Do Generalisation Results Generalise?

Matteo Boglioni[1,2] **Andrea Sgobbi**[1] **Gabriel Tavernini**[1] **Francesco Rita**[1]
Marius Mosbach[2,3] **Tiago Pimentel**[1]

[1]ETH Zürich, [2]Mila - Quebec Artificial Intelligence Institute, [3]McGill University

mboglioni @ethz.ch, marius.mosbach@mila.quebec, tiago.pimentel@inf.ethz.ch

Abstract

A large language model's (LLM's) out-of-distribution (OOD) generalisation performance is crucial to its deployment. Previous work assessing LLMs' generalisation performance, however, typically relies on a single out-of-distribution dataset. This approach may fail to precisely evaluate the capabilities of the model, as the data shifts encountered during deployment are much more diverse. In this work, we investigate whether OOD generalisation results generalise. More specifically, we evaluate a model's performance across multiple OOD testsets throughout a finetuning run; we then evaluate the partial correlation of performances across these testsets, regressing out in-domain performance. This tells us how correlated OOD generalisation performances are, once in-domain performance is controlled for. Analysing `OLMo2`, `OPT` and `SmolLM`, we observe no overarching trend in generalisation results, with the existence of a positive or negative correlation between any two OOD testsets depending on the specific choice of model analysed.

ilboglions/GeneralisationGeneralises

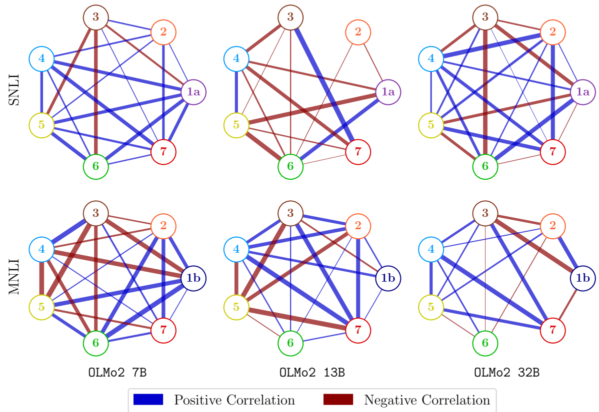


Figure 1: `OLMo2`'s partial OOD correlations on `SNLI` (top) and `MNLI` (bottom). No clear trends are observed. Edge thickness increases with absolute correlation value. Legend: `1a.MNLI`, `1b.SNLI`, `2.WNLI`, `3.SciTail`, `4.RTE`, `5.HANS`, `6.ANLI`, and `7.PAWS`.

1 Introduction

A large language model's (LLM's) out-of-distribution (OOD) generalisation performance is an essential property for its deployment in the wild.[1] This is because, upon deployment, its users will inevitably submit a diverse range of prompts to it, which will fall outside the model's training data distribution. Not surprisingly, generalisation has received increased attention from the community (Xu et al., 2021; Hupkes et al., 2023; Yang et al., 2023, 2024, 2025; Yuan et al., 2023; Ye, 2024).

Most work evaluating generalisation, however, quantifies it using a single out-of-distribution testset per task (Mosbach et al., 2023; Joshi and He, 2022; Bhargava et al., 2021).[2] When a model achieves high scores on such an OOD testset, authors then typically assume it has found a good solution for the task and that the model does not rely on spurious features to solve it. There is, however, no *a priori* reason why a model which generalises in one OOD testset should also generalise in testsets created under different distribution shifts.[3]

In this paper, we investigate whether generalisation results generalise. To this end, we analyse how a model's generalisation performances in different OOD testsets correlate across a single finetuning run. However, generalisation performances are bound to be trivially correlated due to their dependency on a common factor: the model's in-domain performance. We control for that factor by computing **partial OOD correlations** instead, regressing out in-domain performance. These partial correlations quantify how strongly generalisation performances correlate beyond their dependence on in-domain performance.

[1]Throughout this work, generalisation always refers to *out-of-distribution* generalisation.

[2]Two notable exceptions are discussed in §2.

[3]Furthermore, even within a single OOD testset, generalisation performance can be unstable, presenting large fluctuations across training (Mosbach et al., 2023).

Findings of the Association for Computational Linguistics: ACL 2026, pages 25763–25783
July 2-7, 2026 ©2026 Association for Computational Linguistics

Empirically, we show that whether generalisation performance will transfer across OOD testsets is a complex phenomenon. While a specific model's generalisation performance may be strongly correlated across two OOD testsets, it might present negative correlations under another pair. Similarly, while two OOD testsets may present positive partial correlations under one model, they may present negative correlations under another. This large variance in generalisation performance highlights that a proper generalisation evaluation *must* span multiple OOD testsets.

2 Evaluating OOD Generalisation

Robust generalisation beyond the training distribution has long been a challenge in natural language processing (NLP). In the quest to improve OOD generalisation, researchers face an important problem: how do we evaluate it in the first place?

How to evaluate OOD generalisation? Assessing generalisation performance is an intricate game of cat-and-mouse: as models tend to saturate on existing benchmarks, new ones are released to expose new weaknesses. McCoy et al. (2019), for instance, adversarially constructed an OOD testset (HANS) to reveal LLMs' reliance on superficial cues to solve a natural language inference (NLI) task. Similarly, Nie et al. (2020) constructed an OOD dataset (ANLI) in rounds to continually fool NLI models. Further, Liu et al. (2022) used models trained on MNLI (Williams et al., 2018) to generate their own synthetic (adversarial) datasets.

Do finetuned models generalise? Given all these benchmarks, we should have a good idea about how well language models generalise. However, the effect of finetuning on a model's OOD generalisation remains a little-understood topic. Kumar et al. (2022), for instance, show that finetuning models with randomly initialised classifier heads can lead to distorted features and hence poor generalisation. Recent empirical work (Mosbach et al., 2023; Yang et al., 2024), however, shows strong generalisation of finetuned models on OOD data. Both these works, however, use pattern-based fine-tuning (Schick and Schütze, 2021) instead of a randomly initialised classifier head, being thus not directly comparable to the findings of Kumar et al. A few prior works investigate multiple OOD testsets. E.g., Gupta et al. (2024) evaluate which OOD testsets still represent a challenge for finetuned models. Closest to our work is Sun et al. (2023), who compare the rankings achieved by finetuned models on a number of OOD testsets. More specifically, they compute the correlations across OOD testsets of the rankings achieved by several pretrained models when finetuned to perform NLI. Their analyses, however, do not control for either the models' in-domain performance, or the used pretrained models' size and quality. Instead, we will analyse partial correlations within each training run, controlling for both factors.

3 Measuring Correlations between OOD Generalisations

In this paper, we aim to assess how robust generalisation results are to a specific choice of OOD testset. We can quantify this by analysing how correlated generalisation results are across testsets. Language models with better in-domain performance, however, are also likely to perform better out-of-domain (Yang et al., 2023). Naively computing OOD correlations, thus, is likely to mostly capture this trivial (and arguably uninteresting) source of correlation. To control for this, we measure **partial OOD correlations** instead: the correlation between two OOD performances once in-domain performance has been regressed out.

Let $p_{\boldsymbol{\theta}}$ be a language model, which we finetune on a specific (in-domain) training set. While finetuning this model, we measure its **in-domain performance**, denoted $s_t^{\texttt{ind}}$, on an in-domain testset at several checkpoints t; this gives us a vector of performances $\mathbf{s}^{\texttt{ind}}$. Simultaneously, we measure this model's **out-of-domain performance**, denoted $s_t^{\texttt{ood}:d}$, on several out-of-domain testsets, d, using the same checkpoints; this gives us a vector of performances $\mathbf{s}^{\texttt{ood}:d}$. If we simply wanted to examine the correlation between OOD performances, we would evaluate: $\texttt{corr}(\mathbf{s}^{\texttt{ood}:d_1}, \mathbf{s}^{\texttt{ood}:d_2})$.

Computing the partial correlation between two OOD datasets, however, requires regressing out in-domain performance. To do this, we train a regression model $f^d : \mathbb{R} \to \mathbb{R}$ for each OOD dataset d, which, given an in-domain performance measurement, predicts that checkpoints' OOD performance: $s_t^{\texttt{ood}:d} \approx f^d(s_t^{\texttt{ind}})$.[4] Given this model, we compute a residual $e_t^d = s_t^{\texttt{ood}:d} - f^d(s_t^{\texttt{ind}})$, which quantifies how much better or worse a model performs on d

[4]When fitting these regressors, we discard the performance of the first two checkpoints in each fine-tuning run. This mitigates the impact of the initial, highly unstable training phases on the regressor and on the subsequent analyses.

Model	Size	Finetuned on MNLI							
		MNLI‡	SNLI	WNLI	SciTail	RTE	HANS	ANLI	PAWS
OPT	2.7b	*81.6 ± 5.9*	72.7 ± 17.8	49.9 ± 0.7	65.8 ± 5.9	62.5 ± 2.1	51.7 ± 2.9	50.5 ± 0.5	46.3 ± 1.2
	6.7b	*84.7 ± 6.7*	83.7 ± 12.6	50.7 ± 1.8	70.7 ± 10.9	64.3 ± 1.0	55.5 ± 7.8	49.2 ± 1.9	47.3 ± 0.8
	13b	*87.3 ± 5.6*	83.9 ± 13.4	50.9 ± 2.6	71.3 ± 7.6	67.9 ± 1.6	57.0 ± 7.3	52.3 ± 1.5	48.5 ± 2.5
	30b	*89.0 ± 5.9*	86.8 ± 14.2	50.7 ± 1.1	**74.7 ± 3.2**	**71.2 ± 3.2**	59.3 ± 6.2	53.0 ± 1.0	48.6 ± 1.5
OLMo2	7B	*75.1 ± 1.9*	67.1 ± 15.8	55.7 ± 2.6	55.0 ± 8.1	63.1 ± 5.6	52.7 ± 2.2	55.7 ± 4.7	59.1 ± 5.9
	13B	*61.2 ± 1.5*	56.7 ± 5.9	52.4 ± 1.0	57.0 ± 4.8	54.0 ± 1.9	51.3 ± 2.5	50.9 ± 0.5	54.7 ± 2.5
	32B	*87.4 ± 12.0*	81.8 ± 24.8	**68.2 ± 12.5**	53.7 ± 9.5	69.2 ± 4.0	61.5 ± 7.4	**68.3 ± 5.4**	66.9 ± 2.6
SmolLM	3B	***93.7 ± 0.3***	**94.8 ± 0.6**	57.7 ± 1.3	52.6 ± 3.6	68.4 ± 3.5	**63.4 ± 0.2**	62.9 ± 1.1	**68.7 ± 2.2**

Model	Size	Finetuned on SNLI							
		SNLI‡	MNLI	WNLI	SciTail	RTE	HANS	ANLI	PAWS
OPT	2.7b	*94.2 ± 0.2*	78.6 ± 3.1	50.4 ± 0.5	74.1 ± 2.8	66.4 ± 0.7	51.4 ± 1.4	50.6 ± 1.1	50.6 ± 3.8
	6.7b	*94.3 ± 1.3*	78.2 ± 6.4	52.2 ± 0.6	68.8 ± 13.3	66.0 ± 0.9	54.9 ± 2.4	51.8 ± 3.1	49.5 ± 2.3
	13b	*95.3 ± 0.4*	82.0 ± 4.0	49.9 ± 0.4	70.2 ± 4.2	65.3 ± 1.3	53.4 ± 3.0	51.8 ± 1.0	48.6 ± 2.3
	30b	*96.1 ± 0.1*	86.0 ± 3.9	52.0 ± 1.4	**76.8 ± 2.3**	70.7 ± 2.5	58.8 ± 5.6	53.0 ± 1.3	49.7 ± 4.1
OLMo2	7B	*90.6 ± 5.0*	70.7 ± 11.3	59.3 ± 3.4	56.9 ± 4.3	61.0 ± 4.3	61.8 ± 3.1	56.3 ± 3.7	64.6 ± 3.9
	13B	*80.4 ± 2.3*	61.4 ± 2.8	54.9 ± 0.2	54.7 ± 2.7	55.8 ± 0.7	57.3 ± 3.2	52.3 ± 1.9	54.8 ± 1.7
	32B	***98.0 ± 0.1***	84.1 ± 5.1	**73.9 ± 1.0**	60.7 ± 5.8	70.9 ± 1.2	66.7 ± 2.7	**65.4 ± 2.6**	69.6 ± 0.9
SmolLM	3B	*97.3 ± 0.1*	**90.5 ± 2.5**	66.0 ± 2.2	62.6 ± 4.4	**73.9 ± 2.3**	**68.8 ± 2.3**	65.0 ± 1.7	**76.3 ± 0.8**
Chance performance		50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0

Table 1: Accuracies for models trained on MNLI (top) and SNLI (bottom) with 128 examples over 3 independent runs. Measurements are taken using the checkpoint with the highest in-domain performance. ‡ in-domain dataset.

than what would be expected given its in-domain performance.[5] Doing this for all checkpoint steps t, we get a vector of residuals $\mathbf{e}^d$.

Finally, we compute the **partial correlations** we are interested in as:

$$\rho^{d_1,d_2} = \texttt{corr}(\mathbf{e}^{d_1}, \mathbf{e}^{d_2}) \,. \tag{1}$$

Since our focus is on observing joint improvement, we choose to capture simple linear correlations between these residuals, measuring Pearson's correlations as the $\texttt{corr}(\cdot)$ function. Throughout the paper, we present results using GAM regressors f^d.[6]

4 Experimental Setup

Task. We focus on natural language inference (NLI; Dagan and Glickman, 2004; Putra et al., 2024), as generalisation performance on this task has received considerable interest (Bhargava et al., 2021; Zhou and Tan, 2021; Mosbach et al., 2023; Gupta et al., 2024). NLI consists in determining the logical relationship between a pair of sentences. More specifically, each entry in this task consists of a pair of sentences, a premise and a hypothesis; the task is then to determine if the premise entails, contradicts, or is neutral about the hypothesis.

Models. We experiment with three different model families: OPT (Zhang et al., 2022), OLMo2 (OLMo et al., 2024) and SmolLM (Bakouch et al., 2025). We choose these models due to them being publicly available in multiple sizes, and due to their popularity in recent years for a broad range of NLP tasks. Beyond that, OPT also makes our experiments more easily comparable to previous work (e.g., Mosbach et al., 2023; Srinivasan et al., 2024). Following Mosbach et al. (2023), we finetune these models using: a few-shot setting, with 128, 64 and 32 examples; low-rank adaptation (LoRA; Hu et al., 2022); and pattern-based finetuning (Schick and Schütze, 2021; Gao et al., 2021), reusing the pretrained LM head instead of using a randomly initialised one. More details can be found in App. A.

Data. We select 8 different NLI datasets: **SNLI** (Bowman et al., 2015), **MNLI** (Williams et al., 2018), **SciTail** (Khot et al., 2018), **WNLI** and **RTE** (Wang et al., 2018), **PAWS** (Zhang et al., 2019), **HANS** (McCoy et al., 2019), **ANLI** (Nie et al., 2020). We finetune our models on either SNLI or MNLI and evaluate them on the 7 other OOD datasets. More details can be found in App. B.

[5]Interestingly, these residuals were previously analysed by Taori et al. (2020) under the name of **effective robustness**.

[6]We experiment with both linear and GAM (Hastie and Tibshirani, 1986) regressors. Fig. 7 shows the fit of in-domain vs. out-of-domain curves for all regressors, and Fig. 10 to 12 show partial correlations using linear regressors (both in App. D).

5 Results

Finetuned models don't generalise everywhere. Table 1 presents the generalisation performance, across all testsets, of the models finetuned on MNLI and SNLI. These tables present performances for a single checkpoint per model, where checkpoints were selected based on in-domain performance. Our results show that no testset is challenging for all models: every testset has at least one model that generalises successfully. Furthermore, it also shows that finetuning produces models that often perform well across a range of OOD testsets. However, for any given model, there is always at least one testset at which it underperforms. For instance, when finetuned on SNLI, the same OPT 30B checkpoint achieves 86.0% accuracy on MNLI, but 49.7% on PAWS. This variability highlights a key limitation of single-testset evaluations. Additionally, naïvely looking at these tables might lead one to conclude that generalisation results are mostly robust: OPT 30B, OLMo2 32B, and (perhaps more surprisingly) SmolLM 3B consistently achieve the best in-domain and out-of-domain performance across all setups. This conclusion, however, is not necessarily warranted, as both model size and in-domain performance act as strong confounders. We now look at how the generalisation performance of each of these models fluctuates throughout training, as a way to control for the effect of model size on results.

OPT's generalisation performance oscillates, but OLMo2's doesn't. Fig. 2 presents OPT 30B's and OLMo2 32B's OOD generalisation performances across training. (Results for smaller OPT and OLMo2 models are in Fig. 6, in App. D.) Overall, these figures reproduce an important result in Mosbach et al. (2023), showing that OPT's generalisation performance is unstable throughout training, with large (mostly unpredictable) oscillations. Interestingly, OLMo2's performance does not present the same oscillations. Perhaps more important for our research question though, we find that generalisation in some OOD testsets tracks the others; this is most obvious for the results of OPT 30B trained on SNLI. In-domain performance, however, also tracks OOD generalisation in these results—at least to some extent. Next, we thus move to analysing partial correlations as introduced in §3.

Generalisation's generalisation is complicated. Fig. 3 presents the partial correlations across OOD testsets for OPT, SmolLM and OLMo2 models. We observe that OOD generalisation is a highly complex property for which no clear trend emerges across testsets. While for a model two OOD testsets might present strong positive partial correlations, for another model this correlation might be negative. Additional intuition can also be drawn from Fig. 16 (in App. D), which shows that partial correlations do not seem to strengthen with model size or with a particular choice of finetuning dataset (i.e., MNLI vs. SNLI). These findings underscore the importance of conducting a comprehensive evaluation when making claims about a model's generalisation capabilities, an often-lacking aspect in the current literature.

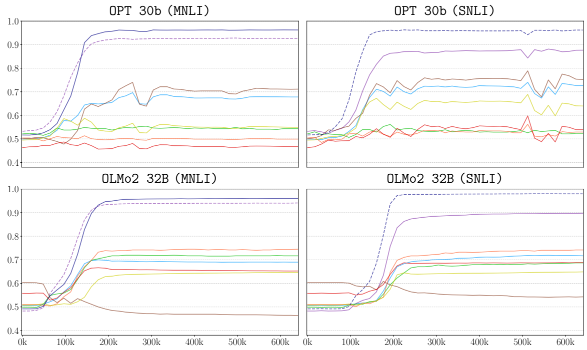


Figure 2: Accuracy (y-axis) across training steps (x-axis) of OPT (top) and OLMo2 (bottom) for a single finetuning run on MNLI (left) and SNLI (right). Legend: MNLI, SNLI, WNLI, RTE, SciTail, ANLI, HANS and PAWS.

More stable generalisation in newer models. Fig. 4 compares the average partial correlation scores achieved by OPT and SmolLM. Interestingly, this figure shows that SmolLM not only outperforms previous generations in terms of accuracy (as seen previously in Table 1), but also exhibits consistently stronger partial correlations. These higher correlations indicate that SmolLM possesses more stable generalisation patterns, presenting more correlated trends across OOD testsets.

Regressor choice matters. Finally, we present the average Mean Absolute Error (MAE) and R^2 of our regressors in Tables 2 and 3. We find that GAM regressors achieve the strongest fit and most effectively model the relationship between in-domain vs. OOD performances. Furthermore, unlike the linear models, GAM regressors' residuals show no clear trends as a function of in-domain performance (see Fig. 5 as an example) achieving the desired goal of controlling for this confounding factor.

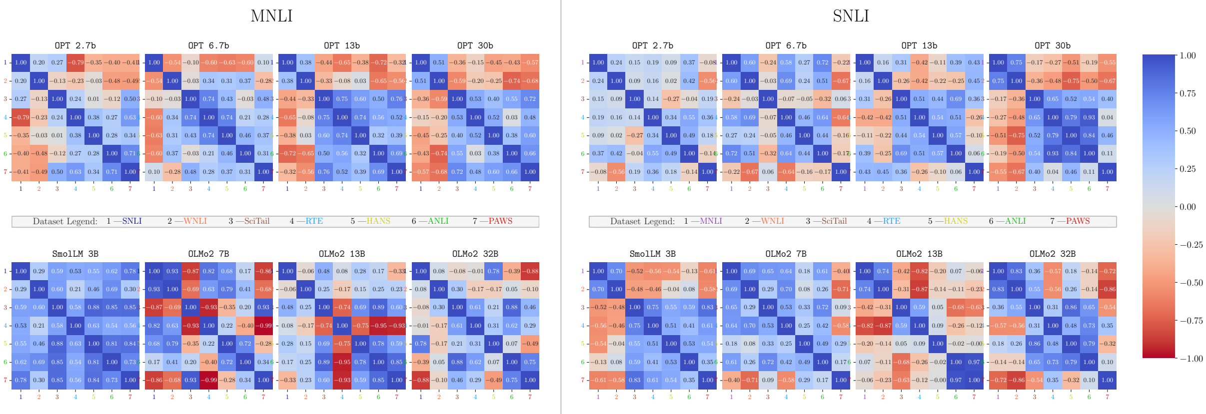


Figure 3: Partial correlations of `OPT` (top), `SmolLM` (bottom) and `OLMo2` (also bottom) across model sizes (ordered from left to right) trained on `MNLI` (left) and `SNLI` (right). All these correlations are obtained by fitting a `GAM` regressor over 3 independent training runs. See a larger version of this plot in Fig. 13 (in App. D).

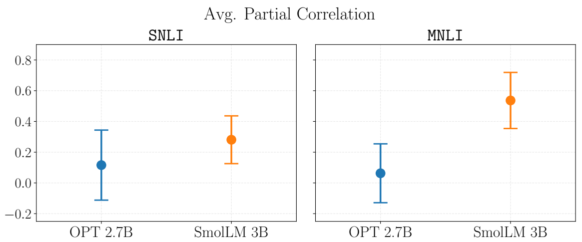


Figure 4: Average partial correlations of `OPT` and `SmolLM`. `SmolLM` consistently achieves higher (positive) correlations, indicating more stable OOD generalisation patterns across testset. Results for 128-shot training runs; see Fig. 18 for 32- and 64-shot models.

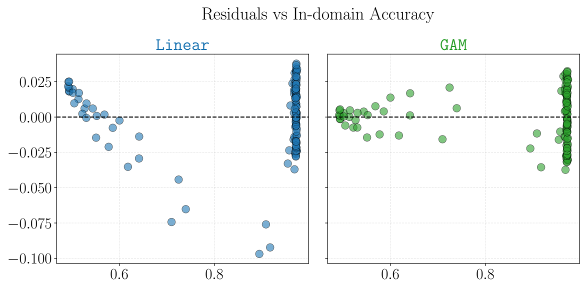


Figure 5: Residuals of `Linear` and `GAM` regressors vs. in-domain performance. Results shown for regressor trained on `RTE` with a `SmolLM 3B` model finetuned with 128-shot on `SNLI`.

	Mean Absolute Error (MAE)			R^2 Scores		
Regressor	**Mean**	**Median**	**Std**	**Mean**	**Median**	**Std**
`Linear`	0.0193	0.0152	0.0151	0.446	0.391	0.337
`Ridge`	0.0220	0.0183	0.0164	0.332	0.188	0.328
`GAM`	**0.0136**	**0.0102**	0.0106	**0.624**	**0.646**	0.287

Table 2: Summary statistics for regressor quality across all fitted regressors. Lower MAE indicates better fit, while higher R^2 indicates better uncertainty modeling.

	`MNLI`	`WNLI`	`SciTail`	`RTE`	`HANS`	`ANLI`	`PAWS`
MAE	3.950	0.770	0.742	0.355	0.620	0.420	0.630
R^2	0.230	0.833	0.492	0.857	0.789	0.622	0.607

Table 3: MAE ($\times 100$ for clarity) and R^2 scores of `GAM` regressors for a specific model: `OLMo2 32B` trained on 32-shot `SNLI`, averaged by OOD dataset. The `MNLI` regressor exhibits much poorer performance than others.

6 Conclusions

In this paper, we analyse whether generalisation results generalise across OOD testsets by investigating their partial correlations. Notably, we do not observe clear trends in these partial correlations: no clear trends arise when comparing different training datasets, model families or model sizes. In fact, the partial correlation on a pair of OOD testsets appears to not be an intrinsic property even of the testset pair itself, depending on the specific model and training dataset considered. Our results thus highlight that, to ensure accurate evaluations, generalisation research must use several OOD testsets.

Limitations

Our paper suffers from a series of limitations, which we highlight here. First, and perhaps most importantly, our experiments focus exclusively on NLI. Unfortunately, the lack of dedicated OOD testsets for other tasks makes it difficult to study the extent to which our findings are NLI-specific. We hope future work will generalise our analyses to other tasks. Second, it is possible that our studied models (`OPT`, `OLMo2`, `SmolLM`) could have been exposed to the analysed testsets during pretraining. We conducted preliminary experiments using `Min-k%++` (Zhang et al., 2025) and `Time Travel in LLMs` (Golchin and Surdeanu, 2024) to investigate such data contamination, but these experiments were inconclusive in most cases.[7] Despite the negative results, though, our analyses suggest that the models have not outright memorised the OOD testsets, which would allow them to trivialise the task. Third, we investigated the relationship between the regressors' R^2 scores across datasets (detailed in Table 3) and the resulting partial OOD-OOD correlations. Specifically, computing the Pearson correlation between these R^2 scores and the partial correlation coefficients yields a positive correlation ($r = 0.4$, significant at $p < 0.01$). At face value, this may indicate that an improvement to the regressors could also improve partial correlations. We refrain from making any strong claims on this topic, as R^2 scores are also discounted by the target variable's variance (in this case the OOD performance), and thus do not solely depend on regressor quality. Although our results on aggregate regressor quality support our choice of GAM regressors, the results across datasets make a strong case for further exploration of the relationship between regressor quality, variance in OOD performance and partial OOD-OOD correlations. While this topic is not explored here, we consider it an interesting direction for future work. Finally, the largest models analysed here had roughly 30B parameters. It would be interesting to investigate if the lack of trends observed here would carry to other model families and to larger sizes.

References

Elie Bakouch, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Lewis Tunstall, Carlos Miguel Patiño, Edward Beeching, Aymeric Roucher, Aksel Joonas Reedi, Quentin Gallouédec, Kashif Rasul, Nathan Habib, Clémentine Fourrier, Hynek Kydlicek, Guilherme Penedo, Hugo Larcher, Mathieu Morlon, Vaibhav Srivastav, Joshua Lochner, and 4 others. 2025. SmolLM3: smol, multilingual, long-context reasoner. https://huggingface.co/blog/smollm3.

Prajjwal Bhargava, Aleksandr Drozd, and Anna Rogers. 2021. Generalization in NLI: Ways (not) to go beyond simple heuristics. In *Proceedings of the Second Workshop on Insights from Negative Results in NLP*, pages 125–135, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Ido Dagan and Oren Glickman. 2004. Probabilistic textual entailment: Generic applied modeling of language variability. *Learning Methods for Text Understanding and Mining*, 2004(26-29):2–5.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Shahriar Golchin and Mihai Surdeanu. 2024. Time travel in LLMs: Tracing data contamination in large language models. In *The Twelfth International Conference on Learning Representations*.

Ashim Gupta, Rishanth Rajendhran, Nathan Stringham, Vivek Srikumar, and Ana Marasovic. 2024. Whispers of doubt amidst echoes of triumph in NLP robustness. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5533–5590, Mexico City, Mexico. Association for Computational Linguistics.

Trevor Hastie and Robert Tibshirani. 1986. Generalized additive models. *Statistical Science*, 1(3):297–310.

Edward J. Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, Dennis Ulmer, Florian Schottmann, Khuyagbaatar Batsuren, Kaiser Sun, Koustuv Sinha, Leila Khalatbari, Maria Ryskina, Rita Frieske, Ryan Cotterell, and Zhijing Jin. 2023. A taxonomy and review of generalization research in NLP. *Nature Machine Intelligence*, 5(10):1161–1174.

Nitish Joshi and He He. 2022. An investigation of the (in)effectiveness of counterfactually augmented data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3668–3681, Dublin, Ireland. Association for Computational Linguistics.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. Scitail: A textual entailment dataset from science question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

[7] Although OLMo2 is trained on open datasets, directly testing for contamination on such a big dataset was a bigger challenge than anticipated, with most solutions relying on massive indexes for lookups (Vu et al., 2023).

Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. 2022. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. WANLI: Worker and AI collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Marius Mosbach, Tiago Pimentel, Shauli Ravfogel, Dietrich Klakow, and Yanai Elazar. 2023. Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12284–12314, Toronto, Canada. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, and 21 others. 2024. 2 OLMo 2 furious. *Preprint*, arXiv:2501.00656.

Buu Phan, Marton Havasi, Matthew J. Muckley, and Karen Ullrich. 2024. Understanding and mitigating tokenization bias in language models. In *ICML 2024 Workshop on Theoretical Foundations of Foundation Models*.

Tiago Pimentel and Clara Meister. 2024. How to compute the probability of a word. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18358–18375, Miami, Florida, USA. Association for Computational Linguistics.

I Made Suwija Putra, Daniel Siahaan, and Ahmad Saikhu. 2024. Recognizing textual entailment: A review of resources, approaches, applications, and challenges. *ICT Express*, 10(1):132–155.

Mario Sanz-Guerrero, Minh Duc Bui, and Katharina von der Wense. 2025. Mind the gap: A closer look at tokenization for multiple-choice question answering with LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19584–19594, Suzhou, China. Association for Computational Linguistics.

Timo Schick, Helmut Schmid, and Hinrich Schütze. 2020. Automatically identifying words that can serve as labels for few-shot text classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5569–5578, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Krishna Prasad Varadarajan Srinivasan, Prasanth Gumpena, Madhusudhana Yattapu, and Vishal H Brahmbhatt. 2024. Comparative analysis of different efficient fine tuning methods of large language models (LLMs) in low-resource setting. *arXiv preprint arXiv:2405.13181*.

Kaiser Sun, Adina Williams, and Dieuwke Hupkes. 2023. The validity of evaluation results: Assessing concurrence across compositionality benchmarks. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 274–293, Singapore. Association for Computational Linguistics.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. 2020. Measuring robustness to natural distribution shifts in image classification. In *Advances in Neural Information Processing Systems*, volume 33, pages 18583–18599. Curran Associates, Inc.

Thuy Vu, Xuanli He, Gholamreza Haffari, and Ehsan Shareghi. 2023. Koala: An index for quantifying overlaps with pre-training corpora. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 90–98.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Runxin Xu, Fuli Luo, Zhiyuan Zhang, Chuanqi Tan, Baobao Chang, Songfang Huang, and Fei Huang. 2021. Raise a child in large language model: Towards effective and generalizable fine-tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9514–9528.

Haoran Yang, Yumeng Zhang, Jiaqi Xu, Hongyuan Lu, Pheng-Ann Heng, and Wai Lam. 2024. Unveiling the generalization power of fine-tuned large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 884–899, Mexico City, Mexico. Association for Computational Linguistics.

Linyi Yang, Shuibai Zhang, Libo Qin, Yafu Li, Yidong Wang, Hanmeng Liu, Jindong Wang, Xing Xie, and Yue Zhang. 2023. GLUE-X: Evaluating natural language understanding models from an out-of-distribution generalization perspective. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12731–12750.

Rem Yang, Julian Dai, Nikos Vasilakis, and Martin Rinard. 2025. Evaluating the generalization capabilities of large language models on code reasoning. *arXiv preprint arXiv:2504.05518*.

Qinyuan Ye. 2024. Cross-task generalization abilities of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 255–262, Mexico City, Mexico. Association for Computational Linguistics.

Lifan Yuan, Yangyi Chen, Ganqu Cui, Hongcheng Gao, Fangyuan Zou, Xingyi Cheng, Heng Ji, Zhiyuan Liu, and Maosong Sun. 2023. Revisiting out-of-distribution robustness in NLP: Benchmark, analysis, and LLMs evaluations. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Frank Yang, and Hai Li. 2025. Min-K%++: Improved baseline for pre-training data detection from large language models. In *The Thirteenth International Conference on Learning Representations*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: Open pre-trained transformer language models. *Preprint*, arXiv:2205.01068.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308.

Yangqiaoyu Zhou and Chenhao Tan. 2021. Investigating the effect of natural language explanations on out-of-distribution generalization in few-shot NLI. In *Proceedings of the Second Workshop on Insights from Negative Results in NLP*, pages 117–124, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Pattern-based finetuning details

Pattern-based finetuning requires us to specify an input pattern and define a mapping between the answer tokens and the actual labels (Schick et al., 2020). Our experiments with NLI use the following pattern:

```
{premise} Question: {hypothesis}
Yes or No?
```

The target tokens we consider are, respectively, `'_Yes'` for entailment and `'_No'` otherwise.[8]

B NLI Datasets

We use 2 main large-scale datasets for finetuning the models: **SNLI** (Bowman et al., 2015), which contains 570K crowdsourced sentence-pairs based on image captions; and **MNLI** (Williams et al., 2018), which is a set of 433K sentence-pairs meant to cover a large range of genres of spoken and written text. Compared to SNLI, MNLI offers more linguistic diversity and difficulty as it includes representative samples from 10 distinct genres of written and spoken English. We assessed the generalisation capacity of fine-tuned models using 6 NLI testsets. These comprise 3 adversarial datasets—designed especially to evaluate the models' robustness to heuristics—as well as 3 more standard NLI datasets with various but comparable input distributions:

- *Standard*: **SciTail** (Khot et al., 2018) is based on science multiple-choice exams, **WNLI** focuses on identifying the referent of a certain pronoun and **RTE** is a general entailment dataset. These last two are a part of the GLUE Benchmark (Wang et al., 2018).

[8] The underscores indicate a whitespace in the token. This is important to guarantee that the correct token-string representing this character-string is considered (Pimentel and Meister, 2024; Phan et al., 2024), which may impact prompting performance (Sanz-Guerrero et al., 2025).

- *Adversarial*: **PAWS** (Zhang et al., 2019) uses paraphrase adversaries, **HANS** (McCoy et al., 2019) tackles failure cases of 3 simple heuristics and **ANLI** (Nie et al., 2020) finds adversaries via human feedback.

To avoid inconsistencies that can result from different annotation policies among datasets, we removed the neutral-labeled samples, enabling us to more effectively separate the impacts of domain shifts on model performance, and guaranteeing a more consistent assessment framework.

C Resource Usage

We ran our experiments on various machines, depending on memory requirements. Small models were trained on 4x A5000 GPUs (with 24GB each), larger models were trained using 8x A6000 (with 48GB each) or 4x A100 (with 80GB). The total runtime for all the experiments presented here is 5,500 GPU hours.

D Detailed Results

D.1 Performance in the 32 and 64-shot settings on SNLI

Model	Size	SNLI							
		SNLI‡	MNLI	WNLI	SciTail	RTE	HANS	ANLI	PAWS
OPT	2.7b	*65.2 ± 5.1*	58.3 ± 5.9	51.2 ± 0.3	61.2 ± 6.3	52.1 ± 2.0	52.0 ± 1.6	51.3 ± 1.3	51.9 ± 3.2
	6.7b	*69.7 ± 3.8*	59.3 ± 6.3	51.4 ± 0.9	64.4 ± 6.7	54.0 ± 2.7	53.3 ± 1.4	50.3 ± 0.7	51.5 ± 4.3
	13b	*82.9 ± 9.3*	71.9 ± 2.7	51.6 ± 0.6	**66.4 ± 1.7**	62.7 ± 2.8	57.1 ± 4.0	50.2 ± 1.6	53.1 ± 3.2
	30b	*75.8 ± 8.2*	62.5 ± 8.5	51.4 ± 1.2	60.9 ± 10.9	54.5 ± 7.5	53.4 ± 5.5	50.8 ± 1.3	50.9 ± 4.9
OLMo2	7B	*56.7 ± 3.4*	53.6 ± 0.2	51.0 ± 0.6	49.3 ± 5.7	52.1 ± 0.3	49.9 ± 0.6	51.0 ± 0.9	53.1 ± 1.5
	13B	*52.6 ± 1.4*	52.7 ± 5.5	51.3 ± 0.3	56.8 ± 1.1	51.3 ± 0.4	52.5 ± 1.2	50.6 ± 0.4	52.3 ± 0.4
	32B	*67.7 ± 8.8*	59.0 ± 5.5	54.2 ± 3.8	61.5 ± 1.8	53.9 ± 1.5	52.6 ± 2.7	52.9 ± 1.2	57.0 ± 1.7
SmolLM	3B	***89.4 ± 7.4***	**77.4 ± 7.8**	**62.8 ± 3.8**	61.3 ± 3.3	**66.4 ± 5.6**	**66.0 ± 5.4**	**58.9 ± 4.4**	**66.2 ± 7.5**
Chance performance		50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0

Table 4: Accuracy on each OOD dataset for models trained on SNLI with 32 examples. Measurements are taken using the checkpoint with the highest in-domain performance. ‡ in-domain dataset.

Model	Size	SNLI							
		SNLI‡	MNLI	WNLI	SciTail	RTE	HANS	ANLI	PAWS
OPT	2.7b	*87.5 ± 6.2*	71.8 ± 5.8	51.7 ± 0.5	69.9 ± 4.5	59.0 ± 5.7	52.5 ± 1.0	50.4 ± 1.7	51.5 ± 4.2
	6.7b	*88.3 ± 4.7*	72.7 ± 9.0	52.7 ± 1.9	62.1 ± 16.5	61.4 ± 2.8	54.3 ± 2.8	51.3 ± 2.4	49.4 ± 2.9
	13b	*93.5 ± 0.9*	80.8 ± 4.7	50.6 ± 1.0	72.4 ± 5.1	66.1 ± 0.3	54.4 ± 3.9	49.9 ± 0.9	52.1 ± 5.1
	30b	*94.5 ± 1.3*	78.8 ± 4.2	54.1 ± 1.7	**76.3 ± 1.8**	67.2 ± 6.4	64.7 ± 4.0	51.6 ± 2.2	53.0 ± 4.9
OLMo2	7B	*70.3 ± 12.5*	56.3 ± 5.1	52.8 ± 0.9	52.3 ± 5.6	53.5 ± 1.6	52.4 ± 2.1	51.8 ± 0.7	56.7 ± 2.9
	13B	*59.7 ± 5.2*	54.5 ± 5.0	52.8 ± 1.0	54.7 ± 4.1	52.2 ± 0.3	53.6 ± 1.0	50.9 ± 0.4	52.4 ± 1.7
	32B	*92.7 ± 4.0*	72.1 ± 8.7	61.8 ± 7.2	61.0 ± 4.4	61.1 ± 3.6	60.7 ± 1.7	57.7 ± 4.4	61.7 ± 3.6
SmolLM	3B	***96.8 ± 0.1***	**88.5 ± 2.3**	**67.0 ± 2.0**	63.9 ± 6.4	**73.9 ± 1.6**	**70.2 ± 2.1**	**62.7 ± 2.1**	**75.2 ± 1.1**
Chance performance		50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0

Table 5: Accuracy on each OOD dataset for models trained on SNLI with 64 examples. Measurements are taken using the checkpoint with the highest in-domain performance. ‡ in-domain dataset.

D.2 Performance in the 32 and 64-shot settings on MNLI

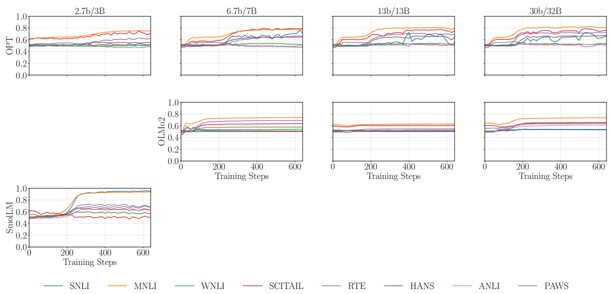
Model	Size	MNLI							
		MNLI‡	SNLI	WNLI	SciTail	RTE	HANS	ANLI	PAWS
OPT	2.7b	*58.8 ± 4.6*	52.8 ± 4.4	51.4 ± 0.6	59.1 ± 4.2	52.0 ± 2.8	51.4 ± 0.3	50.3 ± 2.3	49.7 ± 4.3
	6.7b	*65.1 ± 6.9*	58.0 ± 9.3	50.4 ± 1.4	59.3 ± 1.1	52.7 ± 3.4	52.1 ± 1.2	51.0 ± 2.0	51.4 ± 5.1
	13b	*68.1 ± 10.0*	59.6 ± 15.8	49.9 ± 0.3	**64.1 ± 6.4**	55.4 ± 8.0	53.4 ± 0.9	50.0 ± 1.9	52.1 ± 6.0
	30b	*68.2 ± 6.3*	**60.5 ± 15.3**	51.2 ± 0.5	57.5 ± 4.2	54.5 ± 4.6	52.3 ± 3.8	52.4 ± 3.2	52.9 ± 4.2
OLMo2	7B	*57.9 ± 4.7*	50.8 ± 1.6	50.7 ± 0.6	52.6 ± 3.9	51.0 ± 0.8	51.2 ± 0.6	50.9 ± 0.8	53.3 ± 0.4
	13B	*54.0 ± 5.5*	50.3 ± 0.5	51.4 ± 0.4	56.1 ± 0.7	52.0 ± 0.9	52.2 ± 0.5	48.6 ± 1.4	50.5 ± 1.7
	32B	***70.6 ± 7.5***	60.3 ± 8.4	**55.6 ± 4.2**	56.0 ± 1.9	**58.4 ± 6.0**	**55.0 ± 3.1**	**57.5 ± 5.1**	**58.8 ± 3.8**
SmolLM	3B	*58.5 ± 7.0*	53.4 ± 4.9	52.8 ± 1.1	59.4 ± 4.8	55.1 ± 3.1	52.0 ± 1.1	51.0 ± 0.5	55.2 ± 2.4
Chance performance		50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0

Table 6: Accuracy on each OOD dataset for models trained on MNLI with 32 examples. Measurements are taken using the checkpoint with the highest in-domain performance. ‡ in-domain dataset.

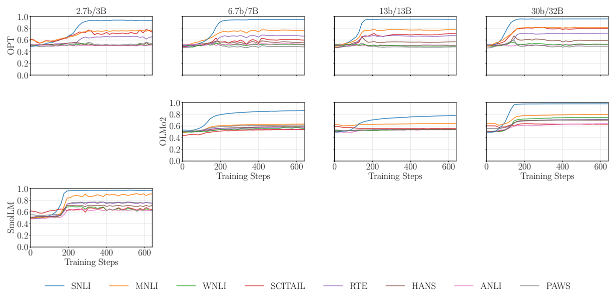
Model	Size	MNLI							
		MNLI‡	SNLI	WNLI	SciTail	RTE	HANS	ANLI	PAWS
OPT	2.7b	*67.2 ± 2.9*	59.4 ± 7.7	50.9 ± 0.3	59.6 ± 6.0	54.5 ± 2.1	51.7 ± 1.5	50.0 ± 0.9	48.3 ± 3.7
	6.7b	*74.4 ± 6.6*	66.7 ± 15.0	50.3 ± 1.2	63.8 ± 4.0	57.5 ± 4.7	54.5 ± 3.5	50.6 ± 0.6	50.6 ± 4.7
	13b	*79.7 ± 8.8*	75.6 ± 22.6	51.0 ± 1.7	**73.9** ± **3.1**	63.5 ± 4.2	55.0 ± 2.5	50.1 ± 2.2	50.7 ± 3.1
	30b	*82.9 ± 8.7*	75.0 ± 23.0	51.8 ± 2.1	63.1 ± 7.5	62.0 ± 2.4	57.5 ± 1.7	52.9 ± 1.7	48.7 ± 2.8
OLMo2	7B	*59.9 ± 3.2*	54.0 ± 3.8	50.5 ± 0.1	52.1 ± 4.8	51.9 ± 1.5	51.0 ± 1.7	51.0 ± 1.7	54.2 ± 1.5
	13B	*56.3 ± 4.5*	52.4 ± 2.4	50.3 ± 0.6	57.0 ± 2.1	51.8 ± 1.0	51.6 ± 1.3	50.7 ± 2.1	51.4 ± 2.7
	32B	*82.5 ± 12.5*	76.6 ± 21.4	**64.9** ± **11.1**	55.5 ± 5.1	64.8 ± 9.4	60.4 ± 5.8	**64.1** ± **9.2**	64.2 ± 5.3
SmolLM	3B	***90.3*** ± ***0.3***	**92.2** ± **1.1**	61.4 ± 1.4	53.8 ± 2.1	**68.8** ± **1.8**	**68.1** ± **0.8**	63.5 ± 0.9	**70.1** ± **1.4**
Chance performance		50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0

Table 7: Accuracy on each OOD dataset for models trained on MNLI with 64 examples. Measurements are taken using the checkpoint with the highest in-domain performance. ‡ in-domain dataset.

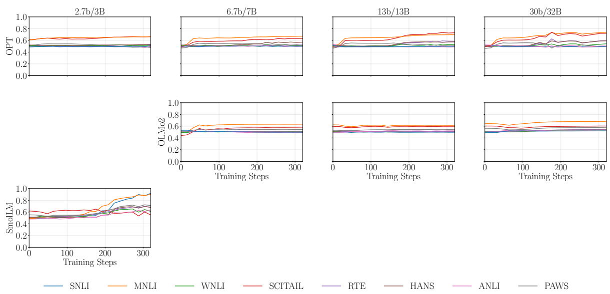
D.3 Performance across Finetuning Runs



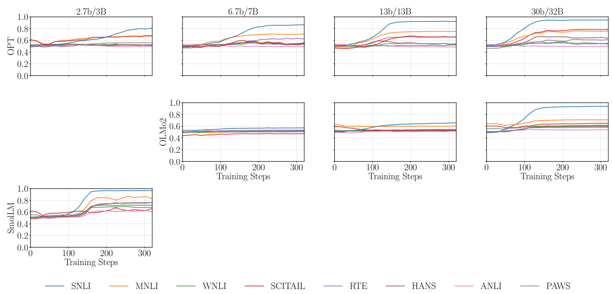
(a) 128-shot on MNLI



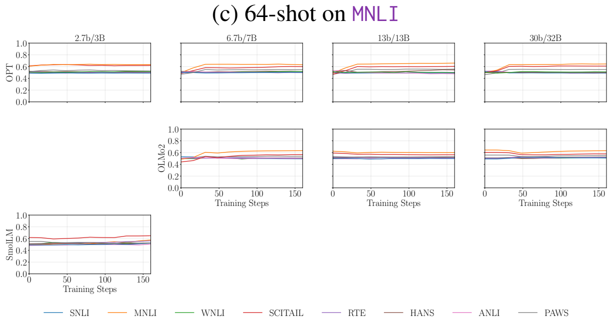
(b) 128-shot on SNLI



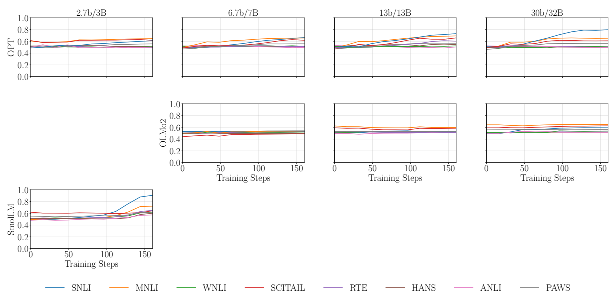
(c) 64-shot on MNLI



(d) 64-shot on SNLI



(e) 32-shot on MNLI



(f) 32-shot on SNLI

Figure 6: Few-shot results throughout a finetuning run on either MNLI or SNLI. OPT OOD performances (first rows) frequently oscillate during training; OLMo2 OOD performances (second rows) are relatively stable across training. SmolLM OOD performances (third rows) show generally stronger generalisation capabilities. Legend: MNLI, SNLI, WNLI, RTE, SciTail, ANLI, HANS and PAWS

D.4 Fit of Regressors Used when Computing Partial Correlations

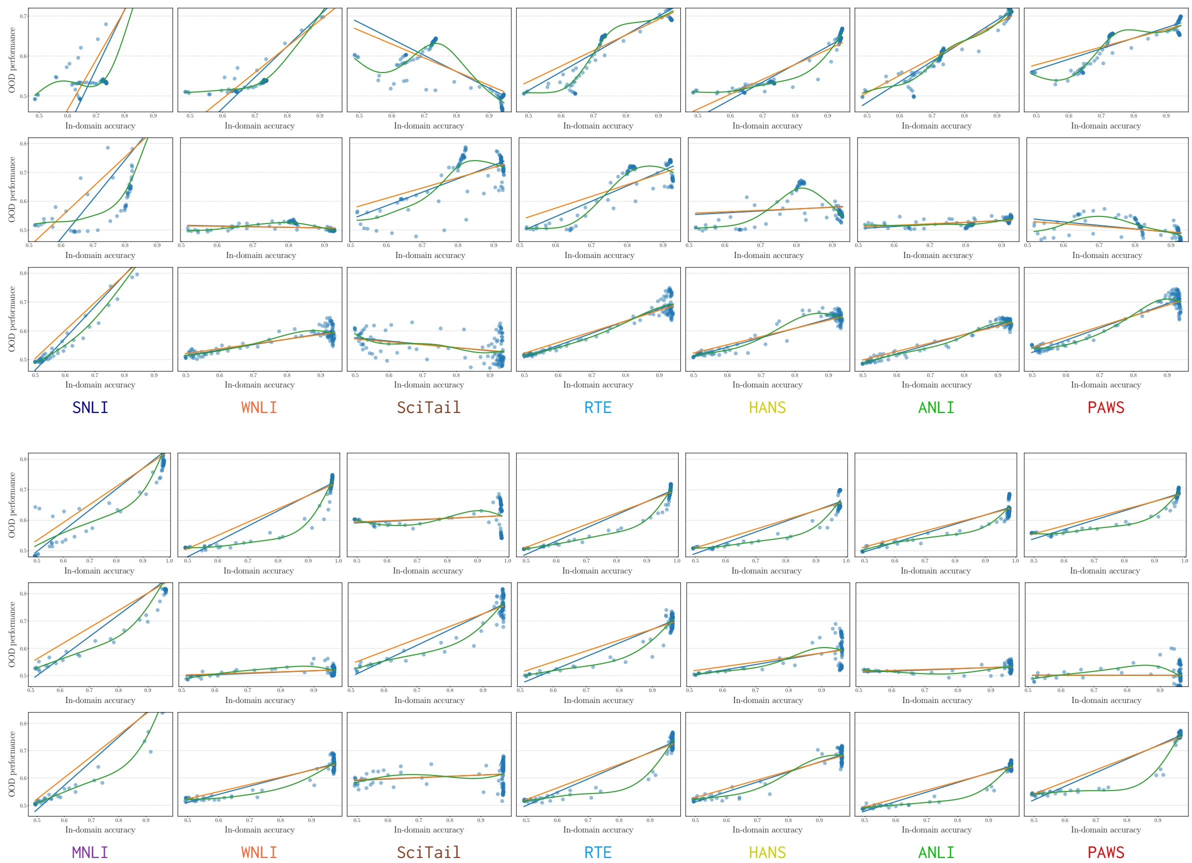


Figure 7: Regressors trained to predict OOD performance for 128-shot models. Models were finetuned on MNLI (top) and SNLI (bottom). Results for OLMo2 32B on first and fourth rows, OPT 30B on second and fifth rows, SmolLM 3B on third and sixth rows. Legend: Linear, Ridge and GAM.

D.5 OPT's Partial OOD Correlation Graphs

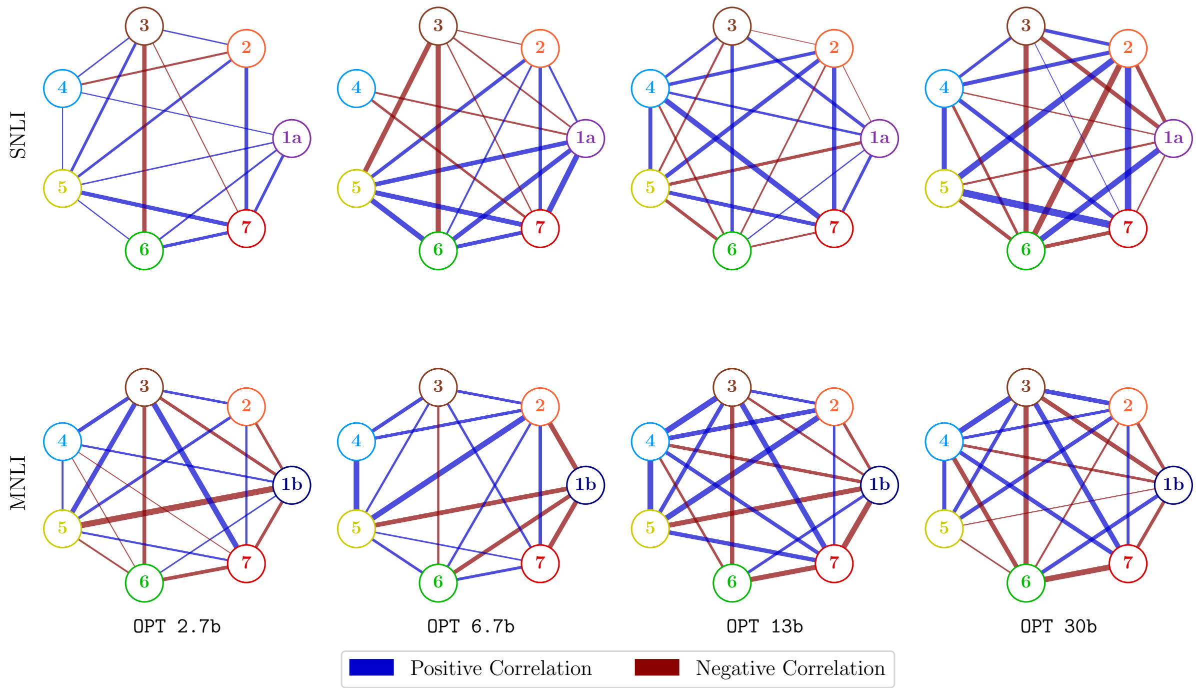


Figure 8: OPT partial OOD correlation graphs on SNLI (top) and MNLI (bottom). Edge thickness increases with absolute correlation value. Legend: 1a.MNLI, 1b.SNLI, 2.WNLI, 3.SciTail, 4.RTE, 5.HANS, 6.ANLI, and 7.PAWS

D.6 SmolLM's Partial OOD Correlation Graphs

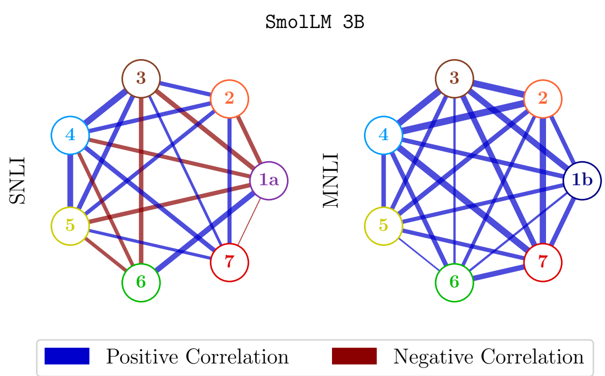


Figure 9: SmolLM partial OOD correlation graphs on SNLI (left) and MNLI (right). Edge thickness increases with absolute correlation value. Legend: 1a.MNLI,1b.SNLI, 2.WNLI, 3.SciTail, 4.RTE, 6.ANLI, 5.HANS, and 7.PAWS

D.7 Heatmaps with Partial OOD Correlations using Linear Regressors

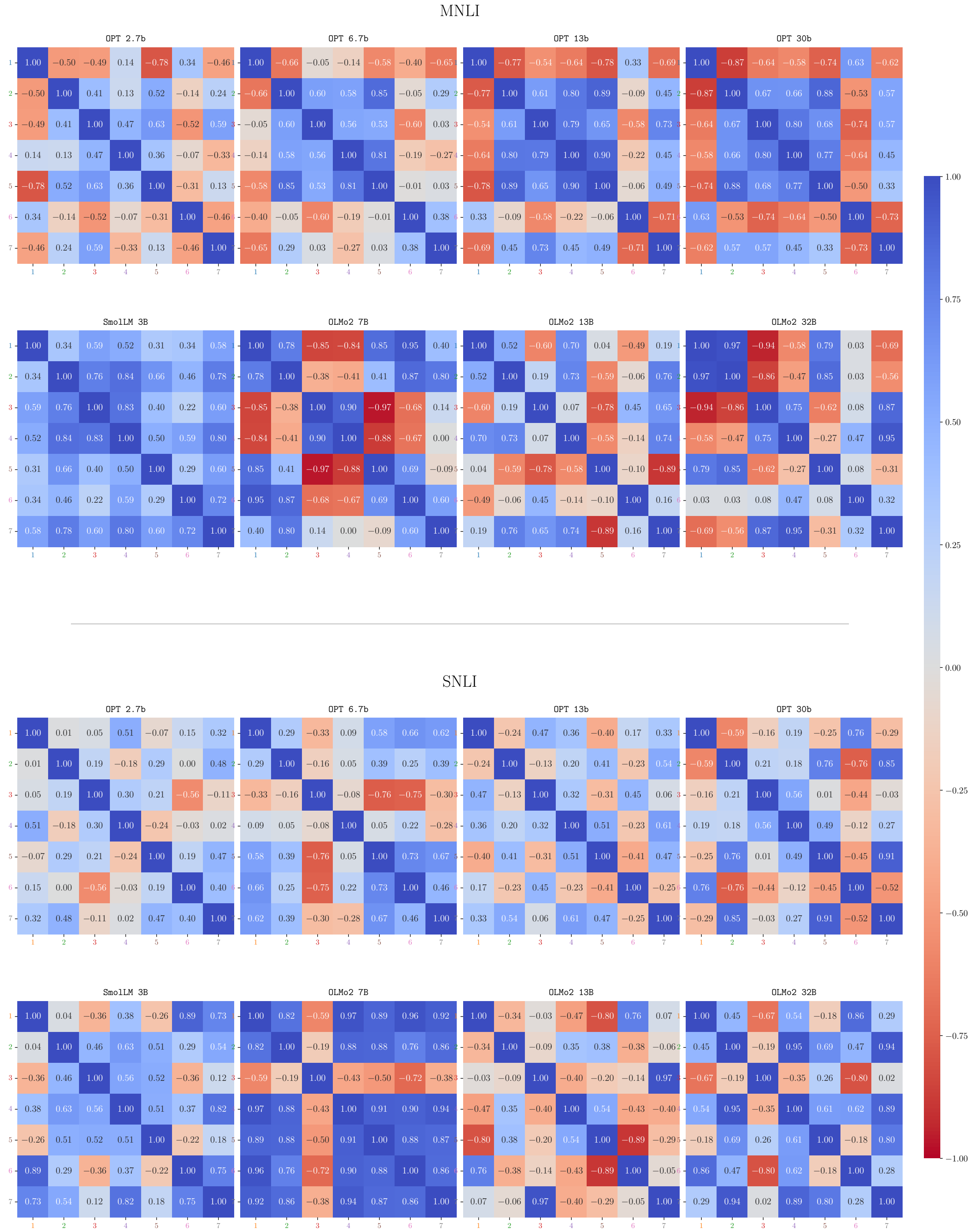


Figure 10: Partial correlations taken with `Linear` regressors of `OPT` (first and third rows), `SmolLM` (second and fourth rows, left) and `OLMo2` (second and fourth rows, center and right) across model sizes (ordered from left to right) trained on `MNLI` (top) and `SNLI` (bottom). Models were fine-tuned with 128-shot. Legend: 1. `SNLI`/`MNLI`, 2. `WNLI`, 3. `SciTail`, 4. `RTE`, 5. `HANS`, 6. `ANLI`, and 7. `PAWS`.

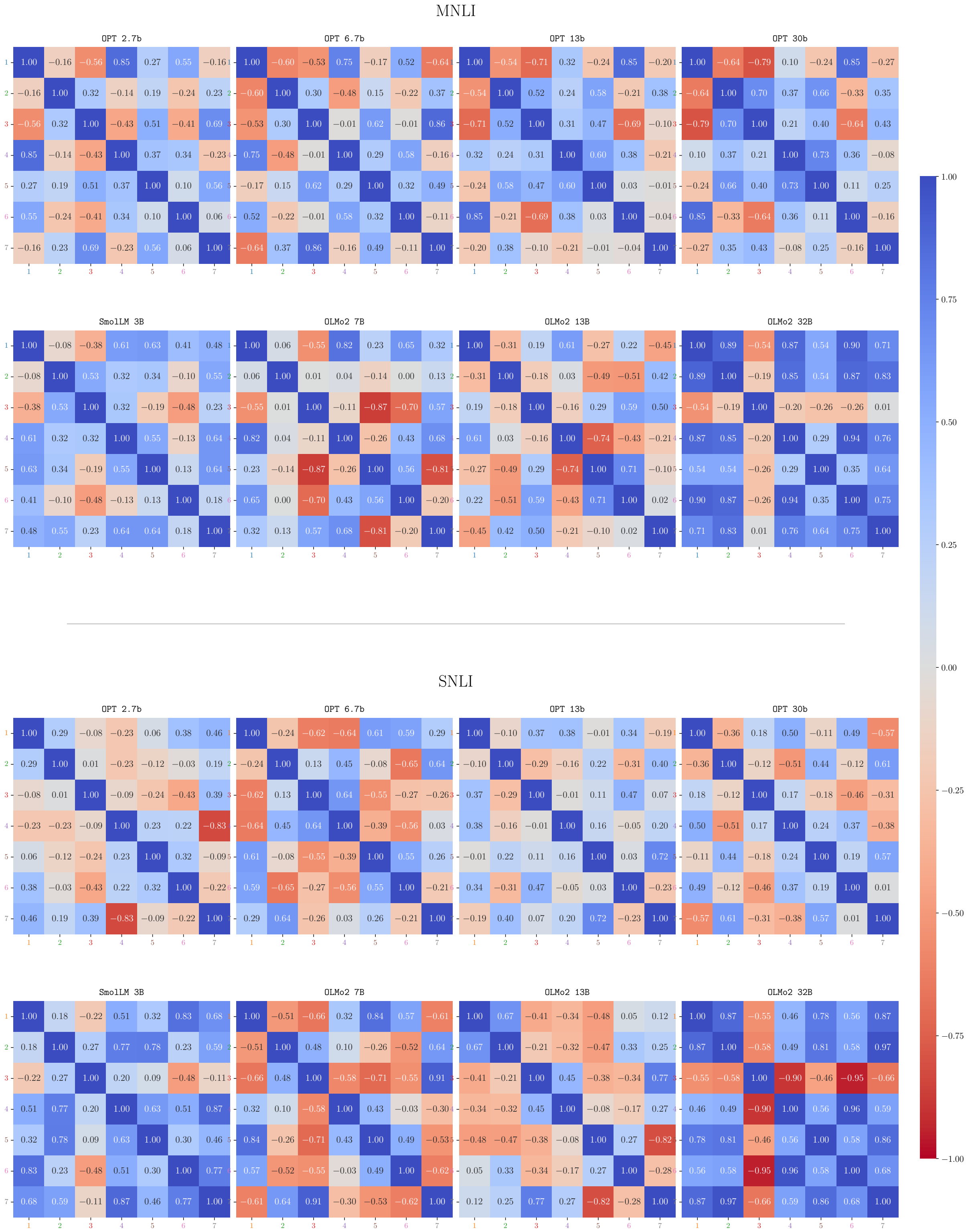


Figure 11: Partial correlations taken with Linear regressors of OPT (first and third rows), SmolLM (second and fourth rows, left) and OLMo2 (second and fourth rows, center and right) across model sizes (ordered from left to right) trained on MNLI (top) and SNLI (bottom). Models were fine-tuned with 64-shot. Legend: 1. SNLI/MNLI, 2. WNLI, 3. SciTail, 4. RTE, 5. HANS, 6. ANLI, and 7. PAWS.

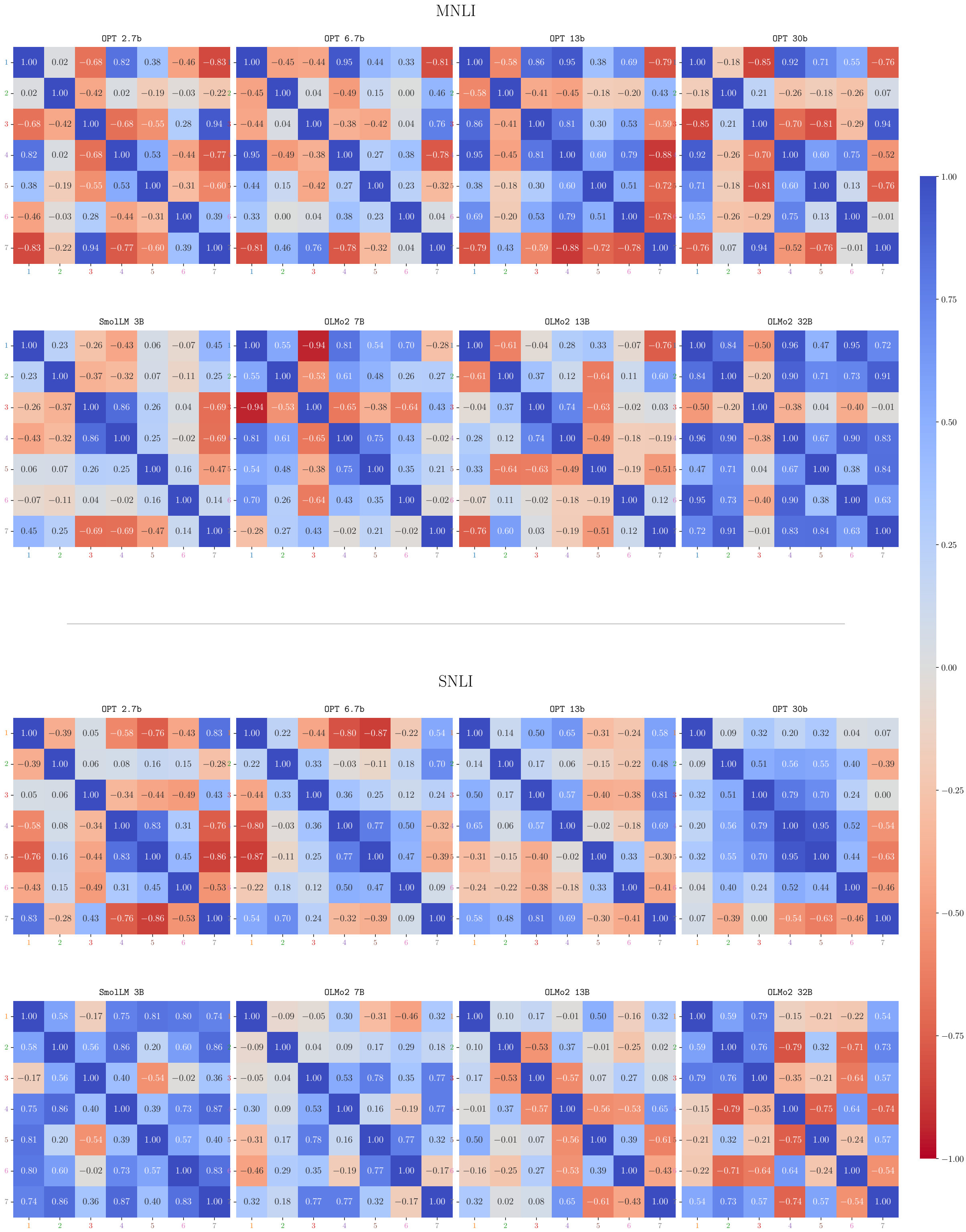


Figure 12: Partial correlations taken with Linear regressors of OPT (first and third rows), SmolLM (second and fourth rows, left) and OLMo2 (second and fourth rows, center and right) across model sizes (ordered from left to right) trained on MNLI (top) and SNLI (bottom). Models were fine-tuned with 32-shot. Legend: 1. SNLI/MNLI, 2. WNLI, 3. SciTail, 4. RTE, 5. HANS, 6. ANLI, and 7. PAWS.

D.8 Heatmaps with Partial OOD Correlations using GAM Regressors

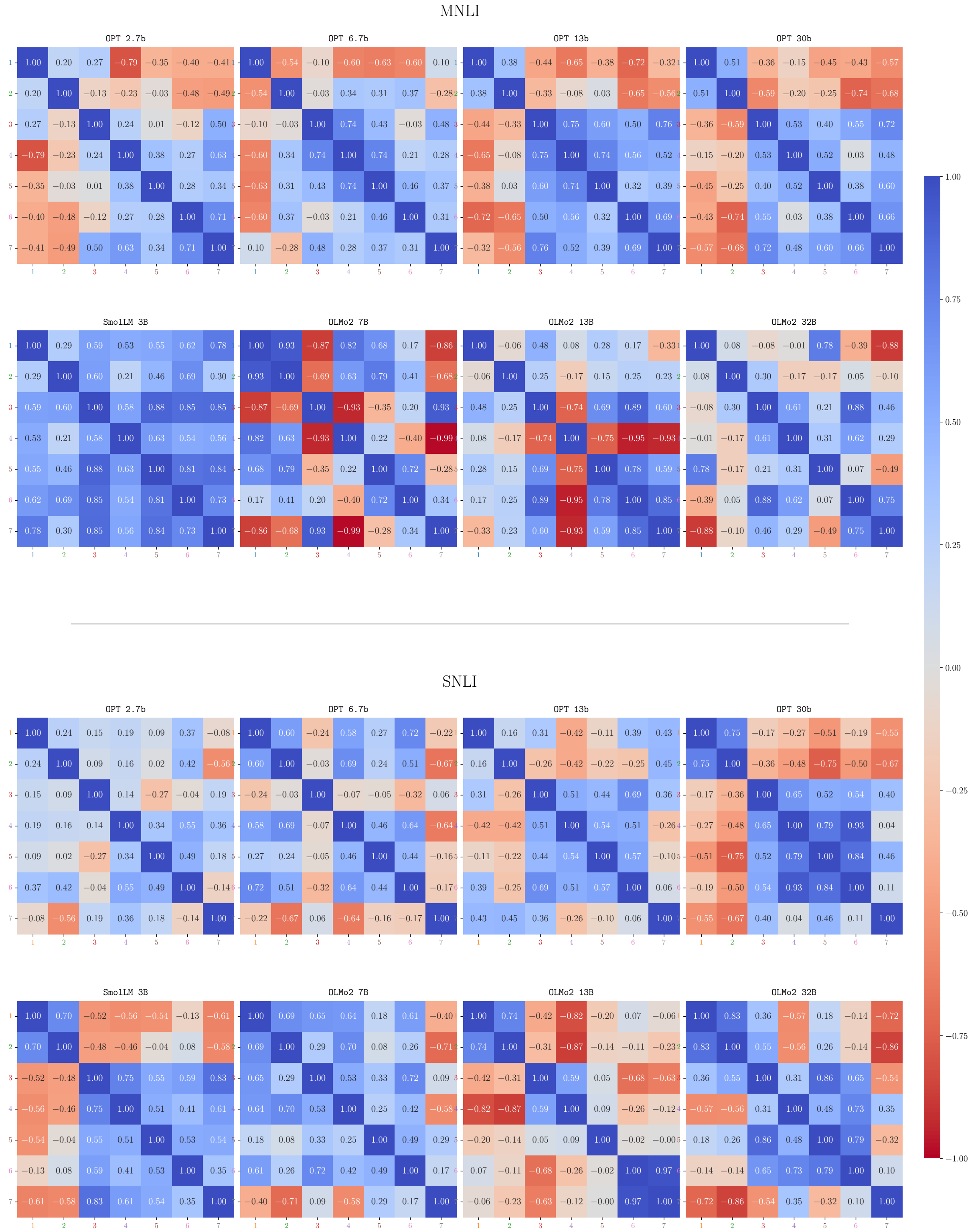


Figure 13: Partial correlations taken with GAM regressors of OPT (first and third rows), SmolLM (second and fourth rows, left) and OLMo2 (second and fourth rows, center and right) across model sizes (ordered from left to right) trained on MNLI (top) and SNLI (bottom). Models were fine-tuned with 128-shot. Legend: 1. SNLI/MNLI, 2. WNLI, 3. SciTail, 4. RTE, 5. HANS, 6. ANLI, and 7. PAWS.

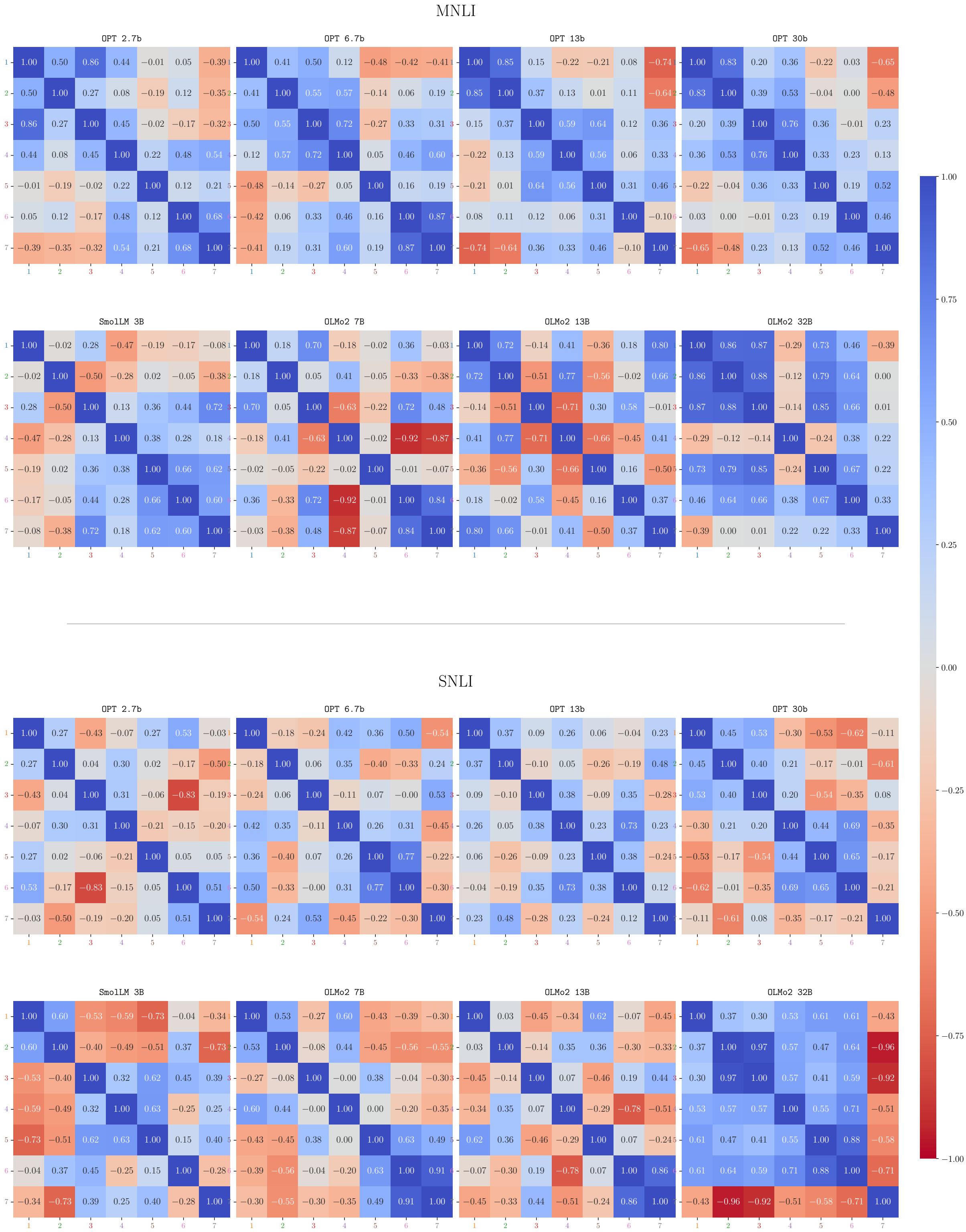


Figure 14: Partial correlations taken with GAM regressors of OPT (first and third rows), SmolLM (second and fourth rows, left) and OLMo2 (second and fourth rows, center and right) across model sizes (ordered from left to right) trained on MNLI (top) and SNLI (bottom). Models were fine-tuned with 64-shot. Legend: 1. SNLI/MNLI, 2. WNLI, 3. SciTail, 4. RTE, 5. HANS, 6. ANLI, and 7. PAWS.

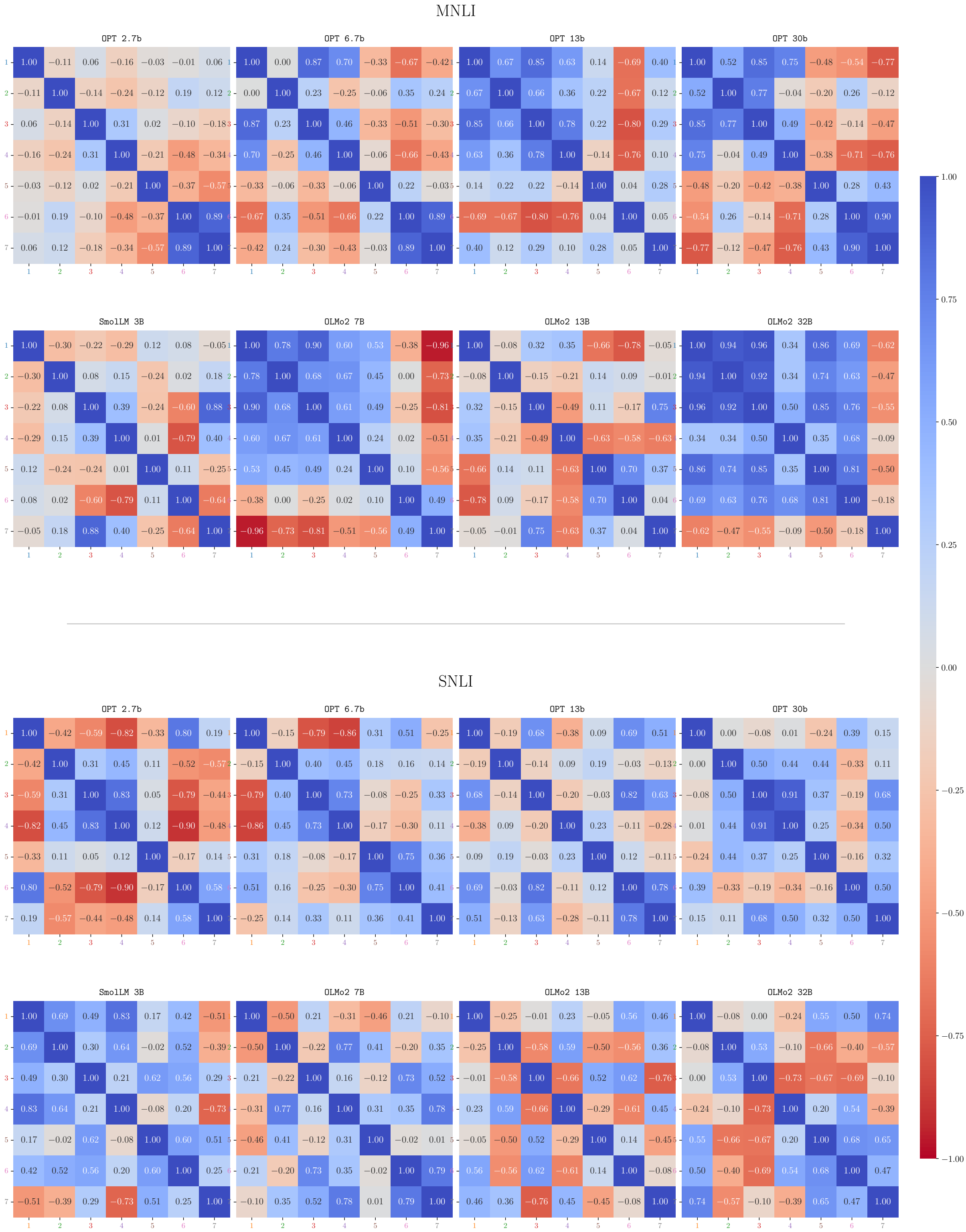


Figure 15: Partial correlations taken with GAM regressors of OPT (first and third rows), SmolLM (second and fourth rows, left) and OLMo2 (second and fourth rows, center and right) across model sizes (ordered from left to right) trained on MNLI (top) and SNLI (bottom). Models were fine-tuned with 32-shot. Legend: 1. SNLI/MNLI, 2. WNLI, 3. SciTail, 4. RTE, 5. HANS, 6. ANLI, and 7. PAWS.

D.9 Average Correlations across Model Sizes

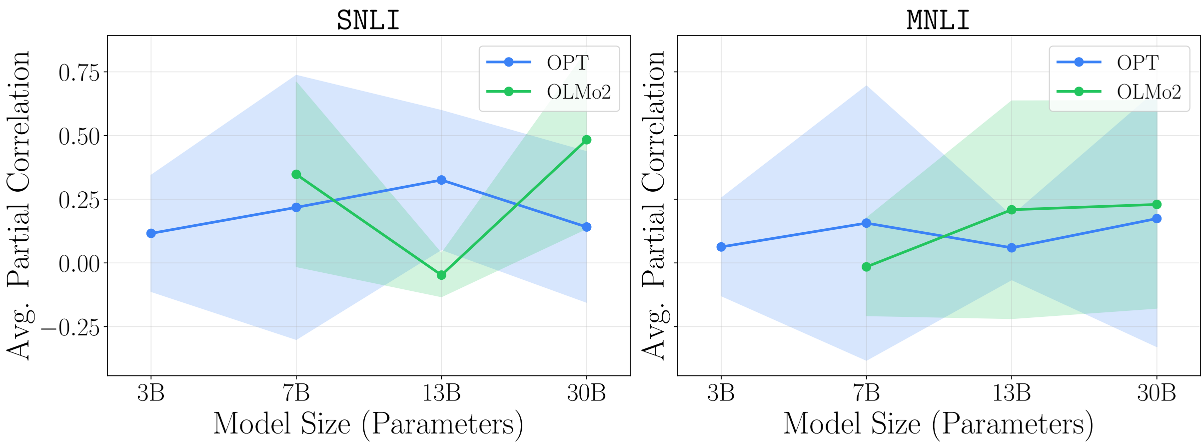


Figure 16: Average partial correlation vs. model size for `OPT` and `OLMo2`. Partial correlations computed with `GAM` regressors and averaged across OOD testsets.

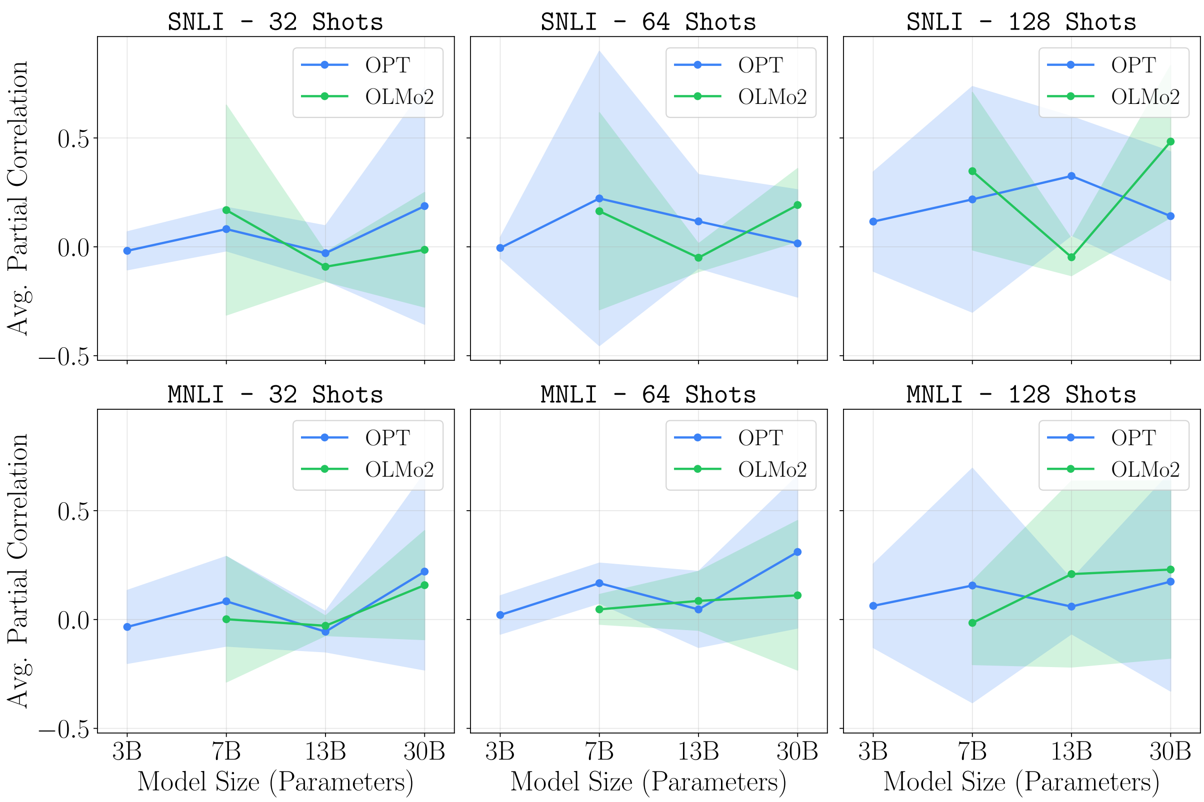


Figure 17: Average partial correlations vs. model size for `OPT` and `OLMo2`. Partial correlations computed with `GAM` regressors and averaged across OOD testsets.

D.10 Average Partial Correlations across Generations of Model Families

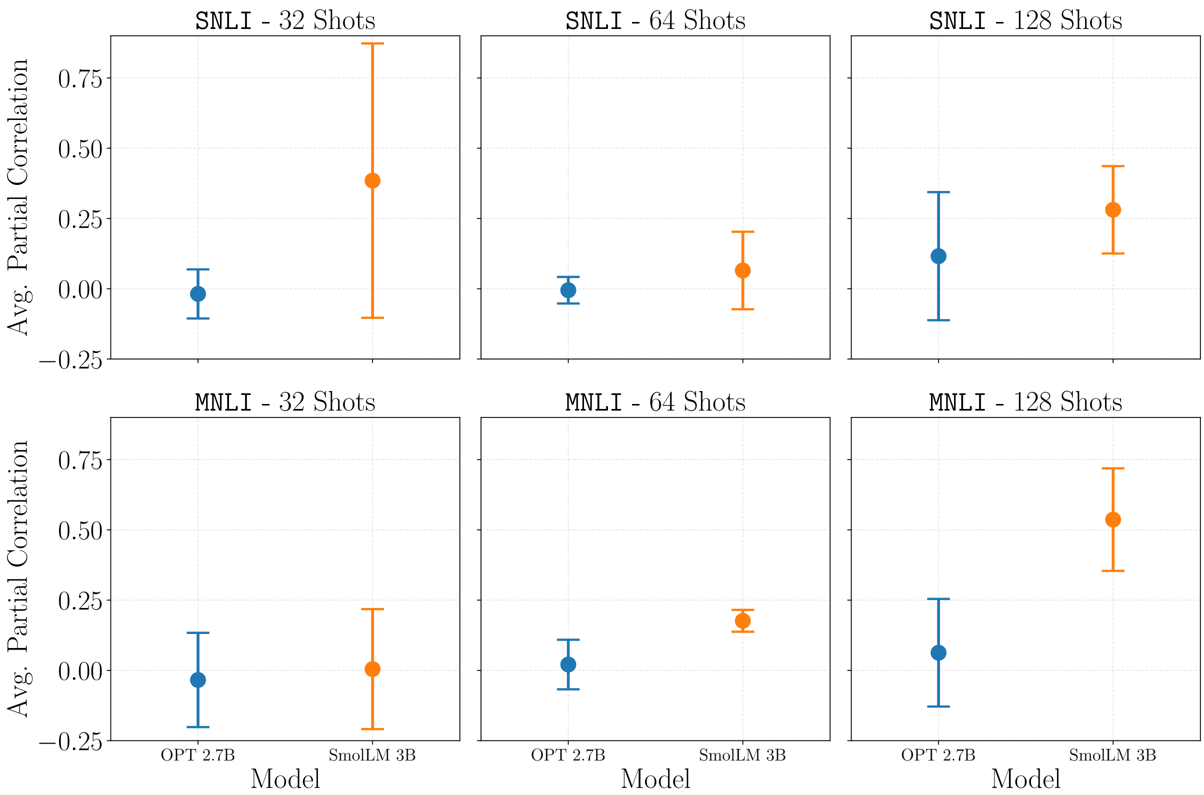


Figure 18: Average partial correlations comparing OPT and SmolLM. Across different training runs, SmolLM consistently demonstrates higher (positive) correlations, indicating more stable OOD generalisation patterns across testset.