Revisiting Generalization Across Difficulty Levels: It's Not So Easy

Yeganeh Kordi[♣] **Nihal V. Nayak**[♢] **Max Zuo**[♣] **Ilana Nguyen**[♣] **Stephen H. Bach**[♣]
[♣]Brown University [♢]Harvard University
{kordi, stephen_bach}@brown.edu

Abstract

We investigate how well large language models (LLMs) generalize across different task difficulties, a key question for effective data curation and evaluation. Existing research is mixed regarding whether training on easier or harder data leads to better results, and whether those gains come on easier or harder test data. We address this question by conducting a systematic evaluation of LLMs' generalization across models, datasets, and fine-grained groups of example difficulty. We rank examples in six datasets using the outputs of thousands of different LLMs and Item Response Theory (IRT), a well-established difficulty metric in educational testing. Unlike prior work, our difficulty ratings are therefore determined solely by the abilities of many different LLMs, excluding human opinions of difficulty. With a more objective, larger-scale, and finer-grained analysis, we show that cross-difficulty generalization is often limited; training on either easy or hard data cannot achieve consistent improvements across the full range of difficulties. These results show the importance of having a range of difficulties in both training and evaluation data for LLMs, and that taking shortcuts with respect to difficulty is risky.[1]

1 Introduction

Can language models trained on data at easier task difficulties generalize to harder tasks, or vice versa? We term this capability *cross-difficulty generalization*. While several recent studies have investigated this question from various perspectives, the findings remain mixed and often contradictory. Hase et al. (2024) found LLMs demonstrate easy-to-hard generalization: LLMs fine-tuned on easy data and LLMs finetuned on hard data often performed comparably on hard problems within the same domain. However, Sun et al. (2024) observed that while LLMs fine-tuned with easy data may *not necessarily* generalize to harder data, reward models can. On the other hand, several works find that LLMs demonstrate hard-to-easy generalization: training LLMs on hard data generalizes better to easier tasks than training on easy data or even on all data (Yang et al., 2024; Pikus et al., 2025). Yang et al. (2024) found that training on hard data improves a model's consistency in its ability to solve easier problems more than if it were trained on easier data and tasked to solve harder problems. Pikus et al. (2025) showed training on the hardest examples using GRPO (Shao et al., 2024) consistently outperforms using all data. Contrary to these works, Ding et al. (2024) observed that the best generalization occurs when train and test data have the same level of difficulty.

As shown in Table 1, despite ongoing research in this area, the relationship between generalization performance and task difficulty remains an open question. We believe that relying on humans' judgment of difficulty (Hase et al., 2024; Yang et al., 2024; Sun et al., 2024; Ding et al., 2024) can be one source of uncertainty in the literature. It also limits the scalability and resolution of difficulty assessment. In this paper, we systematically evaluate the extent to which LLMs exhibit cross-difficulty generalization, where difficulty is estimated based on the models' observed abilities.

Understanding the prevalence of cross-difficulty generalization is crucial for effective data curation and evaluation of LLMs. If performance on hard tasks can be improved by training only on easy tasks, existing datasets might already be sufficient to extend language models' capabilities beyond what they currently demonstrate. Conversely, limited cross-difficulty generalization would place greater importance on curating the right mix of examples based on difficulty during training.

A key technical challenge in studying generalization is accurately estimating task difficulty. Difficulty is not a universally agreed-upon metric; it is inherently relative, subjective, and language model-dependent. Most existing difficulty estimation methods rely on human-based metrics such as grade level and expert ratings or heuristics such as question length and number of reasoning steps required to answer the task (Hase et al., 2024; Sun et al., 2024). However, LLMs can struggle in tasks that are considered easy for humans (e.g., counting), and human-based metrics may not capture the difficulty of a task for LLMs correctly. Some other works proposed using LLM-based metrics such as perplexity, confidence, success rate, and loss (Swayamdipta et al., 2020; Hase et al., 2024; Pikus et al., 2025). However, these approaches usually rely on a small number

[1]Code and data are available at https://github.com/BatsResearch/cross-difficulty

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 7014–7042

Paper	Core Claim	Difficult for whom?	Training Method
Hase et al. (2024)	Training on easy data performs almost as well on the hard test set as training on hard data.	LLM + Human	SFT, ICL, Linear Probing
Sun et al. (2024)	Training only on easy tasks can outperform training on all tasks.	Human	RL
Yang et al. (2024)	Hard data improves the model's consistency on similar questions more effectively than easy data.	Human	SFT, ICL
Pikus et al. (2025)	Training on the hardest data performs best.	LLM	RL
Ding et al. (2024)	Training provides generalization to similar difficulties, but this generalization reduces as training difficulty increases.	LLM + Human	SFT
Our Analysis	Training on only hard or easy data fails to generalize to other difficulty levels. Human-centric difficulty metrics are not well-suited for studying LLMs.	LLM	SFT

Table 1: **Comparison of related work on cross-difficulty generalization.** Prior work focuses on easy-to-hard or hard-to-easy generalization with difficulty metrics obtained from humans, LLMs, or both. Our work focuses on understanding cross-difficulty generalization, i.e., both easy-to-hard and hard-to-easy generalization, using model-based difficulty metrics, and shows that limited easy-to-hard and hard-to-easy generalization occurs in LLMs. Papers highlighted in green focus on the effectiveness of easy training data. Conversely, red papers discuss the gain from hard data. Papers highlighted blue find that generalization across difficulty levels is limited.

of LLMs, and the confidence or accuracy of an individual model can be miscalibrated, as models often cannot accurately estimate their own capabilities. Therefore, analyzing responses and patterns across many LLMs provides a more robust understanding of difficulty.

In this work, we use Item Response Theory (IRT) (Baker, 1985; Lord et al., 1968) to study cross-difficulty generalization. IRT is a tool widely used to quantify the difficulty of questions and the capability of students in standard educational tests (Kingston and Dorans, 1982; McKinley and Kingston, 1987; Cook et al., 1985). Here, we treat large language models as the students. Some metrics for difficulty, such as question or correct response length, examine solely the features intrinsic to benchmark problems themselves. Other model-centric metrics, such as model loss or average model accuracy, typically only incorporate model performance on one question at a time. IRT jointly optimizes question difficulty and LLM ability to better understand the role each plays in model performance. To estimate the IRT parameters, such as task difficulty and the ability of LLMs, we need to evaluate their performance on the target tasks. Given that running inference with thousands of models and evaluating them would be prohibitively expensive, we collect existing evaluation results for thousands of large language models from the Open LLM Leaderboard (Beeching et al., 2023; Fourrier et al., 2024), a popular hub in the community for benchmark results. After collecting evaluation results, we use IRT to estimate LLM-based difficulty scores for each of six datasets (see §3.1). We find that human-based difficulty metrics diverge substantially from IRT difficulty scores, which are calculated based on LLM abilities, underscoring the value of accurate LLM-based difficulty estimation (§3.4). Finally, we divide each dataset into ten equal-sized bins, ordered by increasing difficulty, to systematically study generalization.

We train LLMs on each difficulty bin individually and evaluate them on all the other difficulty bins to characterize generalization. Our experiments show that cross-difficulty generalization is far from a pattern. First, we find that training language models on one of the difficulty bins can sometimes generalize to easier or harder bins, but the extent of generalization is limited (§5). Neither the easy nor the hard training data tends to achieve consistent generalization across the full range of difficulties. Second, we observe the best generalization between the adjacent bins, and the generalization decreases as we increase the gap between train and test difficulty (§5).

Our main contribution is a comprehensive **analysis of cross-difficulty generalization** in LLMs: We estimated difficulty scores for each example in six widely used benchmarks and provided a systematic study of generalization across various difficulty levels. This analysis reveals two key findings:

- **LLMs exhibit weak cross-difficulty generalization:** We demonstrate that for each dataset, across all the models, training solely on easy or hard bins fails to generalize consistently across difficulty levels.
- **Increasing difficulty gap weakens generalization:** We found out that increasing the difficulty gap between the train and test can degrade the performance of the model, sometimes even below the zero-shot baseline.

2 Related Work

Easy-to-hard Generalization. Easy-to-hard generalization is a fast-growing area that studies model generalization on hard tasks by training only on easy examples (Lee et al., 2025; Sun et al., 2024). Hase et al. (2024) has demonstrated that training on easy data can perform almost as well as training on hard data on hard test sets. However, another line of works (Yang et al., 2024; Chen et al., 2024) suggests that training on hard data can show better generalization and consistency compared to using easy data or full data. For instance, Pikus et al. (2025) shows that training on the hardest examples yields the largest performance gains, while training on easy examples provides minimal gains. While these studies highlight an ongoing debate in the community regarding easy-to-hard generalization, their evaluations are often limited to coarse-grained difficulty splits such as easy or hard splits (Hase et al., 2024; Sun et al., 2024; Pikus et al., 2025), or focus on rather simple tasks such as digit multiplication (Lee et al., 2025). In our work, we address these limitations by studying cross-difficulty generalization with fine-grained difficulty splits created using item-response theory.

Our work is closely related to Ding et al. (2024), but differs in key ways. Ding et al. (2024) primarily benchmarked easy-to-hard generalization using three datasets on the Open LLM Leaderboard and estimated the difficulty of each question using an IRT model with a human-aligned approach. They then used a greedy search algorithm to select a subset of LLMs whose difficulty rankings best matched human annotators' difficulty judgments. In contrast, we analyze cross-difficulty generalization using purely model-based difficulty scores, computed directly from model behavior without human calibration.

Difficulty Estimation. Understanding what makes examples difficult for language models is critical for robust training and evaluation, and prior work reflects a wide range of approaches, from human annotations to computational metrics. Many approaches rely on human-based difficulty, using proxies like educational grade level (Clark et al., 2018; Hendrycks et al., 2021c), expert ratings (Ding et al., 2024), required cognitive skills (Bloom et al., 1956; Hase et al., 2024), or indicators such as question length, answer length (Hase et al., 2024), and number of reasoning steps to assign difficulty (Fu et al., 2023). In Section 3.4, we show that these metrics do not consistently correlate with each other and model-based difficulties.

Other works suggested using model-based approaches (Swayamdipta et al., 2020; Ethayarajh et al., 2022; Varshney et al., 2022; Voita and Titov, 2020; Perez et al., 2021; Ding et al., 2024) to estimate the difficulty of questions for LLMs. Swayamdipta et al. (2020) uses training dynamics to estimate the difficulty of data points and provides insights about data quality. However, it is computationally expensive and requires training the LLM, limiting its scalability. Ethayarajh et al. (2022) propose an information-theoretic approach using V-usable information and pointwise V-information (PVI) to measure question difficulty for a particular model. Despite offering more interpretability, calculating V-usable information is also computationally intensive and reflects how informative the data are for a given model class rather than the intrinsic difficulty of the task. Varshney et al. (2022) analyzes difficulty and helps identify trivial or mislabeled examples; however, it relies on model confidence as a proxy for difficulty, which can misrepresent difficulty when a model's confidence does not estimate its actual ability correctly. Overall, these methods are either computationally expensive, rely on a limited set of models, or fail to fully reflect the task difficulty as perceived by LLMs. Instead, our approach uses responses from thousands of LLMs on existing benchmarks to efficiently estimate the difficulty of questions using IRT.

3 Cross-Difficulty Generalization with Item Response Theory

We describe how we constructed datasets used to study cross-difficulty generalization with item-response theory (IRT). First, we describe the IRT model we use to estimate the difficulty of an example in a dataset (§3.1). Next, we outline a scalable method for collecting evaluation data from a large number of language models, which is then passed to the IRT model to estimate example difficulty (§3.2). Next, we validate the IRT difficulty scores with the zero-shot performance of several large language models (§3.3). Finally, we compare the difficulty estimates from IRT with those from existing difficulty metrics (§3.4).

3.1 Difficulty Estimation using IRT

To obtain fine-grained difficulty estimates for each example in our datasets, we use Item Response Theory. In particular, we use the Rasch (1PL) model (Rasch, 1960) to determine the example's difficulty. We are given a dataset $\mathcal{D} = \{x_1, \ldots, x_N\}$ with N examples and a set $\mathcal{S} = \{s_1, \ldots, s_M\}$ representing M subjects. In our work, we treat the different language models as the subjects. Let $r_{ij} \in \{0, 1\}$ be the observed response for the i-th task from the j-th language model; $r_{ij} = 1$ indicates a correct response. Our goal is to estimate the task difficulty $\{\beta_i\}_{i=1}^{N}$ and the ability of the language model $\{\theta_j\}_{j=1}^{M}$, that best explain the observed data. In the 1PL model, the probability of a language model s_j correctly answering the task x_i is given by:

$$P(r_{ij}|\theta_j, \beta_i) = \frac{1}{1 + e^{-(\theta_j - \beta_j)}}.$$

We estimate posterior distributions over the latent parameters of the 1PL model using stochastic variational inference and take their expected values as point estimates. More details on the 1PL model implementation are provided in the `py-irt` package by Lalor and Rodriguez (2023).

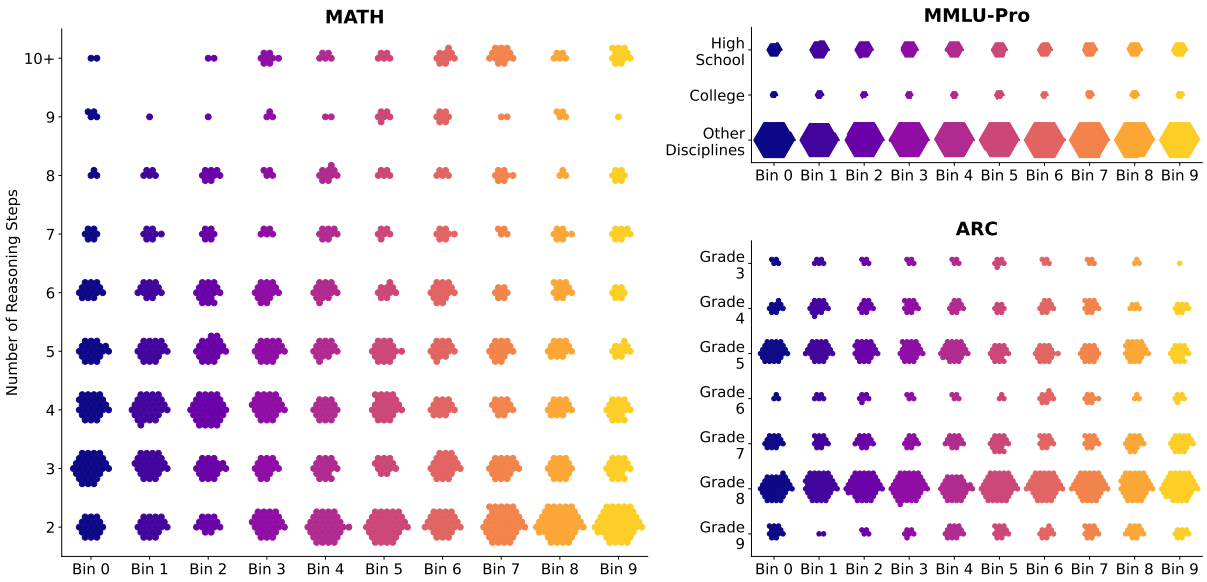

Figure 1: Comparison of human-defined and IRT difficulty estimates for three datasets. Each dot represents one question. **Left**: MATH question distribution by number of reasoning steps (Hendrycks et al., 2021b). **Top right**: MMLU-Pro question distribution by grade level (Wang et al., 2024), with questions lacking assigned grades grouped as "Other Disciplines". **Bottom right**: ARC question distribution by grade level (Clark et al., 2018). All distributions are shown across IRT difficulty score bins.

Dataset	# Examples	# Models
ARC	1,170	5,611
BBH	5,250	4,354
GSM8k	1,319	5,870
MMLU-Pro	12,032	4,359
MATH	1,324	4,437
MuSR	756	4,433

Table 2: **Statistics for all the datasets.** # examples is the number of examples from each dataset on Open LLM Leaderboard. # models is the number of LLMs that has been evaluated on the dataset.

3.2 Data Collection

We estimate the difficulties for examples from six datasets: ARC (Clark et al., 2018), GSM8k (Cobbe et al., 2021), MMLU-Pro (Wang et al., 2024), BBH (Suzgun et al., 2023a), MATH (Hendrycks et al., 2021b), and MuSR (Sprague et al., 2024) (see more details about the datasets in Appendix D).

We collect responses from the Open LLM Leaderboard (Fourrier et al., 2024; Beeching et al., 2023) to compute difficulties using the IRT model. Running inference and evaluation with a large number of models can be prohibitively expensive and time-consuming. For this reason, similar to Ding et al. (2024), we web scrape the evaluations for all models in the leaderboard to scalably collect responses.

Table 2 shows the statistics for the datasets. For each dataset in the leaderboard, we collect responses for all the language models evaluated. Since the evaluations are typically done on the test set, the number of examples in our dataset corresponds to the test set. We also observed that not all the models are evaluated on all tasks. Given these responses, we use the IRT model (§3.1) to estimate the difficulty of examples in each dataset separately. After computing the difficulty scores of each example in the dataset, we sort them and then divide them into ten equally sized bins to study cross-difficulty generalization.

3.3 Difficulty Validation

To further assess the validity of our estimated difficulty scores, we evaluated models from the Qwen3 family (Yang et al., 2025), which were not part of the models used in computing the original IRT-based difficulties. We conducted zero-shot evaluations across all difficulty bins. As shown in Figure 5 in Appendix E, model accuracy consistently decreases with increasing bin difficulty, confirming that the estimated scores generalize reasonably well even to models outside the set used for difficulty estimation.

3.4 Differences between IRT and Human-Based Metrics

We investigate the correlation between difficulty as measured by IRT and human-based difficulty scores across multiple datasets, revealing key differences between these two approaches.

Following Hase et al. (2024), we examined several human-based difficulty metrics. For all datasets, we examine difficulty indicators, such as answer length (in characters) and question length (in words). We also include the number of reasoning steps for math datasets

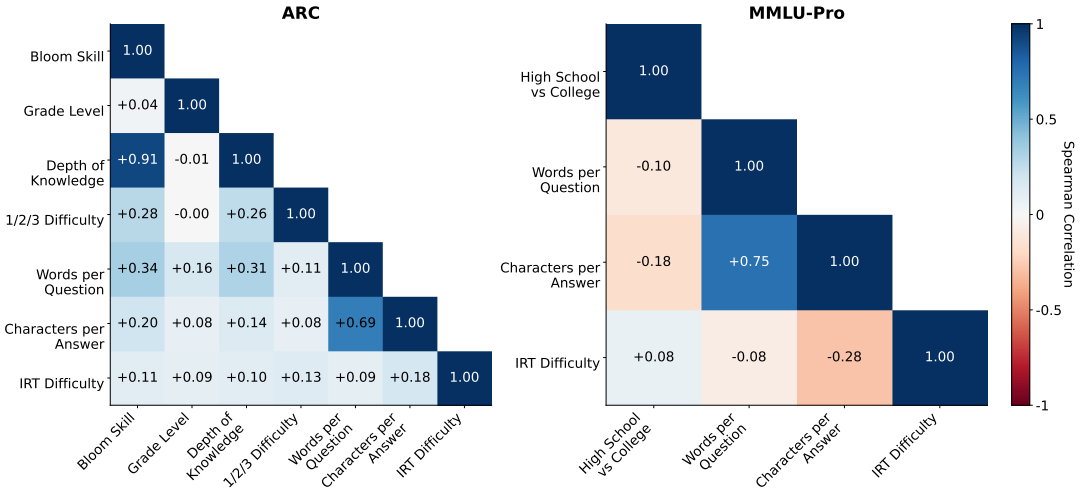

Figure 2: Heatmaps showing Spearman correlations between IRT difficulty scores and human metrics. Colors indicate correlation strength from negative (red) to positive (blue). ARC shows weak positive correlations across all metrics, while MMLU-Pro demonstrates mostly no or negative correlation between IRT difficulty and common human metrics for difficulty.

(GSM8K & MATH) and original dataset annotations for school grade/level information for the ARC, MMLU-Pro, and MuSR datasets. Hase et al. (2024) additionally labels the ARC dataset with cognitive skill requirements based on Bloom's taxonomy (Bloom et al., 1956), Depth of Knowledge (Webb, 2002), and 1/2/3 expert difficulty ratings.

The Bloom metric measures the type of cognitive skill a question requires, from basic recall to higher-order reasoning: (1) Remembering, (2) Understanding, (3) Applying, (4) Analyzing, and (5) Evaluating (Bloom et al., 1956). The Depth of Knowledge metric classifies questions by the cognitive process depth needed to arrive at an answer, ranging from Level 1 (Recall and Reproduction) to Level 4 (Extended Thinking) (Webb, 2002). Depth of Knowledge complements Bloom by emphasizing task complexity and reasoning depth rather than the type of cognitive skill. Finally, the 1/2/3 difficulty rating shows expert annotators' overall judgment of question difficulty (easy, medium, hard).

To visualize these relationships, Figure 1 presents each question as a dot plotted against its IRT-derived difficulty score and corresponding human-based metric. At each human-based difficulty level, the number of questions in each IRT difficulty bin is roughly uniform, suggesting there is very little correlation between these two approaches. (See Appendix G for some examples from each dataset.)

We quantify this relationship by calculating the Spearman rank correlation coefficient between each pair of difficulty metrics. Figure 2 shows the Spearman correlation between the IRT and human-based metrics for ARC and MMLU-Pro (See Figure 19 in Appendix F for other datasets).

We find that most datasets show very little to no positive correlation between IRT difficulty scores and human-based metrics. The fact that most human-based metrics weakly correlate with IRT suggests that what makes a task difficult for language models differs from what humans consider a hard question. We include examples of samples that are difficult for LLMs (high IRT scores, incorrect responses across model families) but easy for humans (low human difficulty scores), as well as the reverse in G. Across all datasets and metrics, the highest positive correlations with IRT are observed for answer length in MATH (0.56) and the number of reasoning steps in GSM8K (0.49). However, these correlations are inconsistent: the number of reasoning steps negatively correlates with IRT (-0.08) in MATH, and answer length negatively correlates with IRT in MMLU-Pro (-0.28), MuSR (-0.13), IFEval (-0.14), and GPQA-Extended (-0.17). Moreover, on average, answer length exhibits little to no positive correlation to most datasets evaluated, contradicting the intuition that longer answers indicate harder problems (Muennighoff et al., 2025).

The relatively stronger correlations between IRT difficulty scores and reasoning steps or question length in some datasets show some support for alignment between those human-based metrics and what is actually difficult in practice for LLMs. However, the weak correlations for most of the human-based metrics suggest that what makes a task difficult for language models differs from human-based ways of measuring difficulty.

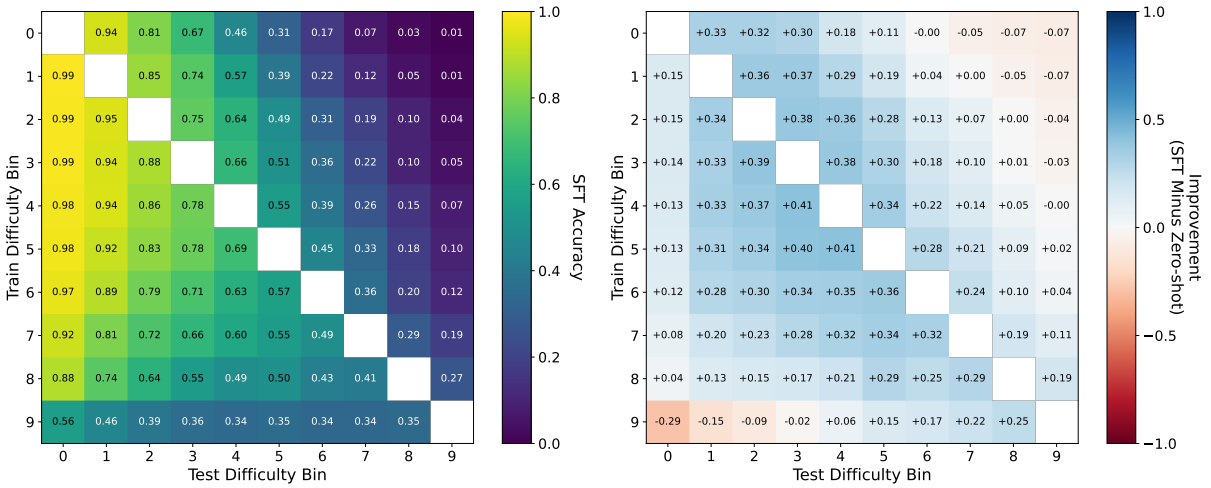

Figure 3: Cross-difficulty generalization heatmaps for Qwen2.5 14B Instruct on MMLU Pro dataset. **Left**: Performance when training on a difficulty bin (y-axis) and testing on another difficulty bin (x-axis). **Right**: Improvement from finetuning on each bin compared to the zero-shot performance of the model on that bin. Diagonal elements are masked as they represent the same train and test data.

4 Experimental Setup

We experiment with seven instruction-tuned LLMs from the Qwen 2.5 (Qwen et al., 2025) and Llama 3 (Dubey et al., 2024) model families. Here we present the experiments with Qwen2.5 14B Instruct and include experiments and evaluations for Qwen2.5 1.5B/3B/7B Instruct, Llama 3.2 1B/3B, and Llama 3.1 8B in Appendix E. Across model families and sizes, the results are broadly consistent with the results presented in the main paper. Implementation details, hyperparameters, and evaluation configurations are described in Appendix A and Appendix B for full reproducibility.

Training. We train the language models using supervised fine-tuning exclusively on a single difficulty bin and repeat this process for all bins in the dataset. We format the datasets into instruction-response templates. All model parameters are trained on the response tokens for five epochs. We use HuggingFace transformers and the TRL packages to train the model.

Evaluation. We use `lm-eval-harness` to evaluate all the datasets. We follow the standard evaluation protocols and metrics from `lm-eval-harness` to evaluate both the zero-shot model and the trained models. For a given dataset, we compute the accuracy across all difficulty bins (except the training bins). We report the difference in performance as heatmaps to isolate the effect of cross-difficulty generalization over the zero-shot baseline (see Figure 3 for a brief illustration). This evaluation enables us to assess how well a model performs beyond the specific difficulty it was initially exposed to during fine-tuning.

5 Findings

We report results for Qwen2.5 14B Instruct trained on single difficulty bins and evaluated across all ten bins of six benchmarks. (See Appendix E for additional Qwen and Llama models.) Across settings, we observe limited cross-difficulty generalization, i.e., models trained on easier data fail to generalize consistently to harder data, and those trained on harder data do not generalize consistently to easier data. We further find that this generalization capability decreases as the train–test difficulty gap increases, with the strongest performance concentrated near the diagonal, where the difficulties between train and test sets are similar. Finally, we observe that these patterns are consistent across model families and sizes, suggesting that they stem from properties of the data distribution rather than from the model.

Easy-to-hard and hard-to-easy generalization are limited. Figure 4 presents the cross-difficulty generalization results for the Qwen2.5 14B Instruct model. The rows indicate the training difficulty bins, while the columns represent the test difficulty bins, with both axes ordered from easiest (bin 0) to hardest (bin 9). The values in each cell are the difference in accuracy between the fine-tuned model and the zero-shot performance. Positive values suggest good generalization from the corresponding training bin to the test bin, and negative values show poor generalization. Moving horizontally to the right from each diagonal cell shows easy-to-hard generalization, and moving left shows hard-to-easy generalization.

Our results show that models trained only on easier bins fail to generalize to harder ones. In the heatmap, this appears as a sharp performance decline when moving right from the diagonal. For instance, in Figure 3, which shows the results for the MMLU-Pro dataset, models trained on the easiest bins (0) perform well on neighboring easy bins but quickly degrade when evaluated on bins five and above. This pattern directly challenges previous claims that easy-only supervision can

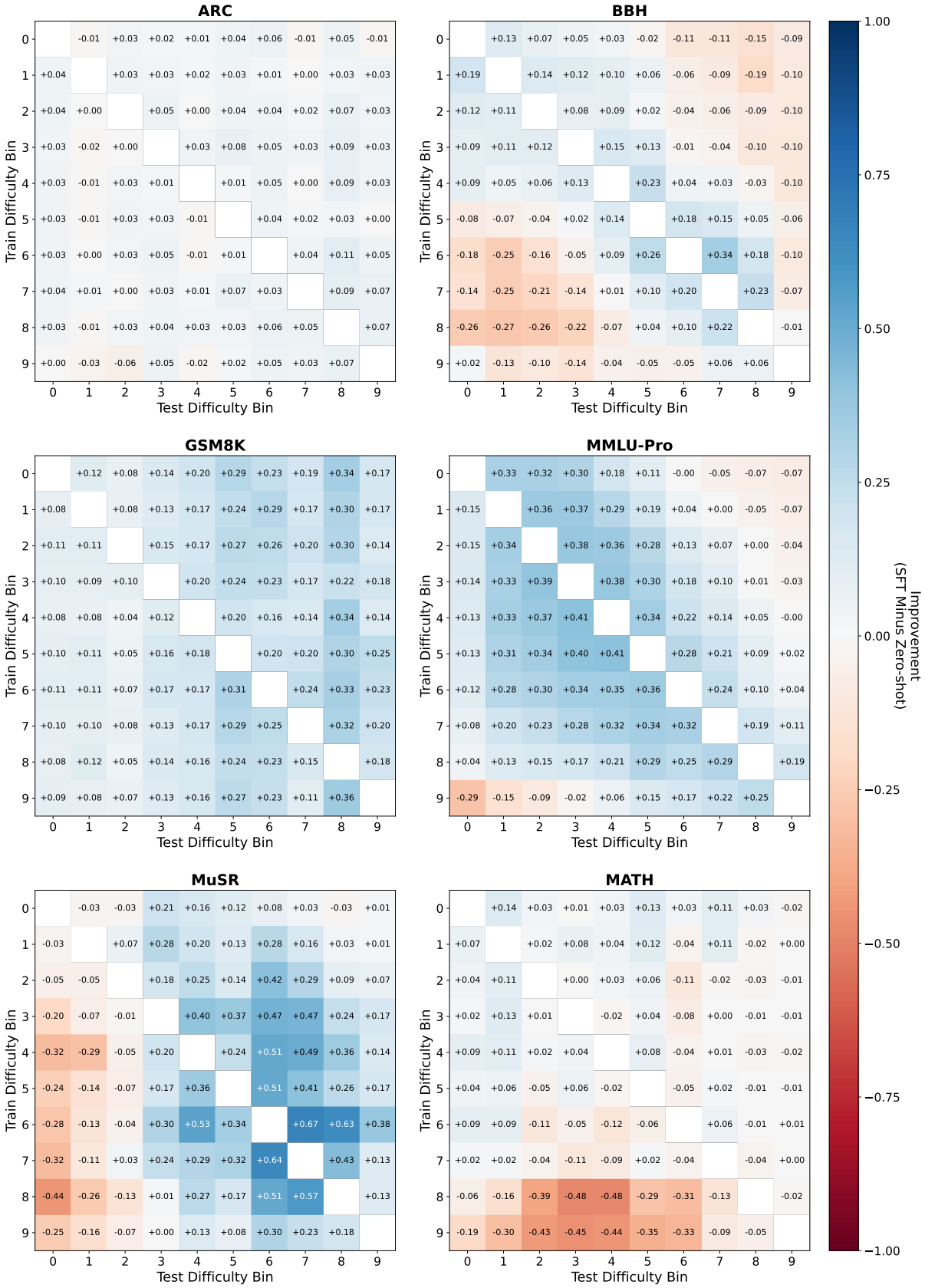

Figure 4: **Improvement analysis for Qwen2.5 14B Instruct showing the difference between SFT and zero-shot performance**. Blue indicates positive improvements (SFT better than zero-shot), red indicates negative improvements (SFT worse than zero-shot).

recover performance on hard tasks (Hase et al., 2024; Sun et al., 2024), while remaining consistent with Hase et al. (2024)'s observation that supervision gains may decline once the train-test difficulty gap becomes sufficiently large.

We also observe limited hard-to-easy generalization in our results. In datasets such as BBH, models trained on the hardest bins actually perform worse on easy questions, producing negative values throughout the lower triangle (Figure 4). This suggests that training exclusively on hard data does not generalize to easier data.

ARC and GSM8K results further demonstrate how cross-difficulty generalization can differ across datasets and model families. ARC shows almost no cross-difficulty generalization, with near zero gain across train and test bins. GSM8K shows moderate generalization in Qwen2.5 models, suggesting better cross-difficulty generalization, but but this pattern doesn't hold for Llama models (see Appendix E), where we don't see any generalization for most of the bins.

Overall, the results demonstrate that cross-difficulty generalization is narrow in scope. These results challenge the notion that simply training on either easy or hard data can achieve broad generalization across difficulty levels.

Larger train–test difficulty gaps lead to weaker generalization. Across datasets, we observe that the strongest generalization values cluster tightly around the diagonal of Figure 4, where the training and test bins are close in difficulty. This suggests that models primarily generalize to data of comparable difficulty rather than across significant differences in difficulty. As the gap between the training and test difficulty increases, accuracy declines, eventually dropping below the zero-shot baseline in both directions. This pattern of generalization explains why both easy-to-hard and hard-to-easy results drop in Section 5, suggesting that the model's ability to generalize is constrained by the difficulty level distribution of its training data rather than by its overall capacity.

Diverse test difficulty is essential. When the model is trained on harder examples, we often find that the performance on harder examples improves, but performance on easier examples sees no benefit or even decreases (see Figures 4, 6, and 9). Hence, if a model performs well on challenging benchmark problems, we cannot assume that it also performs well on easy benchmark problems, as we cannot assume hard-to-easy generalization. This limitation is especially relevant for benchmarks that target only the most challenging problems, such as AIME (Mathematical Association of America, 2025) and HLE (Phan et al., 2025), and provide limited information about models' behavior on easier questions. Therefore, as language models are trained to solve increasingly complex problems, it is essential to continue benchmarking models' capabilities across a broad spectrum of difficulties.

Patterns are consistent across models and datasets. In Appendix E, we show that the limited cross-difficulty generalization patterns persist across model sizes and families. We evaluate instruction-tuned LLMs ranging from the 1B-parameter to the 14B-parameter in Qwen and Llama model families and observe similar patterns of weak generalization across the models as seen in Figure 4. While larger models achieve higher accuracy in absolute numbers, they exhibit weak cross-domain generalization compared to their zero-shot performance. The consistency in results suggests that our findings will hold across different model sizes.

6 Discussion

We revisit the growing claims of cross-difficulty generalization in LLMs. Our study show that cross-difficulty generalization is limited and reduces as we increase the gap between the difficulty of train and test data. Generalization is highest where training and testing difficulties are similar, and this pattern remains consistent across model families, scales, and datasets.

These results suggest that prior works (Yang et al., 2024; Hase et al., 2024; Sun et al., 2024; Pikus et al., 2025) may have overestimated easy-to-hard and hard-to-easy generalization. Several factors may explain this difference. First, we quantify difficulty using IRT and based on models' capabilities and performance rather than human-based metrics or heuristics. As shown in Figure 1 and Section 3.4, human-based difficulty metrics don't correlate with the IRT results. The hardest bin for models may still contain some of the easiest questions for humans, meaning that human-based metrics fail to properly isolate difficulty and produce overlapping distributions. Second, we analyze ten distinct difficulty bins rather than two or three splits. This allows us to train on one difficulty level and evaluate on problems significantly harder or easier than the training data, giving us a better sense of the true extent of cross-difficulty generalization than if we only compared against neighboring difficulty levels. Finally, we evaluate across six benchmarks that span reasoning, factual recall, and instruction following, thereby reducing dataset-specific biases rather than testing a narrow set of skills (Sun et al., 2024).

Given the limited cross-difficulty generalization in LLMs, training and evaluation datasets should explicitly account for difficulty alongside other desirable properties such as diversity and coverage. An open question remains whether a curriculum structured by model-based difficulty, such as IRT scores, can lead to cross-difficulty generalization. Future work should explore training objectives and selection strategies that explicitly target performance stability across difficulty levels.

Taken together, these findings reveal a gap in how the field conceptualizes generalization. Addressing this gap through precise difficulty measurement, systematic analysis, and difficulty-aware data design is essential for developing models that can extend their reasoning

beyond the training distributions.

7 Conclusion

We presented an analysis of cross-difficulty generalization in language models using fine-grained, model-based difficulty estimates from Item Response Theory. By training and evaluating models from multiple families and sizes across six benchmarks and ten difficulty bins, we showed that cross-difficulty generalization is limited and highly dependent on the gap between training and evaluation difficulty. We find that, across model families, model scales, and datasets, as we increase the train-test difficulty gap, generalization from both easy-to-hard data and hard-to-easy data decreases significantly. These findings challenge the common assumption that training an LLM on either easy or hard data can generalize to data with other difficulty levels. Our results motivate a reevaluation of how generalization is measured and improved in LLMs. A systematic, difficulty-aware perspective will be essential for building models that can perform reliably beyond their training distributions.

Limitations

Our analysis relies on publicly available benchmarks (ARC, GSM8K, MMLU-Pro, MATH, MuSR, and BBH) and on fine-grained difficulty scores estimated with Item Response Theory. All of these datasets are in English, so our conclusions may not directly extend to tasks in other languages or multilingual settings, or to domains where reliable ground-truth labels are hard to obtain.

We estimate difficulty from model response patterns rather than human annotation. While this provides a model-centric understanding of difficulty, shifts in model capabilities, data distributions, or evaluation practices could change the difficulty of each task and affect the strength of the observed patterns.

As with most work on open-source LLMs, we cannot guarantee that evaluation questions were entirely unseen during pretraining. Undetected data overlap might inflate accuracy and partially reduce true generalization. Although our focus on relative difficulty and cross-bin comparisons can help with this concern, eliminating it would require test sets collected after the pretraining cut-offs of all evaluated models.

Finally, we concentrate on single-bin training to isolate difficulty effects. This design clarifies the role of difficulty gaps but does not cover all possible training curricula, such as mixtures of bins or adaptive sampling strategies. Exploring how such curricula interact with fine-grained difficulty remains an important direction for future work.

Acknowledgments

The authors would like to thank the anonymous reviewers for their constructive feedback. We especially thank Ellie Pavlick, Chen Sun, Francisco Piedrahita-Velez, Zheng-Xin Yong, Reza Esfandiarpoor, Yik Siu Chan, Zhenke Liu, Tianze Hua, Apoorv Khandelwal, and other members of the BATS research group and AI Superlab at Brown University for their generous feedback on this work. This material is based upon work supported by the National Science Foundation under Grant No. RISE-2425380 and Grant No. IIS-2433429. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Disclosure: Stephen Bach is an advisor to Snorkel AI, a company that provides software and services for data-centric artificial intelligence.

References

Frank B. Baker. 1985. *The Basics of Item Response Theory*. Heinemann.

Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. 2023. Open llm leaderboard (2023-2024). https://huggingface.co/spaces/open-llm-leaderboard-old/open_llm_leaderboard.

Benjamin S. Bloom, Max D. Engelhart, Edward J. Furst, Walker H. Hill, and David R. Krathwohl. 1956. *Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook 1: Cognitive Domain*. McKay, New York.

Xuxi Chen, Zhendong Wang, Daouda Sow, Junjie Yang, Tianlong Chen, Yingbin Liang, Mingyuan Zhou, and Zhangyang Wang. 2024. Take the bull by the horns: Hard sample-reweighted continual training improves llm generalization. *CoRR*, abs/2402.14270.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *Preprint*, arXiv:1803.05457.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Linda L. Cook, Michael L. Eignor, and Nancy S. Petersen. 1985. A study of the temporal stability of irt item parameter estimates for the sat. Ets research report, Educational Testing Service.

Mucong Ding, Chenghao Deng, Jocelyn Choo, Zichu Wu, Aakriti Agrawal, Avi Schwarzschild, Tianyi Zhou, Tom Goldstein, John Langford, Anima Anandkumar, and Furong Huang. 2024. Easy2hard-bench: Standardized difficulty labels for profiling LLM performance and generalization. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 82 others. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5988–6008. PMLR.

Clémentine Fourrier, Nathan Habib, Alina Lozovskaya, Konrad Szafer, and Thomas Wolf. 2024. Open llm leaderboard v2. `https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard`.

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2023. Complexity-based prompting for multi-step reasoning. In *The Eleventh International Conference on Learning Representations (ICLR 2023)*.

Peter Hase, Mohit Bansal, Peter Clark, and Sarah Wiegreffe. 2024. The unreasonable effectiveness of easy training data for hard tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7002–7024, Bangkok, Thailand. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the math dataset. In *NeurIPS Datasets and Benchmarks*.

Dan Hendrycks, Mantas Mazzeo, Rishi Bommasani, Aditi Raghunathan, Chia-Wei Liu, and Jacob Steinhardt. 2021c. Measuring massive multitask language understanding. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021)*. Curran Associates, Inc.

Neal M. Kingston and Neil J. Dorans. 1982. The feasibility of using item response theory as a psychometric model for the gre aptitude test. Ets research report, Educational Testing Service.

John P. Lalor and Pedro Rodriguez. 2023. py-irt: A scalable item response theory library for python. *INFORMS Journal on Computing*, 35(1):5–13.

Nayoung Lee, Ziyang Cai, Avi Schwarzschild, Kangwook Lee, and Dimitris Papailiopoulos. 2025. Self-improving transformers overcome easy-to-hard and length generalization challenges. In *Forty-second International Conference on Machine Learning*.

Frederic M. Lord, Melvin R. Novick, and Allan Birnbaum. 1968. *Statistical Theories of Mental Test Scores*. Addison-Wesley.

Mathematical Association of America. 2025. American invitational mathematics examination (aime). American Invitational Mathematics Examination - AIME 2025.

Robert L. McKinley and Neal M. Kingston. 1987. Exploring the use of irt equating for the gre subject test in mathematics. Ets research report, Educational Testing Service.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candes, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20286–20332, Suzhou, China. Association for Computational Linguistics.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. Rissanen data analysis: Examining dataset characteristics via description length. In *International Conference on Machine Learning (ICML)*, pages 8500–8513. PMLR.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, Michael Choi, Anish Agrawal, Arnav Chopra, Adam Khoja, Ryan Kim, Richard Ren, Jason Hausenloy, Oliver Zhang, Mantas Mazeika, and 1090 others. 2025. Humanity's last exam. *Preprint*, arXiv:2501.14249.

Benjamin Pikus, Pratyush Ranjan Tiwari, and Burton Ye. 2025. Hard examples are all you need: Maximizing grpo post-training under annotation budgets. *Preprint*, arXiv:2508.14094.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Georg Rasch. 1960. *Probabilistic Models for Some Intelligence and Attainment Tests*. Danish Institute for Educational Research, Copenhagen. Reprinted 1980 by The University of Chicago Press.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024.

Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Zayne Rea Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. 2024. MuSR: Testing the limits of chain-of-thought with multistep soft reasoning. In *The Twelfth International Conference on Learning Representations*.

Zhiqing Sun, Longhui Yu, Yikang Shen, Weiyang Liu, Yiming Yang, Sean Welleck, and Chuang Gan. 2024. Easy-to-hard generalization: Scalable alignment beyond human supervision. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. 2023a. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada. Association for Computational Linguistics.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. 2023b. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada. Association for Computational Linguistics.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Neeraj Varshney, Swaroop Mishra, and Chitta Baral. 2022. ILDAE: Instance-level difficulty analysis of evaluation data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3412–3425, Dublin, Ireland. Association for Computational Linguistics.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online. Association for Computational Linguistics.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024. MMLU-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Norman L. Webb. 2002. Depth-of-knowledge levels for four content areas. Technical report, Wisconsin Center for Education Research, Madison, WI. ERIC Document ED414305.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Zhe Yang, Yichang Zhang, Tianyu Liu, Jian Yang, Junyang Lin, Chang Zhou, and Zhifang Sui. 2024. Can large language models always solve easy problems if they can solve harder ones? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1531–1555, Miami, Florida, USA. Association for Computational Linguistics.

Lucia Zheng, Joel Niklaus, Daniel E. Ho, Christopher D. Manning, and Percy Liang. 2023. Ifeval: A benchmark for evaluating instruction following abilities of large language models. *arXiv preprint arXiv:2311.07911*.

A Training Configuration

We fine-tuned a series of instruction-tuned models using full-parameter fine-tuning. The models include:

- Llama3.2 1B Instruct
- Llama3.2 3B Instruct
- Llama3.1 8B Instruct
- Qwen2.5 1.5B Instruct
- Qwen2.5 3B Instruct
- Qwen2.5 7B Instruct
- Qwen2.5 14B Instruct

All models were fine-tuned on difficulty-based subsets of datasets described in Appendix D.

Optimization. Training was conducted with a learning rate of `5e-6` using the `paged_adamw_8bit` optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.99$, weight decay = 0.1). We used cosine learning rate decay with a 10% warmup ratio and clipped gradients to a maximum L2 norm of 0.1. Each model was trained for 5 epochs with a batch size of 2 per device across 4 GPUs, resulting in an effective batch size of 8. Other parameters were kept at defaults.

Precision and Sequence Length. Training used mixed precision with `bf16` enabled and `fp16` disabled. The maximum input sequence length was 4096 tokens. All random seeds (Python, NumPy, and PyTorch) were set to 42 for reproducibility.

Infrastructure. Experiments were conducted using `DeepSpeed ZeRO Stage 3` with single-node distributed training via `accelerate`. We trained on heterogeneous clusters using combinations of NVIDIA A100, A6000, B200, and RTX 3090 GPUs, depending on availability. All runs used 8 GPUs per job, a single process per GPU, and no gradient checkpointing. Host memory was at least 128GB, and CUDA version 12.4 was used.

DeepSpeed Configuration. Training was launched using DeepSpeed, and the configuration is summarized below:

- **Zero Stage:** 3
- **Mixed Precision:** bf16
- **Gradient Accumulation Steps:** 1
- **Gradient Clipping:** 0.1
- **Offload Parameters:** False

Training Dynamics. We monitored training dynamics throughout our experiments and confirmed appropriate learning behavior. Training loss curves show consistent convergence across all difficulty bins, and final training accuracy averaged 0.84 across models and datasets.

Reproducibility. All training scripts, configurations, and requirements files will be released with the code repository to enable exact reproduction of all reported results.

B Evaluation Setup

All models were evaluated using `lm-eval-harness` with `vLLM` for efficient inference, and models are loaded in `bfloat16` precision. We used greedy decoding (`temperature=0`, `top_p=1.0`, `max_new_tokens=256`) and applied each model's chat template during formatting. Metrics were computed using the official harness implementations, reporting mean accuracy across difficulty bins and evaluation tasks. Evaluation was performed on A6000 and A100 GPUs with identical configurations to the training setup.

C IRT Model Selection

In conducting this work, we analyzed multiple IRT formulations, including the 1PL with guessing, 2PL, 3PL, and 4PL models. In IRT, these models differ by the parameters they estimate:

- 1PL (Rasch): models only question difficulty (β_i).
- 1PL with guessing: adds a fixed guessing lower bound while still modeling only difficulty, allowing a non-zero probability of success even for low-ability models.
- 2PL: adds a question discrimination parameter (α_i), allowing some questions to separate high- and low-ability models more sharply.
- 3PL: introduces a guessing parameter (c_i), accounting for a non-zero lower bound (for example, multiple-choice guessing probability).
- 4PL: adds an upper asymptote (d_i), allowing even the best models to have less than 100% success probability.

While these models introduce additional parameters such as discrimination and guessing, we found that they often led to unstable or counterintuitive estimates because they can explain the same poor model performance with different parameters. For example, under the 4PL model, questions that no model could answer were placed near the middle of the difficulty scale (bins 4–5) rather than at the hardest end, despite being infeasible with zero probability of success. Similarly, incorporating a guessing parameter caused easy questions to be mislabeled as "guessable" and assigned artificially high difficulty, or vice versa.

To avoid such artifacts, we opted for the simpler 1PL model, which captures difficulty as a single, consistent factor and achieves lower error and more interpretable scores across tasks.

Stability of IRT Difficulty Estimation. To further assess robustness, we ran an ablation in which the IRT model was fit using random subsets containing 25%, 50%, and 75% of the available LLMs, and compared the resulting difficulty rankings to those obtained using the full model set. Across all datasets, the inferred question difficulty ordering exhibits near-perfect agreement with the full-model baseline, with Spearman rank correlations ranging from 0.998 to 1.000. This shows that

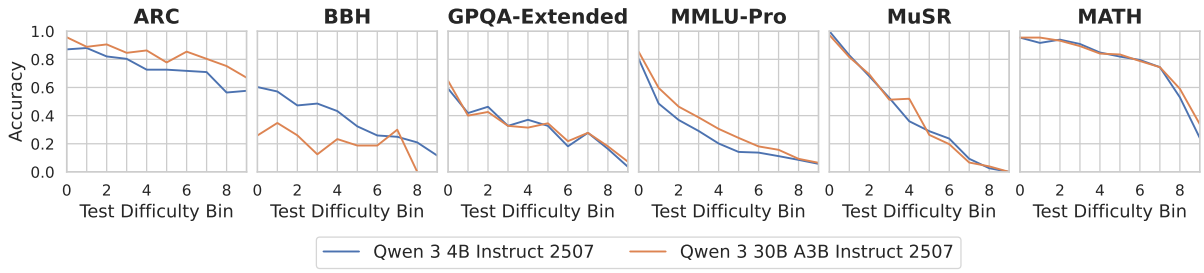

Figure 5: Zero-shot performance of Qwen 3 4B Instruct 2507 and Qwen 3 30B-A3B Instruct 2507 on the same benchmarks we evaluate against, divided by IRT difficulty bins. These models exhibit lower performance on more difficult bins, despite not being calibrated using their model responses.

reducing the model count does not meaningfully change the relative ordering of question difficulties, indicating that the IRT-based difficulty rankings are stable.

D Datasets

To analyze cross-difficulty generalization in language models, we use eight publicly available datasets that span a wide range of reasoning skills and subject areas. Together, they cover domains from elementary-level science and math to advanced expert knowledge, instruction following, and specialized reasoning challenges, providing a natural spectrum of task difficulty.

ARC. The AI2 Reasoning Challenge (ARC) dataset (Clark et al., 2018) consists of grade-school science questions that test a model's ability to apply commonsense reasoning and scientific knowledge.

GSM8K. GSM8K (Cobbe et al., 2021) is a collection of 8.5K linguistically diverse grade school math word problems. Each problem requires multi-step reasoning and arithmetic calculations, making it a benchmark for evaluating mathematical reasoning.

MMLU-Pro. The MMLU-Pro dataset (Wang et al., 2024) is an enhanced version of the Massive Multitask Language Understanding benchmark (Hendrycks et al., 2021a). It assesses LLMs across a broad range of subjects, including STEM, humanities, and social sciences, using multiple-choice questions requiring expert-level knowledge.

IFEval. The Instruction Following Evaluation (IFEval) dataset (Zheng et al., 2023) measures how well models follow natural language instructions. Unlike traditional benchmarks, IFEval is structured as a test set and uses an evaluation algorithm instead of gold labels. Because our framework requires training and evaluation on the same set to avoid distribution shift, we constructed gold labels for this dataset by annotating one correct answer for each question. These labels allow us to integrate IFEval into our pipeline while preserving its role as a challenging test of instruction-following ability.

GPQA-Extended. The GPQA-Extended benchmark (Rein et al., 2024) consists of graduate-level multiple-choice questions across physics, biology, and chemistry. The questions are designed to be challenging even for advanced models, making GPQA-Extended a strong test of reasoning capabilities.

BBH. The BIG-Bench Hard (BBH) dataset (Suzgun et al., 2023b) is a subset of the BIG-Bench benchmark focusing on tasks that are particularly difficult for LLMs. It includes diverse reasoning and comprehension problems that test generalization under challenging conditions.

These six datasets provide a comprehensive evaluation suite for studying cross-difficulty generalization. In the next subsection, we assign fine-grained difficulty scores to each instance in these datasets using Item Response Theory (IRT).

MATH. The MATH dataset (Hendrycks et al., 2021b) consists of competition-level problems from high school and undergraduate mathematics domains. It covers four major categories: Algebra, Number Theory, Counting and Probability, and Geometry. Problems and solutions are consistently formatted using LaTeX, allowing for the flexible encoding of mathematical expressions. Following the setting used in the Open LLM Leaderboard, we use only Level 5 problems, which represent the hardest difficulty level in the dataset.

Directly finetuning on the final answers or the provided solutions in the MATH dataset resulted in significant decreases from zero-shot performance for all models. To mitigate this, we created training samples specific to both Qwen and Llama models. Specifically, we collect responses from Qwen 2.5 14B and Llama 3.1 8B at several temperatures: 0, 0.1, 0.2, and 0.7, collecting 16 samples for each prompt. For every correct response, we replace the dataset's given solution with our sampled solution. We further sample these models with few-shot prompting at higher temperatures (0.7, 1.0, and 1.2), collecting 32 samples for each prompt. We prompt using three examples, each consisting of: (a) the question, (b) the correct final answer, and (c) a correct response sampled from the model itself. Any correct response from these prompts also replaced the

MATH dataset's provided solution. This resulted in two datasets: one MATH dataset with solutions that more closely resembled Qwen 2.5 14B style text, and another with solutions that more closely resembled Llama 3.1 8B style text. We finetune all our Qwen 2.5 and Llama 3.1 models with these datasets.

MuSR. MuSR (Sprague et al., 2024) evaluates multistep soft reasoning in long-form natural language narratives. Instances are complex stories (e.g., 1000-word murder mysteries), and models must use clues to answer the questions about the narrative.

Licensing. All of these datasets are publicly available, and we use them in accordance with with their official licenses: GSM8K (MIT), ARC (CC BY-SA 4.0), MMLU-Pro (MIT), IFEval (Apache 2.0), GPQA-Extended (CC BY 4.0), BBH (MIT), MATH (MIT), and MuSR (MIT). To the best of our knowledge, based on the documentation of the datasets and our checks, none of these datasets contain personally identifiable information (PII) or intentionally offensive content.

Label Quality Analysis. We manually reviewed samples in the hardest difficulty bin to assess whether label noise contributes to high IRT difficulty scores. Across datasets, we did not find evidence of systematic mislabeling in the hardest bin. For example, in GSM8K, we identified five mislabeled questions in the hardest bin, representing less than 4% of that subset, which is insufficient to explain the observed difficulty distribution.

E Supplementary Results

Due to space limitations, the main paper presents results only for the Qwen2.5 14B Instruct model. Here, we provide the corresponding heatmaps for the remaining six models.

We also include the zero-shot results for the Qwen3 family discussed in Section 3.3. The full visualization is shown in Figure 5, which complements the main-text analysis.

We also report experiments on the IFEval and GPQA-Extended datasets. As shown in Table 2, both datasets are relatively small, each containing fewer than 550 examples. Consequently, after splitting by difficulty, each fine-tuning bin included fewer than 55 examples. Under these conditions, the models didn't show meaningful evidence of learning the tasks, likely due to the limited number of examples, and we observed no consistent patterns of cross-difficulty generalization. For completeness, we include their results here but exclude them from the main paper.

F Correlation

In the main paper, we summarized the correlation between IRT-based difficulty scores and human-defined metrics (§3.4). Here, we include the full set of correlation heatmaps across all eight evaluation datasets for completeness. These extended results provide a more detailed view of how IRT-based difficulty correlates with various human annotations (e.g., grade level, reasoning steps, question and answer length). As shown in Figure 19, the overall correlations remain weak, consistent with the trends discussed in the main text.

G Examples

Table 3 shows some examples where the dataset's difficulty annotation disagrees with the IRT-based difficulty estimation, along with the models' performance on those instances. Also, tables 4 and 5 show additional examples from each IRT difficulty bin.

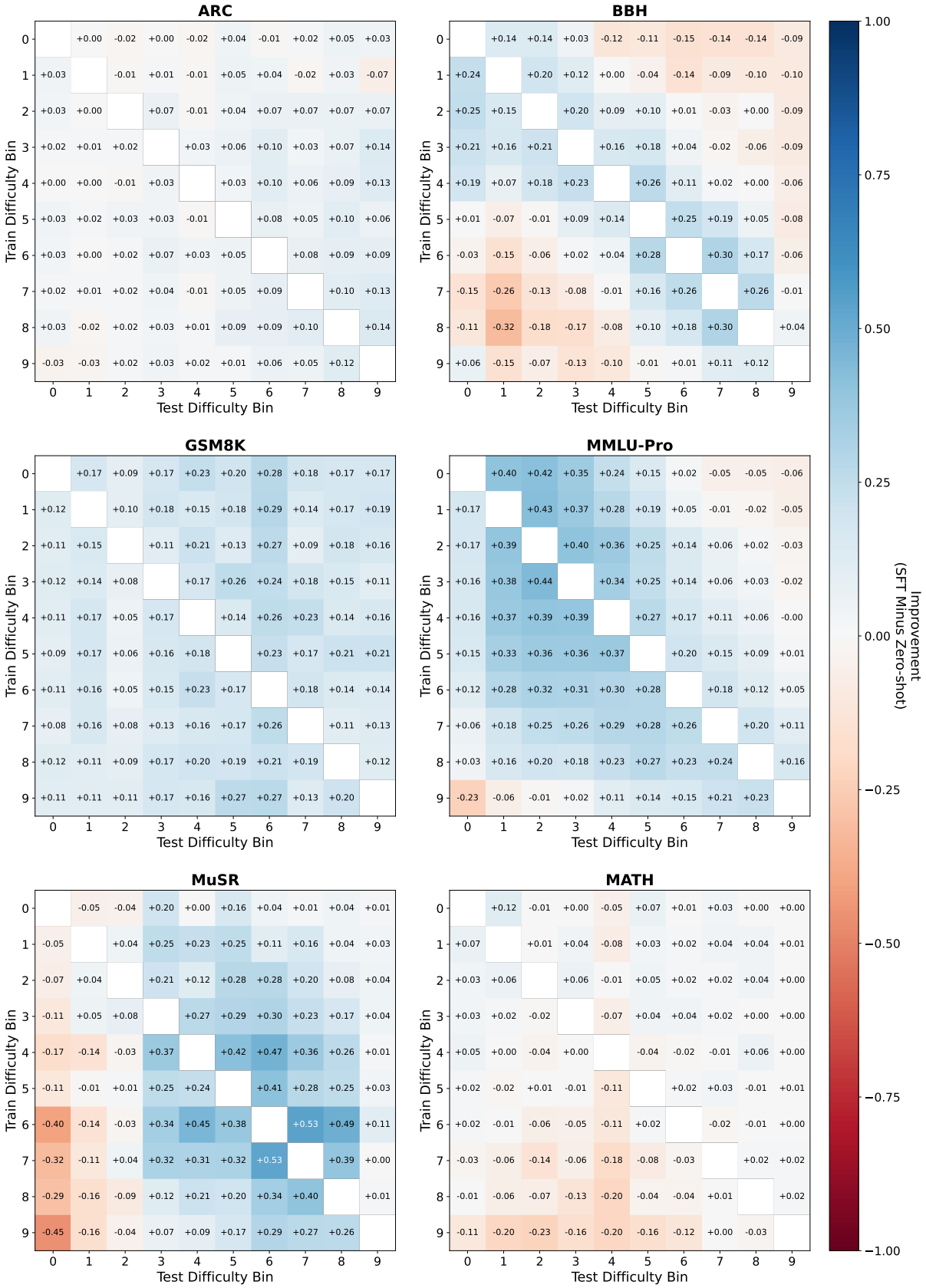

Figure 6: **Improvement analysis for Qwen2.5 7B Instruct showing the difference between SFT and zero-shot performance.** Blue indicates positive improvements (SFT better than zero-shot), red indicates negative improvements (SFT worse than zero-shot).

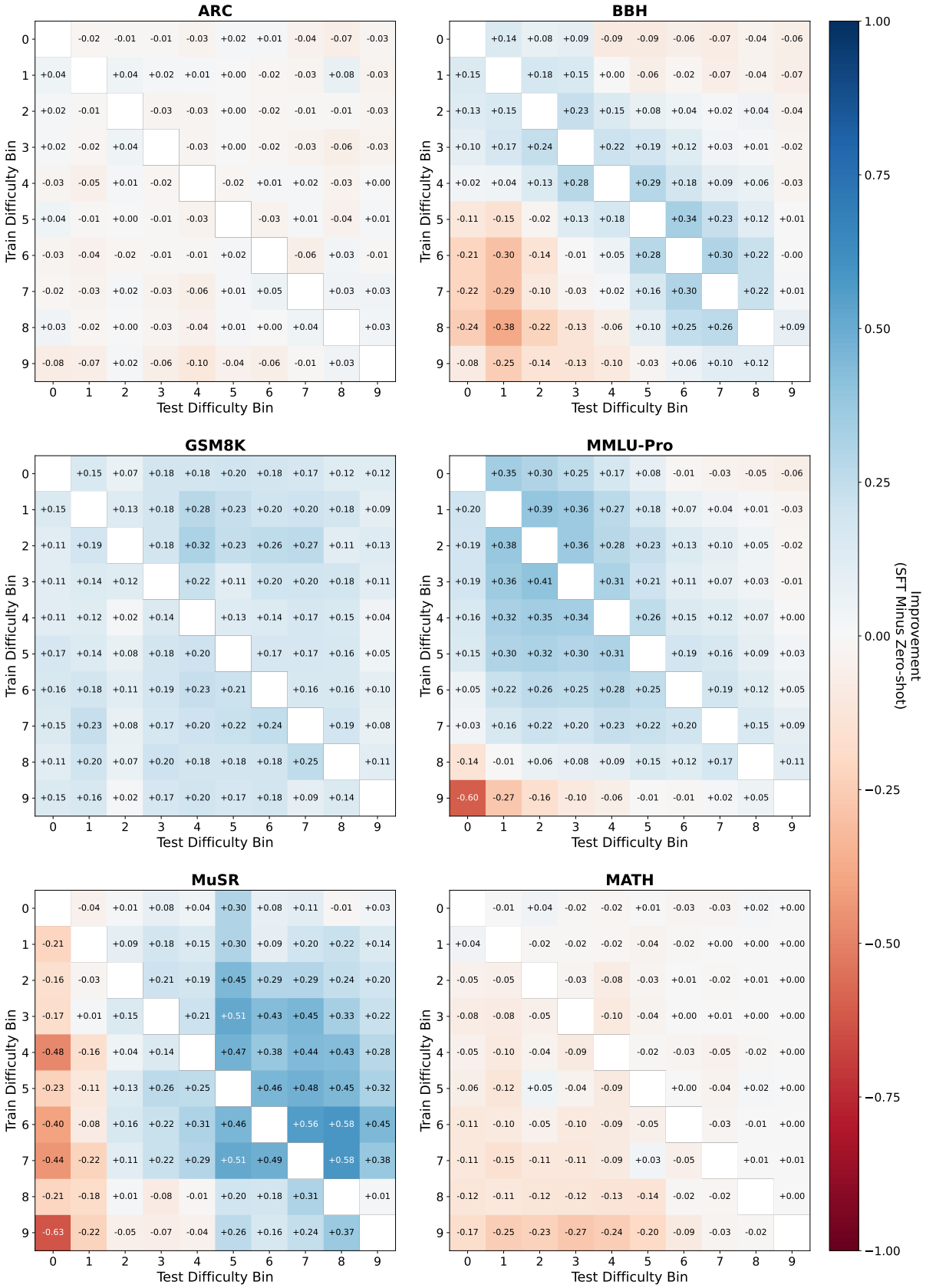

Figure 7: **Improvement analysis for Qwen2.5 3B Instruct showing the difference between SFT and zero-shot performance.** Blue indicates positive improvements (SFT better than zero-shot), red indicates negative improvements (SFT worse than zero-shot).

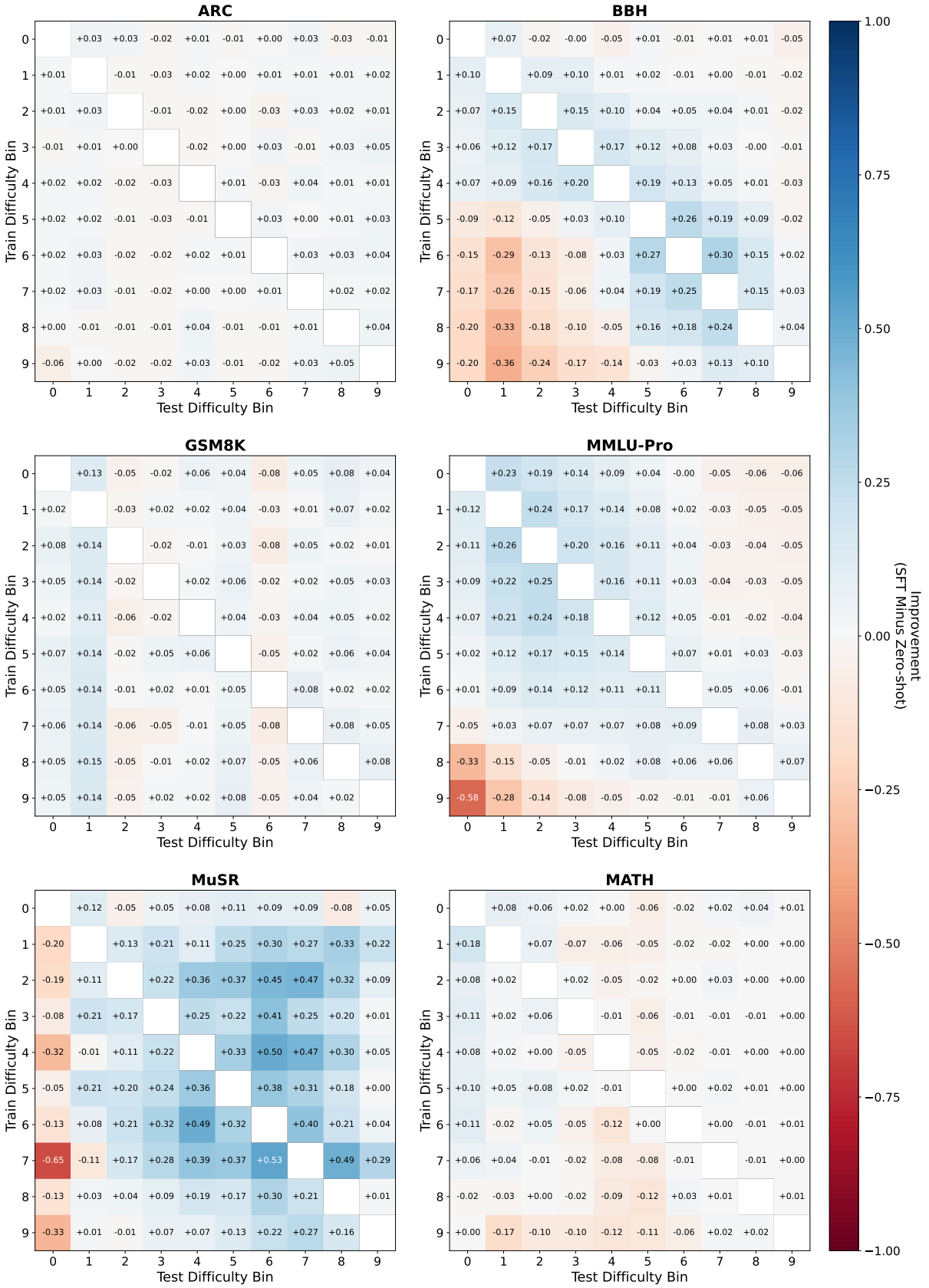

Figure 8: **Improvement analysis for Qwen2.5 1.5B Instruct showing the difference between SFT and zero-shot performance.** Blue indicates positive improvements (SFT better than zero-shot), red indicates negative improvements (SFT worse than zero-shot).

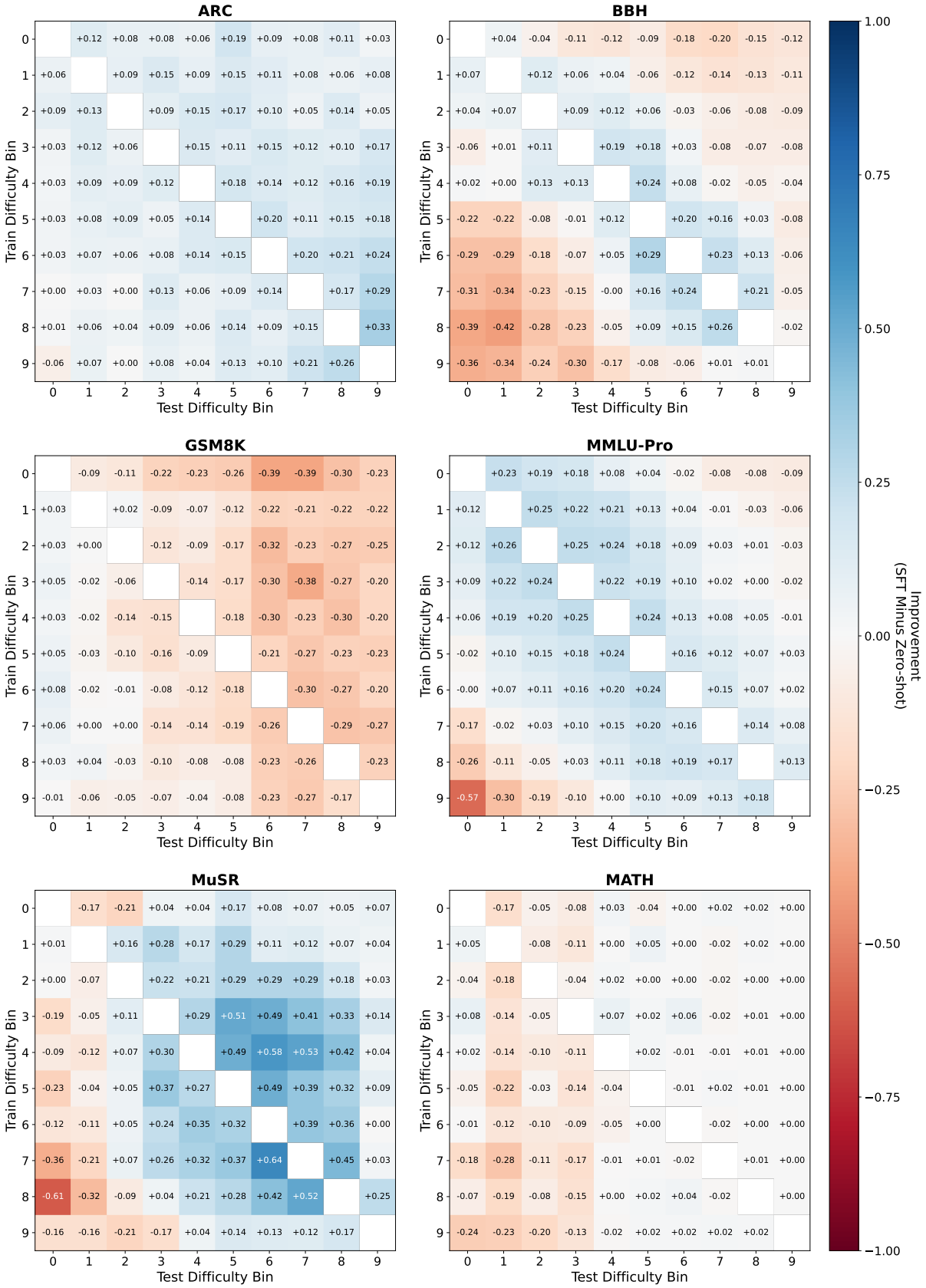

Figure 9: **Improvement analysis for Llama3.1 8B Instruct showing the difference between SFT and zero-shot performance.** Blue indicates positive improvements (SFT better than zero-shot), red indicates negative improvements (SFT worse than zero-shot).

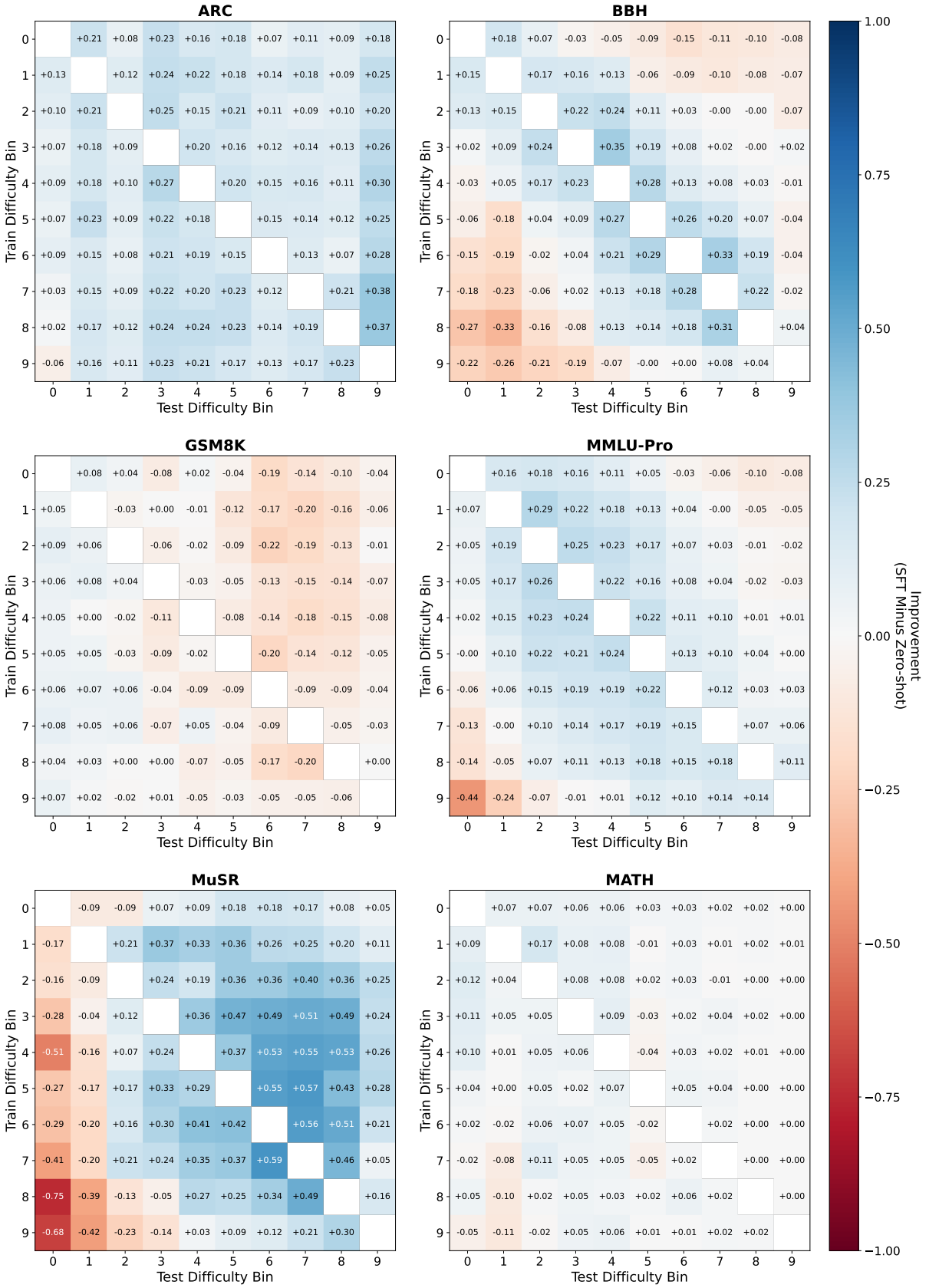

Figure 10: **Improvement analysis for Llama3.2 3B Instruct showing the difference between SFT and zero-shot performance.** Blue indicates positive improvements (SFT better than zero-shot), red indicates negative improvements (SFT worse than zero-shot).

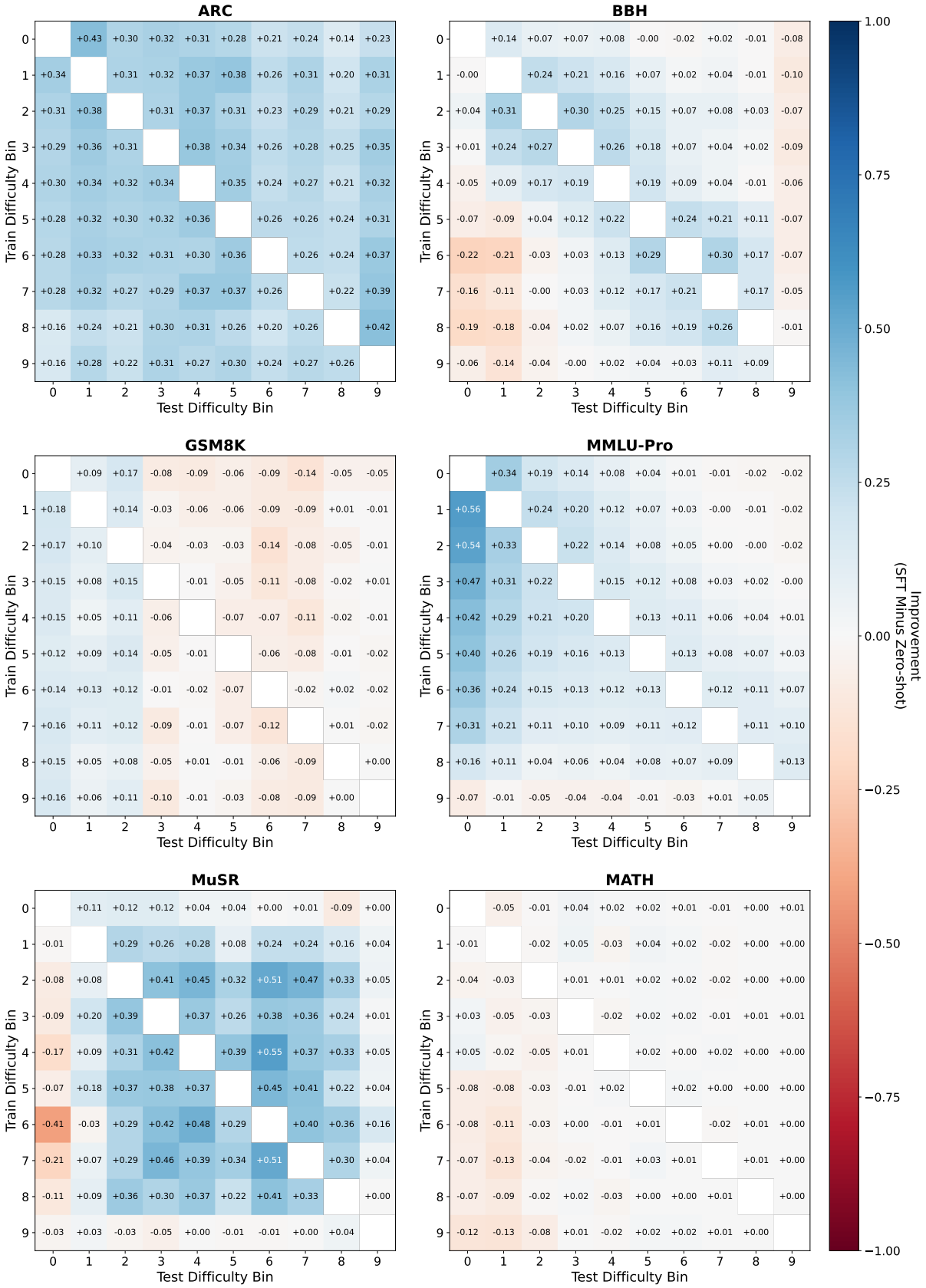

Figure 11: **Improvement analysis for Llama3.2 1B Instruct showing the difference between SFT and zero-shot performance.** Blue indicates positive improvements (SFT better than zero-shot), red indicates negative improvements (SFT worse than zero-shot).

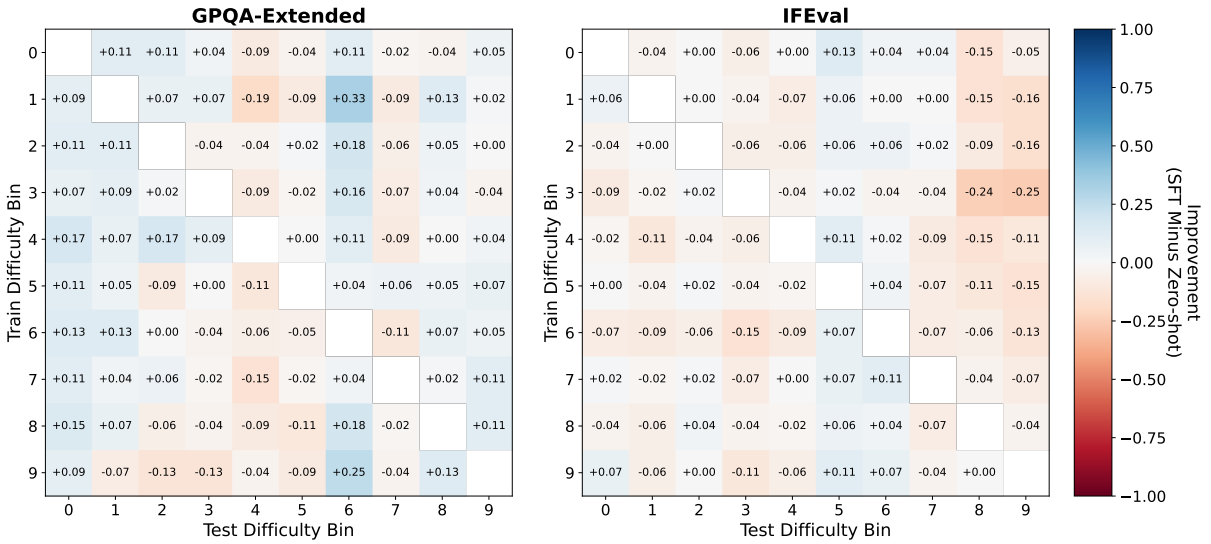

Figure 12: **Improvement analysis on IFEval and GPQA-Extended for Qwen2.5 14B Instruct showing the difference between SFT and zero-shot performance.** Blue indicates positive improvements (SFT better than zero-shot), red indicates negative improvements (SFT worse than zero-shot).

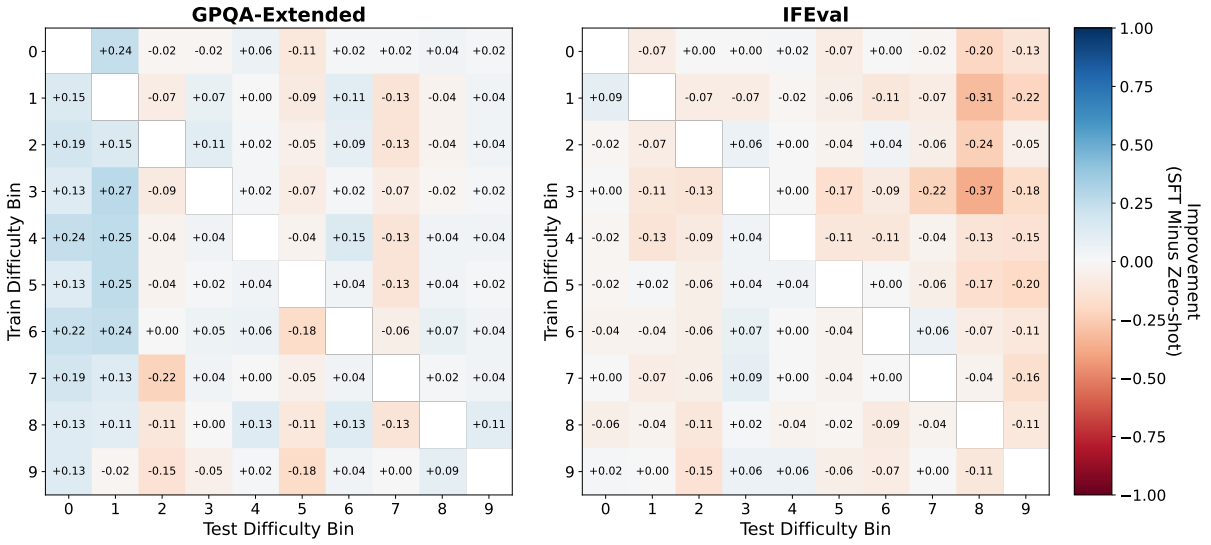

Figure 13: **Improvement analysis on IFEval and GPQA-Extended for Qwen2.5 7B Instruct showing the difference between SFT and zero-shot performance.** Blue indicates positive improvements (SFT better than zero-shot), red indicates negative improvements (SFT worse than zero-shot).

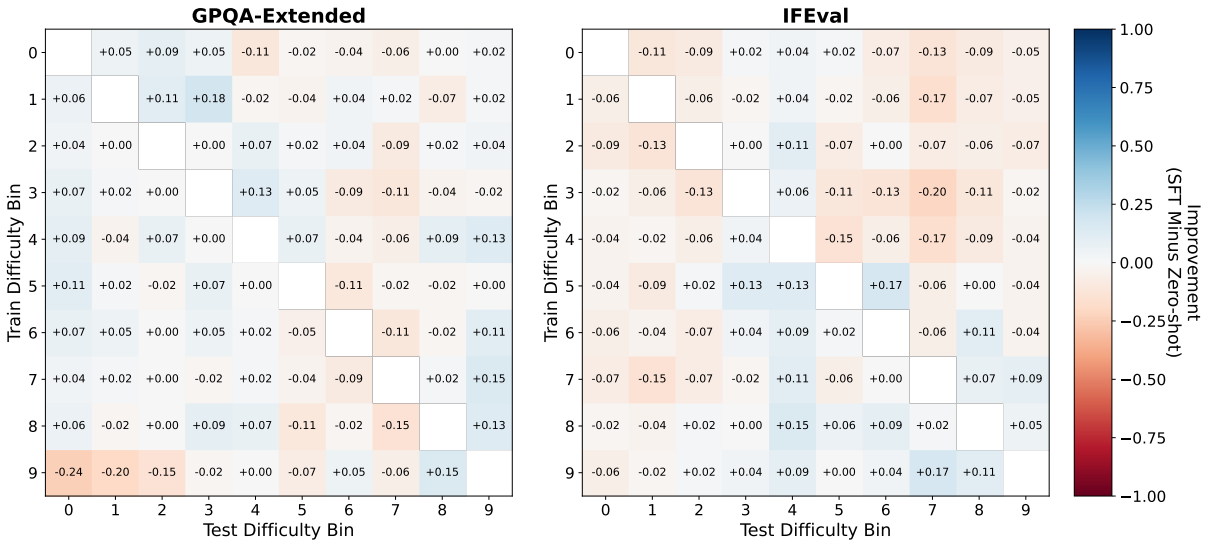

Figure 14: **Improvement analysis on IFEval and GPQA-Extended for Qwen2.5 3B Instruct showing the difference between SFT and zero-shot performance.** Blue indicates positive improvements (SFT better than zero-shot), red indicates negative improvements (SFT worse than zero-shot).

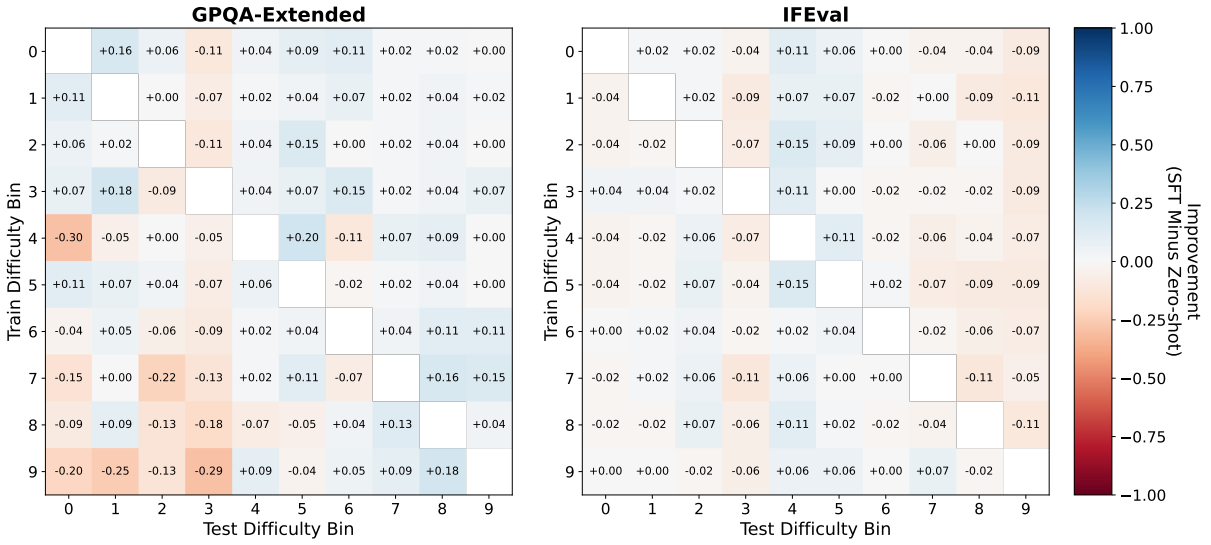

Figure 15: **Improvement analysis on IFEval and GPQA-Extended for Qwen2.5 1.5B Instruct showing the difference between SFT and zero-shot performance.** Blue indicates positive improvements (SFT better than zero-shot), red indicates negative improvements (SFT worse than zero-shot).

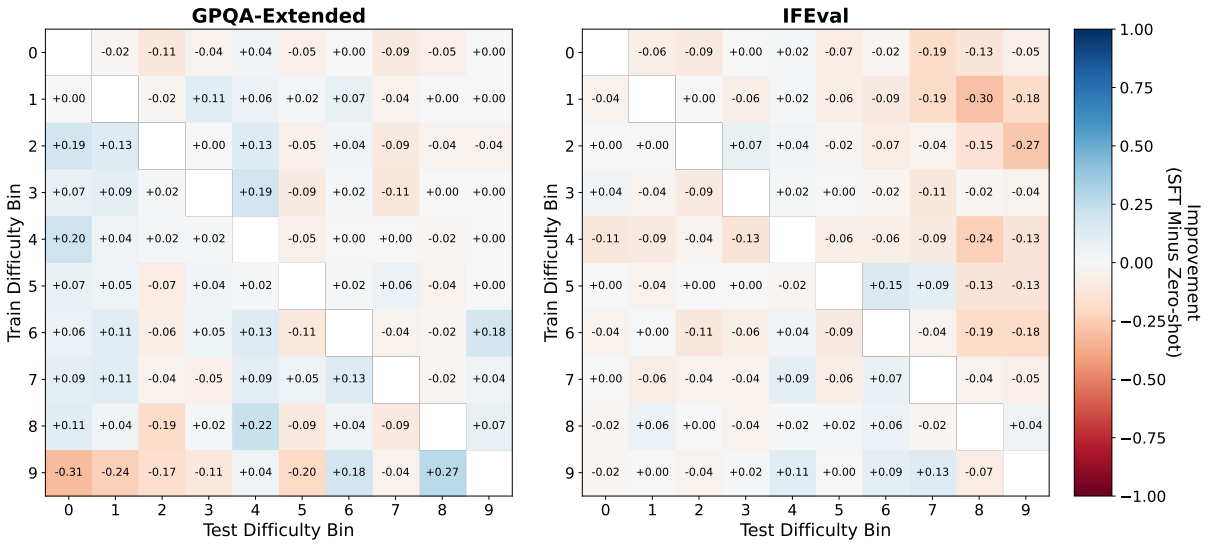

Figure 16: **Improvement analysis on IFEval and GPQA-Extended for Llama3.1 8B Instruct showing the difference between SFT and zero-shot performance.** Blue indicates positive improvements (SFT better than zero-shot), red indicates negative improvements (SFT worse than zero-shot).

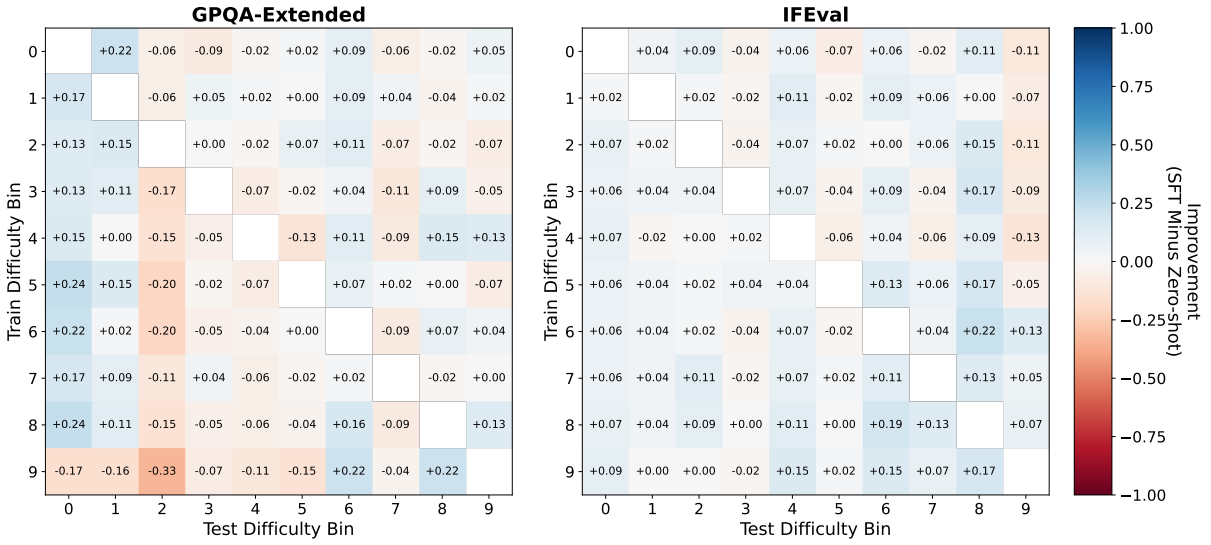

Figure 17: **Improvement analysis on IFEval and GPQA-Extended for Llama3.2 3B Instruct showing the difference between SFT and zero-shot performance.** Blue indicates positive improvements (SFT better than zero-shot), red indicates negative improvements (SFT worse than zero-shot).

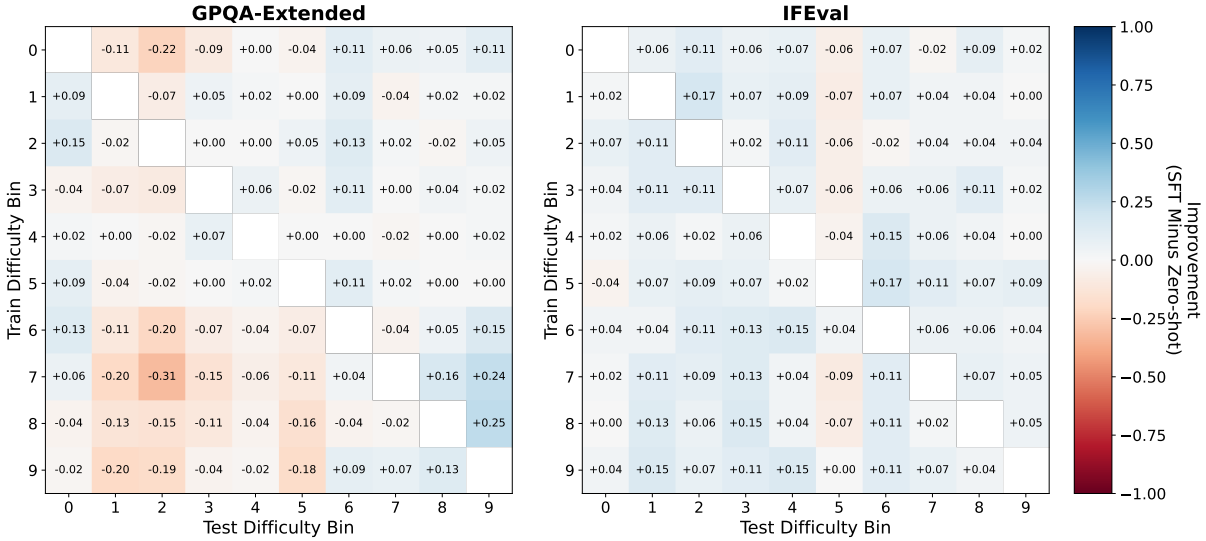

Figure 18: **Improvement analysis on IFEval and GPQA-Extended for Llama3.2 1B Instruct showing the difference between SFT and zero-shot performance.** Blue indicates positive improvements (SFT better than zero-shot), red indicates negative improvements (SFT worse than zero-shot).

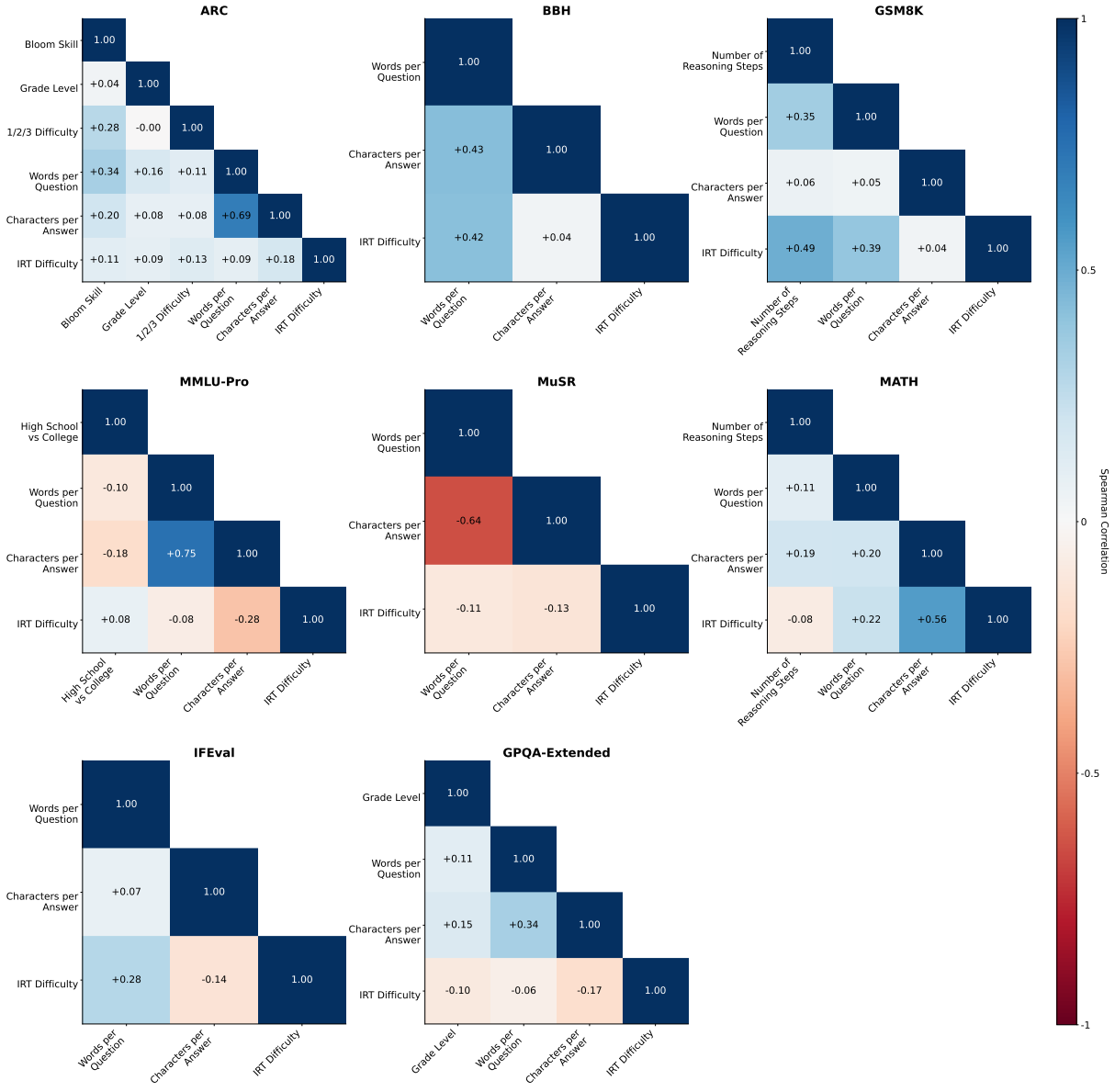

Figure 19: **Correlation heatmaps between IRT-based difficulty scores and human-defined metrics across all eight evaluation datasets.** Each heatmap shows Spearman rank correlation coefficients, with colors ranging from red (negative correlation) to blue (positive correlation). Metrics vary by dataset based on available annotations: ARC includes grade level and Bloom's taxonomy; GSM8K includes reasoning steps; MMLU-Pro includes high school vs. college labels. Across all datasets, most correlations remain weak ($|\rho| < 0.3$), with only GSM8k's reasoning steps ($\rho = 0.49$) and question length in BBH ($\rho = 0.42$) showing moderate positive correlations. Answer length consistently shows negative correlations, while expert-assigned difficulty ratings and educational levels show minimal alignment with model performance patterns.

Question	Answer	Difficulty ratings	Model checklist
Question: A ball is rolling on the ground. A force pushes the ball in the same direction that it is moving. What happens to the ball? **Options:** - A) It stops moving. - B) It moves slower. - C) It moves faster in the same direction it was moving. - D) It keeps moving in the same speed and direction.	C	- ARC-Challenge: 3rd Grade *(easy)* - IRT: bin 9 *(hard)*	- Llama3 70B: ✗ - Qwen2.5 72B: ✗ - Mixtral 8x7B v0.1: ✗
Question: A major polling organization wants to predict the outcome of an upcoming national election (in terms of the proportion of voters who will vote for each candidate). They intend to use a 95% confidence interval with margin of error of no more than 2.5%. What is the minimum sample size needed to accomplish this goal? **Options:** - A) 2048 - B) 1000 - C) 39 - D) 1536 - E) 2000 - F) 40 - G) 1024 - H) 1537 - I) 4096 - J) 500	H	- MMLU-Pro: High School *(easy)* - IRT: bin 9 *(hard)*	- Llama3 70B: ✗ - Qwen2.5 72B: ✗ - Mixtral 8x7B v0.1: ✗
Question: A ligand binds to a deep beta-barrel cleft of an enzyme. Based on the 2.5 Angstrom resolution of the X-ray diffraction crystal structure, the amino acids in contact with the ligand at the active site seem to be H34, T48, S62, and G128. You would like to perform PCR mutagenesis of the enzyme's sequence to verify the mechanism of the ligand-receptor binding. Which of the point mutations below will most likely affect the ligand-receptor interaction? **Options:** - A) 98A → G - B) 128G → C - C) 142A → G - D) 186T → C	C	- GPQA-Extended: Easy Undergrad *(easy)* - IRT: bin 9 *(hard)*	- Llama3 70B: ✗ - Qwen2.5 72B: ✗ - Mixtral 8x7B v0.1: ✗
Question: A potential negative impact of building a dam on a river is that the dam **Options:** - A) prevents sediment from flowing downstream. - B) increases the amount of water available to farms. - C) prevents seasonal downstream flooding. - D) increases the rate of water loss from a lake.	A	- ARC-Challenge: 8th Grade *(hard)* - IRT: bin 0 *(easy)*	- Llama3 70B: ✓ - Qwen2.5 72B: ✓ - Mixtral 8x7B v0.1: ✓
Question: A new enzyme is found in a transgenic mice that participates in synthesis of an unknown product using two reactants. When using radiolabeled compounds to study the enzyme, it is found that the enzyme catalyzes a process that switches a nitrogen group on one reactant to the other reactant. Which of the following categories would this new enzyme fall under? **Options:** - A) Ligase - B) Hydrolase - C) Transferase - D) Synthetase - E) Phosphatase - F) Lyase - G) Oxidoreductase - H) Decarboxylase - I) Kinase - J) Isomerase	C	- MMLU-Pro: College *(hard)* - IRT: bin 0 *(easy)*	- Llama3 70B: ✓ - Qwen2.5 72B: ✓ - Mixtral 8x7B v0.1: ✓
Question: What is a distinctive feature of thin film deposition using High Power Impulse Magnetron Sputtering (HiPIMS)? **Options:** - A) HiPIMS primarily relies on chemical vapor deposition (CVD) mechanisms for film formation. - B) HiPIMS generates short, intense plasma pulses to enhance ionization and film growth. - C) HiPIMS utilizes continuous DC power for a consistent deposition rate. - D) HiPIMS operates at lower vacuum pressures compared to traditional sputtering techniques.	B	- GPQA-Extended: Post Graduate *(hard)* - IRT: bin 0 *(easy)*	- Llama3 70B: ✓ - Qwen2.5 72B: ✓ - Mixtral 8x7B v0.1: ✓

Table 3: Examples of Disagreement Between Dataset Difficulty Label and IRT Difficulty

Question	Answer	Dataset	Difficulty ratings
Question: (not True) and True and False is	False	BBH	IRT: bin 0
Question: There are 220 castles in Scotland. 40 percent of them are ruins, and half of the ruined castles are unmanned. How many unmanned ruined castles are there in Scotland?	Number of ruined castles = $0.40 \times 220 = 88$ Half of the ruined castles are unmanned. Number of unmanned ruined castles = $\frac{88}{2} = 44$	GSM8K	IRT: bin 1
Question: Which is most responsible for recycling dead plants and animals in an ecosystem? Options: - A) water - B) bacteria - C) fish - D) insects	B	ARC	IRT: bin 2
Question: Here is the summary of a research paper on the effect of VHF radio waves on a certain type of bacteria: "Our results show that VHF radio waves have no effect on the growth of bacteria." Can you help me rewrite this summary in a more formal way, using APA format? Do not use words "ours" or "have".	Here's a rewritten summary in a more formal tone, following APA format: "The results of this study indicate that exposure to VHF radio waves does not affect the growth of bacteria." Alternatively, you could also phrase it as: "No significant effect on bacterial growth was observed following exposure to VHF radio waves." Both of these rewritten summaries maintain a formal tone and adhere to APA format guidelines.	IFEval	IRT: bin 3
Question: In the vibrant realm of music, the responsibility of delegating roles fell upon my shoulders. My ensemble consisted of three distinct personalities - Alice, Michael, and Henry. Each brought their own unique flair and abilities to the table, making the task of assigning roles a captivating puzzle. The primary roles to be allocated were that of an instrumentalist and a singer. The unique attributes of each band member meant that the decision required careful consideration and balance. Alice was a complex character. Though she was not one to hold eye contact with Michael during performances, she carried with her a passion for guitar that few could rival. Daily, she spent at least an hour dedicated solely to practicing her guitar skills. Her devotion was evident from the fact that she had been part of her school band as a guitarist. However, while Alice was eager to contribute, her voice would often falter during moments of singing. She'd lose her voice within a short span of singing, probably due to lack of formal vocal training. Henry, on the other hand, was quite a handful. He often interrupted Alice, largely ignoring her efforts to contribute to their shared passion. Henry might have been skilled, but he showed little interest in bettering his abilities on the bass guitar. His lack of practice was evident during the last gig when he was unable to keep up with the tempo, stopping midway through the performance. Perhaps his singing suffered as a result. Even though Henry had never taken singing lessons, his voice tended to crack often when he sang, indicating a lack of vocal control. Michael, whilst being a firecracker in band meetings, had his shortcomings as well. He had a knack for creativity which Henry often unjustly dismissed. Surprisingly, Michael failed to deliver during the last performance where he was unable to hit the high notes during his solo. He also visibly struggled to maintain the rhythm of his drum solos and often forgot his lines mid-song. Michael's odd trait was his tendency to forget his lines when Alice was around, an unfortunate condition considering they were bandmates. This became painfully clear when he failed to remember a simple drum solo he was assigned. The unraveling of the intricacies of each band member, their dynamics with one another, their strengths and weaknesses were all fascinating to say the least. Assigning them to either singing or instrumental duties was a task fraught with careful analysis. But through it all, one thing was for certain - the spark of artistry was present in each one of them. The contours of the task lay ahead with an air of anticipation, offering an unique challenge that I was ready to surmount. The music played on, and so did the story of Alice, Michael and Henry. Given the story, how would you uniquely allocate each person to make sure both tasks are accomplished efficiently?	Instrumentalist: Alice, Singer: Henry and Michael	MuSR	IRT: bin 4

Table 4: Examples of Questions from IRT difficulty bins 0–4

Question	Answer	Dataset	Difficulty ratings
Question: Solve for x: $$\frac{66-2^x}{2^x+3}=\frac{4-2^x}{2^{x+1}+6}$$	First, we recognize that $2^{x+1}+6=2(2^x+3)$: $$\frac{2(66-2^x)}{2(2^x+3)}=\frac{4-2^x}{2(2^x+3)}$$ Then, we expand and collect like terms: $$\frac{128-2^x}{2(2^x+3)}=0$$ This equation can only be true when $2^x=128$, which indicates that $x=\boxed{7}$.	MATH	IRT: bin 5
Question: A researcher interested in examining the potential impact of parent alcoholism on child and family development recruits 12-year-olds (n = 100), 13-year-olds (n = 100), and 14-year-olds (n = 100)-half of whom have an alcoholic parent and half of whom do not-into a multiple-year longitudinal study assessing various outcomes. This study is best characterized as: Options: - A) A cross-sectional design - B) A correlational study - C) A pretest-posttest design - D) A cross-sequential cohort design - E) A quasi-experiment - F) A natural experiment - G) A cross-sectional cohort design - H) A true experiment - I) A case-control study - J) A longitudinal cohort design	D	MMLU-Pro	IRT: bin 6
Question: methyl (E)-but-2-enoate is treated with quinuclidine and acetone, forming product 1. 1 is treated with excess methylmagnesium bromide, forming product 2. how many chemically distinct non-exchanging hydrogen signals will there be in the 1H nmr spectrum of product 2? (There may be signals that practically would have very close chemical shifts, but the answer should be the number that are in principle distinguishable.) Options: - A) 8 - B) 6 - C) 3 - D) 4	D	GPQA-Extended	IRT: bin 7
Question: Two jokers are added to a 52 card deck and the entire stack of 54 cards is shuffled randomly. What is the expected number of cards that will be strictly between the two jokers?	Each card has an equal likelihood of being either on top of the jokers, in between them, or below the jokers. Thus, on average, $1/3$ of them will land between the two jokers. Multiplying this by the 52 yields our answer of $\boxed{\frac{52}{3}}$.	MATH	IRT: bin 8
Question: Alice, Bob, Claire, Dave, Eve, Fred, and Gertrude are dancers at a square dance. At the start of a song, they each have a partner: Alice is dancing with Ophelia, Bob is dancing with Melissa, Claire is dancing with Jamie, Dave is dancing with Sam, Eve is dancing with Patrick, Fred is dancing with Rodrigo, and Gertrude is dancing with Karl. Throughout the song, the dancers often trade partners. First, Dave and Claire switch partners. Then, Alice and Eve switch partners. Then, Eve and Bob switch partners. Then, Claire and Bob switch partners. Then, Fred and Eve switch partners. Then, Gertrude and Dave switch partners. Finally, Dave and Alice switch partners. At the end of the dance, Fred is dancing with Options: - A) Ophelia - B) Melissa - C) Jamie - D) Sam - E) Patrick - F) Rodrigo - G) Karl	B	BBH	IRT: bin 9

Table 5: Examples of Questions from IRT difficulty bins 5-9

H Use of Large Language Models

We used AI Assistants such as ChatGPT and Grammarly for spell-checking, fixing minor grammatical mistakes, and polishing the writing. We also use GitHub CoPilot in VSCode to help write our codebase.