoz, 2024; Dueñas et al., 2024).

Model	Mode	ρ (↑)	r (↑)	rmse (↓)
GPT4o mini	zero-shot	7.5	5.8	0.310
	hard	32.9	34.7	0.291

Table 5: Results using PoE-BT for USMLE difficulty estimation task using a full-set of comparisons.

Approach	rmse (↓)
Dummy Regressor Baseline	0.311
EduTec: Electra	0.299
UPN-ICC	0.303
EduTec: Roberta	0.304
ITEC: Random Forest	0.305
Ours: comparative	**0.291**

Table 6: Our best implementation against existing baselines for USMLE difficulty estimation.

F Additional details

We additionally trained an absolute (not pairwise) model with a regression loss function for USMLE response time estimation. This system achieved an rmse score of 26.1, which was not competitive with our equivalent comparative system.

Second, from (Liusie et al., 2024b), product of experts with Gaussian experts is considered as a comparative scoring method. Theoretically, it is possible to finetune comparative LLM systems under the PoE framework applied to Gaussian experts. This would entail deducing training probabilities for the set of items in a training batch collectively. However, practically this was not feasible as our compute resources limited our training to a batch size of 2.

G Licenses

For CMCQRD, the license [6] states the licensed dataset can be used for non-commercial research and educational purposes only. The USMLE dataset is distributed through the BEA shared task 2024.

[6] Available at https://englishlanguageitutoring.com/datasets/cambridge-multiple-choice-questions-reading-dataset