# In Good GRACEs: Principled Teacher Selection for Knowledge Distillation


## Abstract

Knowledge distillation is an efficient strategy to use data generated by large teacher language models to train smaller "capable" student models, but selecting the optimal teacher for a specific student-task combination requires expensive trial-and-error. We propose a lightweight score called GRACE to quantify how effective a teacher will be when post-training a student model to solve math problems. GRACE efficiently measures distributional properties of student gradients, and it can be computed without access to a verifier, teacher logits, teacher internals, or test data. From an information-theoretic perspective, GRACE measures leave-one-out stability in gradient-based algorithms, directly connecting it to the generalization performance of distilled student models. On GSM8K and MATH, GRACE correlates strongly (up to $86\%$) with the performance of the distilled Llama and OLMo students. In particular, training on GRACE-selected teacher provides at least a $6\%$ improvement over naively using the best-performing teacher. We further demonstrate the utility of GRACE in providing guidance on crucial design choices in distillation, including (1) the best temperature to use when generating from the teacher, (2) the best teacher to use given a size constraint, and (3) the best teacher to use within a specific model family. Altogether, our findings demonstrate that GRACE can efficiently and effectively identify the most compatible teacher for a given student and provide fine-grained guidance on how to perform distillation.


## 1 Introduction

Distillation is a powerful way to train small models from large teachers. We study the case of training autoregressive language models on teacher-generated text. Choosing the right teacher is challenging: stronger models are not always better teachers, as shown in both classic classification/regression settings (Mirzadeh et al., 2019; Harutyunyan et al., 2023; Panigrahi et al., 2025) and recent language models (Zhang et al., 2023b,a; Peng et al., 2024; Razin et al., 2025). With many potential teachers, the current guess-and-check approach is costly, requiring collecting generations from teacher and retraining. Further, hyperparameters in both stages strongly influence outcomes, demanding repeated, careful experimentation. As such, the current work seeks to address the following question:

*Given a pool of candidates, can we efficiently identify the best teacher for a given student and task?*

We propose a score "GRACE" (GRAdient Cross-validation Evaluation) that measures the distributional properties of the student's gradients to identify the most compatible teacher efficiently and effectively (Section 2.1). Motivated by prior data selection and distillation works, GRACE unifies data diversity and student-teacher alignment desiderata into a single score that is efficient to compute and does not require access to an external verifier, teacher logits, teacher representations, or test data. Computing GRACE requires relatively few samples from each teacher, because it uses a cross-validation structure. Extensive experiments on math-related datasets show:

- GRACE correlates strongly with the student's distillation performance (Figure 1), outperforming baselines such as G-Vendi (Jung et al., 2025).

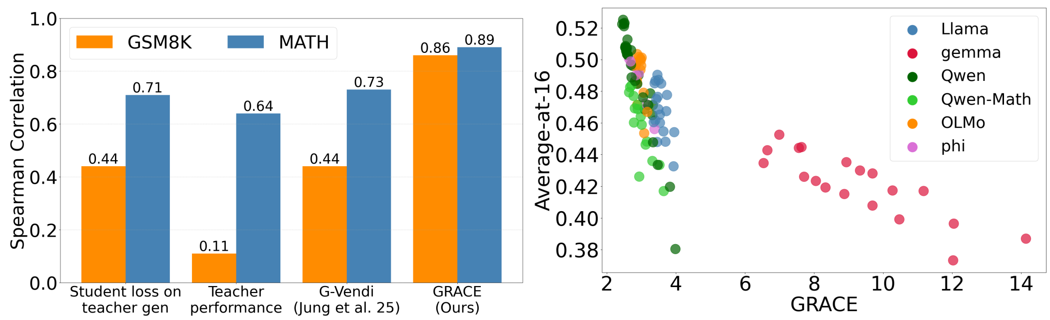



Figure 1: **GRACE correlates most strongly with student performance after distillation on math-related reasoning tasks.** Results in this figure are for a LLaMA-1B-Base student on GSM8K and MATH using 15 teachers of different sizes across the LLaMA, Gemma, Qwen, OLMo, and phi families. (Left) We compare the Spearman correlations between final student performance and four candidate scores: the student's loss on teacher generations, the teacher's performance on the task, G-Vendi (Jung et al., 2025), and our score GRACE. (Right) We plot how our score GRACE compares to the final student performance on GSM8K, measured by the average accuracy of 16 response attempts on each prompt in the test set.

- Selecting teachers using GRACE yields more than $6\%$ improvement in student accuracy compared to using the best-performing teacher, on both GSM8K and MATH. Moreover, students trained on teachers selected by GRACE reach within $1\%$ of the absolute best outcome.
- GRACE offers actionable insights to practitioners. It helps identify 1) the optimal generation temperature for a given teacher model, 2) the best model up to a certain size across model families, and 3) the best size within a model family.

These results indicate that GRACE reliably identifies the most suitable teacher for a given student and offers precise guidance for effective distillation on math-related tasks.

## 2 GRACE: Gradient Cross-Validation Evaluation

We study distillation for fine-tuning a pre-trained student model on downstream tasks. For each of the $N$ prompts $\mathbf{x} \in \mathcal{X}$, we sample $M$ responses $y_1, \ldots, y_M$ from a teacher distribution $\pi_{\mathcal{T}}$ (with sampling temperature). The student is then fine-tuned with the standard autoregressive cross-entropy loss $\mathcal{L}$ on the resulting dataset $\mathcal{D}_{\mathcal{T}}^{\text{distill}}$ of $N \times M$ teacher generations. Student and teacher performance is measured by average accuracy over $k$ sampled responses (average-at-$k$). We denote the student as $\pi_{\mathcal{S}}$, with parameters $\Theta\mathcal{S} \in \mathbb{R}^D$.

**Gradient-Based Scores.** Selecting a teacher for distillation parallels data selection: choosing a teacher's generations is equivalent to selecting a subset from the union of all teachers, with the constraint that each subset comes from one teacher. While data selection methods for language models often use first- or second-order gradient information to score individual datapoints, our goal is to evaluate entire data distributions (teachers). Thus, we adapt gradient-based measures to capture distributional quality. For a teacher $\pi_{\mathcal{T}}$, we assume access only to a subsampled evaluation set $\mathcal{D}_{\mathcal{T}}^{\text{eval}} \subset \mathcal{D}_{\mathcal{T}}^{\text{distill}}$ of size $n \times m$, which may be much smaller than the full $N \times M$ (in our experiments, $60\times$ smaller; see Section 3).

We establish some useful notations first. Let $\boldsymbol{g}(\mathbf{x}, \mathbf{y}) := \nabla \mathcal{L}(\mathbf{y}|\mathbf{x}; \Theta_{\mathcal{S}})$ be the student's gradient on the response $\mathbf{y}$ conditioned on prompt $\mathbf{x}$. Since all gradients are computed with respect to the student model's parameters, we omit the explicit dependency on $\Theta_{\mathcal{S}}$ for notational clarity. For efficiency, we use a random projection $\Pi \in \{\pm 1/\sqrt{D}\}^{d \times D}$ (Park et al., 2023) and rescale by $\log |\mathbf{y}|$ to correct the $\sim 1/\log T$ decay of gradient norms with sequence length (Xia et al., 2024). The processed gradient is $\boldsymbol{h}(\mathbf{x}, \mathbf{y}) := \log(|\mathbf{y}|) \cdot \Pi \boldsymbol{g}(\mathbf{x}, \mathbf{y})$.

For a dataset $\mathcal{D}$ of generations, we also define the matrix consisting of processed gradients ($\boldsymbol{h}$) as $\boldsymbol{G}(\mathcal{D}) \in \mathbb{R}^{nm \times d}$ and processed and *normalized* gradients ($\tilde{\boldsymbol{h}} = \boldsymbol{h}/\|\boldsymbol{h}\|$) as $\tilde{\boldsymbol{G}}(\mathcal{D}) \in \mathbb{R}^{nm \times d}$. We then define $\tilde{\boldsymbol{\Sigma}}(\mathcal{D}) := \frac{1}{nm} \tilde{\boldsymbol{G}}(\mathcal{D})^\top \tilde{\boldsymbol{G}}(\mathcal{D})$, and $\mu(\mathcal{D}) := \frac{1}{nm} \boldsymbol{G}(\mathcal{D})^\top \mathbf{1}$.

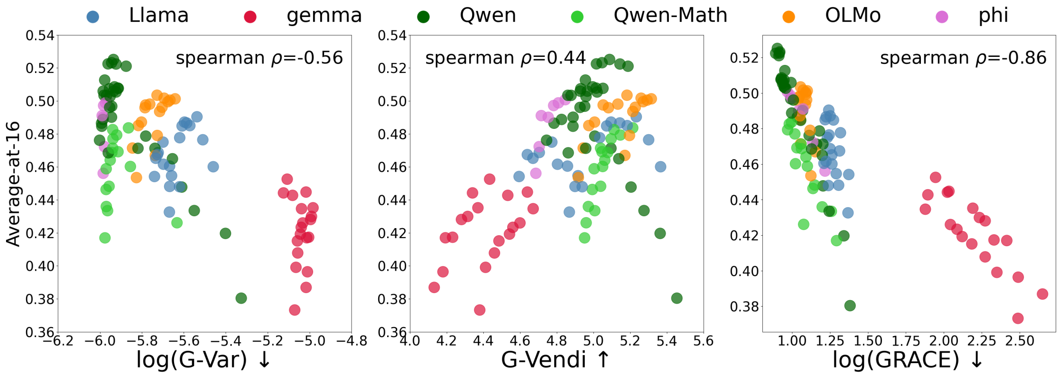



Figure 2: **GRACE achieves** $86\%$ **Spearman correlation to Llama-1B's post-distillation performance on GSM8K**, much higher than G-Var (55%) and G-Vendi (44%).

**Directional coverage and variance.** We consider two key distributional properties for optimization. The first is a notion of diversity, or *directional coverage* of the gradients, which measures how spread out the spectrum of $\tilde{\mathbf{\Sigma}}$ is. For example, the G-Vendi score proposed in (Jung et al., 2025) computes the effective rank of the $\tilde{\mathbf{\Sigma}}$, where a larger score is better. However, we find that G-Vendi is not sufficient for selecting a teacher. For instance, self-distillation leads to a high G-Vendi score but poor performance. This motivates us to consider additional properties orthogonal to the notion of coverage. We find *gradient variance* to be importance, which we denote by G-Var; a smaller G-Var is better. Empirically, we find that G-Var provides different distinctions than G-Vendi, but it is insufficient itself. This motivates us to combine the two into a unified score. We provide more details for G-Vendi and G-Var in Appendix D, and introduce the unified score next.

### 2.1 The GRACE Score

GRAdient Cross-validation Evaluation, or GRACE, computes the gradient variance weighted under the spectrum of the normalized gradient Gram matrix.

**GRACE.** For a dataset $\mathcal{D}$ of teacher generations containing $n \times m$ prompt-generation pairs, and a choice of hyperparameter $C$, construct $C$ partitions of the prompts in the dataset $\mathcal{D}$, denoted $\{\mathcal{D}_i\}_{i=1}^C$, each containing $n/C$ prompts and their generations. Let $\mathcal{D}_{-i}$ denote the concatenation of all partitions except the partition $\mathcal{D}_i$. Then, GRACE is defined as

$$\mathrm{GRACE}(\mathcal{D}) = \frac{1}{nm}\sum_{i=1}^{C} \mathrm{Tr}\left(\boldsymbol{G}_\mu(\mathcal{D}_i)\boldsymbol{M}(\mathcal{D}_{-i})^{-1}\boldsymbol{G}_\mu(\mathcal{D}_i)^\top\right) \tag{1}$$

$$= \frac{1}{nm}\sum_{i=1}^{C}\sum_{(\mathbf{x},\mathbf{y})\in\mathcal{D}_i} \|\boldsymbol{M}(\mathcal{D}_i)^{-1/2}(\boldsymbol{h}(\mathbf{x},\mathbf{y}) - \mu(\mathcal{D}))\|^2, \tag{2}$$

where $\boldsymbol{M}(\mathcal{D}_{-i}) = \tilde{\Sigma}(\mathcal{D}_{-i}) + \frac{\nu}{d}I$ with smoothing parameter $\nu > 0$ for numerical stability.

A smaller GRACE score indicates a better distillation teacher. GRACE combines the spectral information of G-Vendi with the variance computation in G-Var. In particular, we can interpret GRACE as spectral-weighted gradient variance: for a random partition $(\mathcal{D}_1, \mathcal{D}_2)$, if $\{\lambda_j, \boldsymbol{u}_j\}_{j\in[d]}$ denote the set of eigenvalues and eigenvectors for $\tilde{\Sigma}(\mathcal{D}_2)$, then GRACE computes the following for the given partition, i.e. $\sum_{j\in[d]}\frac{1}{\sigma_j+\frac{\nu}{d}}\left(\frac{1}{|\mathcal{D}_1|}\sum_{(\mathbf{x},\mathbf{y})\in\mathcal{D}_1}(\boldsymbol{h}(\mathbf{x},\mathbf{y})^\top\boldsymbol{u}_j)^2\right)$. A small GRACE score requires the gradients to have a small variance along all eigenvectors of $\tilde{\Sigma}$, and it penalizes the variances in directions where the eigenvalue is small more heavily. Variance along such high-signal directions is more harmful, because even small amounts of noise can induce instability or poor generalization. We consider the spectrum of the normalized gradients, since direction of the gradients is more relevant than scale with the use of adaptive optimizers and normalization layers (Loshchilov & Hutter, 2017; Ba et al., 2016; Li et al., 2022). Please refer to Appendix E, where we show connection

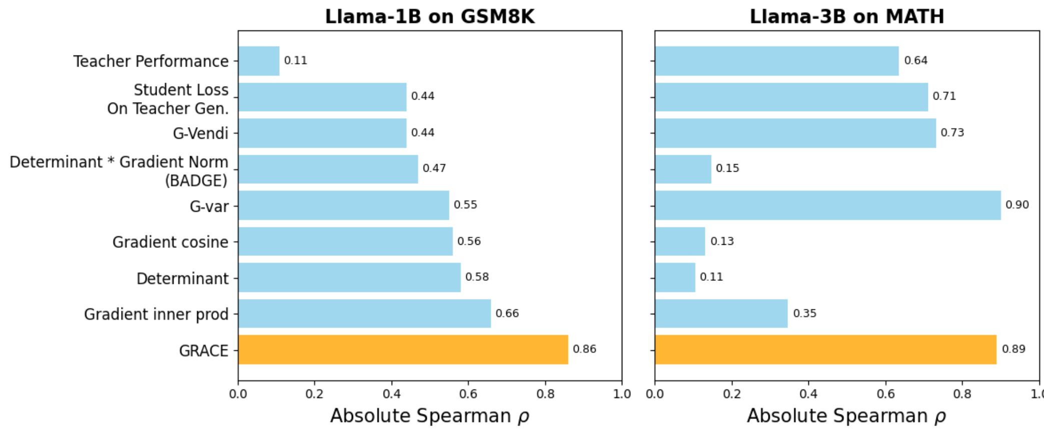



Figure 3: **GRACE achieves $> 80\%$ correlation with the performance of Llama-1B on GSM8K and Llama-3B on MATH.** Teacher performance and the pre-trained student's loss on teacher generations show only weak correlations.

to leave-one-out conditional mutual information, a commonly used tool to study generalization (Xu & Raginsky, 2017; Steinke & Zakynthinou, 2020; Rammal et al., 2022).

# 3 Experiments

**Experiment details:** We compare G-Var, G-Vendi, and GRACE on two math reasoning benchmarks, GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021), which offer many strong teacher models. Performance is measured as average-at-$k$: the mean binary correctness across $k$ sampled responses per prompt.

Student model is taken to be Llama-1B-base or OLMo-1B-base on GSM8K (Cobbe et al., 2021), and Llama-3B-base on MATH (Hendrycks et al., 2021). We compare 15 teachers: Llama-(3.2/3.3) 3/8/70B Instruct models, Qwen-2.5 1.5/3/7/14B Instruct models, Qwen-2.5 Math 1.5/7B Instruct models, Gemma-2 2/9/27B Instruct models, OLMo 7/13B Instruct models, and Phi-4 on both MATH and GSM8K (Dubey et al., 2024; Abdin et al., 2024; Yang et al., 2024; Qwen et al., 2025; Team, 2024). The teacher's generation temperature is varied from $0.3$ to $1.0$ at $0.1$ intervals. Experiment details are in sec.B and details of our additional baselines in sec.G.1.

**Experiment results:** Figure 2 shows that for a Llama-1B model trained on GSM8K, GRACE achieves the best Spearman correlation with the student performance on (0.86) when compared against G-Var (0.55) and G-Vendi (0.44). Additional experiments with an OLMo-1B model trained on GSM8K (Figure 4) and with a Llama-3B model trained on MATH (Figure 8) verify the utility of GRACE. In addition to G-Vendi and G-Var, we also compare against other data selection baselines (Figure 3); a full list is provided in Appendix G.1. Among all scores, GRACE is the only one to achieve consistently high correlation ($> 85\%$) with student performance on both GSM8K and MATH.

In contrast, two intuitive baselines fail to reflect the student's distillation performance. The first is the teacher's own performance, measured in terms of its Average-at-16 performance, which only shows a weak correlation of $11\%$ for Llama-1B on GSM8K, in agreement with findings in prior work (Mirzadeh et al., 2019; Harutyunyan et al., 2023; Panigrahi et al., 2025; Zhang et al., 2023b,a; Peng et al., 2024; Razin et al., 2025). As an example, Llama-70B Instruct has the best performance among all teachers, yet a student trained with Llama-70B Instruct reaches only $46\%$ Average-at-k performance. This is a $6.5\%$ gap to the best performing student which has $52.5\%$ accuracy. Similarly, the student's loss on teacher's generations, measured on the base student, is also poorly correlated with the student's post-distillation performance ($44\%$ with Llama-1B training on GSM8K).

**Discussion and Conclusion:** We give additional experiments in Appendix C. Primarily, we show that GRACE can be used to predict (1) the best temperature to use when generating from the teacher, (2) the best teacher to use given a size constraint, and (3) the best teacher to use within a specific model family. Thus, GRACE provides a reliable metric to select the best teacher to distil to a student on math-related tasks.

## References


Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. Phi-4 technical report. *arXiv preprint arXiv: 2412.08905*, 2024.

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, Colin Raffel, Shiyu Chang, Tatsunori Hashimoto, and William Yang Wang. A survey on data selection for language models. *Trans. Mach. Learn. Res.*, 2024, 2024. URL https://openreview.net/forum?id=XfHWcNTSHp.

J. Ash, Chicheng Zhang, A. Krishnamurthy, J. Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. *International Conference on Learning Representations*, 2019.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Dan Busbridge, Amitis Shidani, Floris Weers, Jason Ramapuram, Etai Littwin, and Russ Webb. Distillation scaling laws. *arXiv preprint arXiv: 2502.08606*, 2025.

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. Training verifiers to solve math word problems. *arXiv preprint arXiv: 2110.14168*, 2021.

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 e-prints*, pp. arXiv–2407, 2024.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

Ronen Eldan and Yuanzhi Li. Tinystories: How small can language models be and still speak coherent english? *arXiv preprint arXiv: 2305.07759*, 2023.

Yu Feng and Yuhai Tu. The inverse variance–flatness relation in stochastic gradient descent is critical for finding flat minima. *Proceedings of the National Academy of Sciences*, 118(9):e2015617118, 2021.

Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, Ashima Suvarna, Benjamin Feuer, Liangyu Chen, Zaid Khan, Eric Frankel, Sachin Grover, Caroline Choi, Niklas Muennighoff, Shiye Su, Wanjia Zhao, John Yang, Shreyas Pimpalgaonkar, Kartik Sharma, Charlie Cheng-Jie Ji, Yichuan Deng, Sarah Pratt, Vivek Ramanujan, Jon Saad-Falcon, Jeffrey Li, Achal Dave, Alon Albalak, Kushal Arora, Blake Wulfe, Chinmay Hegde, Greg Durrett, Sewoong Oh, Mohit Bansal, Saadia Gabriel, Aditya Grover, Kai-Wei Chang, Vaishaal Shankar, Aaron Gokaslan, Mike A. Merrill, Tatsunori Hashimoto, Yejin Choi, Jenia Jitsev, Reinhard Heckel, Maheswaran Sathiamoorthy, Alexandros G. Dimakis, and Ludwig Schmidt. Openthoughts: Data recipes for reasoning models. *arXiv preprint arXiv: 2506.04178*, 2025.

Hrayr Harutyunyan, Ankit Singh Rawat, Aditya Krishna Menon, Seungyeon Kim, and Sanjiv Kumar. Supervision complexity and its role in knowledge distillation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=8jU7wy7N7mA.

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

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv: 1503.02531*, 2015.

Audrey Huang, Adam Block, Qinghua Liu, Nan Jiang, Akshay Krishnamurthy, and Dylan J. Foster. Is best-of-n the best of them? coverage, scaling, and optimality in inference-time alignment. *arXiv preprint arXiv: 2503.21878*, 2025.

Aref Jafari, Mehdi Rezagholizadeh, Pranav Sharma, and Ali Ghodsi. Annealing knowledge distillation. *arXiv preprint arXiv: 2104.07163*, 2021.

Jaehun Jung, Seungju Han, Ximing Lu, Skyler Hallinan, David Acuna, Shrimai Prabhumoye, Mostafa Patwary, Mohammad Shoeybi, Bryan Catanzaro, and Yejin Choi. Prismatic synthesis: Gradient-based data diversification boosts generalization in llm reasoning. *arXiv preprint arXiv:2505.20161*, 2025.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

Krishnateja Killamsetty, Durga Sivasubramanian, Ganesh Ramakrishnan, and Rishabh Iyer. Glister: Generalization based data subset selection for efficient and robust learning. *arXiv preprint arXiv: 2012.10630*, 2020.

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv: 2309.05463*, 2023.

Zhiyuan Li, Srinadh Bhojanapalli, Manzil Zaheer, Sashank Reddi, and Sanjiv Kumar. Robust training of neural networks using scale invariant architectures. In *International Conference on Machine Learning*, pp. 12656–12684. PMLR, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and H. Ghasemzadeh. Improved knowledge distillation via teacher assistant. *AAAI Conference on Artificial Intelligence*, 2019. doi: 10.1609/AAAI.V34I04.5963.

Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. *arXiv preprint arXiv: 1906.01827*, 2019.

Abhishek Panigrahi, Bing Liu, Sadhika Malladi, Andrej Risteski, and Surbhi Goel. Progressive distillation induces an implicit curriculum. *International Conference on Learning Representations*, 2025.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. Trak: Attributing model behavior at scale. *arXiv preprint arXiv:2303.14186*, 2023.

Hao Peng, Xin Lv, Yushi Bai, Zijun Yao, Jiajie Zhang, Lei Hou, and Juanzi Li. Pre-training distillation for large language models: A design space exploration. *arXiv preprint arXiv: 2410.16215*, 2024.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. *Advances in Neural Information Processing Systems*, 33: 19920–19930, 2020.

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, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Mohamad Rida Rammal, Alessandro Achille, Aditya Golatkar, Suhas Diggavi, and Stefano Soatto. On leave-one-out conditional mutual information for generalization. *arXiv preprint arXiv: 2207.00581*, 2022.

Noam Razin, Zixuan Wang, Hubert Strauss, Stanley Wei, Jason D. Lee, and Sanjeev Arora. What makes a reward model a good teacher? an optimization perspective. *arXiv preprint arXiv: 2503.15477*, 2025.

Dhruv Rohatgi, Adam Block, Audrey Huang, Akshay Krishnamurthy, and Dylan J. Foster. Computational-statistical tradeoffs at the next-token prediction barrier: Autoregressive and imitation learning under misspecification. *arXiv preprint arXiv: 2502.12465*, 2025.

Yash Savani, Asher Trockman, Zhili Feng, Avi Schwarzschild, Alexander Robey, Marc Finzi, and J. Zico Kolter. Antidistillation sampling. *arXiv preprint arXiv: 2504.13146*, 2025.

Yuda Song, Gokul Swamy, Aarti Singh, J. Andrew Bagnell, and Wen Sun. The importance of online data: Understanding preference fine-tuning via coverage. *arXiv preprint arXiv: 2406.01462*, 2024.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/7b75da9b61eda40fa35453ee5d077df6-Abstract-Conference.html.

Thomas Steinke and Lydia Zakynthinou. Reasoning about generalization via conditional mutual information. *arXiv preprint arXiv: 2001.09122*, 2020.

Gemma Team. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118.

Chong Wang, Xi Chen, Alexander J Smola, and Eric P Xing. Variance reduction for stochastic gradient optimization. *Advances in neural information processing systems*, 26, 2013.

Hao Wang, Yizhe Huang, Rui Gao, and Flavio Calmon. Analyzing the generalization capability of sgld using properties of gaussian channels. *Advances in Neural Information Processing Systems*, 34:24222–24234, 2021.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: selecting influential data for targeted instruction tuning. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=PG5fV50maR.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. *Advances in neural information processing systems*, 30, 2017.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Bill Yuchen Lin, and Radha Poovendran. Stronger models are not stronger teachers for instruction tuning. *arXiv preprint arXiv: 2411.07133*, 2024.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024.

Chen Zhang, Dawei Song, Zheyu Ye, and Yan Gao. Towards the law of capacity gap in distilling language models. *arXiv preprint arXiv: 2311.07052*, 2023a.

Chen Zhang, Yang Yang, Jiahao Liu, Jingang Wang, Yunsen Xian, Benyou Wang, and Dawei Song. Lifting the curse of capacity gap in distilling language models. *Annual Meeting of the Association for Computational Linguistics*, 2023b. doi: 10.48550/arXiv.2305.12129.

Kaixiang Zheng and En-Hui Yang. Knowledge distillation based on transformed teacher matching. *arXiv preprint arXiv: 2402.11148*, 2024.

# A Related work

**Knowledge distillation** Knowledge distillation is a classic method used to improve the optimization and generalization of a small model (Hinton et al., 2015). A counterintuitive finding is that a better-performing model is not necessarily a better teacher, which has been observed in both classic classification or regression settings (Mirzadeh et al., 2019; Jafari et al., 2021; Harutyunyan et al., 2023) and more recently in language models (Zhang et al., 2023a,b; Xu et al., 2024; Panigrahi et al., 2025). For language models, one can distill from either the logits of the teacher or the generated texts. [1] While the former can lead to better student performance, it is more computationally costly, requires higher access, and is less flexible due to tokenizer choices. We hence focus on distilling from generated texts (Eldan & Li, 2023; Li et al., 2023; Busbridge et al., 2025). Recent work by Guha et al. (2025) supports our findings: they demonstrate that a weaker teacher can yield a stronger distilled model, that distillation benefits from increased sample size, and that filtering has little impact on the resulting student's performance.

**Data selection** For text-based distillation, selecting the best teacher can be considered as the problem of choosing the most useful subset of samples from the generations of all teachers. This aligns with the broad task of *data selection*, which aims to identify subsets of data that maximize certain utility (Sorscher et al., 2022; Albalak et al., 2024). Many approaches leverage gradient information (Mirzasoleiman et al., 2019; Killamsetty et al., 2020; Pruthi et al., 2020; Xia et al., 2024), including some that directly rely on notions of coverage (Ash et al., 2019; Jung et al., 2025). Directional coverage also ties to the notion of coverage in reinforcement learning. Specifically, autoregressive training on teacher generations can be viewed as a form of behavior cloning, for which increasing the coverage is provably beneficial (Song et al., 2024; Huang et al., 2025; Rohatgi et al., 2025). Despite these similarities, distillation differs from standard data selection in that it allows generating new data and offers a richer design space (Peng et al., 2024). An effective teacher-selection score should therefore be versatile and broadly applicable across scenarios, a property that GRACE demonstrates as shown in Appendix C.1.

# B Additional experiment details

To compute our scores, we use a subset of $n = 512$ randomly selected training prompts from the training set, with $m = 4$ generations per prompt. For GRACE, we use $C = 10$-way cross validation. The student gradients are randomly projected to dimension $d = n = 512$; we provide ablation results on these hyperparameter choices in Appendix C.2.

Each distillation training run uses learning rate[2] $10^{-5}$ and 4 epochs over the training set. We use the cosine learning rate schedule with $5\%$ warmup, 0 weight decay, and batch size $64$. We generate $M = 16$ responses per prompt from each teacher and fine-tune the student on all generations without filtering for correctness of the final answer.[3] We compare correlations of our metric to average-at-16 performance for the trained student model when responses are generated at temperature $1.0$. [4] We discuss later in Appendix C.2 the results change when we look at other performance metrics. The computation costs for computing GRACE are provided in Appendix G.3.

# C Additional Experiments

**Teacher selection requires balancing directional coverage and variance.** As a case study, we compare different teachers under a fixed generation temperature of $0.6$ (Figure 5). G-Var clearly separates Qwen-Instruct from Llama-Instruct teachers but fails to distinguish between Qwen, Phi-Instruct, and Qwen-Math-Instruct, suggesting that a low gradient variance alone is insufficient to identify the best teacher. On the other hand, although G-Vendi provides better separation among teachers with low G-Var, it also assigns higher scores to sub-optimal teachers, indicating that directional coverage by

[1]We consider generations following standard next-token distributions, as opposed to antidistillation sampling (Savani et al., 2025).

[2]We searched over learning rates $\{5 \times 10^{-5}, 10^{-5}, 5 \times 10^{-6}\}$ and found $10^{-5}$ to be consistently the best.

[3]Surprisingly, our ablations in Appendix H.1 show that our results are not significantly affected if we filter by correctness.

[4]Results for greedy decoding is included in Figure 11 in appendix.

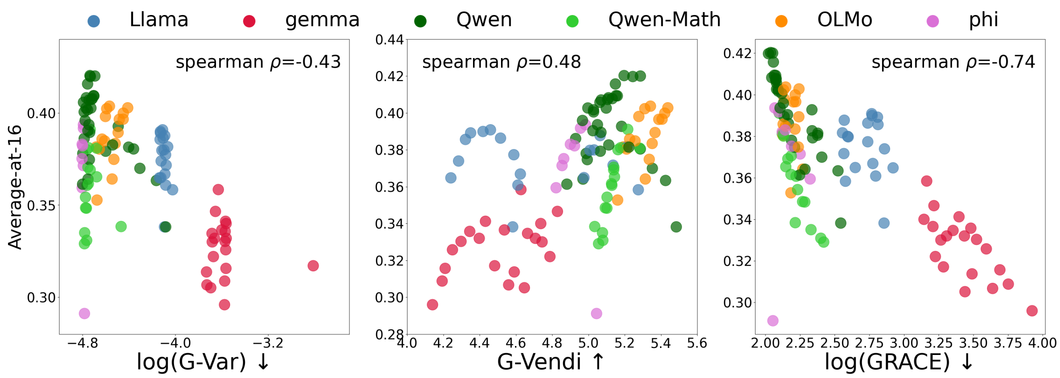



Figure 4: **GRACE achieves** $74\%$ **Spearman correlation to OLMo-1B's post-distillation performance on GSM8K**, significantly outperforming G-Var (43%) and G-Vendi (48%).

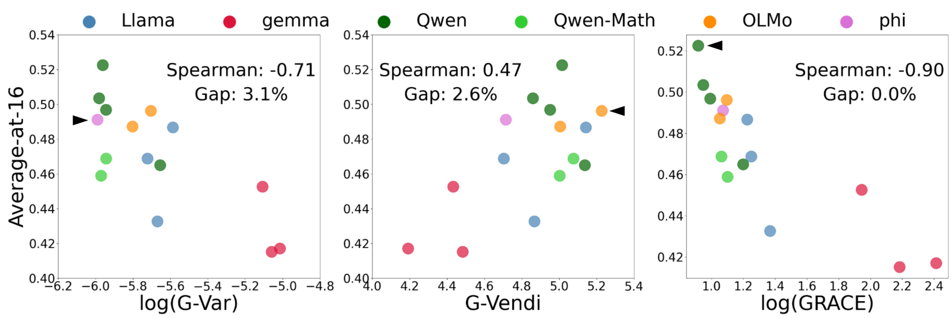



Figure 5: **GRACE can effectively correlate with student performance when compared across different teacher choices.** Here, we report Llama-1B performance on GSM8K across different teacher choices at a generation temperature $0.6$. GRACE achieves 90% correlation with student performance after training, while also predicting Qwen-3B-Instruct to be the optimal teacher. The black triangles mark the best teacher selected by each score. Gap denotes the performance gap between the best performing student and student trained under the best teacher selected by each score.

itself is also inadequate. In contrast, GRACE achieves the strongest correlation ($92\%$) and correctly identifies Qwen-3B-Instruct as the optimal teacher.

### C.1 Guiding distillation practice with GRACE

GRACE can go beyond identifying the best teacher and inform distillation practices. Below we discuss how GRACE provides guidance under common scenarios.

**Selecting generation temperature.** The temperature $\tau$ used to rescale the teacher's logits when generating responses is known to have a strong influence on student performance after distillation (Zheng & Yang, 2024; Peng et al., 2024). However, there hasn't been a principled approach to choose the temperature. We show in Figure 6 that GRACE can identify such a good generation temperature for two Qwen teachers: it closely predicts the optimal generation temperature for Llama-1B training, which are $0.8$ (vs. predicted $0.9$) with the 3B teacher and $0.4$ (vs. predicted $0.5$) with the 1.5B teacher. In comparison, G-Var and G-Vendi tend to increase monotonically with the temperature, even though the student's performance shows an inverse U-shape in temperature. In Figure 7 (left), when averaged across all temperatures, we find that GRACE achieves 75% correlation with the student performance, outperforming the 53% and 59% correlations by G-Var and G-Vendi.

**Selecting a teacher under a size budget**. In practice, one common resource constraint for distillation is the compute required to locally host open-source teachers. Motivated by this, we test whether GRACE can be used to select a teacher under a given size. Specifically, we evaluate three scale constraints: (1) 3B and below, (2) 10B and below, and (3) 30B and below. As shown in Figure 7 (right), GRACE is highly effective, reaching more than $75\%$ correlations and consistently identifying the best teacher under all three size budgets, while the baseline scores are much less reliable. Such

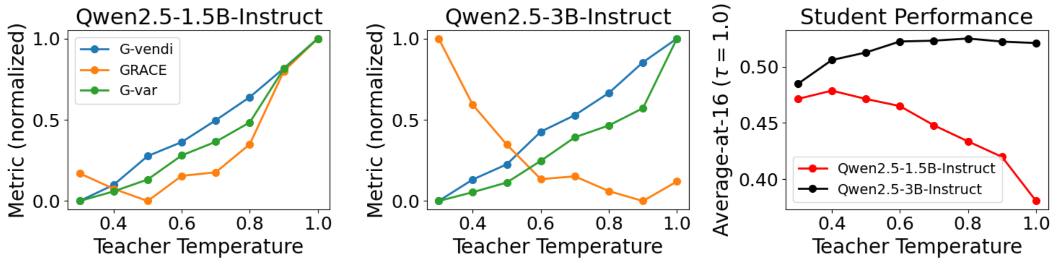



Figure 6: **GRACE can identify a good generation temperature.** Results are shown for Llama-1B trained with Qwen-2.5-1.5B-Instruct and Qwen-2.5-3B-Instruct teachers on GSM8K. GRACE correctly identifies that (1) a lower temperature is optimal for Qwen-2.5-1.5B-Instruct, and (2) a higher temperature is effective for Qwen-2.5-3B-Instruct. In contrast, G-Var can only identify (1) and G-Vendi can only identify (2).

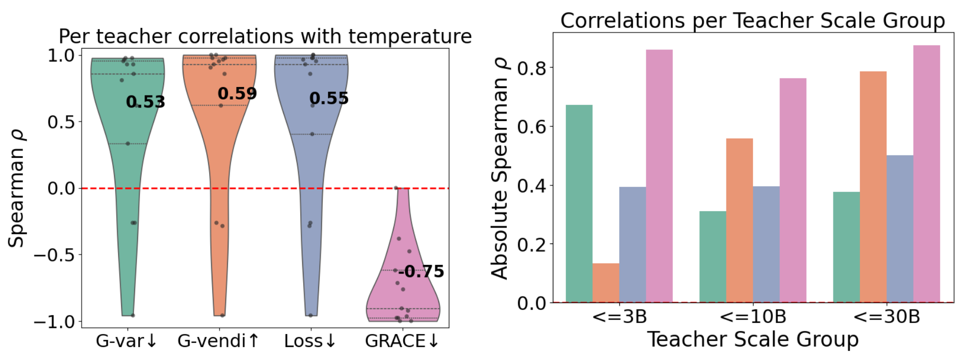



Figure 7: **GRACE is effective at predicting behavior of student performance with teacher generation temperature (left) and the best teacher up to a certain size (right)).** Results are for Llama-1B on GSM8K. (Left) When varying the generation temperature for a fixed teacher, GRACE gets a consistent strong negative correlation (75%). In contrast, all other scores do not show consistent trends across teachers. Violin plots show the distribution over teachers. (Right) GRACE achieves high correlation (75% and above) to performance for teachers under various size constraints.

difference is also reflected by the performance gap between the student trained by the ground truth best teacher, and the student trained by the teacher selected by each score. The gaps for GRACE are under 1% across all groups, indicating that it is often close to selecting the optimal teacher, whereas G-Vendi and G-Var can induce performance gaps of at least 5% for teacher sizes below 10B.

**Selecting teachers within a model family** Another practical limitation is the family of models that one can access, motivating us to test GRACE against models within each model family. We split the teacher models by model family and consider all generation temperatures. Since some families include only a small number of teachers, the Spearman correlations can be unreliable. We hence report the performance gap between learning from the true best teachers and from the teacher selected by a score. As shown in Figure 13, when averaged across all families, GRACE achieves a gap of just 1%, whereas other metrics yield average gaps of at least 3% or more. Moreover, we note that it is not always preferred to choose teacher from the same family as the student. For example, a Llama-1B base student learns better from a Qwen-Instruct teacher than any of Llama-Instruct teachers.

### C.2 Ablations

We test the effect of various hyperparameters used in the GRACE computation. We vary the number of prompts ($n$), the number of generations per prompt ($m$), and the dimension of the gradient random projection ($d$). For the Llama-1B student on GSM8K, we find that GRACE is generally robust to these hyperparameter choices, and the default values ($m = d = 512$, $m = 4$) work well (see details in Appendix H.3). We also vary the number of cross-validation splits used in GRACE. For both GSM8K and MATH, the correlation with student performance remains fairly stable once $C >= 6$ (Figure 20), so we set $C = 10$ for our experiments.

To test the robustness with respect to teacher selection, we evaluate correlations on random subsets of teachers. In addition to the case studies in Appendix C.1, we repeatedly compute scores over random subsets of teachers. As shown in Figure 22 and Figure 23, GRACE consistently maintains high correlations across these subsets (see details in Appendix H.5).

We further examine how correlations change when replacing Average-at-k with other evaluation metrics. For GSM8K, we find that Spearman correlation drops when switching from Average-at-k to either greedy or best-of-k accuracy, even though GRACE still identifies the best teacher model (Figures 11 and 12). Greedy reflects performance from a single generation at temperature $0.0$, and best-of-k measures whether the student answers correctly at least once over $k$ responses at generation temperature $1.0$. A deeper investigation into the discrepancy between Average-at-k and these discrete performance metrics is left to future work.

# D Key properties: G-Vendi and G-Var

**G-Vendi (Jung et al., 2025).** One natural distributional measure of data quality is diversity. Along these lines, Jung et al. (2025) propose the G-Vendi score, which measures the directional coverage of $\mathcal{D}$ as the entropy of the eigenvalues of the gradient Gram matrix.

$$\text{G-Vendi}(\mathcal{D}) := \text{Entropy}(\lambda(\tilde{\mathbf{\Sigma}}(\mathcal{D}))) = -\sum_{\lambda \in \lambda(\tilde{\mathbf{\Sigma}}(\mathcal{D}))} \lambda \log \lambda, \tag{3}$$

where $\lambda(\tilde{\mathbf{\Sigma}}(\mathcal{D}))$ denotes the eigenvalues of the normalized gradient gram matrix with $|\lambda(\tilde{\mathbf{\Sigma}}(\mathcal{D}))| = \min\{nm, d\}$. A larger G-Vendi score is better. Jung et al. (2025) use G-Vendi to select an optimal subset of training data $\mathcal{D}$ from a full dataset generated by a single teacher. However, using G-Vendi to select a teacher out of many candidates may yield suboptimal choices. For example, when performing self-distillation, where the student serves as its own teacher, we find that the G-Vendi score for GSM8K (5.93) is higher than all other teacher models, even though the resulting student's performance is as low as 4%. This observation leads us to investigate another gradient-based distributional score.

**G-Var.** Prior works have shown that reducing gradient variance can boost generalization performance (Wang et al., 2013; Keskar et al., 2016; Wang et al., 2021; Feng & Tu, 2021). As such, we also compute the gradient variance (G-Var) as

$$\text{G-Var}(\mathcal{D}) := \frac{1}{nm} \operatorname{Tr}\left(\boldsymbol{G}_\mu(\mathcal{D}) \boldsymbol{G}_\mu(\mathcal{D})^\top\right) = \frac{1}{nm} \sum_{(\mathbf{x},\mathbf{y}) \in \mathcal{D}} \|\boldsymbol{h}(\mathbf{x},\mathbf{y}) - \mu(\mathcal{D})\|^2, \tag{4}$$

where $\boldsymbol{G}_\mu(\mathcal{D}) = \boldsymbol{G}(\mathcal{D}) - \mathbf{1}\mu(\mathcal{D})^\top$ denotes the centered processed gradient matrix. A smaller G-Var score is considered better. Though G-Var alone is also insufficient. For example, on GSM8K, G-Var's value is largely determined by the model family and not reflecting the student's performance (Figure 2).

G-Var and G-Vendi together capture complementary distributional properties and can sometimes trend in different directions. For instance, we find that increasing the teacher's generation temperature increases G-Var, suggesting that higher temperatures induce worse data, but also increases G-Vendi, indicating higher diversity (Figure 6).

# E Connecting GRACE to leave-one-out CMI

GRACE connects naturally to leave-one-out conditional mutual information (CMI), a frequently used concept in studying generalization (Xu & Raginsky, 2017; Steinke & Zakynthinou, 2020; Rammal et al., 2022). At a high level, CMI captures how much gradient updates are sensitive to removal of a sample and how much of this sensitivity can be tracked to the dropped sample. A higher sensitivity suggests necessary memorization to reduce loss on the training set $\mathcal{D}$, which can lead to low generalization to unseen test examples. Under this framework, we show that GRACE successfully unifies G-Var and G-Vendi.

Formally, we overload $\boldsymbol{g}(\mathcal{D}; \Theta) = \frac{1}{|\mathcal{D}|} \sum_{(\mathbf{x},\mathbf{y}) \in \mathcal{D}} \boldsymbol{g}(\mathbf{x}, \mathbf{y}; \Theta)$ to denote the gradient update on a dataset $\mathcal{D}$. To keep our discussion general, we consider $\boldsymbol{g}(\mathcal{D}; \Theta)$ that uses gradients and a preconditioner matrix $\mathbf{M}$:

$$\boldsymbol{g}(\mathcal{D},\Theta) = \mathbf{M}(\mathcal{D};\Theta)\boldsymbol{g}(\mathcal{D};\Theta) + \epsilon,$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ denotes the gradient noise. Setting $\mathbf{M}$ as identity recovers gradient descent, and setting $\mathbf{M}$ as a function of gradient second moments recovers various adaptive algorithms in practice.

Let $\Theta'_{\mathcal{D}}$ denote the resulting parameters after a gradient update with $\mathcal{D}$, and $\Theta'_{\mathcal{D}\setminus\{(\mathbf{x},\cdot)\}}$ denote the parameters from a set where all training data connected to a uniformly sampled prompt $\mathbf{x}$ are dropped from the training set $\mathcal{D}$, then CMI measures the mutual information between the parameters $\Theta'_{\mathcal{D}\setminus\{(\mathbf{x},\cdot)\}}$ and the dropped prompt $\mathbf{x}$. We show that CMI can be bounded as follows:

**Lemma E.1** (Informal)**.** *Let $C = n$, then for any $\mathcal{D}'$, take $\mathbf{M}(\Theta, \mathcal{D}') := \Sigma(\mathcal{D}')^{-1/2}$, then CMI is bounded by $\frac{1}{\sigma^2 n^2}$ GRACE($\mathcal{D}$).*

**Choice of $\mathbf{M}$ for GRACE:** We defined GRACE based on a particular choice of the pre-conditioner matrix in the definition of CMI. This is motivated by the adaptive optimization algorithms used in practice (Kingma, 2014; Loshchilov & Hutter, 2017; Duchi et al., 2011). In principle, one could obtain sharper predictions by choosing $\mathbf{M}$ optimally. We leave a more thorough exploration of this direction to future work.

(TODO: smooth out the writing here)

### E.1 Informal discussion

**All our discussion assumes that we don't apply a pre-processing function $h$ and we look into the original gradient space in this section.**

Suppose the parameters of the student model are denoted by $\Theta_{\mathcal{S}} \in \mathbb{R}^D$. For theoretical presentation purposes, we collect 1 response per prompt from the teacher on $n$ prompts, forming the training set $\mathcal{D}$. Our theoretical statements can be generalized to the case, where we collect multiple responses for each prompt. We will use $\hat{\mathbb{E}}$ as the empirical mean. Let $U = \mathcal{U}(\mathbf{x} \in \mathcal{D})$ be a random variable that selects a prompt $\hat{\mathbf{x}}$ uniformly at random and removes all prompt–response pairs associated with it. The resulting dataset is

$$\mathcal{D}_U := \mathcal{D} \setminus \{(\hat{\mathbf{x}}, \hat{\mathbf{y}})\}.$$

We then perform a single gradient update with a preconditioner matrix $\mathbf{M}$ that can depend on the training set $\mathcal{D}_U$:

$$\Theta_{ft;U} \;\leftarrow\; \Theta_{\mathcal{S}} - \eta\, \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}_U}\left[\mathbf{M}(\mathcal{D}_U;\Theta_{\mathcal{S}})\nabla\mathcal{L}(\mathbf{y}|\mathbf{x};\Theta_{\mathcal{S}})\right] + \epsilon, \tag{5}$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ denotes Gaussian noise.

We measure the CMI between the updated parameters $\Theta_{ft;U}$ and the random variable $U$, defined as $I(\Theta_{ft;U}; U \mid \mathcal{D})$. This quantifies how much information about the omitted prompt $\hat{\mathbf{x}}$ can be inferred from the updated parameters after training. For simplicity of notation, we define the following notations, following our notation on GRACE:

$$\mu(\mathcal{D} \setminus \{(\hat{\mathbf{x}}, \hat{\mathbf{y}})\}) = \hat{\mathbb{E}}_{\mathcal{D}\setminus\{(\hat{\mathbf{x}},\hat{\mathbf{y}})\}}\nabla\mathcal{L}(\mathbf{y}|\mathbf{x};\Theta_{\mathcal{S}})$$
$$\tilde{\Sigma}(\mathcal{D} \setminus \{(\hat{\mathbf{x}}, \hat{\mathbf{y}})\}) = \frac{1}{(n-1)m}\tilde{\mathbf{G}}^\top\tilde{\mathbf{G}}$$

where $\tilde{\mathbf{G}}$ contains normalized gradients from examples in the set $\mathcal{D} \setminus \{(\hat{\mathbf{x}}, \hat{\mathbf{y}})\}$.

**Lemma E.2** (Informal)**.** *Under the one-step update rule on the parameters $\Theta$ (Equation (5)),*

$$I(\Theta_{ft;U}; U|S) \lesssim \frac{2\eta^2}{\sigma^2 n^2} \quad \hat{\mathbb{E}}_{(\hat{\mathbf{x}},\hat{\mathbf{y}})} \quad \|\mathbf{M}(\mathcal{D} \setminus \{(\hat{\mathbf{x}}, \hat{\mathbf{y}})\}; \Theta_{\mathcal{S}})\, \bar{\mathbf{g}}_{\hat{\mathbf{x}},\hat{\mathbf{y}}}\|_2^2$$

If we use gradient descent and set $\mathbf{M}$ as $\mathbf{I}$, we get G-Var that uses mean shifted gradients. If instead we choose $\mathbf{M}$ as the inverse normalized gradient covariance matrix, i.e. $\mathbf{M}_{\mathcal{D}\setminus\{(\hat{\mathbf{x}},\hat{\mathbf{y}})\}} = \tilde{\Sigma}(\mathcal{D} \setminus \{(\hat{\mathbf{x}}, \hat{\mathbf{y}})\})^{-1/2}$, we recover GRACE.

The lemma indicates that GRACE evaluates the stability of a one-step gradient update when few prompts are removed from the batch. Importantly, the outcome of this update depends on the optimization method, since gradient descent and preconditioned updates can behave differently. In our setting, the preconditioner matrix is closely related to the one used in AdaGrad (Duchi et al., 2011). Since adaptive optimizers are the de facto choice for training language models, it is essential to incorporate this preconditioning effect in our analysis. In principle, one could obtain sharper predictions by choosing $\mathbf{M}$ optimally. This might require a short warm-up training phase of the student model and setting $\mathbf{M}$ as a function of the optimizer states during the warm-up training, akin to Xia et al. (2024). We leave a more thorough exploration of this direction to future work.

**Note on theoretical limitations:** Our current analysis only establishes a connection between GRACE and leave-one-out conditional mutual information. Prior work by Rammal et al. (2022) shows that this quantity upper-bounds the generalization gap in terms of the gap between train and test loss. By contrast, our experiments focus on tracking the student model's test performance using GRACE. Empirically, we find that GRACE serves as a reliable predictor of student performance, even though it fails to correlate with loss-based quantities. This gap highlights the need for a stronger theoretical framework to fully explain the behavior of GRACE, which we leave to future work.

# F Proof of Lemma E.2

We will slightly simplify notations for presentation. We will use

$$\begin{aligned}
\mathbf{M}_{-\{(\hat{\mathbf{x}},\hat{\mathbf{y}})\}} &:= \mathbf{M}\left(\mathcal{D} \setminus \{(\hat{\mathbf{x}},\hat{\mathbf{y}})\}; \Theta_{\mathcal{S}}\right) \\
\mu_{-\{(\hat{\mathbf{x}},\hat{\mathbf{y}})\}} &:= \mu\left(\mathcal{D} \setminus \{(\hat{\mathbf{x}},\hat{\mathbf{y}})\}; \Theta_{\mathcal{S}}\right).
\end{aligned}$$

Then, a more formal version of Lemma E.2 is given as follows:

**Lemma F.1** (Bounds for Pre-conditioned Gradient Descent)**.** *Under the one-step update rule on the parameters $\Theta$ (Equation (5)),*

$$\begin{aligned}
I(\Theta_{ft;U}; U|S) \lesssim & \frac{3\eta^2}{\sigma^2 n^2} \quad \hat{\mathbb{E}}_{(\hat{\mathbf{x}},\hat{\mathbf{y}})} \quad \left\|\mathbf{M}_{-\{(\hat{\mathbf{x}},\hat{\mathbf{y}})\}} \bar{\mathbf{g}}_{\hat{\mathbf{x}},\hat{\mathbf{y}}}\right\|_2^2 \\
& + \frac{3\eta^2}{\sigma^2} \hat{\mathbb{E}}_{(\hat{\mathbf{x}},\hat{\mathbf{y}})} \left\| \left(\mathbf{M}_{-\{(\hat{\mathbf{x}},\hat{\mathbf{y}})\}} - \hat{\mathbb{E}}_{(\bar{\mathbf{x}},\bar{\mathbf{y}})} \mathbf{M}_{-\{(\bar{\mathbf{x}},\bar{\mathbf{y}})\}}\right) \mu_{-\{(\hat{\mathbf{x}},\hat{\mathbf{y}})\}} \right\|_2^2 \\
& \textit{where } \bar{\mathbf{g}}_{\hat{\mathbf{x}},\hat{\mathbf{y}}} = \nabla \mathcal{L}(\hat{\mathbf{y}} \mid \hat{\mathbf{x}}; \Theta_{\mathcal{S}}) - \mu_{-\{(\hat{\mathbf{x}},\hat{\mathbf{y}})\}}.
\end{aligned}$$

*Proof.* For any $(\bar{\mathbf{x}},\bar{\mathbf{y}})$ pair, denote the mean parameter update on the training set $\mathcal{D} \setminus (\bar{\mathbf{x}},\bar{\mathbf{y}})$ as $\delta_{-(\bar{\mathbf{x}},\bar{\mathbf{y}})} := \Theta_{\mathcal{S}} - \eta \mathbf{M}_{-\{(\bar{\mathbf{x}},\bar{\mathbf{y}})\}} \mu_{-\{(\bar{\mathbf{x}},\bar{\mathbf{y}})\}}$.

By the definition of CMI,

$$I(\Theta_{ft;U}; U|S) = \hat{\mathbb{E}}_{u\sim U} D_{\mathrm{KL}} \left( p_{\Theta_{ft;u}|\mathcal{D},u} || \hat{\mathbb{E}}_{\bar{u}} p_{\Theta_{ft;\bar{U}}|\mathcal{D},\bar{u}} \right),$$

where $p_{\Theta_{ft;u}|\mathcal{D},u}$ denotes the probability distribution of $\Theta_{ft;u}$ conditioned on dropping prompts from $\mathcal{D}$ according to the random variable $u$. Note that there is a one-to-one correspondence between the variable $u$ and the random prompt $\hat{\mathbf{x}}$ that we drop. Thus, one can write

$$I(\Theta_{ft;U}; U|S) = \hat{\mathbb{E}}_{\hat{\mathbf{x}}} D_{\mathrm{KL}} \left( p_{\Theta_{ft;-\hat{\mathbf{x}}}|\mathcal{D},\hat{\mathbf{x}}} || \hat{\mathbb{E}}_{\bar{\mathbf{x}}} p_{\Theta_{ft;-\bar{\mathbf{x}}}|\mathcal{D},\bar{\mathbf{x}}} \right),$$

where $p_{\Theta_{ft;-\hat{\mathbf{x}}}|\mathcal{D},\hat{\mathbf{x}}}$ denotes the probability distribution of $\Theta_{ft;-\hat{\mathbf{x}}}$ conditioned on dropping prompts from $\hat{\mathbf{x}}$ from the training set.

The update rule for any set $\mathcal{D} \setminus (\bar{\mathbf{x}},\bar{\mathbf{y}})$ is given by

$$\Theta_{ft;-\bar{\mathbf{x}}} \leftarrow \Theta_{\mathcal{S}} - \delta_{-(\bar{\mathbf{x}},\bar{\mathbf{y}})} + \epsilon := \delta_{-(\bar{\mathbf{x}},\bar{\mathbf{y}})} + \epsilon.$$

Because of the gaussian noise $\epsilon$,

$$\Theta_{ft;-\hat{\mathbf{x}}} \sim \mathcal{N}\left(\delta_{-(\bar{\mathbf{x}},\bar{\mathbf{y}})}, \sigma^2 \mathbf{I}\right).$$

Then, using the properties of gaussian distribution;

$$
\begin{aligned}
I(\Theta_{ft;U}; U \mid \mathcal{D}) &= \hat{\mathbb{E}}_{\hat{\mathbf{x}}} D_{\mathrm{KL}}\left(p_{\Theta_{ft;-\hat{\mathbf{x}}}|\mathcal{D},\hat{\mathbf{x}}} \,\middle\|\, \hat{\mathbb{E}}_{\bar{\mathbf{x}}} p_{\Theta_{ft;-\bar{\mathbf{x}}}|\mathcal{D},\bar{\mathbf{x}}}\right) \\
&= \hat{\mathbb{E}}_{\hat{\mathbf{x}}} \hat{\mathbb{E}}_{X \sim \mathcal{N}\left(\delta_{-(\hat{\mathbf{x}},\hat{\mathbf{y}})}, \sigma^2 \mathbf{I}\right)} \left(\log\left(\tfrac{1}{Z} e^{-\left\|X-\delta_{-(\hat{\mathbf{x}},\hat{\mathbf{y}})}\right\|_2^2/2\sigma^2}\right) - \log \hat{\mathbb{E}}_{\bar{\mathbf{x}}}\left(\tfrac{1}{Z} e^{-\left\|X-\delta_{-(\bar{\mathbf{x}},\bar{\mathbf{y}})}\right\|_2^2/2\sigma^2}\right)\right) \\
&\leq \hat{\mathbb{E}}_{\hat{\mathbf{x}}} \hat{\mathbb{E}}_{X \sim \mathcal{N}\left(\delta_{-(\hat{\mathbf{x}},\hat{\mathbf{y}})}, \sigma^2 \mathbf{I}\right)} \left(\log\left(\tfrac{1}{Z} e^{-\left\|X-\delta_{-(\hat{\mathbf{x}},\hat{\mathbf{y}})}\right\|_2^2/2\sigma^2}\right) - \hat{\mathbb{E}}_{\bar{\mathbf{x}}} \log\left(\tfrac{1}{Z} e^{-\left\|X-\delta_{-(\bar{\mathbf{x}},\bar{\mathbf{y}})}\right\|_2^2/2\sigma^2}\right)\right) \\
&= \frac{1}{2\sigma^2} \hat{\mathbb{E}}_{\hat{\mathbf{x}}} \hat{\mathbb{E}}_{X \sim \mathcal{N}\left(\delta_{-(\hat{\mathbf{x}},\hat{\mathbf{y}})}, \sigma^2 \mathbf{I}\right)} \left(-\left\|X - \delta_{-(\hat{\mathbf{x}},\hat{\mathbf{y}})}\right\|_2^2 + \hat{\mathbb{E}}_{\bar{\mathbf{x}}}\left\|X - \delta_{-(\bar{\mathbf{x}},\bar{\mathbf{y}})}\right\|_2^2\right) \\
&= \frac{1}{2\sigma^2} \hat{\mathbb{E}}_{\hat{\mathbf{x}}} \hat{\mathbb{E}}_{\bar{\mathbf{x}}} \hat{\mathbb{E}}_{X \sim \mathcal{N}\left(\delta_{-(\hat{\mathbf{x}},\hat{\mathbf{y}})}, \sigma^2 \mathbf{I}\right)} \left(-\left\|X - \delta_{-(\hat{\mathbf{x}},\hat{\mathbf{y}})}\right\|_2^2 + \left\|X - \delta_{-(\bar{\mathbf{x}},\bar{\mathbf{y}})}\right\|_2^2\right) \\
&= \frac{1}{2\sigma^2} \hat{\mathbb{E}}_{\hat{\mathbf{x}}} \hat{\mathbb{E}}_{\bar{\mathbf{x}}} \left\|\delta_{-(\bar{\mathbf{x}},\bar{\mathbf{y}})} - \delta_{-(\hat{\mathbf{x}},\hat{\mathbf{y}})}\right\|_2^2 \\
&= \frac{1}{\sigma^2} \hat{\mathbb{E}}_{\hat{\mathbf{x}}} \left\|\delta_{-(\hat{\mathbf{x}},\hat{\mathbf{y}})} - \hat{\mathbb{E}}_{\bar{\mathbf{x}}} \delta_{-(\bar{\mathbf{x}},\bar{\mathbf{y}})}\right\|_2^2
\end{aligned}
$$

In the second step, we simply use the CDF formulation of gaussian distribution, where $Z = (2\pi e)^{-D}$. The third step applies a jensen's inequality.

Using the definition of $\delta$, we have

$$
I(\Theta_{ft;U}; U \mid \mathcal{D}) \leq \frac{1}{\sigma^2} \hat{\mathbb{E}}_{\hat{\mathbf{x}}} \left\|\mathbf{M}_{-(\hat{\mathbf{x}},\hat{\mathbf{y}})} \mu_{-\{(\hat{\mathbf{x}},\hat{\mathbf{y}})\}} - \hat{\mathbb{E}}_{\bar{\mathbf{x}}} \mathbf{M}_{-\{(\bar{\mathbf{x}},\bar{\mathbf{y}})\}} \mu_{-\{(\bar{\mathbf{x}},\bar{\mathbf{y}})\}}\right\|_2^2
$$

**Warmup: When the pre-conditioner is identity matrix** Then for any $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ pair, we have $\mathbf{M}_{-\{(\bar{\mathbf{x}},\bar{\mathbf{y}})\}} = \mathbf{I}$. Then, the formulation simplifies to

$$
\begin{aligned}
I(\Theta_{ft;U}; U \mid \mathcal{D}) &\leq \frac{\eta^2}{\sigma^2} \hat{\mathbb{E}}_{\hat{\mathbf{x}}} \left\|\mu_{-\{(\hat{\mathbf{x}},\hat{\mathbf{y}})\}} - \hat{\mathbb{E}}_{\bar{\mathbf{x}}} \mu_{-\{(\bar{\mathbf{x}},\bar{\mathbf{y}})\}}\right\|_2^2 \\
&= \frac{\eta^2}{\sigma^2} \hat{\mathbb{E}}_{\hat{\mathbf{x}}} \left\|\frac{n}{n-1}\mu(\mathcal{D}) - \frac{1}{n-1}\nabla\mathcal{L}(\hat{\mathbf{y}} \mid \hat{\mathbf{x}}; \Theta_{\mathcal{S}}) - \hat{\mathbb{E}}_{\bar{\mathbf{x}}}\left(\frac{n}{n-1}\mu(\mathcal{D}) - \frac{1}{n-1}\nabla\mathcal{L}(\bar{\mathbf{y}} \mid \bar{\mathbf{x}}; \Theta_{\mathcal{S}})\right)\right\|_2^2 \\
&= \frac{\eta^2}{\sigma^2(n-1)^2} \hat{\mathbb{E}}_{\hat{\mathbf{x}}} \left\|\nabla\mathcal{L}(\hat{\mathbf{y}} \mid \hat{\mathbf{x}}; \Theta_{\mathcal{S}}) - \hat{\mathbb{E}}_{\bar{\mathbf{x}}}\nabla\mathcal{L}(\bar{\mathbf{y}} \mid \bar{\mathbf{x}}; \Theta_{\mathcal{S}})\right\|_2^2 \\
&= \frac{\eta^2}{\sigma^2(n-1)^2} \hat{\mathbb{E}}_{\hat{\mathbf{x}}} \left(1 - \frac{1}{n}\right)^2 \left\|\nabla\mathcal{L}(\hat{\mathbf{y}} \mid \hat{\mathbf{x}}; \Theta_{\mathcal{S}}) - \hat{\mathbb{E}}_{\mathcal{D}\setminus\bar{\mathbf{x}}}\nabla\mathcal{L}(\bar{\mathbf{y}} \mid \bar{\mathbf{x}}; \Theta_{\mathcal{S}})\right\|_2^2 \\
&= \frac{\eta^2}{\sigma^2 n^2} \hat{\mathbb{E}}_{\hat{\mathbf{x}}} \left\|\nabla\mathcal{L}(\hat{\mathbf{y}} \mid \hat{\mathbf{x}}; \Theta_{\mathcal{S}}) - \hat{\mathbb{E}}_{\mathcal{D}\setminus\{\bar{\mathbf{x}},\bar{\mathbf{y}}\}}\nabla\mathcal{L}(\bar{\mathbf{y}} \mid \bar{\mathbf{x}}; \Theta_{\mathcal{S}})\right\|_2^2
\end{aligned}
$$

The first step follows from the fact that $\mu(\mathcal{D}) = \hat{\mathbb{E}}_{\hat{\mathbf{x}}\sim\mathcal{D}}\mathcal{L}(\hat{\mathbf{y}} \mid \hat{\mathbf{x}}; \Theta_{\mathcal{S}})$.

**General pre-conditioner $\mathbf{M}$:** We follow similar steps as above:

$$
\begin{aligned}
I(\Theta_{ft;U}; U \mid \mathcal{D}) &\leq \frac{\eta^2}{\sigma^2} \hat{\mathbb{E}}_{\hat{\mathbf{x}}} \left\| \mathbf{M}_{-\{(\hat{\mathbf{x}},\hat{\mathbf{y}})\}} \mu_{-\{(\hat{\mathbf{x}},\hat{\mathbf{y}})\}} - \hat{\mathbb{E}}_{\bar{\mathbf{x}}} \mathbf{M}_{-\{(\bar{\mathbf{x}},\bar{\mathbf{y}})\}} \mu_{-\{(\bar{\mathbf{x}},\bar{\mathbf{y}})\}} \right\|_2^2 \\
&= \frac{\eta^2}{\sigma^2} \hat{\mathbb{E}}_{\hat{\mathbf{x}}} \Big\| \frac{n}{n-1} \mathbf{M}_{-\{(\hat{\mathbf{x}},\hat{\mathbf{y}})\}} \mu(\mathcal{D}) - \frac{1}{n-1} \mathbf{M}_{-\{(\hat{\mathbf{x}},\hat{\mathbf{y}})\}} \nabla \mathcal{L}(\hat{\mathbf{y}} \mid \hat{\mathbf{x}}; \Theta_{\mathcal{S}}) \\
&\qquad - \hat{\mathbb{E}}_{\bar{\mathbf{x}}} \left( \frac{n}{n-1} \mathbf{M}_{-\{(\bar{\mathbf{x}},\bar{\mathbf{y}})\}} \mu(\mathcal{D}) - \frac{1}{n-1} \mathbf{M}_{-\{(\bar{\mathbf{x}},\bar{\mathbf{y}})\}} \nabla \mathcal{L}(\bar{\mathbf{y}} \mid \bar{\mathbf{x}}; \Theta_{\mathcal{S}}) \right) \Big\|^2 \\
&= \frac{\eta^2}{\sigma^2} \hat{\mathbb{E}}_{\hat{\mathbf{x}}} \Big\| \frac{n}{n-1} \left( \mathbf{M}_{-\{(\hat{\mathbf{x}},\hat{\mathbf{y}})\}} - \hat{\mathbb{E}}_{\bar{\mathbf{x}}} \mathbf{M}_{-\{(\bar{\mathbf{x}},\bar{\mathbf{y}})\}} \right) \mu(\mathcal{D}) \\
&\qquad - \frac{1}{n-1} \left( \mathbf{M}_{-\{(\hat{\mathbf{x}},\hat{\mathbf{y}})\}} \nabla \mathcal{L}(\hat{\mathbf{y}} \mid \hat{\mathbf{x}}; \Theta_{\mathcal{S}}) - \hat{\mathbb{E}}_{\bar{\mathbf{x}}} \left( \mathbf{M}_{-\{(\bar{\mathbf{x}},\bar{\mathbf{y}})\}} \nabla \mathcal{L}(\bar{\mathbf{y}} \mid \bar{\mathbf{x}}; \Theta_{\mathcal{S}}) \right) \right) \Big\|^2 \\
&= \frac{\eta^2}{\sigma^2} \hat{\mathbb{E}}_{\hat{\mathbf{x}}} \Big\| \frac{n}{n-1} \left( \mathbf{M}_{-\{(\hat{\mathbf{x}},\hat{\mathbf{y}})\}} - \hat{\mathbb{E}}_{\bar{\mathbf{x}}} \mathbf{M}_{-\{(\bar{\mathbf{x}},\bar{\mathbf{y}})\}} \right) \mu(\mathcal{D}) \\
&\qquad - \frac{1}{n-1} \mathbf{M}_{-\{(\hat{\mathbf{x}},\hat{\mathbf{y}})\}} \left( \nabla \mathcal{L}(\hat{\mathbf{y}} \mid \hat{\mathbf{x}}; \Theta_{\mathcal{S}}) - \hat{\mathbb{E}}_{\bar{\mathbf{x}}} \nabla \mathcal{L}(\bar{\mathbf{y}} \mid \bar{\mathbf{x}}; \Theta_{\mathcal{S}}) \right) \\
&\qquad - \frac{1}{n-1} \hat{\mathbb{E}}_{\bar{\mathbf{x}}} \left( \mathbf{M}_{-\{(\hat{\mathbf{x}},\hat{\mathbf{y}})\}} - \mathbf{M}_{-\{(\bar{\mathbf{x}},\bar{\mathbf{y}})\}} \right) \nabla \mathcal{L}(\bar{\mathbf{y}} \mid \bar{\mathbf{x}}; \Theta_{\mathcal{S}}) \Big\|^2 \\
&\leq \frac{3\eta^2}{\sigma^2} \hat{\mathbb{E}}_{\hat{\mathbf{x}}} \left( \frac{n}{n-1} \right)^2 \hat{\mathbb{E}}_{\hat{\mathbf{x}}} \left\| \left( \mathbf{M}_{-\{(\hat{\mathbf{x}},\hat{\mathbf{y}})\}} - \hat{\mathbb{E}}_{\bar{\mathbf{x}}} \mathbf{M}_{-\{(\bar{\mathbf{x}},\bar{\mathbf{y}})\}} \right) \mu(\mathcal{D}) \right\|_2^2 \\
&\quad + \frac{3\eta^2}{\sigma^2} \frac{1}{(n-1)^2} \hat{\mathbb{E}}_{\hat{\mathbf{x}}} \left\| \hat{\mathbb{E}}_{\bar{\mathbf{x}}} \left( \mathbf{M}_{-\{(\hat{\mathbf{x}},\hat{\mathbf{y}})\}} - \mathbf{M}_{-\{(\bar{\mathbf{x}},\bar{\mathbf{y}})\}} \right) \nabla \mathcal{L}(\bar{\mathbf{y}} \mid \bar{\mathbf{x}}; \Theta_{\mathcal{S}}) \right\|_2^2 \\
&\quad + \frac{3\eta^2}{\sigma^2} \frac{1}{(n-1)^2} \hat{\mathbb{E}}_{\hat{\mathbf{x}}} \left\| \mathbf{M}_{-\{(\hat{\mathbf{x}},\hat{\mathbf{y}})\}} \left( \nabla \mathcal{L}(\hat{\mathbf{y}} \mid \hat{\mathbf{x}}; \Theta_{\mathcal{S}}) - \hat{\mathbb{E}}_{\bar{\mathbf{x}}} \nabla \mathcal{L}(\bar{\mathbf{y}} \mid \bar{\mathbf{x}}; \Theta_{\mathcal{S}}) \right) \right\|_2^2 \\
&\leq \frac{3\eta^2}{\sigma^2} \hat{\mathbb{E}}_{\hat{\mathbf{x}}} \left( \frac{n}{n-1} \right)^2 \hat{\mathbb{E}}_{\hat{\mathbf{x}}} \left\| \left( \mathbf{M}_{-\{(\hat{\mathbf{x}},\hat{\mathbf{y}})\}} - \hat{\mathbb{E}}_{\bar{\mathbf{x}}} \mathbf{M}_{-\{(\bar{\mathbf{x}},\bar{\mathbf{y}})\}} \right) \mu(\mathcal{D}) \right\|_2^2 \\
&\quad + \frac{3\eta^2}{\sigma^2} \frac{1}{(n-1)^2} \hat{\mathbb{E}}_{\hat{\mathbf{x}}} \left\| \mathbf{M}_{-\{(\hat{\mathbf{x}},\hat{\mathbf{y}})\}} \left( \nabla \mathcal{L}(\hat{\mathbf{y}} \mid \hat{\mathbf{x}}; \Theta_{\mathcal{S}}) - \hat{\mathbb{E}}_{\bar{\mathbf{x}}} \nabla \mathcal{L}(\bar{\mathbf{y}} \mid \bar{\mathbf{x}}; \Theta_{\mathcal{S}}) \right) \right\|_2^2 + \mathcal{O}\left(1/n^4\right).
\end{aligned}
$$

Here, we assume that $\mathbf{M}$ is a well conditioned matrix, and so the second term is a small term of order $\frac{1}{n^4}$. This can be ensured by a small smoothing term. The first term looks at the sensitivity of the pre-conditioned matrix $\mathbf{M}$ when a sample is dropped. The second term looks at the change in gradient with a drop in sample.

$\square$

When $\mathbf{M}$ is set as $\tilde{\Sigma}^{-1/2}$, we find there are two terms in the bound above: how much $\tilde{\Sigma}^{-1/2}$ changes with a drop in sample and second, how much the gradients change with respect to the $\tilde{\Sigma}^{-1/2}$ matrix, which is related to the GRACE term. We find that $\tilde{\Sigma}^{-1/2}$ is extremely stable in our experiments, and the first term is $5 - 10x$ smaller compared to the second term. This gives us the rough bound that the CMI is bounded by GRACE.

# G Additional results

Here, we report the performance when we allow more computation for the computation of GRACE. We use highher $d$ than the ones reported in Figures 2 and 4. We use $d = 1024$ and $n = 512$. The correlation improves for both the models on GSM8K; however it hurts on MATH.

## G.1 More baselines

We consider the following baselines:

1. Student Loss on the teacher's generations;

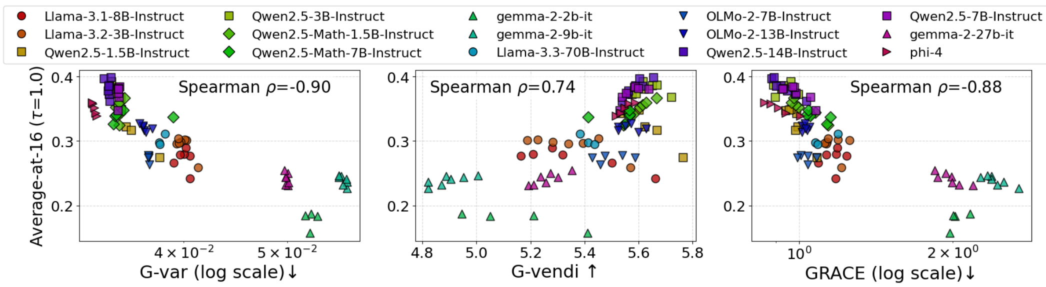



Figure 8: **GRACE achieves** $88\%$ **correlation to Llama-3B performance after training on MATH, across all teacher, generation temperature combinations.** G-Var and G-Vendi can achieve $90\%$ and $74\%$ correlation respectively. Here, $n = 512, d = 512$ are used to compute all metrics.

2. G-Var (Equation (4));
3. G-Vendi (Equation (3));
4. Determinant
5. Determinant $\times$ gradient norm, corresponding to BADGE (Ash et al., 2019), which captures both the diversity and magnitude of gradients;
6. Gradient inner product, which is another way to capture gradient diversity: Given gradients from the training set $\mathcal{D}$, we compute pairwise inner product between the normalized gradients of generations for the same prompt:

$$\mathbb{E}_{\mathbf{x}}\mathbb{E}_{(\mathbf{x},\mathbf{y}_1),(\mathbf{x},\mathbf{y}_2)\sim\mathcal{D}}\left[\frac{\mathbf{g_1}}{\|\mathbf{g}_1\|_2}\right]^\top \frac{\mathbf{g_2}}{\|\mathbf{g}_2\|_2},$$
$$\text{where } \mathbf{g}_1 = \nabla\mathcal{L}_{CE}(\mathbf{x},\mathbf{y}_1;\pi_{\mathcal{S}}),$$
$$\mathbf{g}_2 = \nabla\mathcal{L}_{CE}(\mathbf{x},\mathbf{y}_2;\pi_{\mathcal{S}}).$$

7. Gradient inner product with norm, which is similar to the above but additionally considering gradient magnitude: Here, we compute pairwise inner product between the gradients of generations from the same prompt.
8. Average Probabilities (per token): this computes the average probability per token of the student on the teacher's generations, averaged over all generations and all prompts.
9. Best average probabilities per prompt: we compute the average probability per token for each generation, and take the highest average probability (i.e. the most probable) across all generations of the same prompt. We then take an average across all prompts.
10. Correct average probabilities: Here, we simply compute the average probabilities of tokens in correct generations for each prompt and take the average across all prompts.
11. Incorrect average probabilities: Same as above, but over incorrect generations.
12. Different average probabilities per prompt: For each prompt, we compute the average per-token probabilities for correct and incorrect generations respectively, and take the difference of the two. We then average over all prompts.

As mentioned in Section 3, naive metrics are not useful for identifying the best teachers.

### G.2 Performance gap with GRACE selected teacher v/s the absolute best teacher

In addition to spearman correlations that we reported in the main paper, we also report the performance gap of the student trained with the teacher that is judged to be the best w.r.t. a metric, and the performance of the absolute best student. We report this metric for the following two cases: first, when we look at teachers constrained to a some size, and second, when we look at teachers constrained to a particular model family (from our discussion in Appendix C.1). We observe that in both cases, across different groups, GRACE returns the least performance gap. Please see Figures 13 and 14.

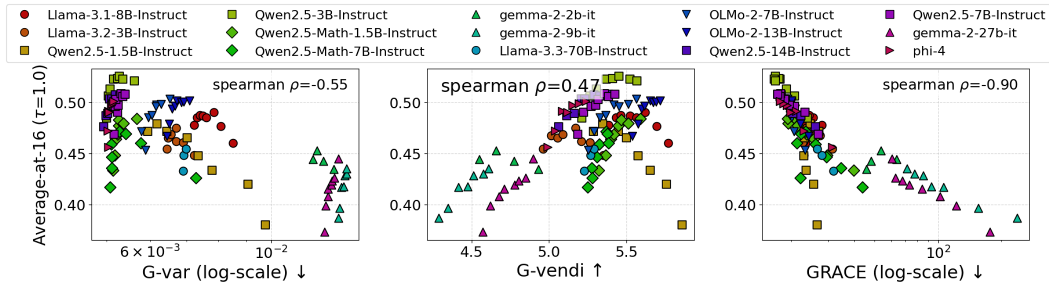



Figure 9: Repeated experiment from Figure 2 but with $d = 1024$. **GRACE achieves** $90\%$ **correlation to Llama-1B performance after training on GSM8K, across all teacher, generation temperature combinations.** G-Var and G-Vendi can only achieve 55% and 47% correlation respectively.

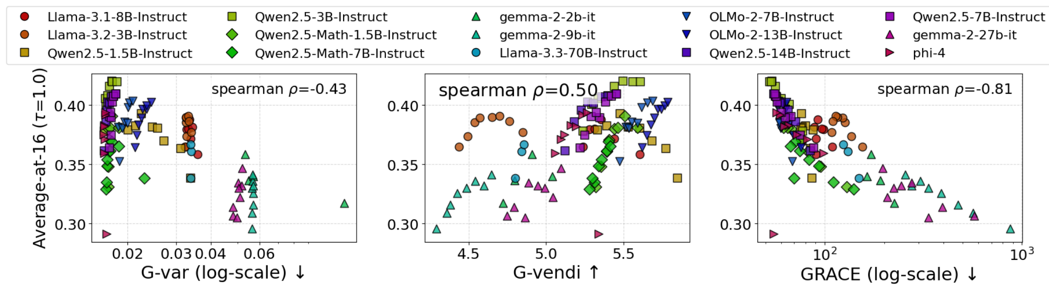



Figure 10: Repeated experiment from Figure 4 but with $d = 1024$. **GRACE achieves** $81\%$ **correlation to Llama-1B performance after training on GSM8K, across all teacher, generation temperature combinations.** G-Var and G-Vendi can only achieve 43% and 50% correlation respectively.

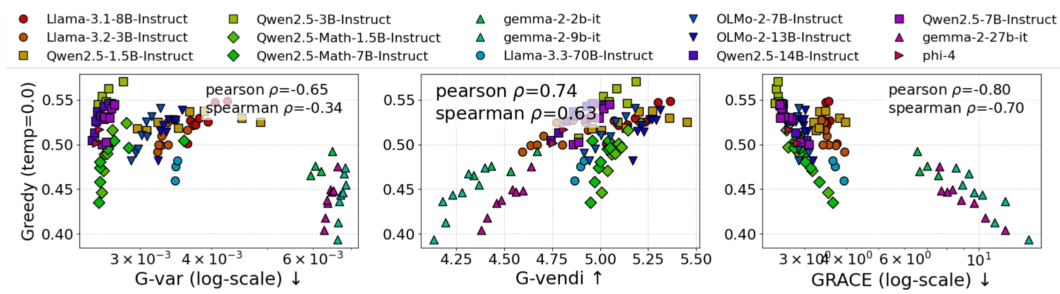



Figure 11: Repeated experiment from Figure 2 but greedy performance of trained student model. **GRACE achieves only** $70\%$ **correlation to Llama-1B performance after training on GSM8K, across all teacher, generation temperature combinations.** This is a sharp reduction from $90\%$ correlation to Average-at-16. However, GRACE still predicts the optimal teacher.

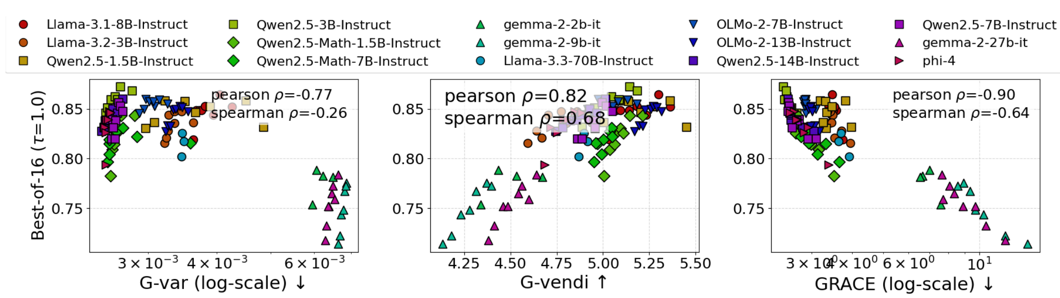



Figure 12: Repeated experiment from Figure 2 but best-of-16 performance of trained student model. **GRACE achieves only** $64\%$ **correlation to Llama-1B performance after training on GSM8K, across all teacher, generation temperature combinations.** This is a sharp reduction from $90\%$ correlation to Average-at-16. However, GRACE still predicts the optimal teacher.

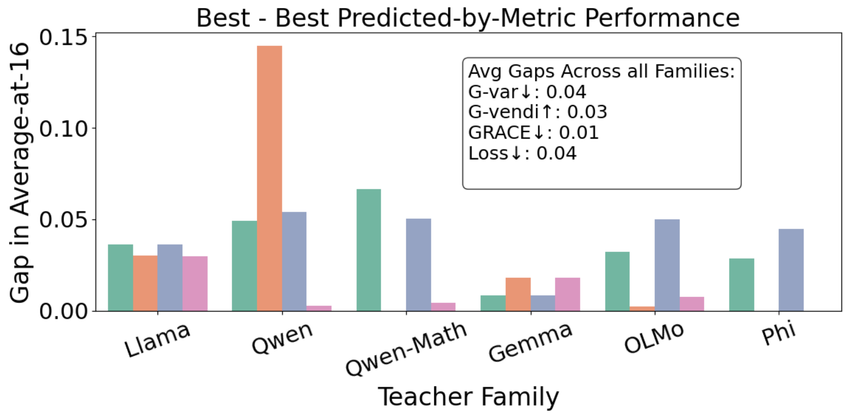



Figure 13: Gaps in the best performing and best predicted student model for each metric across teacher families for Llama-1B training on GSM8K. We observe that on average, GRACE selects a teacher that returns a student within $1\%$ performance to the absolute best performing student from the teachers in a model family. On the other hand, other metrics can select a teacher that can return a student with performance gap atleast $3\%$ w.r.t. the absolute best performing student from the teachers in a model family.

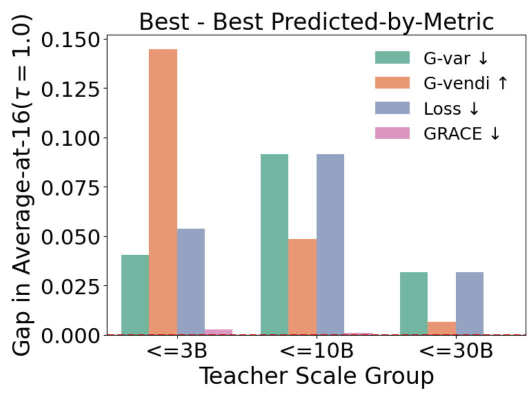



Figure 14: Gaps in the best performing and best predicted student model for each metric across teacher scale groups for Llama-1B training on GSM8K. We observe that across each group, GRACE selects a teacher that returns a student within $1\%$ performance to the absolute best performing student from the teachers in the group. On the other hand, other metrics can select a teacher that can return a student with performance gap atleast $2.5\%$ w.r.t. the absolute best performing student from the teachers in the group.

### G.3 Computational complexity

GRACE is computationally inexpensive to compute. As shown in Table 1, for $m = d = 512$ and $m = 4$, the gradients for each model takes around 10 minutes to compute and around 4.3MB to store.

## H Ablations

### H.1 Filtering v/s No filtering

In our experiments in the main paper, we perform no filtering of the responses from the teacher. Here, we compare to the case when we filter the teacher's responses by correctness. We sample 16 responses from each teacher and remove the incorrect responses. Then, we sample with repetition to get a set of 16 responses to train the model.

First, we find that the student gets worse performance with filtering of correct responses from the teacher (Figure 15). However, we find that when we compare our metrics to the student performance after training, we find that our metrics have slightly higher spearman correlation with the student

| | Gradient Features Computation | Metric Computation |
|---|---|---|
| Computation complexity<br>Running time | $\mathcal{O}(n \cdot m \cdot P \cdot d)$<br>$\approx$ 10 minutes | $\mathcal{O}(n \cdot m \cdot d^2 + d^3)$<br>$<$ 10 seconds |
| Storage Complexity<br>Actual storage | $\mathcal{O}(n \cdot m \cdot d)$<br>4.3 MB | -<br>- |

Table 1: Time complexity to compute GRACE. The running time and the actual storage have been computed on $\tilde{n} = 512$, $m = 4$, $d = 512$ for Llama-1B training on GSM8K, and have been reported as a rough average across all settings. Wall-clock time has been reported on a single H100 (80 GB) GPU. For gradient computation, we use 32 parallel CPU threads following Park et al. (2023). Here, $P$ denotes the number of parameters in the model.

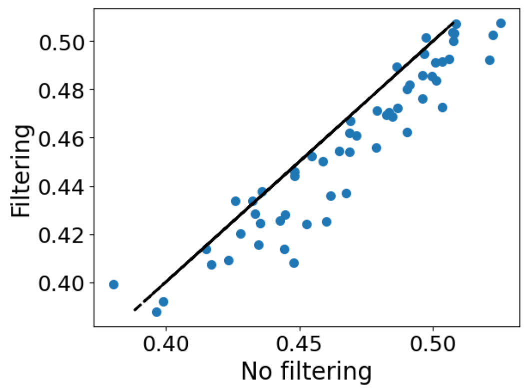



Figure 15: Comparing teachers, when we filter correct responses from the teacher v/s when we don't filter correct responses from the teacher. Here, we train Llama-1B on GSM8K with 15 teachers and generation temperatures 0.4, 0.6, 0.8, 1.0. We compare students trained from teacher without filtering (x-axis) with students trained from teacher with correct answer filtering (y-axis). We find that students trained with no filtering outperforms models trained with filtering.

performance when we train with filtering on teacher responses, compared to student trained with no filtering on the teacher responses (Figure 16).

### H.2 Ablation on training hyperparameters

We observe that a Llama-1B model trained on generations of Llama-70B Instruct models and Gemma-2-27B Instruct models perform badly. We train with learning $1e0^{-5}$ on the 16 generations per prompt of the teacher for 4 epochs. One primary question is whether the small model is over-optimizing on the teacher's generations. To check this, we track the train and test performance of the trained model with varying number of generations (Figure 17) and epochs of training (Figure 18). We observe that the performance of the trained student model improves with increasing number of epochs and number of generations, implying no over-optimization in our training setting.

### H.3 Ablations on the parameters of GRACE

In Figure 19, we show the behavior of GRACE with changing hyperparameters. We take Llama-1B training on GSM8K as a case-study. We vary number of prompts ($n$), number of generations per prompt ($m$), and the projection dimension of gradients ($d$) for computing the GRACE score and compare correlations to the student performance. We observe that (a) GRACE improves with increasing gradient dimension, (b) GRACE gives a good enough estimate with $m = 4$ generations per prompt, (c) GRACE generally increases with number of prompts that we consider but might show a small dip as we increase further.

We additionally vary the number of cross-validation splits used in GRACE. As shown in Figure 20, the correlations to the student performance do not vary much for both GSM8k and MATH for more than $C = 6$ splits. We take $C = 10$ as the default.

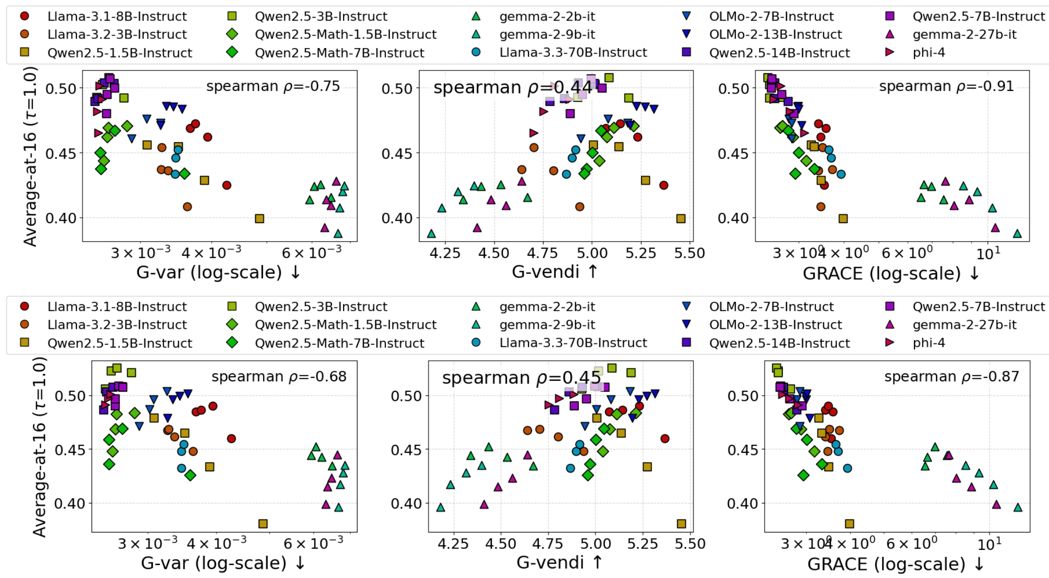



Figure 16: Comparisons between the metrics and the student performance when we filter responses v/s we don't filter correct responses from the teacher. Here, we train Llama-1B on GSM8K with 15 teachers and generation temperatures 0.4, 0.6, 0.8, 1.0. We find that our metrics have slightly higher spearman correlation to the student performance when we filter correct responses from the teacher and train only on them.

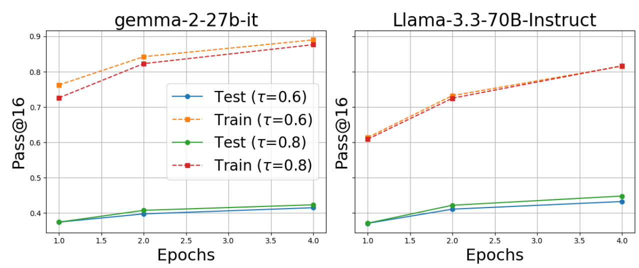



Figure 17: Llama-1B training on GSM8K with 16 responses per prompt of gemma-27b-instruct and llama-70b instruct model. We vary the number of epochs and observe that both train and test performance improves with more epochs of training. Here, the definition of pass@16 on y-axis is identical to Average-at-16.

### H.4 Gradient norm's relation to length

Figure 21 shows that the norm of the gradient on a generation decreases as the generation length grows, roughly following a trend of $1/\log T$ for length-$T$ generations, consistent with observations in Xia et al. (2024). Intuitively, this is likely because longer generations tend to contain a larger fraction of less important tokens that do not contribute much to the overall gradient. This observation motivates the $\log T$ scaling in Section 2.

### H.5 Ablation on robustness of metrics

We check the robustness of each metric by reporting the distributions of the metric values computed over random subsets of teachers. Specifically, we use 100 random draws of subsets consisting of 60% of teachers.

We compare GRACE against the baselines listed in Appendix G.1. Among all candidate metrics, GRACE is the only one showing consistently strong correlations on both datasets.

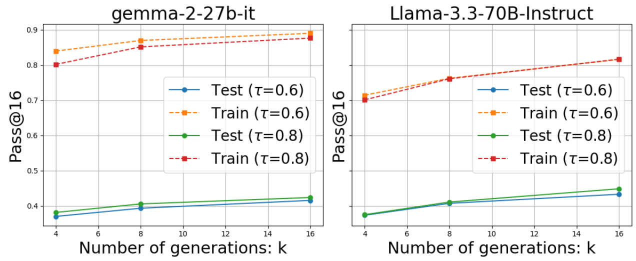
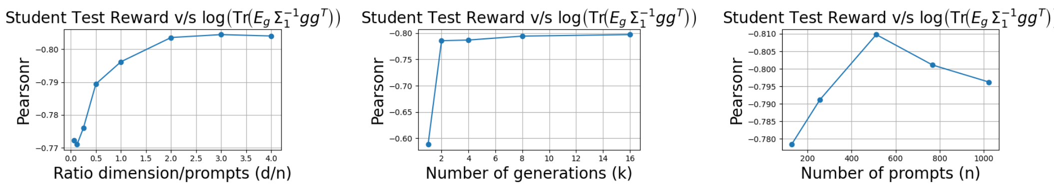



Figure 18: Llama-1B training on GSM8K with varying number of responses per prompt of gemma-27b-instruct and llama-70b instruct model. We observe that both train and test performance improves with more training samples from the teacher. Here, the definition of pass@16 on y-axis is identical to Average-at-16.




(a) When we vary gradient projection dimension $d$ with $n$.

(b) When we vary number of generations $m$ per prompt.

(c) When we vary number of prompts

Figure 19: Varying hyperparameters for GRACE on Llama-1B training on GSM8K at generation temperature $0.8$. We use the base setup as $n = 512$, $m = 16$, and $d = n$. We vary one of them, while fixing the others. Main takeaway: (a) GRACE improves with increasing gradient dimension, (b) GRACE gives a good enough estimate with $m = 4$ generations per prompt, (c) GRACE generally increases with number of prompts that we consider but might show a small dip as we increase further.

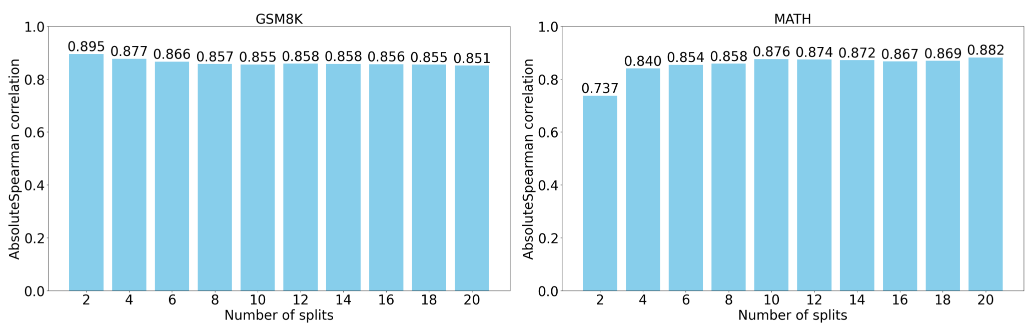



Figure 20: **Varying** number of cross-validation splits on GSM8K (left) and MATH (right).

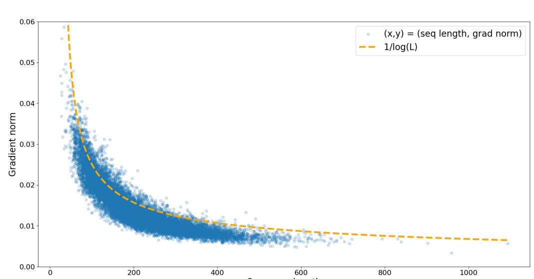



Figure 21: **Gradient norm decreases inversely with** $\log T$, where $T$ is the sequence length. This motivates the gradient scaling in Section 2.

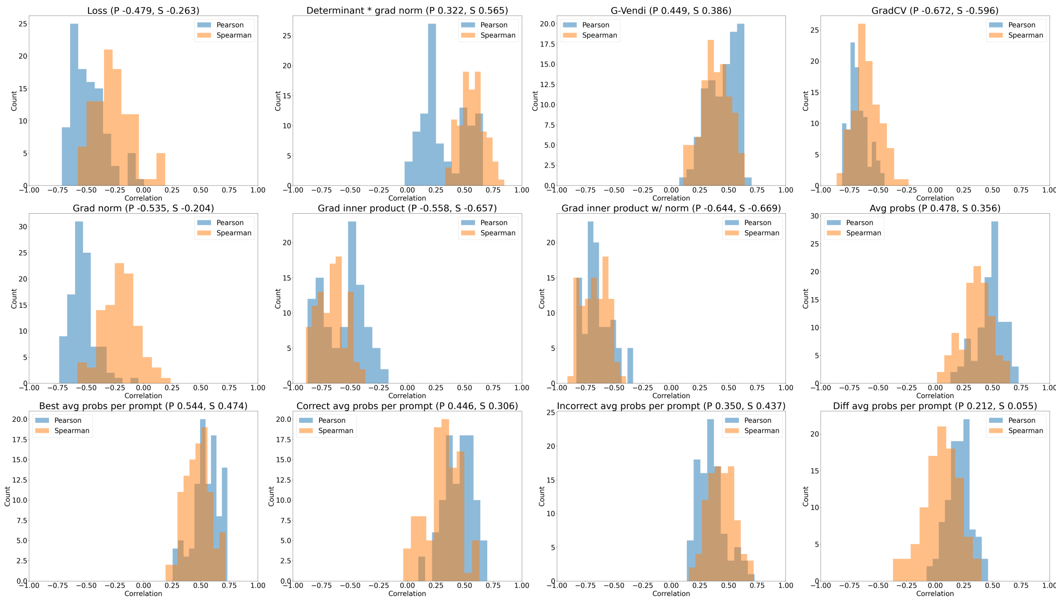



Figure 22: **Robustness of metrics on GSM8k**: we report the distribution of metric values, computed over 100 random subsets of teachers, each consisting of 60% of the full set of teacher-temperature combinations. The proposed metric GRACE consistently shows strong correlations.

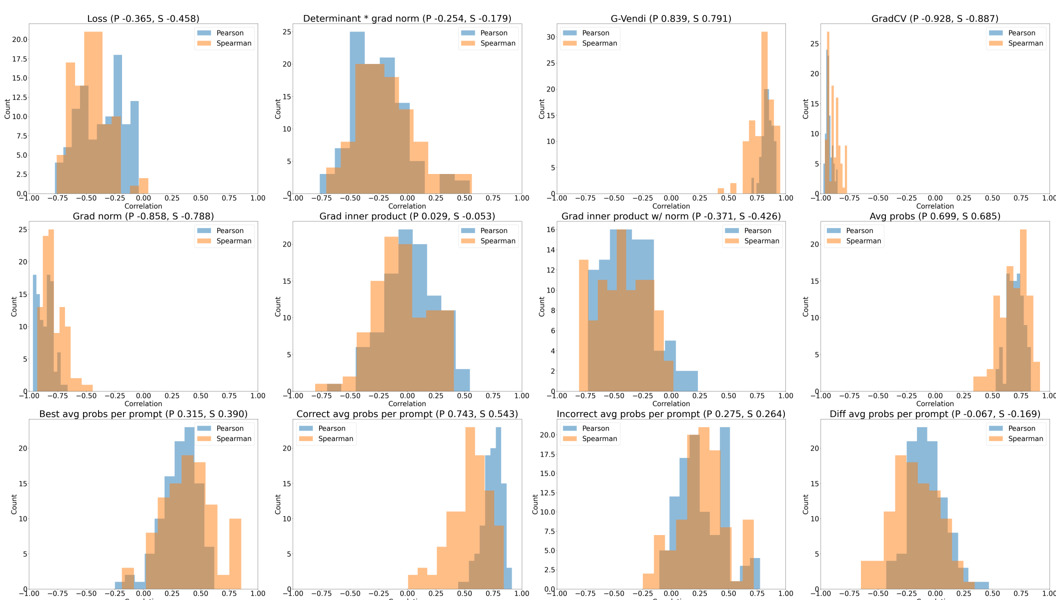



Figure 23: **Robustness of metrics on MATH**: following the same setup as Figure 22, GRACE shows the strongest correlation with smallest variations across random subsets.