Characterizing good teachers for distillation using gradient features

Anonymous Author(s)
Affiliation
Address
email

Abstract

Knowledge distillation is a primary strategy to produce powerful small models, where a "student" learns to mimic the generations of powerful "teacher" models. It is of high practical value to understand what makes a teacher suitable for distillation, so that one can efficiently identify the teacher that leads to the best student from a possibly large set of candidates. In this work, we show that good teachers should both align with the students and provide diverse training signals. Combining both leads to a single metric, GradCV, that strongly correlates with the student's post-distillation performance. We demonstrate the effectiveness of GradCV on GSM8k and MATH with LLama and OLMo student models.

1 Introduction

Distillation is an efficient and effective method to produce capable small models from existing, powerful large models across many settings. In this work, we focus on the specific case of training autoregressive language models on the generations produced by a teacher model. It is difficult to select the right teacher for a given student and task, given the large number of available models and the counterintuitive fact that a better model is not always a better teacher. The current approach of guess-and-check is costly, because it requires collecting generations from a capable teacher and subsequently training a student on those generations. Additionally, the specific hyperparameters used in both phases can dramatically affect the final performance of the student, underscoring the need for careful, repeated testing to select the right teacher. 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?

In this work, we address this question by identifying two factors that strongly correlated with the student's performance after distillation. The first is the *alignment* between the teacher and the student. One example is capacity gap (Mirzadeh et al., 2019), which is known to affect both generalization and learning speed of the student (Mirzadeh et al., 2019; Harutyunyan et al., 2023; Panigrahi et al., 2025). In this work, we use the gradient norm of the student as a proxy of the alignment, where smaller norms are associated with better alignment, basing loosely on intuitions from convex problems. Another factor is the *diversity* of the teacher's generations, which is commonly used in data selection (Ash et al., 2019; Jung et al., 2025; Wang & Gu, 2025).

Based on these observations, we propose a metric, GradCV, that strongly correlates with the student's distillation performance from a teacher. GradCV is efficient to compute, requiring only the gradients of the student at the beginning of training, and a few teacher generations. Empirically, GradCV can be used to guide the selection of both the teacher model and its generation temperature, a key hyperparameter that crucially affects distillation performance (Zheng & Yang, 2024; Peng et al.,

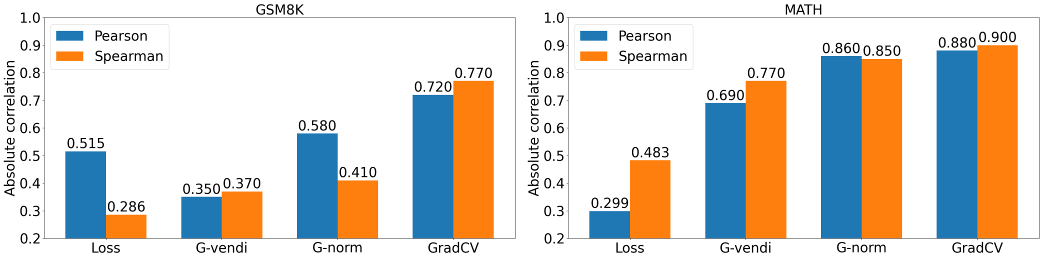


Figure 1: Comparing metrics on GSM8k (left) and MATH (right), in terms of Pearson correlations (blue) and Spearman correlations (orange). The proposed GradCV most strongly correlates with the student's test performance on both datasets.

2024). Results on GSM8k (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021) demonstrate the effectiveness of the metric (Figure 1).

1.1 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).

Alignment and coverage in RL Despite the simplicity of the concept, the design space of distillation can be complex (Peng et al., 2024). The teacher choice unsurprisingly has a big impact, but one counterintuitive finding is that a strong teacher is not always a better teacher (Mirzadeh et al., 2019; Jafari et al., 2021; Harutyunyan et al., 2023; Panigrahi et al., 2025). A widely accepted explanation is the so-called capacity gap, which posits that the student can become worse when the teacher is "far" and hence harder to distill from. This is consistent with our positive correlation between the teacher-student alignment and the student's post-distillation performance.

Alignment also ties to the notion of coverage in reinforcement learning. Indeed, 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).

Diversity in data selection For text-based distillation, selecting the best teacher can be considered as selecting the best subset of samples from the generations of all teachers. This can be considered as a (more structured) special case of *data selection*, whose goal is to identity the most useful samples given a data budget (Sorscher et al., 2022; Xia et al., 2024; Albalak et al., 2024). Among numerous metrics, the most relevant to us gradient-based diversity measures. In particular, Ash et al. (2019) and Jung et al. (2025) both use guide the selection by requiring diverse gradients. Unlike data selection though, distillation is a more controllable process as one can trade off the data quality and diversity of a given teacher by changing the generation temperature.

2 Problem setup and the proposed metric

We consider the setting where the student distills from the teacher by autoregressively training on the teacher's generated texts. Contrasting to logit-based distillation, this enables distillation across architectures with various tokenizer choices. We denote the space of all possible prompts as $\mathcal{X}$ and responses as $\mathcal{Y}$. The set of training prompts is denoted $\mathcal{X}_{\text{train}} \subseteq \mathcal{X}$, with $n := |\mathcal{X}_{\text{train}}|$. For an

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

input $\mathbf{x} \in \mathcal{X}$, the teacher π^* autoregressively generates responses from $\mathcal{Y}$ according to a distribution $P^{\pi^*}(\mathbf{x}; \tau)$, where τ represents temperature used in softmax.

The student π is trained with standard auto-regressive loss $\mathcal{L}_{CE}$ on concatenations of prompts and teacher responses. We will use $\mathcal{L}_{CE}(\mathbf{x}, \mathbf{y}; \pi)$ to denote the auto-regressive loss of the response $\mathbf{y}$ with the prompt $\mathbf{x}$, and $\nabla \mathcal{L}_{CE}(\mathbf{x}, \mathbf{y}; \pi)$ to denote the gradient with respect to the student's parameters. We sample k generations from each prompt, and train the student with a total of $n \times k$ sequences. The pre-trained and fine-tuned students are denoted by π_{pt} and π_{ft}, respectively.

We measure a model's performance by the expected reward score J:

$$J(\pi; \mathcal{X}) = \mathbb{E}_{\mathbf{x} \in \mathcal{X}} \mathbb{E}_{\mathbf{y} \sim P^{\pi}(\mathbf{x}; \tau=1)} r(\mathbf{x}, \mathbf{y}),$$

where r is a reward function that takes a prompt-response pair and returns a score between $[0, 1]$.

Primary goal Given a student model, we aim to select a teacher model from a pool of N teachers $\{\pi_i^*\}_{i=1}^N$ so that the student's post-distillation performance is maximized. One naive strategy is to simply train the student with every teacher and return the best model. This would correctly select the best teacher by definition, but is computationally infeasible. We instead ask:

> *Can we find metrics that can account for the performance of the student, using a small number of forward or backward passes of the teacher and the student?*

Desiderata: alignment and diversity The best teacher for distillation isn't necessarily the best performing one (Mirzadeh et al., 2019; Razin et al., 2025). The intuition is that performance may not be reflective of the quality of the supervision, as there are other desiderata when considering the effect on optimization and generalization. The two main factors we consider are the teacher-student *alignment* and the teacher's generation *diversity*: the former has been shown to be important in distillation (Mirzadeh et al., 2019; Harutyunyan et al., 2023; Busbridge et al., 2025; Panigrahi et al., 2025), and the latter has been proven effective for data selection and active learning (Ash et al., 2019; Jung et al., 2025; Wang & Gu, 2025).

Before presenting our proposed metrics, we discuss several promising candidates. While the metrics may be defined using population quantities, in the experiments, we use empirical estimates computed on a subset of training set prompts $\tilde{n}$ each with k generations. We denote this set as $\mathcal{D}$.

Alignment measure: One candidate is to simply measure the average loss of the student on the teacher's generations, where a smaller loss means better alignment. However, we find that the pre-trained student (π_{pt})'s loss on a teacher's generations do not correlate well with the performance of the distilled student (π_{ft}). What's more, π_{ft}'s loss on teacher's generations on the test set after training is not strongly correlated with performance either. This suggests that the loss is mismatched with the reward and is hence not a reliable metric.

We instead use *gradient norms* to quantify the alignment between a pre-trained student and its teacher. Loosely following intuitions from convex analyses, a smaller gradient norm means smaller distance to a minimum, indicating better alignment. Mathematically, we measure

$$\text{G-Norm}(\mathcal{D}; \pi_{pt}) := \mathbb{E}_{(\mathbf{x},\mathbf{y}) \sim \mathcal{D}} \left\| \nabla \mathcal{L}_{CE}(\mathbf{x}, \mathbf{y}; \pi_{pt}) \right\|_2 . \quad (1)$$

Diversity measure: We use the definition of diversity in gradients from Jung et al. (2025). Given the training set $\mathcal{D}$ of prompt-generation sequences, let $\boldsymbol{G} \in \mathbb{R}^{\tilde{n}k \times d}$ denote the matrix that contains the pre-trained student's gradients on $\mathcal{D}$, which have been projected to be d-dimensional for computational efficiency (Park et al., 2023) and normalized. Define $\Sigma = \boldsymbol{G}^\top \boldsymbol{G}$. Then, the *G-Vendi score* (Jung et al., 2025) is defined as the entropy of spectrum of Σ:

$$\text{G-Vendi}(\mathcal{D}; \pi_{pt}) := \text{Entropy}(\lambda(\Sigma)) = -\sum_{\lambda \in \lambda(\Sigma)} \lambda \log \lambda, \quad (2)$$

where $\lambda(\Sigma)$ denotes the eigenvalues of the gradient matrix Σ, with $|\lambda(\Sigma)| = \min\{\tilde{n}k, d\}$. We perform ablation study on the choice of the projected dimension d in Appendix C.1.

The proposed metric: GradCV A good metric should jointly consider both alignment and diversity. Note that these two can be at odds. For instance, we find that when increasing the teacher's generation temperature τ, G-Norm increases, indicating decreasing alignment with the student, but G-Vendi increases, indicating an desirable increase in diversity. Thus, we want a metric to reflect the trade off of these two measures, with appropriately adjusted scales.

We first adapt the gradient diversity measure. Recall that G-Vendi computes the entropy of the gradient covariance, which measures how spread out the gradients are. One can instead compute the spread by checking how well the eigenspaces computed from two random subsets are aligned. Specifically, let $\mathcal{D}_1, \mathcal{D}_2$ be a partition of $\mathcal{D}$, we compute Σ_1 using the projected and normalized gradients from $\mathcal{D}_1$, and compute projected but unnormalized gradients $\{\boldsymbol{g}\}$ from $\mathcal{D}_2$. We perform such random splits in a cross validation fashion, where we take the expectation over random draws of $\mathcal{D}_1, \mathcal{D}_2$.

Formally, the GRADient-based Cross-Validation (GradCV) metric is given by:

$$\begin{aligned} \text{GradCV}(\mathcal{D}; \pi_{pt}) :=& \log\left(\mathbb{E}_{\mathcal{D}_1,\mathcal{D}_2:\mathcal{D}_1\cup\mathcal{D}_2=\mathcal{D}}\left[\boldsymbol{g}^\top\tilde{\Sigma}_1^{-1}\boldsymbol{g}\right]\right), \\ \text{where } \tilde{\Sigma}_1 =& (1-\lambda)\Sigma_1 + \lambda\frac{\text{Trace}(\Sigma_1)}{d}I. \end{aligned} \tag{3}$$

Here, λ is a smoothing hyperparameter. Since $\boldsymbol{g}$ is unnormalized, GradCV reflects the gradient norm. To see how GradCV captures diversity, note that if the gradients are not diverse and behave differently for each subset, then Σ_1 and $\mathbb{E}_{\boldsymbol{g}\sim\mathcal{D}_2}[\boldsymbol{g}\boldsymbol{g}^\top]$ can misalign (up to normalization), in which case the trace can blow up. Therefore, GradCV effectively combines both the diversity in gradient spectrum and the gradient norm. We also note that GradCV shares similarity to a form of leave-one-out conditional mutual information, a provably effective generalization measure (Rammal et al., 2022).

3 Experiments

We test out all metrics on two common math reasoning datasets, GSM8k (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021). We compare the Pearson and Spearman correlations of between metrics and the student's performance after distillation; the former captures the goodness of linear fit, and the latter captures the correctness of ranking.

Experiment details The student model is taken to be Llama-1B-base or OLMo-1B-base on GSM8k, and Llama-3B-base on MATH. We compare 13 teachers on GSM8k: 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 2/9B Instruct models, and OLMo 7/13B Instruct models. We compare 9 teachers on MATH: Qwen-2.5 Math 1.5/7B Instruct models, Qwen-2.5 1.5/3/14/32B Instruct models, Llama-(3.2/3.3) 3/8/70B Instruct models. The teacher's generation temperature is varied between 0.3 and 1.0.

To compute GradCV, G-Norm, and G-Vendi, we use $\mathcal{D}$ as a randomly selected set of $\tilde{n} = 512$ training prompts from the training dataset. For GradCV, we randomly divide $\mathcal{D}$ into a random 90-10 split, where $\mathcal{D}_1$ and $\mathcal{D}_2$ contain $0.9\tilde{n}$ and $0.1\tilde{n}$ prompts respectively. The cross validation is computed based on 10 draws of $\mathcal{D}_1, \mathcal{D}_2$. We use random projection on gradients to dimension $d = \tilde{n}$.

Each distillation 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 $k = 16$ responses per question (prompt) from each teacher and fine-tune the student on all generations without filtering for correctness.

In the rest of the section, we show that GradCV gives the best correlation in different settings.

3.1 Warmup: measuring correlations at a single temperature of generation

We first consider a constrained case where all teachers use the same generation temperature of 0.8, which is within the common range. Note that naive metrics such as teacher performance are not indicative of distillation performance. Specifically, we consider 1) the teacher performance on the test set and 2) the student's loss on teacher generations; neither provides non-trivial correlation to student performance on GSM8k (Figure 7). Strikingly, this holds true even for the fine-tuned student's loss on teacher generations, showing that loss-based measures are fundamentally unreliable indicators.

[2] We searched over learning rates $\{3 \times 10^{-5}, 10^{-5}, 8 \times 10^{-6}\}$ and found 10^{-5} to be consistently the best.

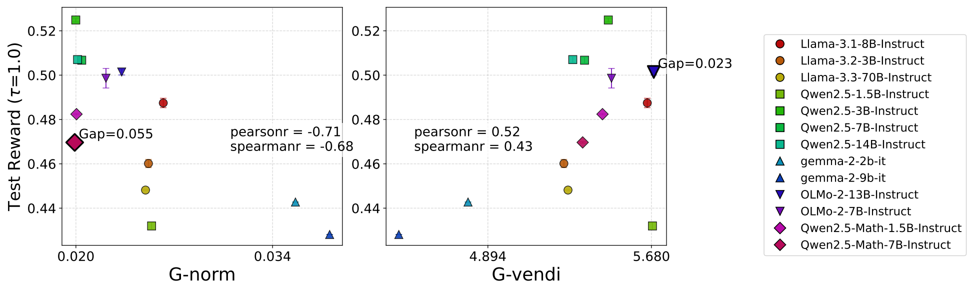


Figure 2: G-Norm and G-Vendi's correlation with Llama-1B performance on GSM8k, based on 16 teacher generations per prompt at temperature 0.8. (Left to right) (a) G-Norm correlates well with student performance but fails to identify the best teacher. (b) G-Vendi highlights a teacher closer to the best-performing one, though its overall correlation with student performance is weaker.

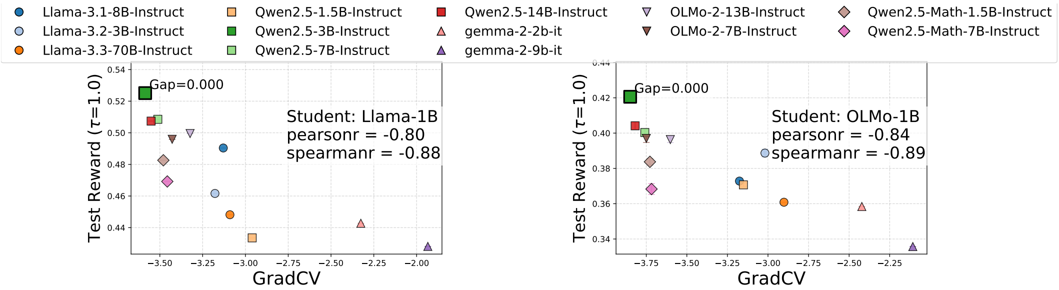


Figure 3: GradCV's correlation with Llama-1B and OLMo-1B performance on GSM8k, based on 16 teacher generations per prompt at temperature 0.8. Main takeaway: GradCV can achieve 80% and 84% pearson correlation with student performance after training, while also predicting Qwen-3B-Instruct to be the optimal teachers for both the student models.

Alignment and diversity matter: Next, we consider the two desiderata discussed in Section 2. Figure 2 shows results for gradient norms as defined by Equation (1), which has improved correlation to the student's performance but fails to predict the best teacher. On the other hand, G-Vendi can predict a teacher closer to the optimal teacher, but has worse correlation to the student performance.

GradCV achieves a good trade-off: By balancing alignment and diversity, GradCV not only demonstrates strong correlation with the student's performance, and also helps select the optimal teacher. As shown in Figure 3, GradCV achieves 80% and 84% Pearson correlation for Llama-1B and OLMo-1B, and correctly predicts Qwen-3B-Instruct as the optimal teachers for both students.

The above observations also hold true for a Llama-3B student on MATH. Similar to GSM8k, the pre-trained student's loss on teacher generations exhibits little correlation with the student's final performance (Figure 10). While G-Vendi and G-Norm show reasonable correlation, GradCV outperforms both with above stronger predictive power, each achieving a correlation of approximately 95% with the student's performance.

3.2 Jointly considering teacher and temperature choices

Up to this point, our comparisons have used a fixed generation temperature across the teachers. As distillation performance can be sensitive to the temperature, a natural question is whether the metric can also account for the additional variation introduced by temperature changes. In this section, we extend to a broader setting where the temperature is varied between 0.3 and 1.0. We use the Llama-1B student for GSM8k and the Llama-3B student for MATH, and compute correlations using all teachers across different temperatures. We find that GradCV continues to show stronger correlation than G-Norm and G-Vendi (Figure 4, Figure 5).

GradCV can identify close-to-optimal generation temperature: A useful metric should ideally inform the selection of key hyperparameters such as the generation temperature. To better understand how metrics capture temperature change, we zoom in to a pair of teachers, Qwen-2.5 3B Instruct and

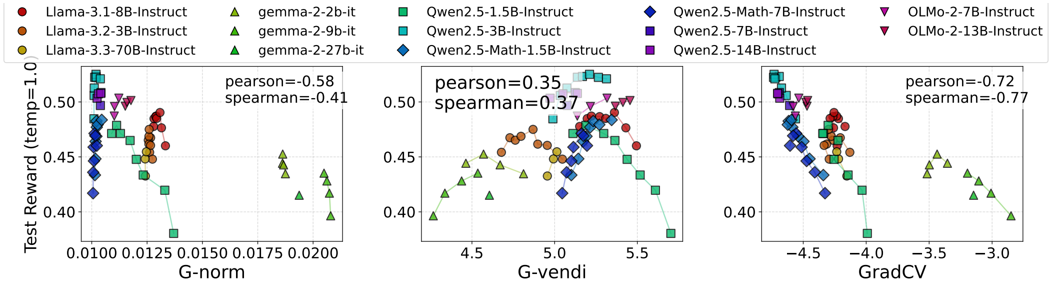


Figure 4: Correlation between GradCV with Llama-1B performance on GSM8k, based on 16 teacher generations per prompt at temperatures $\in \{0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$. GradCV can achieve 77% spearman correlation with student performance after training, while G-Norm and G-Vendi can only achieve 41% and 37% respectively.

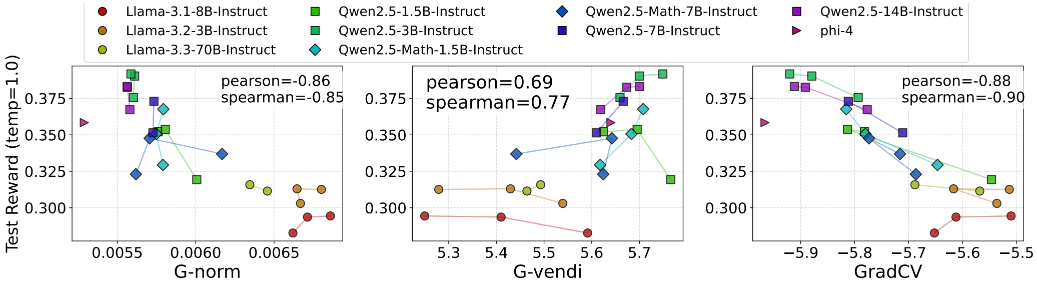


Figure 5: Correlation between GradCV with Llama-3B performance on MATH, based on 16 teacher generations per prompt at temperatures $\in \{0.4, 0.6, 0.8\}$. GradCV can achieve 90% spearman correlation with student performance after training, improving over G-Norm and G-Vendi.

Qwen-2.5 1.5B Instruct, which exhibit strikingly opposite behaviors as τ increases. Figure 6 shows that as temperature increases, the student's G-Norm on teacher generations increases, indicating decreased alignment, while G-Vendi also increases, reflecting higher diversity. As the two metrics provide conflicting signal, they do not give a clear preference on the temperature.

In contrast, GradCV is able to separate the behaviors of the two models: its value decreases with τ for Qwen-2.5 3B Instruct and increases with τ for Qwen-2.5 1.5B Instruct, matching the actual trends in student performance. GradCV correctly pinpoints the optimal training temperatures for Llama-1B, which are 0.8 for Qwen-2.5 3B Instruct and 0.4 for Qwen-2.5 1.5B Instruct, demonstrating its usefulness as a predictive metric.

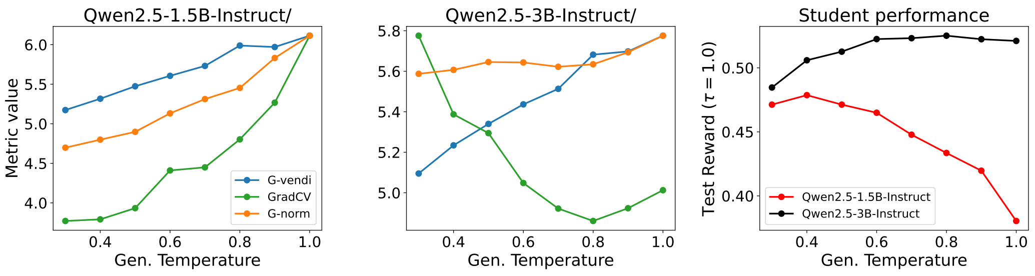


Figure 6: Comparing Llama-1B training with Qwen-2.5-1.5B-Instruct and Qwen-2.5-3B-Instruct teachers for GSM8k across different temperatures of generation. While both G-Norm and G-Vendi show increasing trends across temperature for both teachers, GradCV helps to distinguish between the teachers by showing reverse trends, which also correlates with the student's performance after training.

Robustness to teacher choices One should caution that it is always possible to fit a metric that maximize the correlation on a given set of teachers, which would not be meaningful. It is therefore important to test the robustness of the metrics. We do so by computing metric values repeatedly over random subsets of all teachers. As shown in Figure 13 and Figure 14, GradCV demonstrates consistently high correlations across random subsets. Details and additional baseline metrics are deferred to Appendix C.2.

4 Discussion and Conclusion

In conclusion, we have shown that the gradients of the student offer a reliable lens for identifying the teacher most effective for distillation. By disentangling two fundamental factors—alignment between teacher and student, and diversity in teacher generations—we demonstrated that gradient norm (G-Norm) captures alignment while gradient diversity (G-Vendi) captures diversity. Yet, relying on both separately complicates the prediction of student performance. To overcome this limitation, we introduced GradCV, a unified metric that integrates alignment and diversity into a single measure. Our experiments on GSM8k and MATH establish that GradCV not only enables principled comparison across teachers but also accurately predicts the optimal generation temperature for each teacher. These findings highlight GradCV as a practical and general-purpose tool for guiding teacher selection and generation strategies in student training.

There are several promising directions for future work. First, the definition of GradCV bears a close connection to leave-one-out conditional mutual information, which has been shown to be a provably effective measure of generalization (Rammal et al., 2022). Formalizing this connection could yield deeper theoretical insights. Second, while GradCV provides a natural interpolation between G-Norm and G-Vendi, the role of the gradient dimensionality parameter in mediating this transition remains unclear. Developing a theoretical framework to characterize how this parameter governs the relationship between alignment and diversity would be an important step forward. Finally, GradCV currently does not lead to perfect Spearman correlations, suggesting that some variates are yet to be captured. Possible next steps include incorporating additional properties of the teacher and distribution-specific quantities.

References

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.

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.

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

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.

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.

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.

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.

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.

Yuqing Wang and Shangding Gu. Data uniformity improves training efficiency and more, with a convergence framework beyond the ntk regime. *arXiv preprint arXiv: 2506.24120*, 2025.

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.

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.

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.

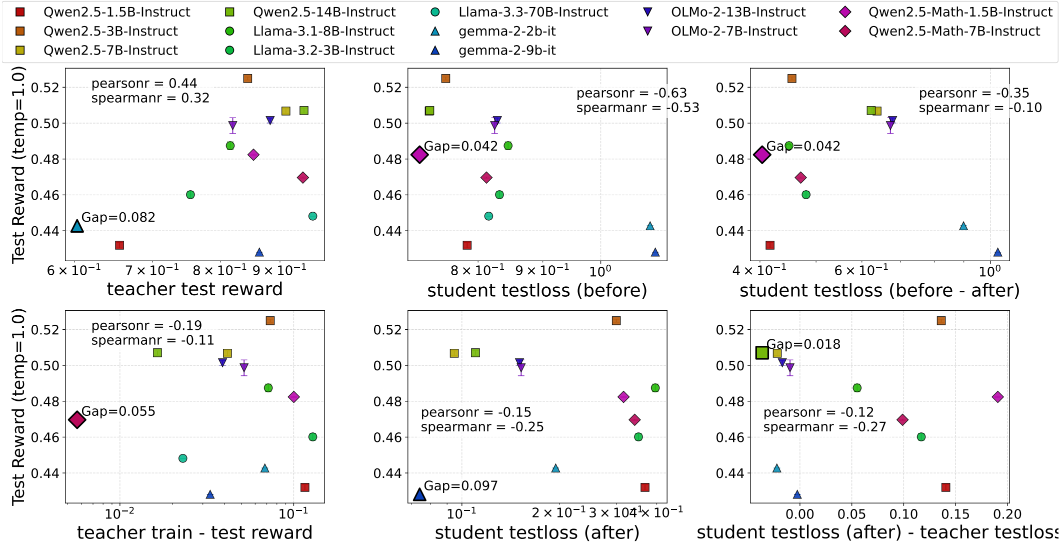


Figure 7: Using teacher performance and student loss to predict Llama-1B performance on GSM8k, based on 16 teacher generations per prompt at temperature 0.8. (Top to bottom, Left to right): For different teachers, we compare (a,b): teachers' test reward and the gap in train, test rewards at $\tau = 0.8$. (c,d): The student's loss on teachers' test prompt generations, before (π_{pt}) and after (π_{ft}) training. (e): Drop in student's loss after training, (f) Gap in student's and teacher's loss after training. Takeaway: **Neither student loss on teacher generations nor teacher performance alone is a reliable indicator of the student's eventual performance after training.**

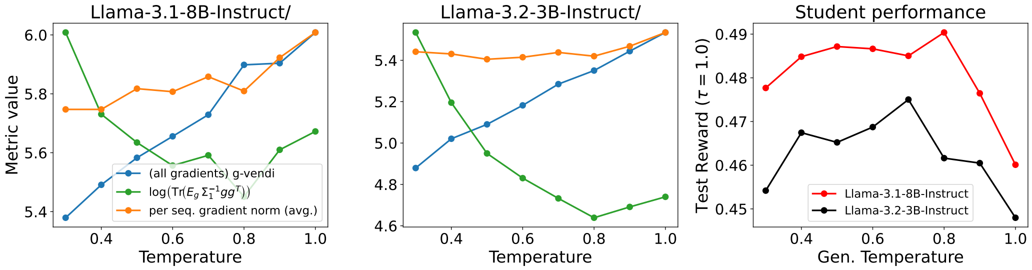


Figure 8: Comparison study between Llama teachers for GSM8k. While both G-Norm and G-Vendi show increasing trends with temperature of generation for both the Llama teachers, GradCV can differentiate different generation temperatures, and predict close-to-optimal generation temperature for each teacher. On Llama-3.1-8B-Instruct, it predicts 0.8 as the optimal temperature, which also corresponds to the optimal temperature in terms of student performance. On Llama-3.2-3B-Instruct, it predicts 0.8 as the optimal temperature, while the optimal temperature in terms of the student performance is 0.7.

A Additional results on GSM8K

A.1 More basic metrics

As mentioned in Section 3, naive metrics are not useful for identifying the best teachers. Figure 7 shows that both the student loss on the teacher's generations and the teacher's performance.

A.2 Student performance with generation temperature

A good metric should be able to guide the selection of key hyperparameters, such as the generation temperature. Figure 8 and Figure 9 show that GradCV is indeed more informative to the temperature choice than G-Vendi and G-Norm.

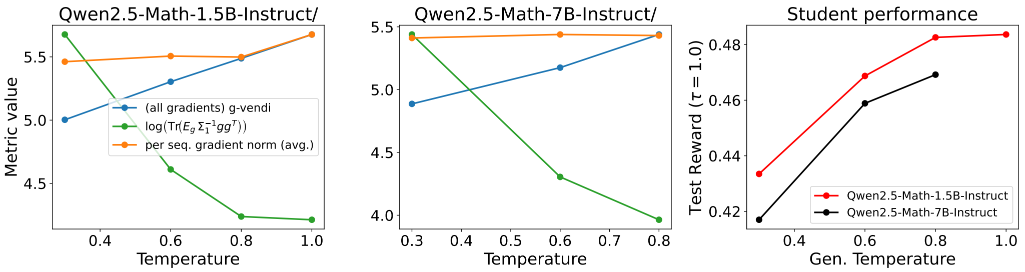


Figure 9: Comparison study between Qwen-Math teachers for GSM8k. Here, GradCV shows a decreasing trend across teachers for both Qwen-Math models, indicating that the performance of the student increases as we increase the generation temperature.

B Additional results on MATH

We repeat our experiments from Figure 2 and Figure 3 on the MATH dataset, where we train a Llama-3B model. We observe that GradCV achieves similar correlation performance to G-Norm, while achieving better correlation performance than both G-Vendi and test loss of the student.

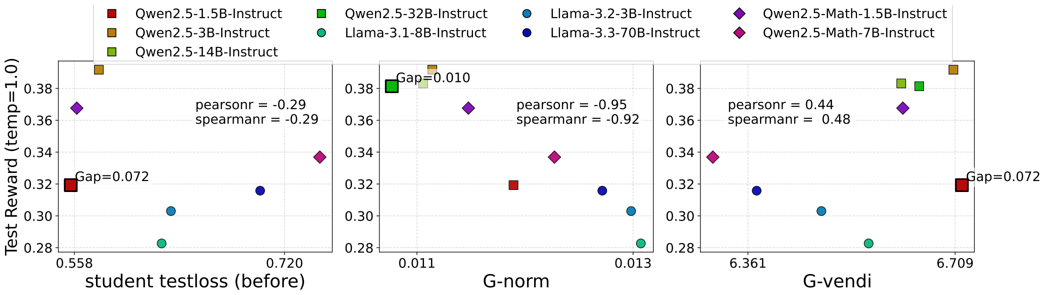


Figure 10: We compare the correlations of the pre-trained student's loss, G-Norm, and G-Vendi with Llama-3B's performance on MATH after training, using 16 teacher generations per prompt at temperature 0.8. While both the student's loss and G-Vendi fail to capture the relationship, G-Norm achieves an almost perfect correlation.

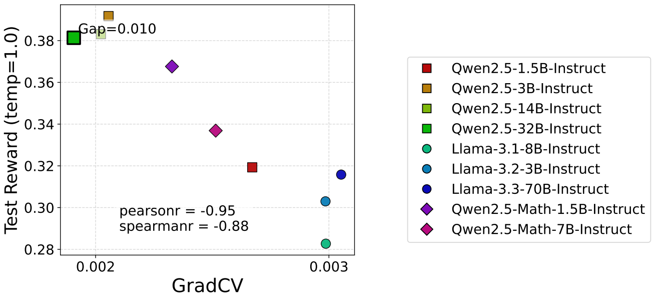
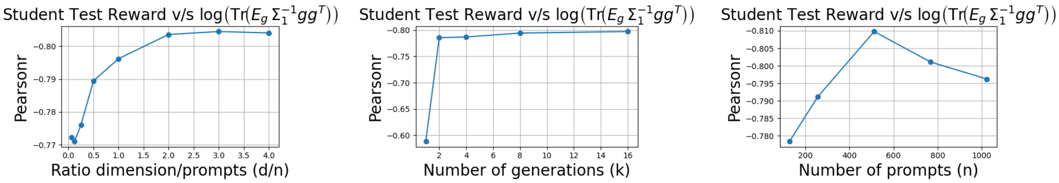


Figure 11: On Llama-3B's training on MATH using 16 teacher generations per prompt at temperature 0.8, GradCV can match the correlation performance on G-Norm.

C Ablations

C.1 Ablations on the parameters of GradCV

In Figure 12, we show the behavior of GradCV 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 (k), and the projection dimension of gradients (d) for computing the GradCV score and compare correlations to the student performance. We observe that (a) GradCV improves with



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

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

(c) When we vary number of prompts

Figure 12: Varying hyperparameters for GradCV on Llama-1B training on GSM8k at generation temperature 0.8. We use the base setup as $n = 512$, $k = 16$, and $d = n$. We vary one of them, while fixing the others. Main takeaway: (a) GradCV improves with increasing gradient dimension, (b) GradCV gives a good enough estimate with $k = 4$ generations per prompt, (c) GradCV generally increases with number of prompts that we consider but might show a small dip as we increase further.

increasing gradient dimension, (b) GradCV gives a good enough estimate with $k = 4$ generations per prompt, (c) GradCV generally increases with number of prompts that we consider but might show a small dip as we increase further.

C.2 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 GradCV with the following metrics:

1. Student Loss on the teacher's generations;
2. G-Norm (Equation (1));
3. G-Vendi (Equation (2));
4. Determinant × gradient norm, corresponding to BADGE (Ash et al., 2019), which captures both the diversity and magnitude of gradients;
5. 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:

$$\mathbb{E}_{\mathbf{x}} \mathbb{E}_{(\mathbf{x},\mathbf{y}_1),(\mathbf{x},\mathbf{y}_2)\sim\mathcal{D}} \left[\frac{\mathbf{g_1}}{\|\mathbf{g_1}\|_2} \right]^\top \frac{\mathbf{g_2}}{\|\mathbf{g_2}\|_2},$$

$$\text{where } \mathbf{g}_1 = \nabla \mathcal{L}_{CE}(\mathbf{x}, \mathbf{y}_1; \pi_{pt}),$$
$$\mathbf{g}_2 = \nabla \mathcal{L}_{CE}(\mathbf{x}, \mathbf{y}_2; \pi_{pt}).$$

6. 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.
7. 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.
8. 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.
9. 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.
10. Incorrect average probabilities: Same as above, but over incorrect generations.
11. 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.

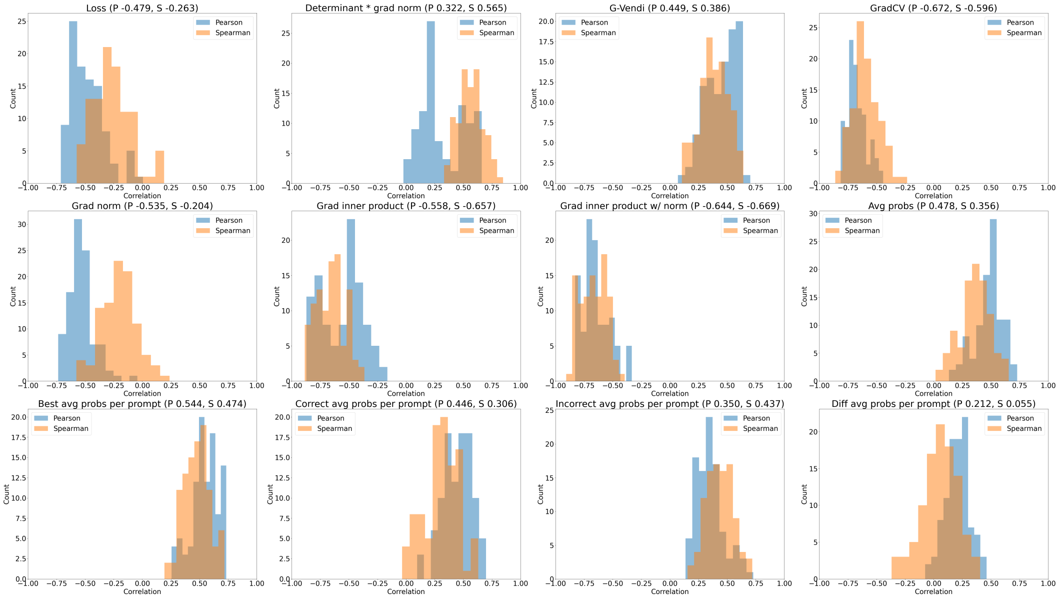


Figure 13: **Robustness of metrics on GSM8k**: we report the distribution of metric values, computed over 100 random subsets of teachers, each consisting of 60% of the full set of teacher-temperature combinations. The proposed metric GradCV consistently shows strong correlations.

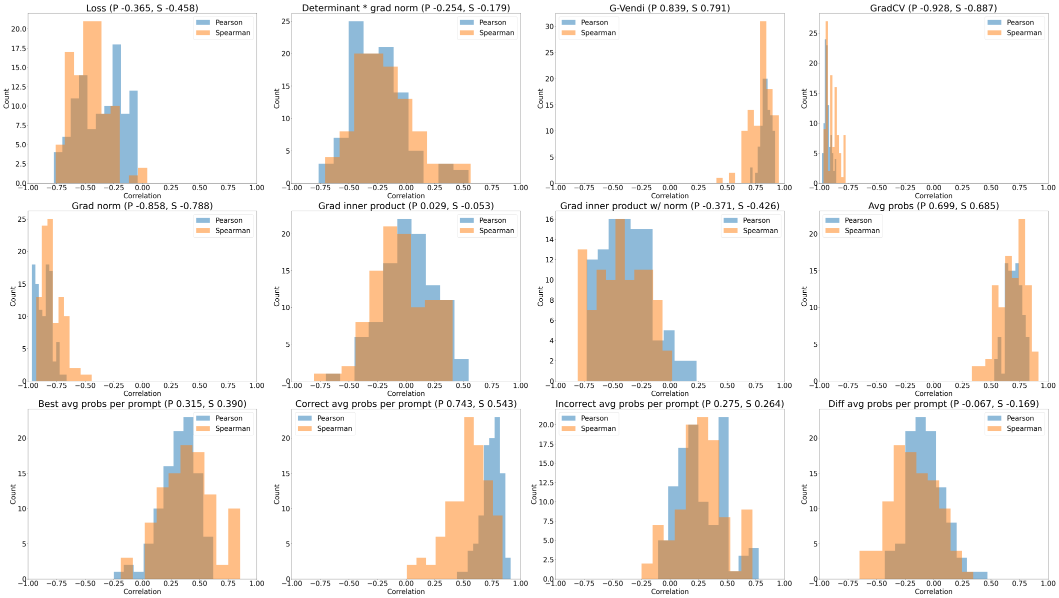


Figure 14: **Robustness of metrics on MATH**: following the same setup as Figure 13, GradCV shows the strongest correlation with smallest variations across random subsets.

Among all candidate metrics, GradCV is the only one showing consistently strong correlations on both datasets.