RECOVERY-ON-THE-LINE: LINEAR TRENDS IN POST-QUANTIZATION PERFORMANCE RECOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Many state-of-the-art large language models exceed tens of billions of parameters. To compress these models, several prior works have proposed *quantizing* the weights and activations of these models, using techniques like GPTQ and QuIP. An important reference for quantized language models is their ability to *recover* performance metrics such as accuracy after quantization; that is, the quantized language models should be as accurate as the original base models. It remains, however, unclear, *which* evaluations are most needed to assess *recovery*. If, for instance, recovery across all tasks is strongly linear (i.e. recovery on task A is a linear function of recovery on task B), then there should exist a dense subset of (independent) evaluations that are most necessary. In this paper, we examine the trends in recovery across pairs of tasks and metrics. Drawing from prior works which have shown that in-distribution and out-of-distribution accuracy often exhibit a strong linear relationship, we show that this relationship holds for recovery of accuracy as well.

1 INTRODUCTION

Post-training quantization (PTQ) has emerged as a popular method to improve the memory footprint and scalability of large language models (LLMs) while maintaining performance (Frantar et al., 2022; Chee et al., 2024; Huang et al., 2024). Typically, performance is measured by evaluating both the base and quantized LLM on a suite of benchmark tasks and assessing the quantized model's ability to *recover* baseline performance metrics (Kurtic et al., 2024; Hong et al., 2024). These evaluations ideally cover a large array of tasks and metrics, such as multiple choice accuracy, calibration error, and safety of generations (Xu et al., 2024).

Running such extensive evaluations, however, is time-consuming and expensive. For instance, running MosaicML's EvalGauntlet on both the LLama-3.1-8B base and it's 4-bit quantized version requires several GPU hours.

Prior work in classification settings has shown that performance (e.g. accuracy) across tasks is often linearly correlated, even between in-distribution (ID) and out-of-distribution (OOD) tasks, a phenomenon known as *accuracy-on-the-line* (Miller et al., 2021; Santurkar et al., 2020; Sanyal et al., 2024; Cohen-Wang et al., 2024; Mania & Sra, 2020). Several works have extended this empirical observation to other metrics as well, including pairwise agreement (Baek et al., 2022; Saxena et al., 2024; Kim et al., 2024), pairwise disagreement Deng et al. (2022), and model invariance and generalization Lee et al. (2023). Others have noted that ID and OOD performance don't always correlate linearly, for instance when comparing model performance across subpopulations (Liang et al., 2023).

We investigate whether such a relationship exists for *recovery* of performance in quantized language models. There are two key reasons to study this relationship:

1. **Understanding evaluations**: The relationship in the recovery of performance, both across tasks (e.g. question-answering, multiple choice, etc.) and metrics (accuracy, calibration, etc.) is poorly understood. Investigating these relationships provides a basis for generating new tasks and metrics which current evaluations fail to cover.

2. **Evaluation efficiency**: For generative models, one can always generate new evaluations, but these evaluations may be redundant. Analyzing the quantization recovery relationship across tasks can find a core subset of tasks that "covers" all key evaluations. For instance, if recovery on one task exactly predicts recovery on all other tasks (e.g. linear relationship), then one evaluation is sufficient to measure recovery.

We empirically show with evaluations across 24 tasks that several state-of-the-art LLMs exhibit strong correlation in recoveries across benchmark datasets and metrics, showing that (a) accuracy recovery generally follows a strongly linear relationship across all tasks and quantization levels, and (b) this relationship seems to break down when evaluating non-accuracy metrics.

2 ANALYSIS

We compress four widely-used, open-source LLMs—`Llama-3.1-8B-Instruct`, `Qwen2.5-7B-Instruct`, `Falcon-3-7B-Instruct`, and `Mistral-7b-instruct-v0.3` using the GPTQ framework (Frantar et al., 2022) at four distinct bitwidth configurations: `W8A8`, `W4A16`, `W3A16`, and `W2A16`. Here, `WxAy` refers to `x`-bit weight and `y`-bit activation quantization respectively. We follow standard hyperparameter settings for GPTQ and quantize each model separately at each bitwidth setting.

We evaluate both base and quantized models on 24 tasks drawn from MosaicML's EvalGauntlet, spanning five categories of large language model evaluations: **(i)** World Knowledge, **(ii)** Language Understanding, **(iii)** Symbolic Problem Solving, **(iv)** Reading Comprehension, and **(v)** Commonsense Reasoning. Across these tasks, we primarily assess performance via standard accuracy-based metrics. We additionally evaluate whether a "core" subset of tasks can reliably predict quantization recovery on the full set. To that end, we measure and compare performance recovery for each quantized model, then investigate how strongly recovery on a small number of tasks correlates with overall recovery across all tasks.

In addition, we also evaluate our models' abilities to recover *non*-accuracy based metrics such as calibration. We evaluate calibration on four open-sourced multiple choice datasets available on HuggingFace: OpenbookQA (Mihaylov et al., 2018), CosmosQA (Huang et al., 2019), MMLU (Hendrycks et al., 2020), and MMLU-Pro (Wang et al., 2024) using three models from the Llama-3 suite (Dubey et al., 2024).

2.1 RESULTS

Recovery of Accuracy We show that recovery of accuracy across various benchmark tasks is highly correlated, with a strong linear fit in most cases. We analyze recovery on every pair of 24 tasks drawn from EvalGauntlet, resulting in 276 total pairwise comparisons.

On average, tasks in this evaluation set were highly correlated in accuracy recovery (mean pairwise $R^2 = 0.84$, median $R^2 = 0.92$). We evaluate the slope (of the line of best fit) and correlation (measured by R^2) for each pair of tasks in EvalGauntlet, similar to the plots shown in Figure 1 (except disaggregated to not be just within categories). The overall distribution of pairwise correlations across tasks, shown in Figure 2, suggests that most pairs of tasks are strongly correlated, with a small subset that are independent of other evaluations. This independent subset includes symbolic problem solving tasks (e.g, mathematical reasoning tasks) (Zhong et al., 2023).

To further investigate these correlations and how they change with quantization levels, we aggregate our suite of 24 tasks into five categories – **commonsense reasoning**, **language understanding**, **symbolic problem solving**, **world knowledge**, and **reading comprehension** – using the categories defined in EvalGauntlet. As shown in Figure 1, most pairs of categories exhibit a strong linear relationship in their recovery of accuracy, with smaller (2-bit) models having significantly poorer recovery than larger ones, which often have recovery close to 1. We again observe that symbolic problem solving tasks, which include various mathematical and analytical reasoning questions, are least correlated in recovery to other tasks. In particular, the least correlated task in our evaluation suite is AGIEvalLSAT Arithmetic Reasoning (mean $R^2 = 0.018$) and most correlated is HellaSwag (mean $R^2 = 0.905$).

Moreover, the slope of the category-to-category linear fit is always bounded between 0.5 and 2 after aggregation [1].

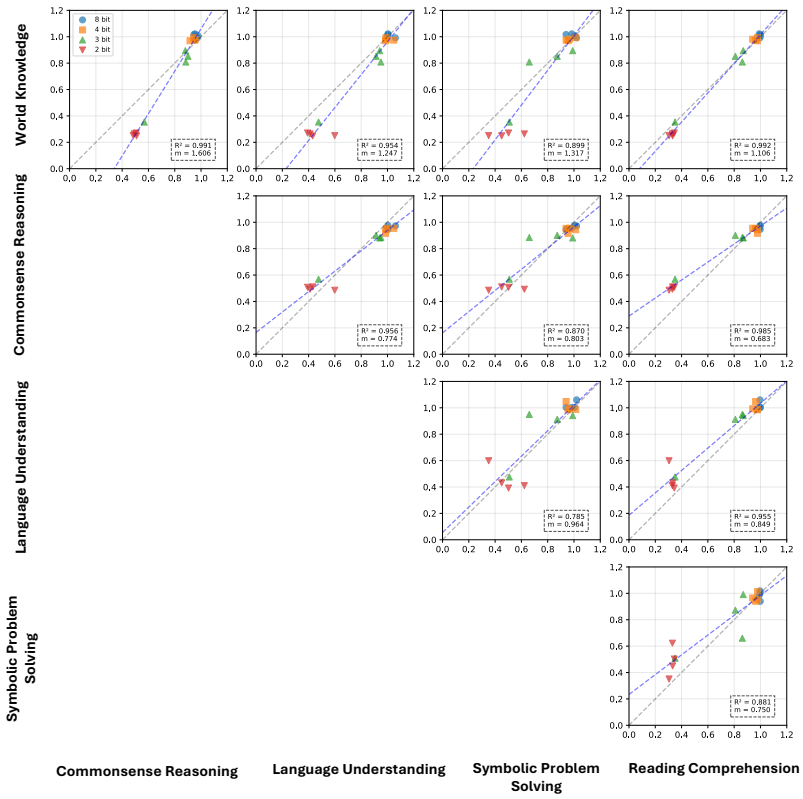

Figure 1: Across all categories of tasks in EvalGauntlet, we see strong linear trends in recovery.

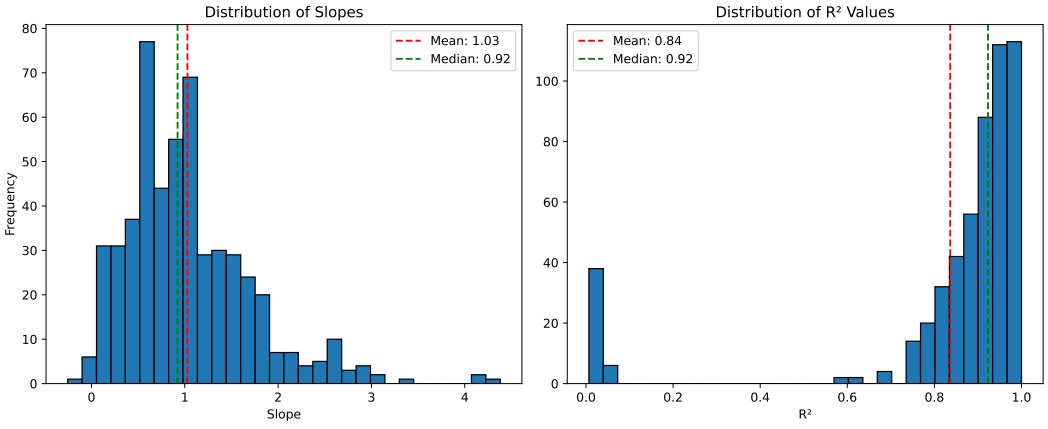

Figure 2: Slope and R^2 distribution of the best-fit line of the pairwise task-recovery plots evaluated across all 24 tasks. Majority of the tasks have high coefficient of determination, indicating the linear trend of performance recovery.

[1]The original accuracy-on-the-line papers use *probit scaling* before estimating their linear fits, to account for the fact that accuracies are bounded in $[0, 1]$. Given that recovery can be > 1, we do not use this rescaling. We additionally tried logit-scaling our plots, which maintained the linear relationship (see Appendix A.5.

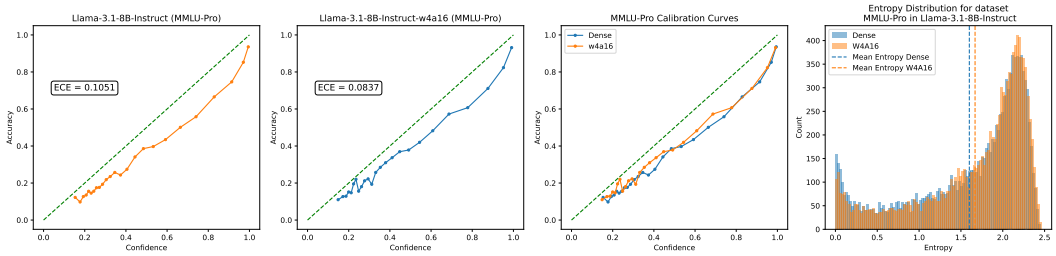

Figure 3: Quantizing models often *improves* calibration error. The base model tends to be overconfident at all confidence levels (left), which is partially mitigated after quantizing to 4 bits (middle). This is evident in the entropy distribution among the multiple choice options, where we see that the quantized model has greater entropy (right).

Recovery of Additional Metrics The strong linear trend in recovery of accuracy fails to hold for other metrics, such as calibration and entropy. The 5 shows the trends in calibration recovery across relevant benchmark datasets.

In particular, we observe that base models are often (very) *overconfident* and that quantizing mildly *corrects* this overconfidence. For instance, as shown in Figure 3 with Llama-3.1-8B-Instruct, the token-level entropy increases post-quantization, which in turn corrects overconfidence in the base model.

3 Discussion and Future Work

Here we showed preliminary evidence that performance recovery post-quantization follows a strongly linear trend, especially in the recovery of accuracy. In the future, we hope to extend these results to find a *robust subset* of evaluations that "covers" all evaluations needed to assess recovery. This includes both extending to new tasks and, more importantly, new metrics, such as perplexity and the proportion of answers which flip after quantization, some of which may be poorly correlated with accuracy (Dutta et al., 2024). A second direction of future work is to analyze the bit-level trends in recovery, i.e. analyzing the scaling of recovery against quantization-level across our dense subset.

References

Christina Baek, Yiding Jiang, Aditi Raghunathan, and J Zico Kolter. Agreement-on-the-line: Predicting the performance of neural networks under distribution shift. *Advances in Neural Information Processing Systems*, 35:19274–19289, 2022.

Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, and Christopher M De Sa. Quip: 2-bit quantization of large language models with guarantees. *Advances in Neural Information Processing Systems*, 36, 2024.

Benjamin Cohen-Wang, Joshua Vendrow, and Aleksander Madry. Ask your distribution shift if pre-training is right for you. *arXiv preprint arXiv:2403.00194*, 2024.

Weijian Deng, Stephen Gould, and Liang Zheng. On the strong correlation between model invariance and generalization. *Advances in Neural Information Processing Systems*, 35: 28052–28067, 2022.

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

Abhinav Dutta, Sanjeev Krishnan, Nipun Kwatra, and Ramachandran Ramjee. Accuracy is not all you need. *arXiv preprint arXiv:2407.09141*, 2024.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*, 2022.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Junyuan Hong, Jinhao Duan, Chenhui Zhang, Zhangheng Li, Chulin Xie, Kelsey Lieberman, James Diffenderfer, Brian Bartoldson, Ajay Jaiswal, Kaidi Xu, et al. Decoding compressed trust: Scrutinizing the trustworthiness of efficient llms under compression. *arXiv preprint arXiv:2403.15447*, 2024.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. *arXiv preprint arXiv:1909.00277*, 2019.

Wei Huang, Yangdong Liu, Haotong Qin, Ying Li, Shiming Zhang, Xianglong Liu, Michele Magno, and Xiaojuan Qi. Billm: Pushing the limit of post-training quantization for llms. *arXiv preprint arXiv:2402.04291*, 2024.

Eungyeup Kim, Mingjie Sun, Christina Baek, Aditi Raghunathan, and J Zico Kolter. Test-time adaptation induces stronger accuracy and agreement-on-the-line. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Eldar Kurtic, Alexandre Marques, Shubhra Pandit, Mark Kurtz, and Dan Alistarh. " give me bf16 or give me death"? accuracy-performance trade-offs in llm quantization. *arXiv preprint arXiv:2411.02355*, 2024.

Donghwan Lee, Behrad Moniri, Xinmeng Huang, Edgar Dobriban, and Hamed Hassani. Demystifying disagreement-on-the-line in high dimensions. In *International Conference on Machine Learning*, pp. 19053–19093. PMLR, 2023.

Weixin Liang, Yining Mao, Yongchan Kwon, Xinyu Yang, and James Zou. Accuracy on the curve: On the nonlinear correlation of ml performance between data subpopulations. In *International Conference on Machine Learning*, pp. 20706–20724. PMLR, 2023.

Horia Mania and Suvrit Sra. Why do classifier accuracies show linear trends under distribution shift? *arXiv preprint arXiv:2012.15483*, 2020.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*, 2018.

John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International conference on machine learning*, pp. 7721–7735. PMLR, 2021.

Shibani Santurkar, Dimitris Tsipras, and Aleksander Madry. Breeds: Benchmarks for subpopulation shift. *arXiv preprint arXiv:2008.04859*, 2020.

Amartya Sanyal, Yaxi Hu, Yaodong Yu, Yian Ma, Yixin Wang, and Bernhard Schölkopf. Accuracy on the wrong line: On the pitfalls of noisy data for out-of-distribution generalisation. *arXiv preprint arXiv:2406.19049*, 2024.

Rahul Saxena, Taeyoun Kim, Aman Mehra, Christina Baek, Zico Kolter, and Aditi Raghunathan. Predicting the performance of foundation models via agreement-on-the-line. *arXiv preprint arXiv:2404.01542*, 2024.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*, 2024.

Zhichao Xu, Ashim Gupta, Tao Li, Oliver Bentham, and Vivek Srikumar. Beyond perplexity: Multi-dimensional safety evaluation of llm compression. *arXiv preprint arXiv:2407.04965*, 2024.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*, 2023.

A Appendix

A.1 High correlation between tasks of same categories

In Figure 4, we show heatmaps illustrating correlation of quantization recovery for tasks *within* each of the five evaluation categories (e.g., World Knowledge, Language Understanding). We observe that tasks within a single category exhibit consistently high correlations, mirroring the strong *cross-category* correlations described in the main text.

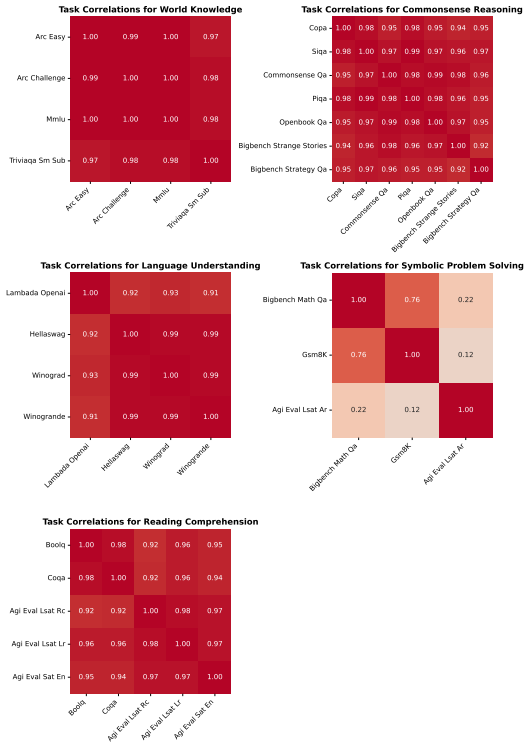

Figure 4: Correlation heatmaps by category, demonstrating that tasks within the same category exhibit high accuracy recovery correlations. Only in Symbolic Problem Solving do the tasks show notably weaker correlations compared to the other categories.

A.2 CALIBRATION RECOVERY IS NOT FULLY CORRELATED WITH ACCURACY RECOVERY

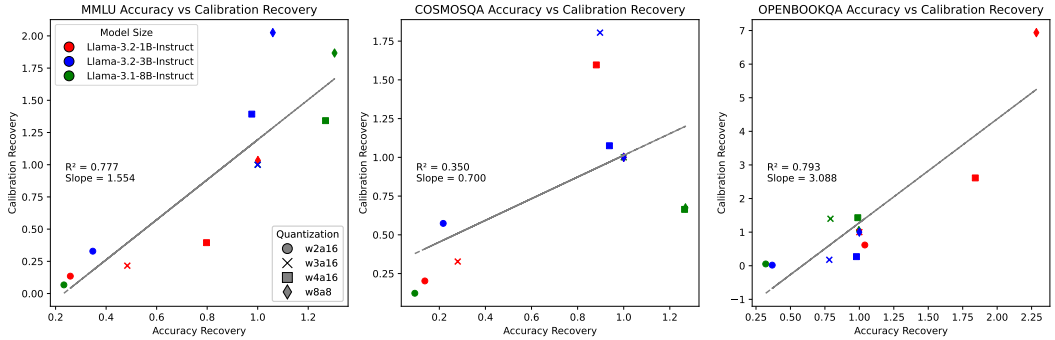

Figure 5: The linear trends we observe in recovery of accuracy across tasks do not hold across metrics, for instance between accuracy and calibration on the same task.

To calculate recovery of calibration, we use the following equation:

$$\text{Recovery}_{\text{ECE}} = \frac{\text{ECE}_{\text{base}}}{\text{ECE}_{\text{quantized}}}$$

We flip the equation relative to the accuracy recovery since a lower calibration error is better.

A.3 TASKS IN MOSAICEVAL GAUNTLET V0.3

Category	Task	N-shot
Common sense reasoning tasks	copa	0
	social_iqa	3
	commonsense_qa	0
	piqa	0
	openbookqa	10
	bigbench_strange_stories_multiple_choice	0
	bigbench_strategyqa_multiple_choice	0
Reading comprehension tasks	boolq	0
	coqa	0
	agieval_lsat_rc	5
	agieval_lsat_lr	5
	agieval_sat_en	5
Symbolic problem solving tasks	bigbench_elementary_math_qa_multiple_choice	1
	bigbench_operators_generate_until	3
	gsm8k	0
	agieval_lsat_ar	5
World knowledge tasks	arc_easy	3
	arc_challenge	3
	mmlu	5
	triviaqa	3
Language understanding tasks	lambada_openai	0
	hellaswag	0
	wsc273	3
	winogrande	5

Table 1: Categories, tasks, and their respective N-shot values in EvalGauntlet.

A.4 ACCURACY RECOVERY TREND BETWEEN TASKS IN DIFFERENT CATEGORIES

Figures 6-10 shows pairwise comparisons of quantization recovery across tasks within and across the five major categories in our evaluation suite. Overall, we observe near-linear correlations in many of these task pairs (e.g., reading comprehension tasks), while more diverse tasks (e.g., symbolic problem solving) exhibit weaker correlations. These findings indicate that some tasks can serve as strong proxies for others when assessing performance recovery, but not all categories share equally strong relationships.

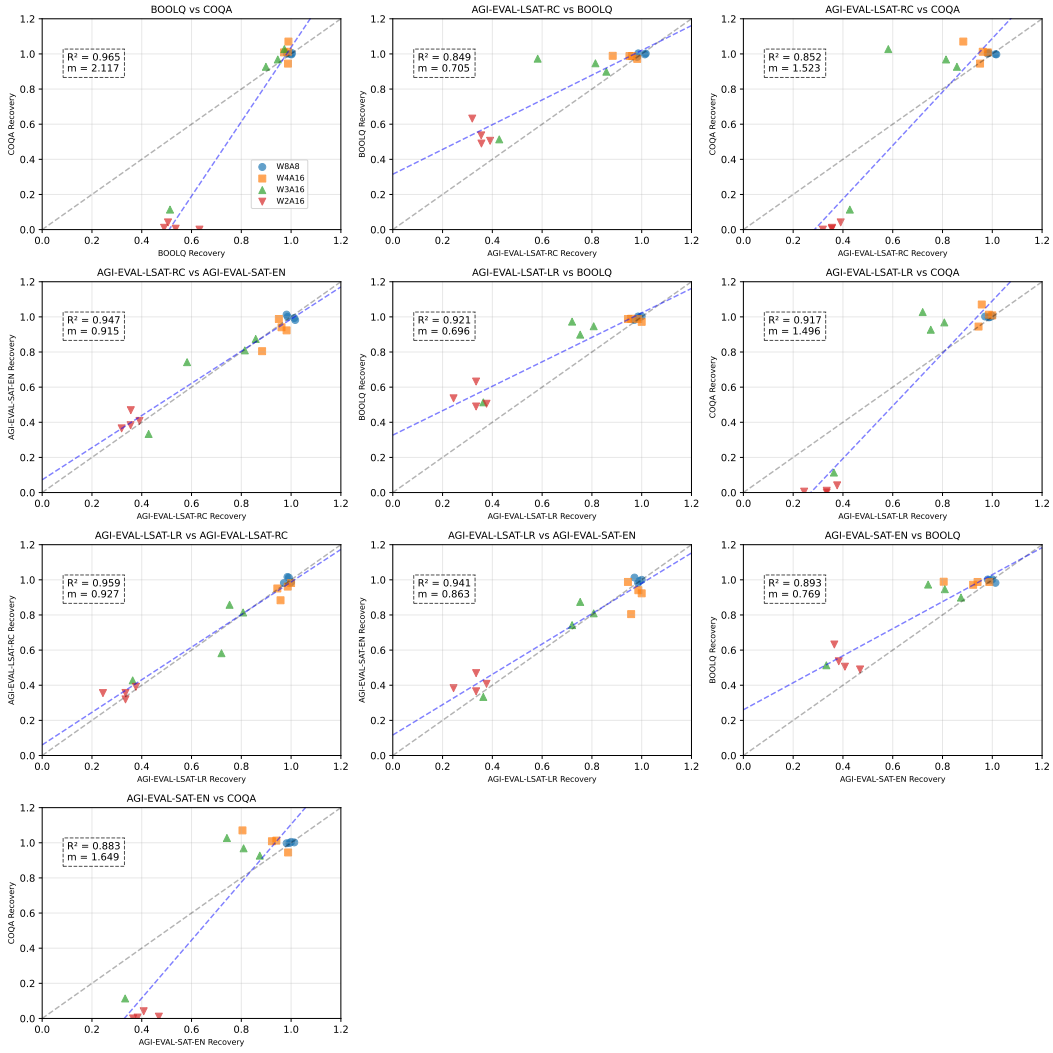

Figure 6: Accuracy recovery in reading comprehension tasks.

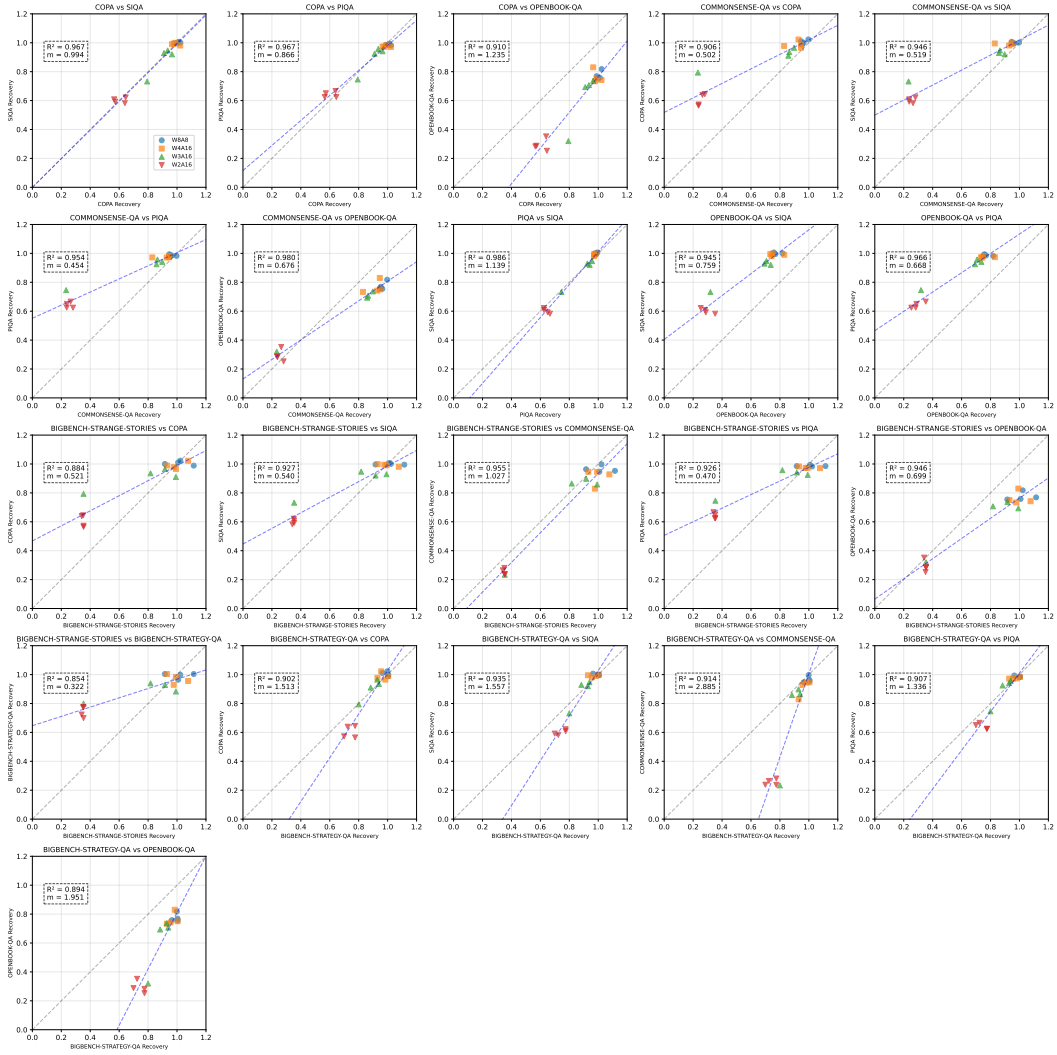

Figure 7: Accuracy recovery in commonsense reasoning tasks.

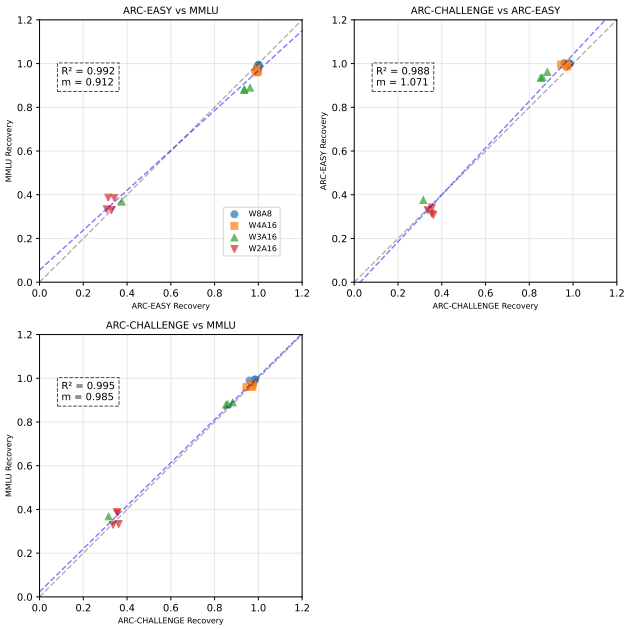

Figure 8: Accuracy recovery in world knowledge tasks.

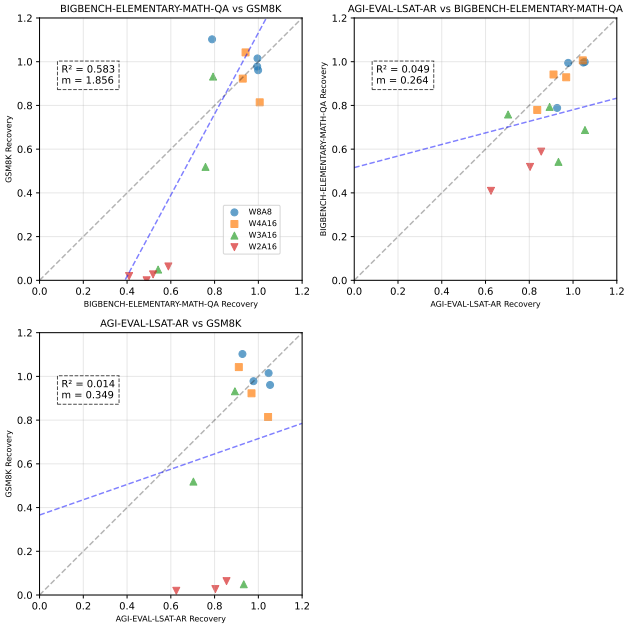

Figure 9: Accuracy recovery in symbolic problem solving tasks.

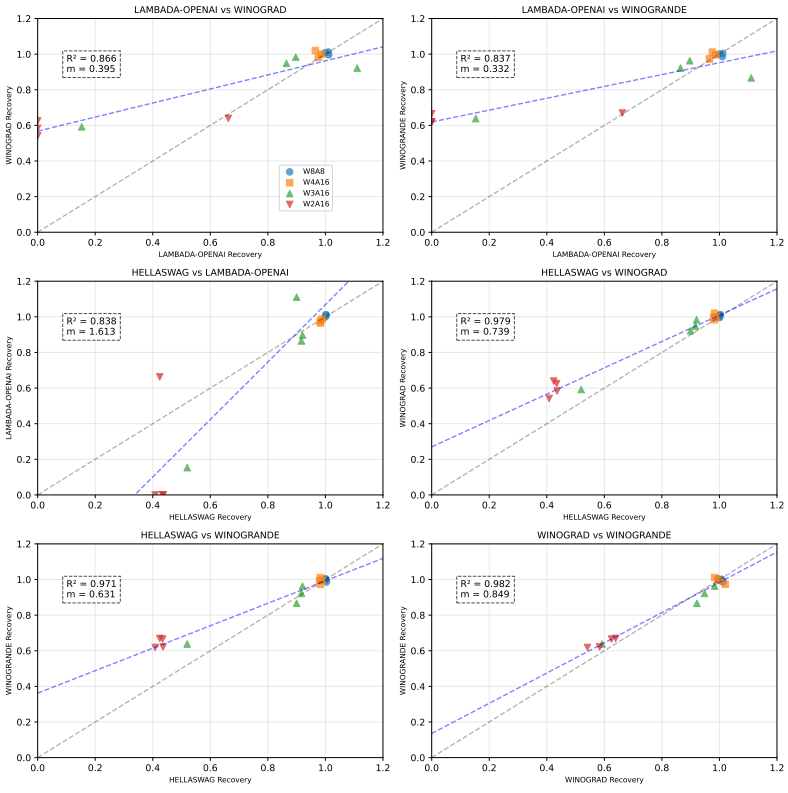

Figure 10: Accuracy recovery in language understanding tasks.

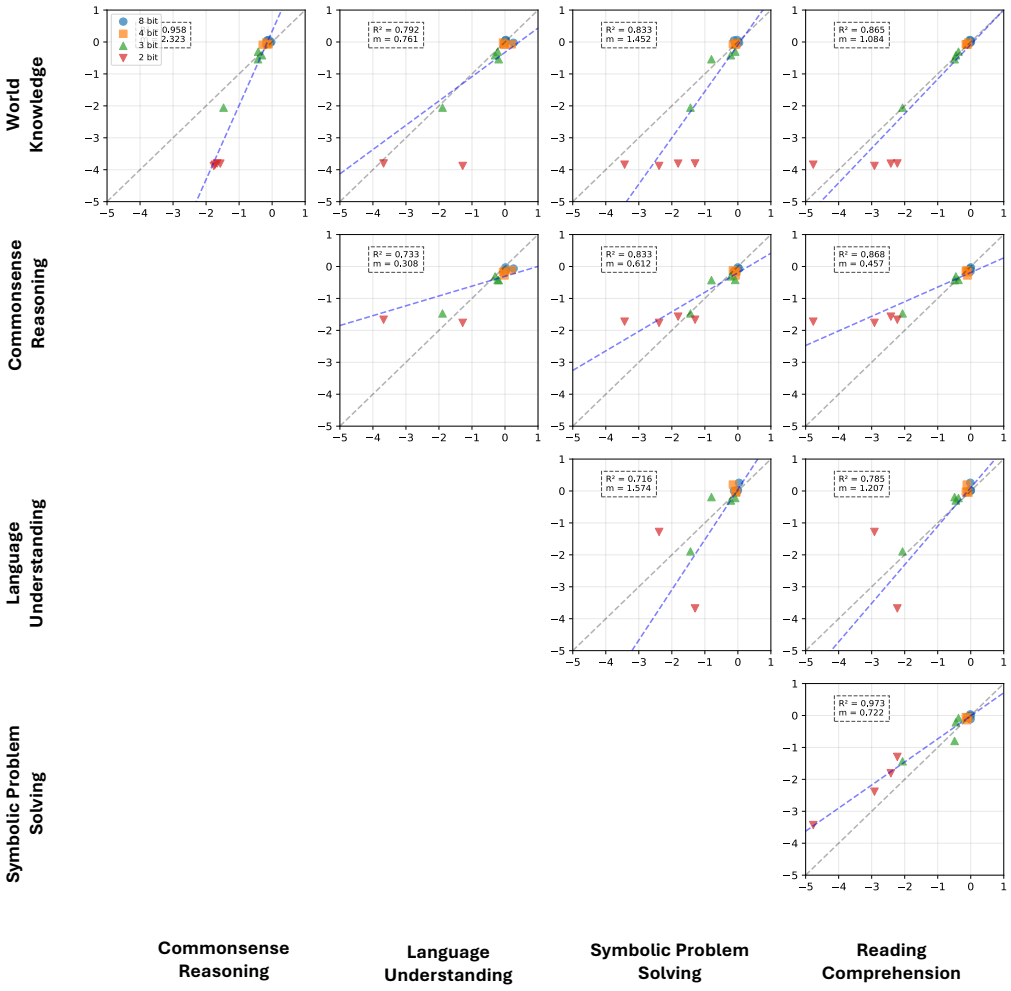

Figure 11: Logit-scaling recovery metrics preserves most linear trends.

A.5 Logit-scaling preserves linear trends in recovery

The original accuracy-on-the-line works applied *probit scaling* to the raw accuracies before fitting a best fit line, as accuracies range in $[0, 1]$. Probit scaling rescales the raw accuracies using the inverse Gaussian CDF, i.e. $\text{Acc}_{\text{rescaled}—} = \Phi^{-1}(\text{Acc}_{\text{original}})$, which stabilizes the linear fits. Probit scaling is not applicable to our setting, as recovery can exceed 1 if the quantized model is more accurate than the base model. Instead, we apply *logit scaling* to our recoveries, where we defined

$$\text{Acc}_{\text{logit-scaled}} = \text{logit}(\text{Acc}_{\text{quantized}}) - \text{logit}(\text{Acc}_{\text{base}})$$

This does not substantially improve the linear fits, and we leave it as future work to find the best rescaling in which the linear trends stabilize.