The Geometries of Truth Are Orthogonal Across Tasks

Waiss Azizian [1] **Michael Kirchhof** [1] **Eugene Ndiaye** [1] **Louis Bethune** [1] **Michal Klein** [1] **Pierre Ablin** [1] **Marco Cuturi** [1]

Abstract

Large Language Models (LLMs) have demonstrated impressive generalization capabilities across various tasks, but their claim to practical relevance is still mired by concerns on their reliability. Recent works have proposed examining the activations produced by an LLM at inference time to assess whether its answer to a question is correct. Some works claim that a "geometry of truth" can be learned from examples, in the sense that the activations that generate correct answers can be distinguished from those leading to mistakes with a linear classifier. In this work, we underline a limitation of these approaches: we observe that these "geometries of truth" are intrinsically task-dependent and fail to transfer across tasks. More precisely, we show that linear classifiers trained across distinct tasks share little similarity and, when trained with sparsity-enforcing regularizers, have almost disjoint supports. We show that more sophisticated approaches (e.g., using mixtures of probes and tasks) fail to overcome this limitation, likely because activation vectors commonly used to classify answers form clearly separated clusters when examined across tasks.

1. Introduction

Large Language Models (LLMs) have seen tremendous success in recent years across a wide range of tasks. However, their widespread deployment is not without risks: from hallucinations (Ji et al., 2023) to outright deception (Park et al., 2024), the complexities underpinning LLM generation can be the root of many issues. These challenges are particularly concerning in high-stakes domains like healthcare, legal advice, and financial analysis, where incorrect or misleading information can lead to serious harm. As a result, several works have suggested leveraging the various activations generated by a model at inference time to understand and assess the truthfulness of its output. Azaria & Mitchell (2023) demonstrated that training a simple classifier on top of the hidden activations of LLMs can help predict whether an LLM has provided a truthful answer to a user-provided question. This finding suggests that models somewhat reveal enough information in their activations at inference time to help users assess whether they are producing correct information. Numerous subsequent works have explored this "geometry of truth" and confirmed that it is approximately linear, in the sense that a linear classifier can distinguish reliably truthful from erroneous answers (Li et al., 2023; Marks & Tegmark, 2024; Xiong et al., 2024; Burns et al., 2023; Kossen et al., 2025). As a token of their relevance, these directions can then be used to steer the LLM towards factual generations (Li et al., 2023; Wang et al., 2024), and are increasingly studied as a cheap and effective proxy to assess the uncertainty of an LLM output on a given task, e.g. (Sky et al., 2024; Zhang et al., 2025; Slobodkin et al., 2023).

But is one hyperplane all it takes? While the literature abounds with examples of the relevance of such probes when tested within a given task or knowledge domain, it remains unclear whether they can generalize across different tasks. Although some works have reported encouraging results in that direction (Azaria & Mitchell, 2023; Marks & Tegmark, 2024; Beigi et al., 2024), others provide a mixed assessment (Slobodkin et al., 2023; Kossen et al., 2025; Zhang et al., 2025), while some works (Orgad et al., 2025; Levinstein & Herrmann, 2024; Sky et al., 2024) show on the contrary that probes completely fail to generalize in some settings. In that context, one might be tempted to increase the complexity of probes and their training procedures, such as the more sophisticated pipeline proposed by Beigi et al. (2024) that incorporates augmentations to build additional representations on top of the activations of multiple layers taken jointly. Taken together, these works paint a mixed picture of whether simple probes can generalize at all, and if they do not, whether their disappointing results are an artifact of their training or an intrinsic limitation.

Our contributions. Our work proposes to answer whether there is any hope to see linear probes transfer reliably *across tasks*. We study the variability of these probes across tasks

[1]Apple. Correspondence to: Michael Kirchhof, Marco Cuturi <contact see website>.

Published at ICML 2025 Workshop on Reliable and Responsible Foundation Models.

and reach the following findings:

Task-specific truthfulness geometries. We first demonstrate that truthfulness geometries are fundamentally task-specific. Through comprehensive cross-task evaluation on seven diverse datasets, we show that linear probes trained on different tasks exhibit distinct "geometries of truth" that fail to generalize. While some task pairs show successful transfer, most combinations result in substantial performance degradation. This systematic analysis shows that generalization success depends critically on task similarity rather than being a universal property of truthfulness probes.

Geometric analysis of orthogonality. We then provide geometric analysis revealing why these failures occur. We demonstrate that truthfulness directions are largely orthogonal across tasks, with a clear correlation between geometric similarity and generalization performance. Using sparse probes, we reveal that probe supports are nearly disjoint across tasks, providing interpretable evidence for orthogonality. Visualizations show that different tasks form distinct clusters in the representation space of the model, confirming our geometric explanation for the failure to transfer.

Orthogonality persists in multi-task settings. Finally, we test whether our orthogonality hypothesis holds when training on mixtures of task. We demonstrate that training on diverse task mixtures fails to resolve generalization problems, and critically show that optimal directions for target tasks cannot be recovered through linear combinations of directions from other tasks. We further show that more complex architectures are also no better than naive parameter summation, suggesting the limitation is intrinsic. Given these findings, we explore conservative deployment strategies using conformal prediction to maintain reliability guarantees in cross-task scenarios.

Our work is structured as follows:

- In Section 2, we review the necessary background on probing and present a detailed survey of the literature on generalization properties of uncertainty probes.
- In Section 3, we introduce our experimental setup (models and datasets) and systematically study cross-task generalization failures through geometric analysis of probe directions and sparse probe supports.
- In Section 4, we examine whether training on mixtures of tasks can overcome these limitations, test more complex architectures, and explore conservative deployment strategies.

2. Background

Linear probing for uncertainty quantification. Uncertainty quantification for LLMs has attracted significant attention recently. Apart from relatively more costly multi-samples methods such as semantic entropy (Farquhar et al., 2024), many efforts have focused on learning simple classifiers whose inputs are activation vectors computed by the LLM at inference time, using labeled pairs of questions and either correct or incorrect answers. Azaria & Mitchell (2023) introduced this approach for uncertainty quantification for LLMs. While the classifier was originally set to be a multi-layer perceptron, follow-up works showed that using a simple linear logistic regression model could achieve similar performance (Li et al., 2023; Orgad et al., 2025; Marks & Tegmark, 2024; Santilli et al., 2025). In other words, and following Marks & Tegmark's terminology, there might be a linear "geometry of truth" that can separates the representations of correct from incorrect outputs.

The Geometry of Truth Hypothesis. Given a user question q, the LLM generates an answer $\hat{a} = (\hat{a}_1, \dots, \hat{a}_T)$ autoregressively. At each step $t \in [T]$, the model produces hidden state vectors $h_{t,\ell} \in \mathbb{R}^d$ where ℓ indexes the layer and t the token position. In this work, we focus on a fixed layer $\ell^\star$ (e.g. 28 and 21 for Qwen models) and extract the representation $h_{T,\ell^\star}$ at the final token of the output (or the token before, $T-1$). To provide supervision, we follow (Farquhar et al., 2024; Santilli et al., 2025) and label the correctness of $\hat{a}$ given the gold answer a using LLM-as-a-judge (Zheng et al., 2023). This yields a label $y \in \{-1, +1\}$ which is positive is the answer $\hat{a}$ is correct and negative otherwise. This provides a dataset $\mathcal{D} = \{(h_i, y_i)\}_{i \in [N]}$ where each point $h_i = h_{T,\ell^\star}$ is the final hidden representation for a generated answer i. It is the internal state right before the model decides what the final answer token will be. We then train a linear probe i.e. a logistic regression classifier to predict correctness from the hidden states:

$$\min_{\substack{\theta \in \mathbb{R}^d, \\ b \in \mathbb{R}}} \frac{1}{N} \sum_{i=1}^{N} \log\left(1 + e^{-y_i(\theta^\top h_i + b)}\right) + \frac{\lambda_2}{2} \|\theta\|_2^2 \qquad (1)$$

The *geometry of truth hypothesis* states that truthful and untruthful generations are linearly separable in model's hidden space, such that a single hyperplane parametrized by (θ, b) can distinguish correct from incorrect outputs. We explore if this is universal or task-specific.

On the Generalization of Uncertainty Probes. The generalization of uncertainty estimates has been studied in several works but there is still no definitive consensus, neither among papers nor within the papers themselves, as many of them conclude with a nuanced assessment. Some of the historically first results were promising:

- Azaria & Mitchell (2023) introduce their own dataset of facts, the *true-false* dataset, made of several splits with different subjects. With their carefully constructed dataset, they show that probes trained on several subjects can accurately predict the correctness of facts of another subject.

- Kapoor et al. (2024) consider finetuning the whole LLM to obtain better uncertainty probes; although they do not conduct systematic evaluations, they report good performance on a couple of datasets that were not seen during training.

Other works apply probes in slightly different contexts and report mixed generalization results.

- Slobodkin et al. (2023) leverage probes to classify unanswerable queries and study their generalization performance on datasets not seen at training (their Figure 6). Although performance decreases, the authors still note that it remains better than probes trained on the first hidden layer on the correct dataset.
- Zhang et al. (2025) use uncertainty probes for reasoning tasks and find that, although these probes generalize across similar datasets, they fail when the type of reasoning required changes.
- Kossen et al. (2025) suggest training probes with semantic entropy (Farquhar et al., 2024) and show that, for generalization to new tasks, this improves robustness in the choice of layer (their Figure 6).
- Kadavath et al. (2022) introduces a finetuning approach trained on a diverse mixture of tasks which is evaluated both on a held-out dataset and when training on only one task: the authors observe a decrease in performance but still note decent generalization.

Then there are paper that report generally negative generalization results.

- Beigi et al. (2024) report significantly worse performance on out-of-distribution data (Table 3) despite using all hidden states of an LLM as an input for an uncertainty estimator.
- Marks & Tegmark (2024) study, in a controlled setting similar to (Azaria & Mitchell, 2023), the geometry of representations of correct and incorrect answers as well as generalization properties of probes across datasets. Though the authors note that training on datasets and their negations helps, the performance of probes on out-of-distribution data remains suboptimal (Figure 5). They also visualize the hidden spaces at different layers to understand when a linear representation of uncertainty emerges.
- Levinstein & Herrmann (2024) reproduce the setting of Azaria & Mitchell (2023) and note that their probes catastrophically fail to generalize under trivial changes like introducing negations (§4.4), with a roughly 20% accuracy loss.
- In Orgad et al. (2025), the authors take a systematic approach: they consider a wide variety of question-answers datasets and find that probes trained on each of these datasets fail to generalize (their Figure 3).
- Sky et al. (2024) consider ensembles of attention-based probes for hallucination detection but notes that they do not generalize, both when trained on one dataset and tested in another (their Table 6) and when trained in two tasks and tested in a third (Table 8). This work raises the question of whether having a larger mixture of tasks night help, which we will answer by the negative in this work.

3. Truthfulness directions across tasks

After having reviewed the necessary background and the literature on the generalization of probes, we now present our experimental setup before starting our study of cross-task generalization.

Training Probes. To obtain uncertainty estimates with linear probes, we follow standard practices (Li et al., 2023; Orgad et al., 2025; Marks & Tegmark, 2024; Santilli et al., 2025) and train the probes using the logistic regression implementation of Pedregosa et al. (2011). These probes are trained with L_2 regularization with the hyperparameter tuned with cross-validation on the training set. We consider two standard token positions t: the stop token of the output and the token before the stop token of the output. Similarly, we base the probes on two embeddings: those of the last hidden layer and those at 75% depth.

Models. We consider the following models: Qwen 2.5 7B Instruct (Qwen authors, 2024), Phi 4 Mini (Phi-4 authors, 2024) and Llama 3.1 8B Instruct (The Llama 3 authors, 2024). In the main text, we consider probes trained via Equation (1) that operate on the stop token at the last layer (layer 28) of Qwen 2.5 7B Instruct. All other combinations give similar results, which we report in the appendix.

Datasets. To study generalization, we use several datasets that are variously related: NQ (Kwiatkowski et al., 2019) and SimpleQA (Wei et al., 2024) are general question-answering datasets. TriviaQA (Joshi et al., 2017), and SQUAD (Rajpurkar et al., 2016) are reading-comprehension datasets with a context, but also on generic topics. BioASQ (Nentidis et al., 2023) is composed of biology questions. SVAMP (Patel et al., 2021), made of simple arithmetic questions, and GSM8K (Cobbe et al., 2021) of more complex math word problems.

Cross-task generalization failures. We train linear probes for each dataset, then systematically evaluate cross-task transfer performance. Our dataset collection enables analysis of both successful and failed generalization cases. Figure 1 confirms widespread generalization failures, in line with Orgad et al. (2025), but reveals a crucial pattern: tasks cluster into semantically coherent groups. Factual recall tasks (TriviaQA, NQ, SimpleQA) show mutual transferability, while specialized domains (BioASQ biol-

ogy, GSM8K/SVAMP mathematics) remain isolated. We thus confirm that probes fail to generalize to new tasks in general but also that the situation is more nuanced. Indeed, this suggests that truthfulness representations adapt to domain-specific reasoning patterns rather than capturing one universal truth signal.

Underlying geometry. While previous work has documented the poor cross-task generalization of truthfulness probes, the geometric basis for this failure remains unexplored. We test whether generalization patterns reflect fundamental differences in how truthfulness is encoded across task domains, rather than mere statistical artifacts of limited training data. We examine cosine similarities between probe weight vectors (Figure 2). Most probe pairs exhibit near-orthogonal directions (cosine similarity < 0.5), with successful transfer occurring only between geometrically aligned probes. Indeed, Figure 3 shows that more similar probes also have better generalization performance in each others task (correlation coefficient of 0.59), showing that this is not just a mere geometrical problem but the cause for the performance drops; similar patterns with even lower cosine similarities are shown in the appendix for different models. These findings challenge the notion of a universal "geometry of truth." Instead, truthfulness directions emerge from task-specific representational structures, with different domains occupying orthogonal subspaces in the model's hidden state space. However, cosine similarity provides only a coarse measure of geometric relationships. To gain deeper insight into which specific features drive these differences, we next examine the support structure of sparse probes.

Sparse probes. To deepen our understanding of the features learned by probes, we consider sparse probes obtained through *sparse* logistic regression. Instead of Equation (1), we consider probes trained with L_1 regularization:

$$\min_{\substack{\theta\in\mathbb{R}^d,\\ b\in\mathbb{R}}} \frac{1}{N}\sum_{i=1}^{N} \log\left(1+e^{-y_i(\theta^\top h_i+b)}\right) + \lambda_1\|\theta\|_1\,. \qquad (2)$$

The ℓ_1-regularization coefficient λ_1 is tuned on a held-out validation set. As shown in Figure 4, this approach does not degrade performance. But it provides us with a visual way of comparing probes in Figure 5: The supports of probes between tasks are nearly disjoint. This can be made formal by computing support overlap, see Figure 6. We observe the same pattern as in Figure 2: some tasks have more support overlap than others and this coincides with the ones where probes generalize better.

Task-specific geometries. The previous experiments indicate that geometric information and generalization (failures) align with semantic task properties. This suggests that truthfulness detection mechanisms are not universal but

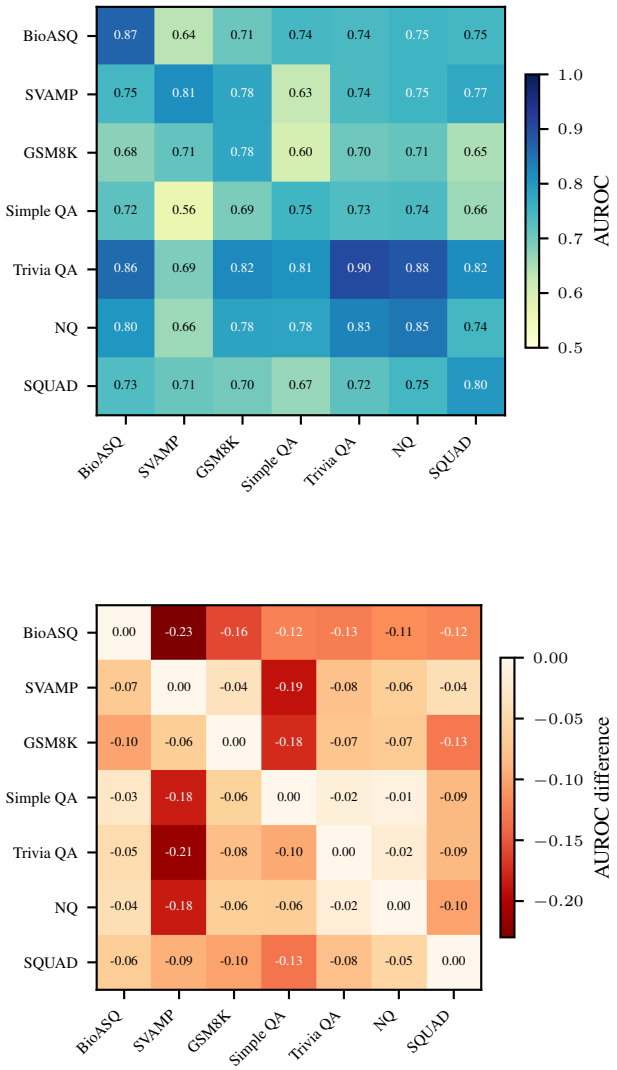


Figure 1. AUROC of probes trained on different tasks on the stop token of the output on last layer. Rows correspond to evaluation tasks while columns correspond to training tasks. The second plot represents the difference between the probe trained on this task and probes trained on the other datasets. Results are averaged over 5 runs.

rather emerge from task-specific representations within the model's hidden states. To visualize this phenomenon, we provide t-SNE plots (Van der Maaten & Hinton, 2008) of the hidden states of the model for each task Figure 7. What we see is that most of the tasks actually form distinct clusters. Similar tasks such as TriviaQA, NQ, SimpleQA are slightly mixed, but BioASQ or mathematical reasoning tasks are clearly separated from each other. The uncertainty signal, represented by the two right or wrong classes, is secondary compared to the distinction between tasks. This confirms

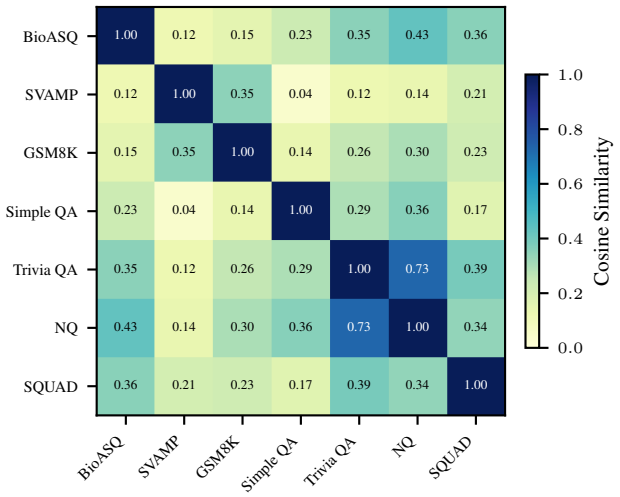


Figure 2. Cosine similarity between probes trained on different datasets using L2 regularization. Results are averaged over 5 runs. The cosine similarity between probe directions is consistently low (less than 0.5). Task pairs with large similarity Trivia QA - NQ (above 0.7) are the ones showing good generalization.

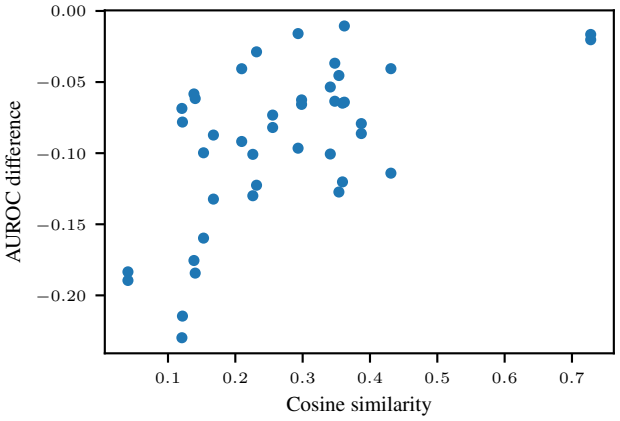


Figure 3. AUROC difference to probe trained on the right dataset as a function of cosine similarity between probes ($r = 0.59$, $R^2 = 0.35$, $p = 3.8 \times 10^{-5}$)

again our hypothesis that the truthfulness information is not universal and is very much task-dependent.

4. Generalization from mixture of tasks

4.1. Orthogonality in mixture of tasks

We now revisit our orthogonality hypothesis in a multi-task setting. We consider learning probes not on only one dataset but on a mixture of them, examining whether the

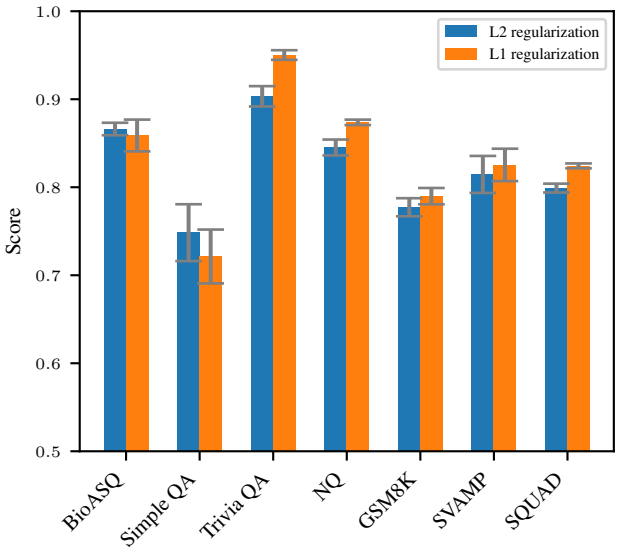


Figure 4. AUROC of linear probes with L1 or L2 regularisation. Results are averaged over 5 runs.

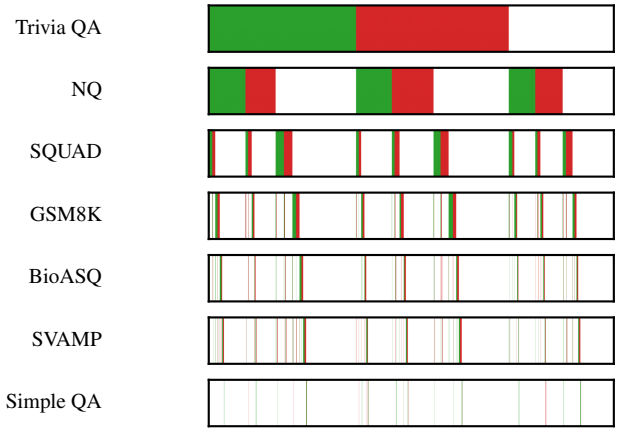


Figure 5. Signed support of sparse probes trained on different datasets at the stop token of the output, using L1 regularization at layer 28. Each row represents one probe trained on the corresponding dataset (y-axis labels). The x-axis shows the 3584 dimensions of the hidden state vector. Green indicates positive coefficients, red indicates negative coefficients, and white indicates zero coefficients (sparsity). Dimensions are sorted by sparsity level across all datasets, with the least sparse dimensions on the left and the most sparse on the right.

geometric relationships we observed persist in this more complex training scenario.

Figure 8 presents our main findings across several training strategies. The purple bars show the best performing probes: task-specific probes trained and tested on the same domain. The blue bars test our central question by showing performance when probes are trained on all tasks except the target

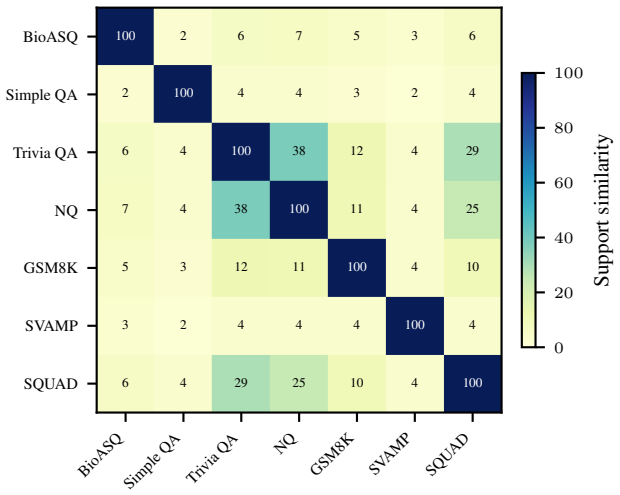


Figure 6. Support overlap between sparse probes trained on different datasets. Darker colors indicate higher overlap percentages. Task pairs with $> 30\%$ overlap (TriviaQA, NQ) correspond to successful cross-task generalization, while most pairs show $< 15\%$ overlap, explaining generalization failure.

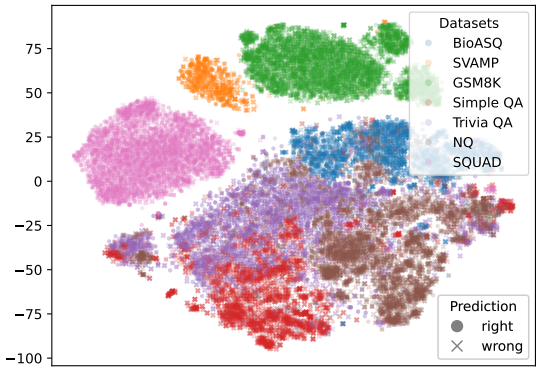


Figure 7. t-SNE plots of the hidden space at layer 28 at the stop token of the output. Different tasks form distinct clusters in representation space, with the correct/incorrect distinction being secondary to task boundaries.

task. Despite this diverse training mixture, we observe substantial performance drops across all domains, suggesting that multi-task training does not resolve the generalization problem.

To test whether this failure stems from our orthogonality hypothesis, we examine whether truthfulness directions from other tasks can be linearly combined to recover the direction from the target task. The orange bars show results from a constrained optimization where we restrict probe coefficients to lie within the subspace spanned by probes from other tasks. Formally, given probes $\theta_1, \ldots, \theta_6$ from non-target tasks, we solve:

$$\min_{\substack{\alpha \in \mathbb{R}^6 \\ b \in \mathbb{R}}} \frac{1}{N} \sum_{i=1}^{N} \log\left(1 + e^{-y_i(\theta_\alpha^\top h_i + b)}\right) + \frac{\lambda_2}{2}\|\alpha\|_2$$

$$\text{where } \theta_\alpha = \sum_{i=1}^{6} \alpha_i \theta_i \,.$$

The suboptimal performance of this constrained approach compared to task-specific probes confirms that target task directions lie outside the subspace generated by directions of the other tasks, providing direct evidence for our orthogonality claim.

The remaining comparisons in Figure 8 further support this interpretation. Training on all datasets simultaneously (red bars) yields performance roughly equivalent to simply summing individually trained probe parameters (green bars). This equivalence is striking: if probe directions overlapped significantly, naive parameter summation would cause destructive interference and degrade performance. Instead, the similar results confirm that probe directions are approximately orthogonal across tasks. However, both approaches still fall slightly short of task-specific training (purple bars), demonstrating that even when all tasks are included in training, the resulting probes cannot match the performance of domain-specific optimization.

4.2. Mixture of probes

Given the failure of linear probes to generalize across tasks, we test whether more complex probe architectures can improve cross-task transfer performance. We experiment with a "mixture of probes" approach inspired by Mixture-of-Experts architectures. Our approach uses a single gating layer with 16 expert probes, where each expert consists of a 2-layer feedforward network that takes the LLM's hidden states as input. The gating mechanism learns to route different inputs to different expert probes based on the hidden state representations.

For each target task, we train this mixture of probes on the six other tasks while using a mixture of these six tasks as a validation set. We perform grid search over hyperparameters including learning rate, weight decay, and auxiliary loss coefficients for the gating mechanism taking inspiration from Fedus et al. (2021).

Figure 9 presents results under two scenarios: an "oracle" setting where hyperparameters are selected using the test task performance, and a realistic setting where hyperparameters are chosen based on validation task performance.

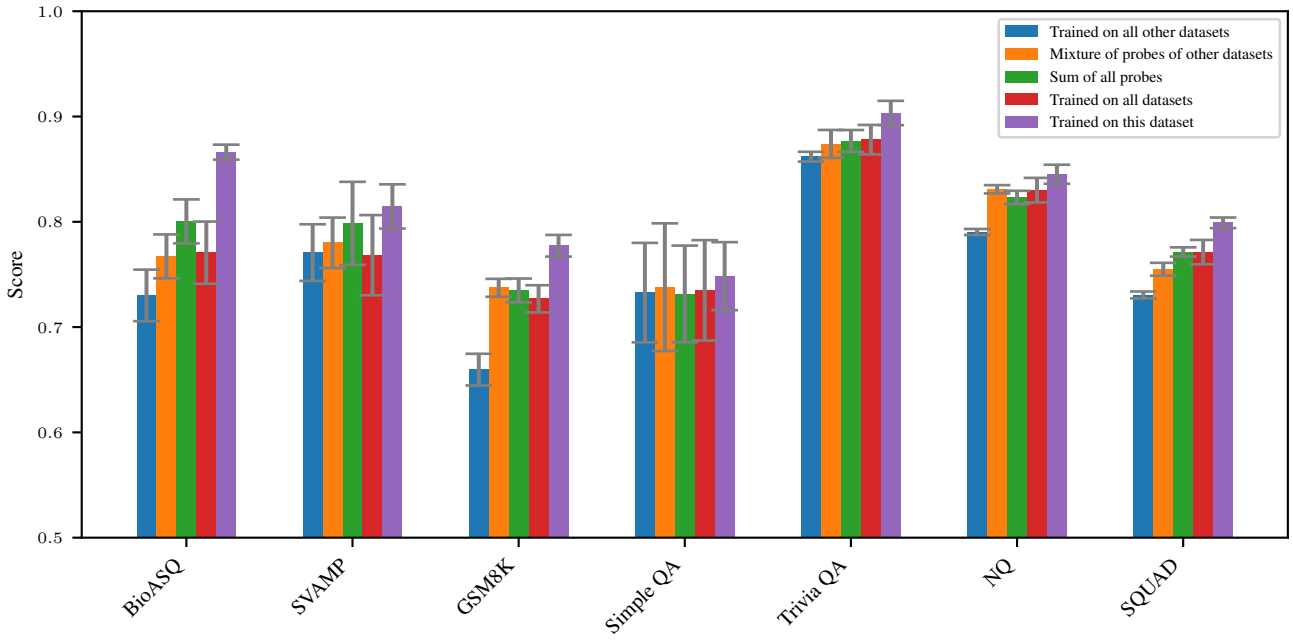


Figure 8. AUROC of linear probes at the stop token of the output in the multi-task setting, using L2 regularization.

In both cases, performance remains below that of linear probes trained directly on the target task. Moreover, the performance of this non-linear model matches that of simple linear probes trained on the same six held-out tasks.

These results show that even sophisticated probe architectures cannot bridge the performance gap with task-specific probes. The equivalence between complex and simple models trained on identical data suggests that the generalization failure stems from the orthogonal task geometries we identified, rather than limitations in model architecture.

4.3. Conservative approaches via conformal prediction

The previous sections have demonstrated that truthfulness representations are inherently task-dependent and fail to generalize across domains. This poses a critical challenge for real-world deployment: how can we maintain reliable uncertainty estimates when the distribution of user queries may differ from training data?

Given that guaranteeing generalization appears unattainable, we explore whether conservative calibration methods can provide reliability guarantees despite poor cross-task transfer. We focus on conformal prediction as a principled approach to control error rates, examining scenarios where avoiding false endorsement of incorrect information is crucial.

We consider the setting of Section 4 where probes are trained on multiple tasks and are evaluated on a new, unseen domain. Specifically, we train probes on all datasets except one test task using Equation (1) and seek to ensure that the false positive rate remains below a threshold $\alpha = 0.3$ on the held-out task. We compare three approaches: plain probes with default thresholds (Plain), standard split conformal prediction (Vovk et al., 2005) (CP), and a variant designed for multi-task settings (Park et al., 2022) (Meta-CP).

Given a trained probe $f(h) = \theta^\top h + b$ obtained by Equation (1), these methods calibrate a threshold τ such that the probe's confidence score must exceed τ before predicting an answer as correct. For new hidden states $h_{t,\ell}$ corresponding to question-answer (q, a) with label y, the following bound on the false positive rate holds:

$$\mathbb{P}\left(f(h_{t,\ell}) > \tau \,|\, y = -1\right) \leq \alpha\,.$$

This ensures that, on average, the false positive rate will be at most α for new questions. We refer to Vovk et al. (2005) and Park et al. (2022) for methodological details. For the multi-task variant, we set both hyperparameters to 0.3 and artificially randomly split calibration tasks into subtasks of size 1000 to match the experimental setup from Park et al. (2022).

The results in Figures 10 and 11 reveal substantial differences between approaches. Plain probes achieve a mean false positive rate of 0.34, exceeding the target threshold of 0.3, with high variability (80th percentile: 0.69). Standard conformal prediction reduces the mean false positive rate to 0.25, approaching but not consistently achieving the target, with the 80th percentile still reaching 0.47. The multi-task variant achieves the strongest false positive rate control,

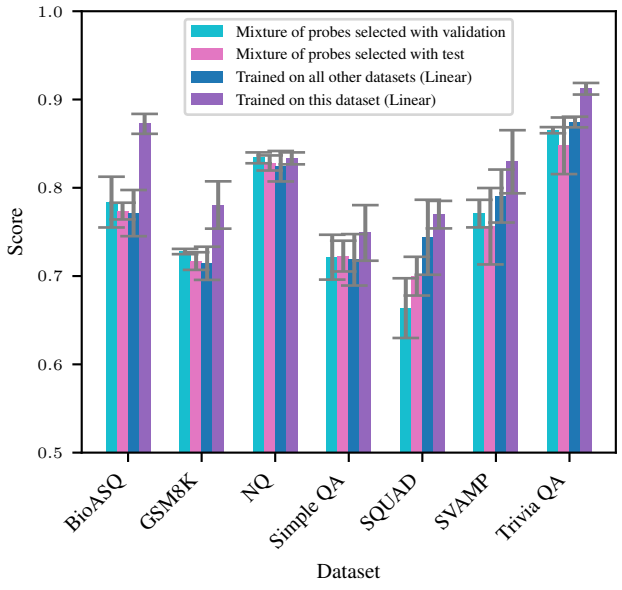
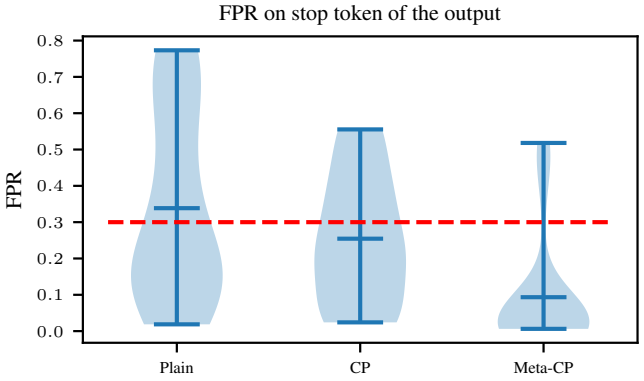
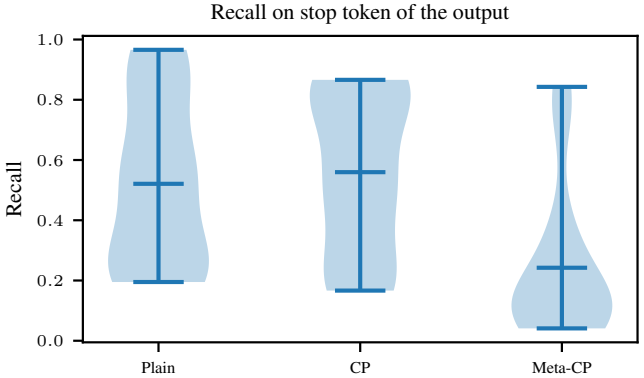


Figure 9. AUROC of different methods in the multi-task setting. Mixture of probes are trained on six non-target tasks and evaluated on the target task. "Validation" uses hyperparameters selected on a validation set from the training tasks; "Test" uses hyperparameters selected on the target task (oracle setting). Linear probe baselines ("Trained on all other datasets" and "Trained on this dataset") reproduce results from Figure 8 for comparison. Results averaged over 3 runs.



(a) Violin plot of the false positive rate (FPR) of the different methods with threshold $\alpha = 0.3$.



(b) Violin plot of the recall of the different methods with threshold $\alpha = 0.3$.

Figure 10. False positive rates (FPR) and recall for thresholds tuned using different methods: standard training (Plain), split conformal prediction (CP), conformal prediction for multi-task settings (Meta-CP). The results are averaged over 5 repetitions and test tasks.

with a mean of 0.09 and 80th percentile of 0.23, successfully staying below the target threshold. However, these improvements in false positive rate control come at substantial cost to recall. While plain probes achieve 0.52 mean recall and standard conformal prediction reaches 0.56, the multi-task variant falls to just 0.24. This dramatic reduction means that this conservative approach correctly identifies only about one-fourth of true positives, illustrating the fundamental trade-off when truthfulness representations fail to generalize. Moreover, this poor performance also reflects a deeper issue: conformal prediction assumes the scoring function $f(h)$ reliably ranks correctness across domains. However, our findings show that truthful and untruthful generations are only linearly separable within task-specific subspaces that vary significantly across tasks. When probes encounter out-of-distribution tasks, they become misaligned with actual correctness labels. To maintain the required false positive rate guarantees, conformal prediction must set extremely conservative thresholds, filtering out many correct answers not because they are ambiguous, but because the probe's direction no longer matches the task's geometry of truth. Our insight is that conformal prediction becomes overly conservative precisely because the underlying geometry fails to generalize.

Method	Mean FPR	Q-80% FPR	Mean Recall
Plain	0.34	0.69	0.52
CP	0.25	0.47	0.56
Meta CP	0.09	0.23	0.24

Figure 11. False positive rates (FPR) and recall for thresholds tuned using different methods: standard training (Plain), split conformal prediction (CP), conformal prediction for multi-task settings (Meta-CP). The results are averaged over 5 repetitions and means and 80% quantiles (Q-80%) are considered over test tasks.

Conclusion. The premise of the "geometry of truth" hypothesis is that one might be able to detect whether an LLM provides a correct answer when prompted with a question. These works claim that a simple classifier, taken as inputs the activations generated by the LLM as it produces its answer, can suffice to predict the correctness of the final

answer. Although extremely appealing, we show in this work that such a promise may not be yet reliable enough as it fails to transfer across domains and tasks, notably if the domain on which the probe was trained is markedly different from that where the performance of the classifier is evaluated. We explain this failure to generalize by noticing that probes trained independently on various tasks have both low similarity and small feature overlap when trained with sparse regularizers. We have explored more advanced classification paradigms, such as mixture-of-probes, which could have been able to handle this heterogeneity, but we were not able to achieve reliable generalization. We conclude that LLMs likely have multiple geometries of truth, but that they are irreconcilable and highly task-dependent.

References

Azaria, A. and Mitchell, T. The internal state of an llm knows when it's lying. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Beigi, M., Shen, Y., Yang, R., Lin, Z., Wang, Q., Mohan, A., He, J., Jin, M., Lu, C.-T., and Huang, L. Internalinspector i2: Robust confidence estimation in llms through internal states. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 12847–12865, 2024.

Burns, C., Ye, H., Klein, D., and Steinhardt, J. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ETKGuby0hcs.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Farquhar, S., Kossen, J., Kuhn, L., and Gal, Y. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Fedus, W., Zoph, B., and Shazeer, N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity.(2021). *arXiv preprint cs.LG/2101.03961*, 2021.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38, 2023.

Joshi, M., Choi, E., Weld, D., and Zettlemoyer, L. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Barzilay, R. and Kan, M.-Y. (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147/.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Kapoor, S., Gruver, N., Roberts, M., Collins, K. M., Pal, A., Bhatt, U., Weller, A., Dooley, S., Goldblum, M., and Wilson, A. G. Large language models must be taught to know what they don't know. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Kossen, J., Han, J., Razzak, M., Schut, L., Malik, S. A., and Gal, Y. Semantic entropy probes: Robust and cheap hallucination detection in LLMs, 2025. URL https://openreview.net/forum?id=YQvvJjLWX0.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M.-W., Dai, A. M., Uszkoreit, J., Le, Q., and Petrov, S. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

Levinstein, B. A. and Herrmann, D. A. Still no lie detector for language models: Probing empirical and conceptual roadblocks. *Philosophical Studies*, pp. 1–27, 2024.

Li, K., Patel, O., Viégas, F., Pfister, H., and Wattenberg, M. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530, 2023.

Marks, S. and Tegmark, M. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=aajyHYjjsk.

Nentidis, A., Katsimpras, G., Krithara, A., López, S. L., Farré-Maduell, E., Gasco, L., Krallinger, M., and Paliouras, G. Overview of bioasq 2023: The eleventh bioasq challenge on large-scale biomedical semantic indexing and question answering. *arXiv preprint arXiv:2307.05131*, 2023.

Orgad, H., Toker, M., Gekhman, Z., Reichart, R., Szpektor, I., Kotek, H., and Belinkov, Y. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=KRnsX5Em3W.

Park, P. S., Goldstein, S., O'Gara, A., Chen, M., and Hendrycks, D. Ai deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(5), 2024.

Park, S., Dobriban, E., Lee, I., and Bastani, O. Pac prediction sets for meta-learning. *Advances in Neural Information Processing Systems*, 35:37920–37931, 2022.

Patel, A., Bhattamishra, S., and Goyal, N. Are NLP models really able to solve simple math word problems? In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y. (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2080–2094, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.168. URL https://aclanthology.org/2021.naacl-main.168/.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Phi-4 authors. Phi-4 technical report. *arXiv preprint arXiv: 2412.08905*, 2024.

Qwen authors. Qwen2.5 technical report. *arXiv preprint arXiv: 2412.15115*, 2024.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, 2016.

Santilli, A., Golinski, A., Kirchhof, M., Danieli, F., Blaas, A., Xiong, M., Zappella, L., and Williamson, S. Revisiting uncertainty quantification evaluation in language models: Spurious interactions with response length bias results. *arXiv preprint arXiv:2504.13677*, 2025.

Sky, C.-W., Van Durme, B., Eisner, J., and Kedzie, C. Do androids know they're only dreaming of electric sheep? In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 4401–4420, 2024.

Slobodkin, A., Goldman, O., Caciularu, A., Dagan, I., and Ravfogel, S. The curious case of hallucinatory (un) answerability: Finding truths in the hidden states of overconfident large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3607–3625, 2023.

The Llama 3 authors. The llama 3 herd of models. *arXiv preprint arXiv: 2407.21783*, 2024.

Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Wang, T., Jiao, X., He, Y., Chen, Z., Zhu, Y., Chu, X., Gao, J., Wang, Y., and Ma, L. Adaptive activation steering: A tuning-free llm truthfulness improvement method for diverse hallucinations categories. *CoRR*, abs/2406.00034, 2024. URL https://doi.org/10.48550/arXiv.2406.00034.

Wei, J., Karina, N., Chung, H. W., Jiao, Y. J., Papay, S., Glaese, A., Schulman, J., and Fedus, W. Measuring short-form factuality in large language models. *arXiv preprint arXiv: 2411.04368*, 2024.

Xiong, M., Santilli, A., Kirchhof, M., Golinski, A., and Williamson, S. Efficient and effective uncertainty quantification for LLMs. In *Neurips Safe Generative AI Workshop 2024*, 2024. URL https://openreview.net/forum?id=QKRLH57ATT.

Zhang, A., Chen, Y., Pan, J., Zhao, C., Panda, A., Li, J., and He, H. Reasoning models know when they're right: Probing hidden states for self-verification, 2025. URL https://arxiv.org/abs/2504.05419.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36: 46595–46623, 2023.

A. Qwen-2.5 7B Instruct

A.1. Layer 28

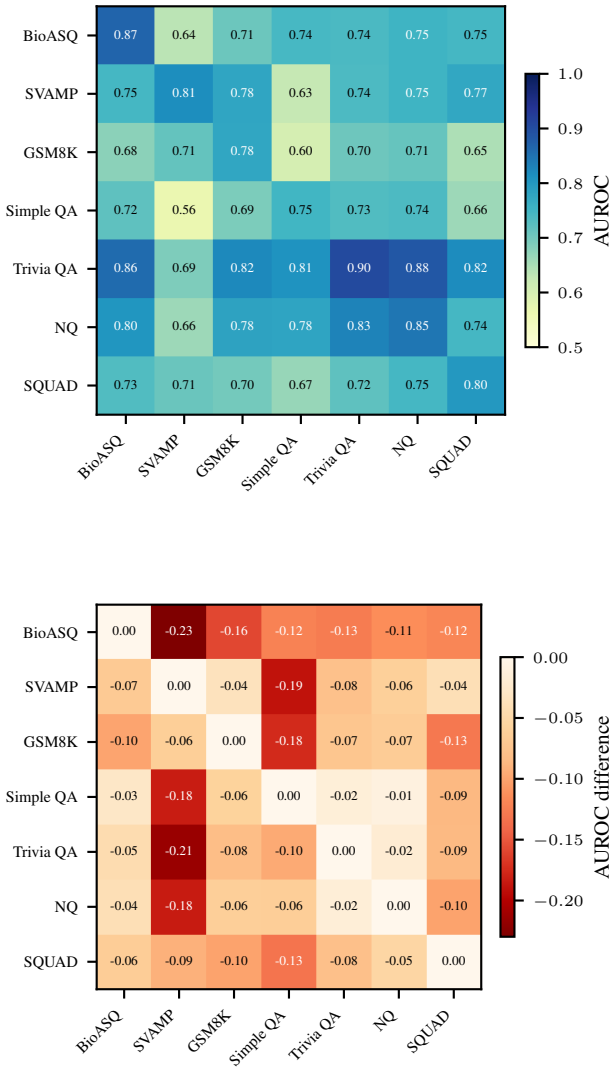


Figure 12. AUROC of probes trained on different tasks on the stop token of the output on last layer. Rows correspond to evaluation tasks while columns correspond to training tasks. The second plot represents the difference between the probe trained on this task and probes trained on the other datasets. Results are averaged over 5 runs.

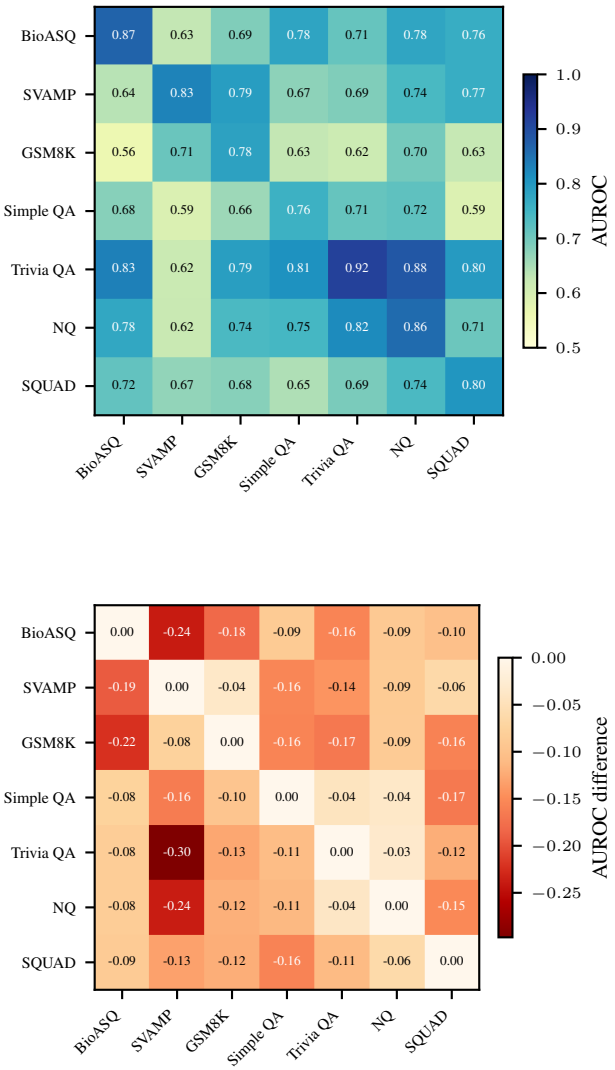


Figure 13. AUROC of probes trained on different tasks on the token before the stop token of the output on last layer. Rows correspond to evaluation tasks while columns correspond to training tasks. The second plot represents the difference between the probe trained on this task and probes trained on the other datasets. Results are averaged over 5 runs.

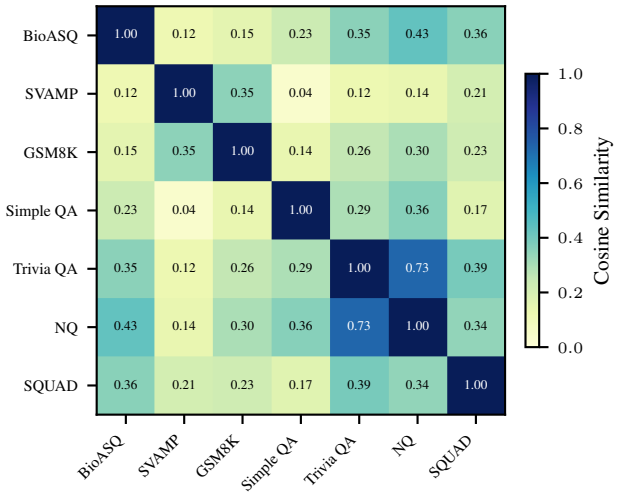


Figure 14. Cosine similarity between probes trained on different datasets using L2 regularization. Results are averaged over 5 runs.

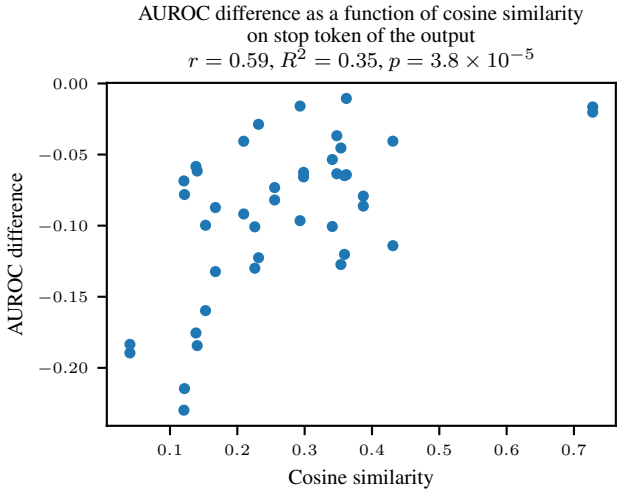


Figure 16. AUROC difference to probe trained on the right dataset as a function of cosine similarity between probes. Results are averaged over 5 runs.

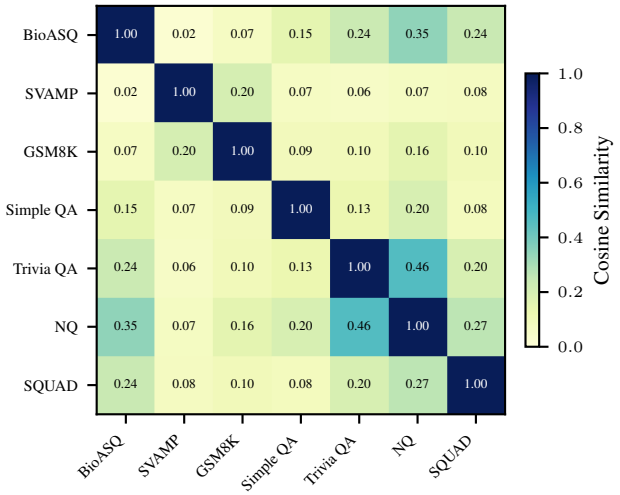


Figure 15. Cosine similarity between probes trained on different datasets using L2 regularization. Results are averaged over 5 runs.

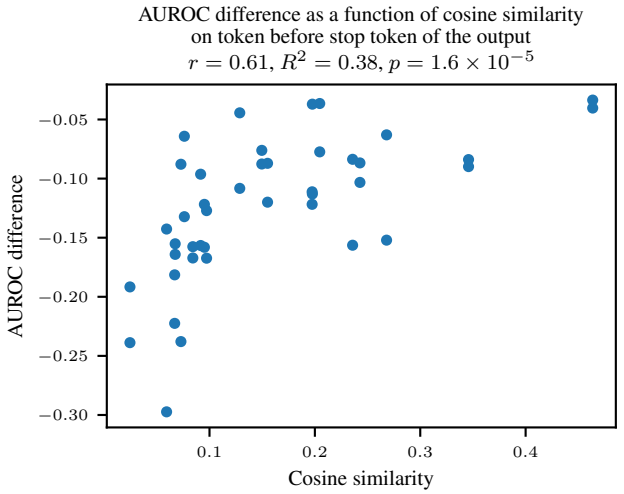


Figure 17. AUROC difference to probe trained on the right dataset as a function of cosine similarity between probes. Results are averaged over 5 runs.

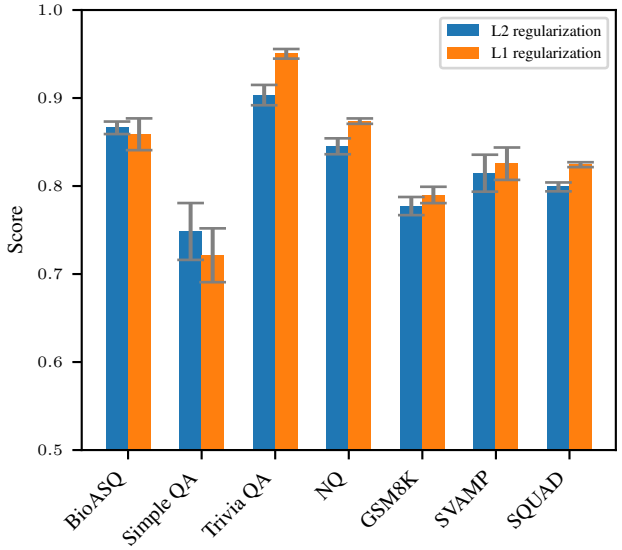


Figure 18. AUROC of linear probes at the stop token of the output, trained with either L1 or L2 regularization.

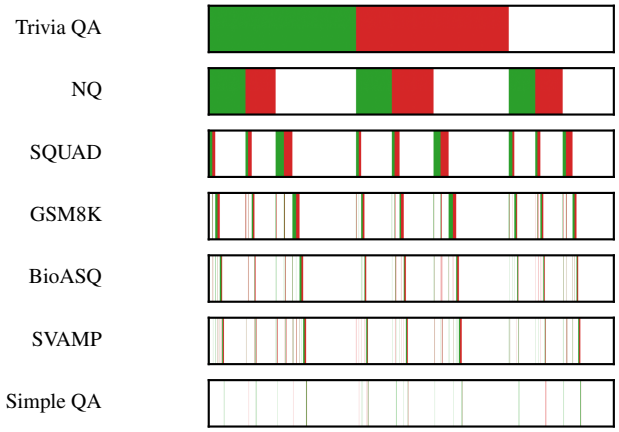


Figure 20. Signed support of sparse probes trained on different datasets at the stop token of the output, using L1 regularization at layer 28. Each row represents one probe trained on the corresponding dataset (y-axis labels). The x-axis shows the 3584 dimensions of the hidden state vector. Green indicates positive coefficients, red indicates negative coefficients, and white indicates zero coefficients (sparsity). Dimensions are sorted by sparsity level across all datasets, with the least sparse dimensions on the left and the most sparse on the right.

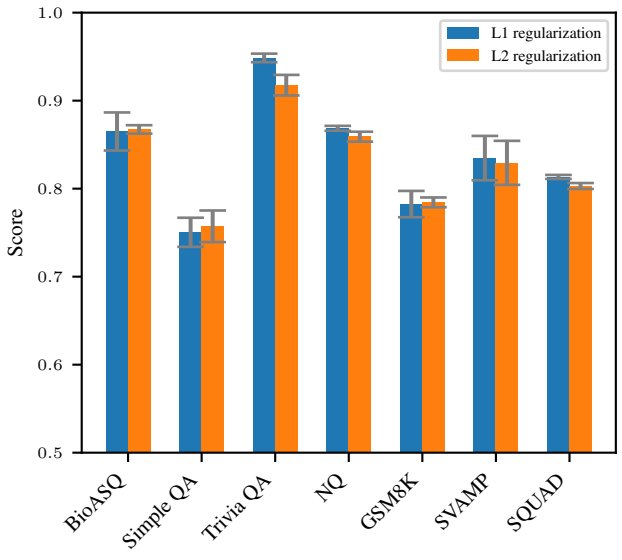


Figure 19. AUROC of linear probes at the token before the stop token of the output, trained with either L1 or L2 regularization.

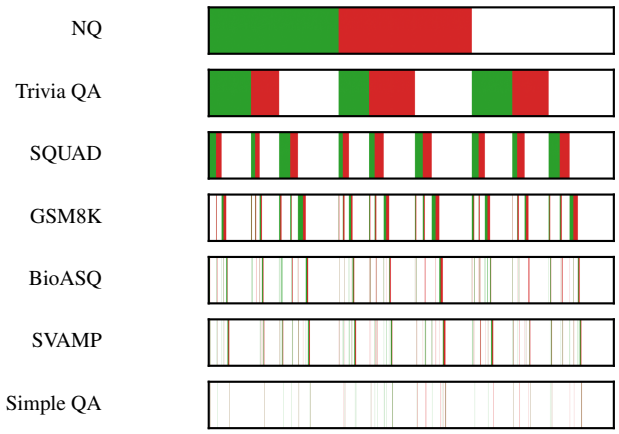


Figure 21. Signed support of sparse probes trained on different datasets at the token before the stop token of the output, using L1 regularization at layer 28. Each row represents one probe trained on the corresponding dataset (y-axis labels). The x-axis shows the 3584 dimensions of the hidden state vector. Green indicates positive coefficients, red indicates negative coefficients, and white indicates zero coefficients (sparsity). Dimensions are sorted by sparsity level across all datasets, with the least sparse dimensions on the left and the most sparse on the right.

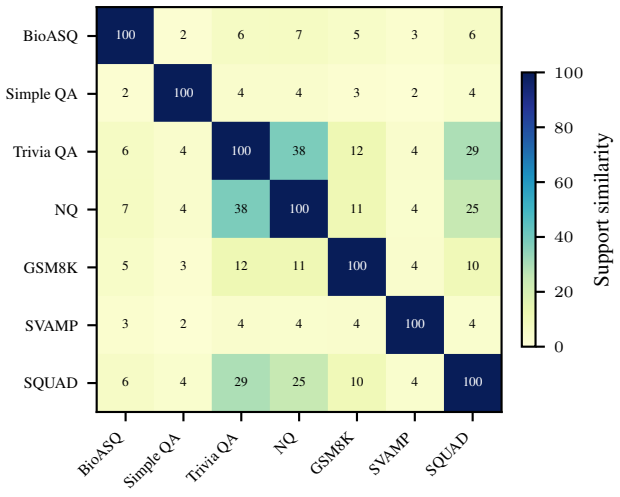


Figure 22. Support overlap between sparse probes trained on different datasets. Darker colors indicate higher overlap percentages. Task pairs with $> 30\%$ overlap (TriviaQA, NQ, SimpleQA) correspond to successful cross-task generalization, while most pairs show $< 15\%$ overlap, explaining generalization failure.

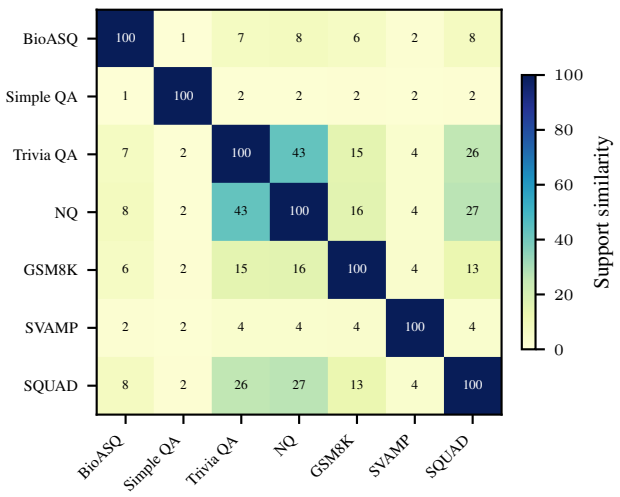


Figure 23. Support overlap between sparse probes trained on different datasets. Darker colors indicate higher overlap percentages. Task pairs with $> 30\%$ overlap (TriviaQA, NQ, SimpleQA) correspond to successful cross-task generalization, while most pairs show $< 15\%$ overlap, explaining generalization failure.

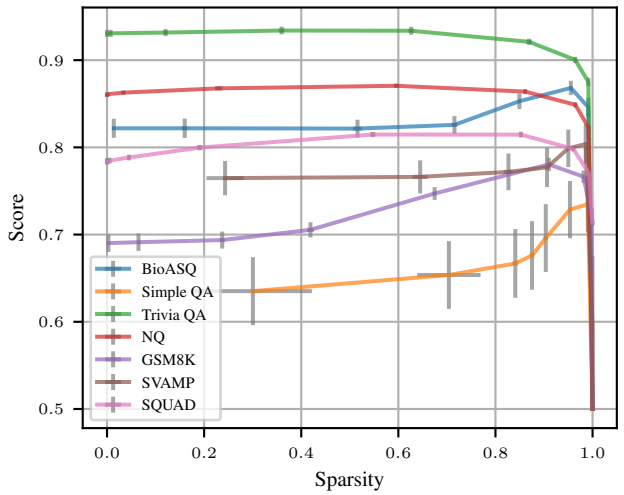


Figure 24. AUROC of probes trained using L1 regularisation as a function of the sparsity level on the stop token of the output.

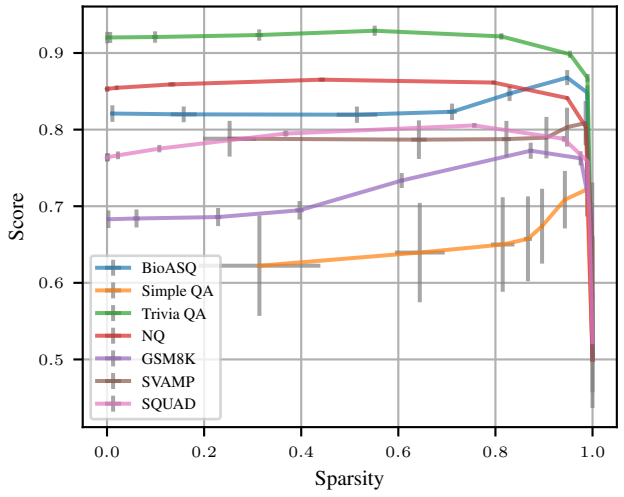


Figure 25. AUROC of probes trained using L1 regularisation as a function of the sparsity level on the token before the stop token of the output.

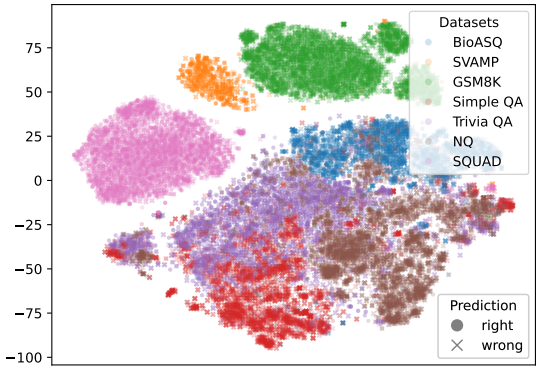


Figure 26. t-SNE plots of the hidden space at layer 28 at the stop token of the output.

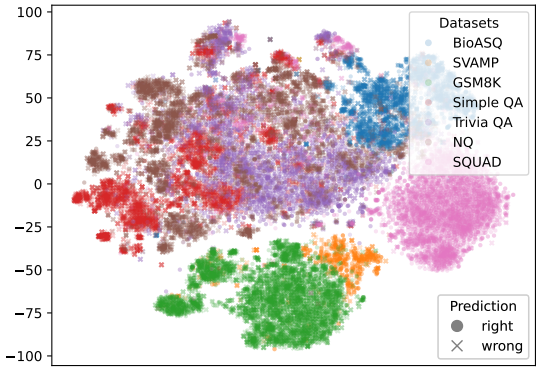


Figure 27. t-SNE plots of the hidden space at layer 28 at the token before stop token of the output.

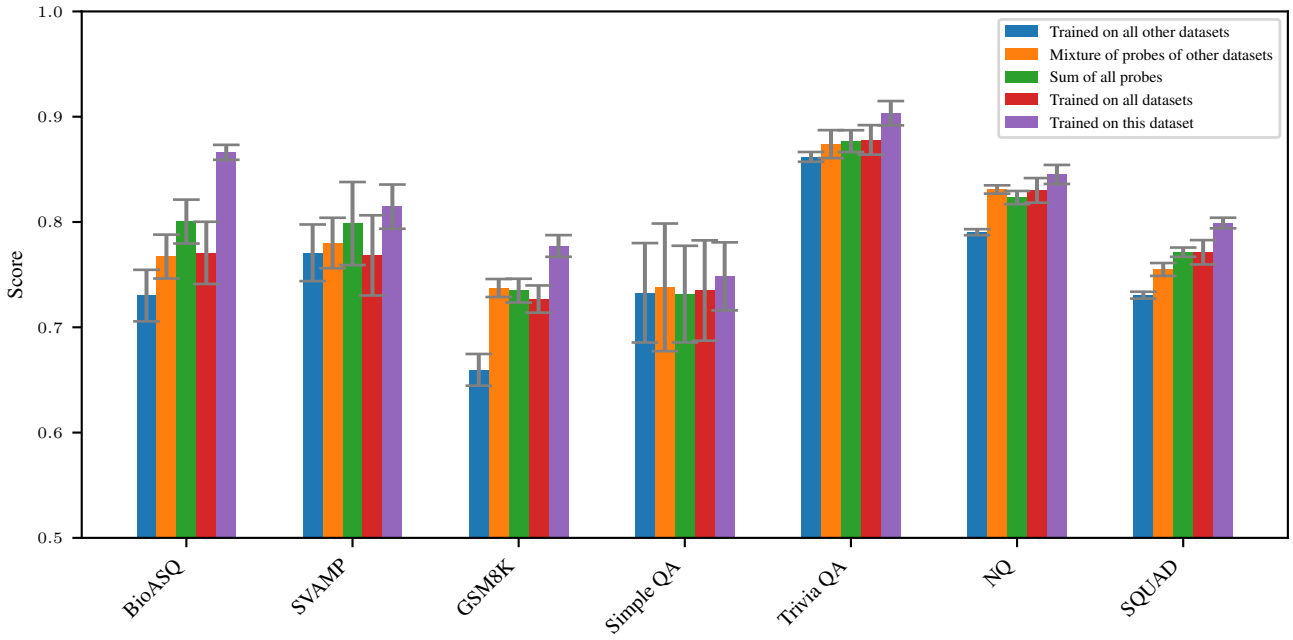


Figure 28. AUROC of linear probes at the stop token of the output in the multi-task setting, using L2 regularization.

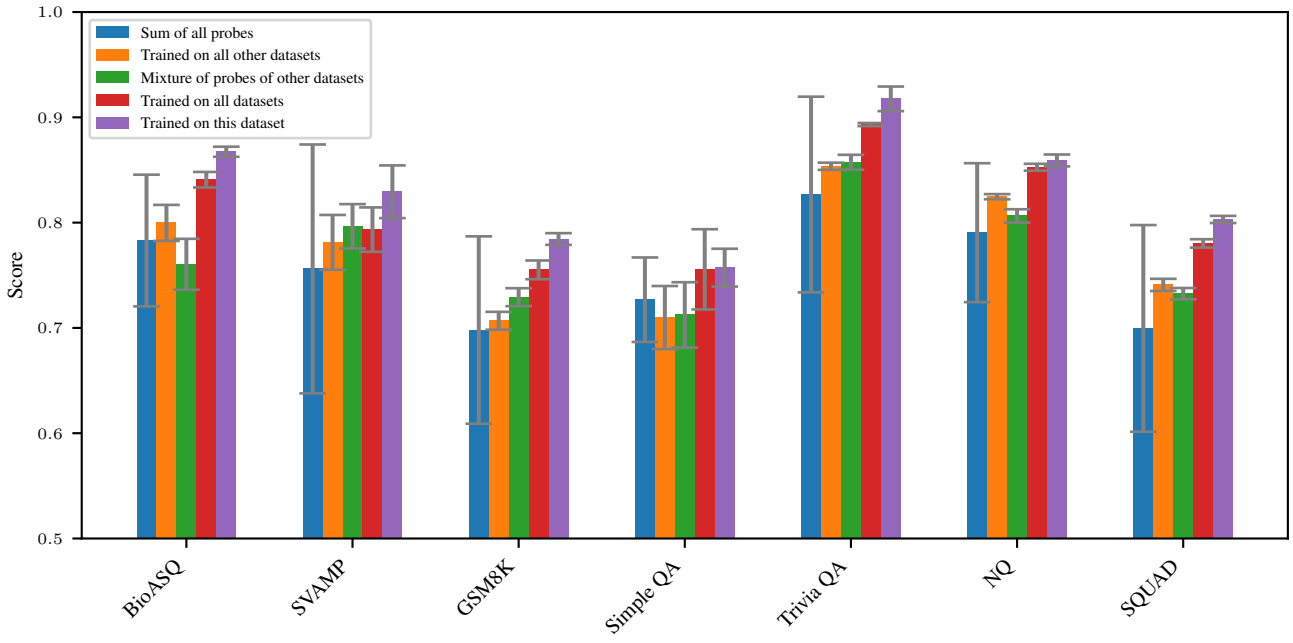


Figure 29. AUROC of linear probes at the token before the stop token of the output in the multi-task setting, using L2 regularization.

A.2. Layer 21

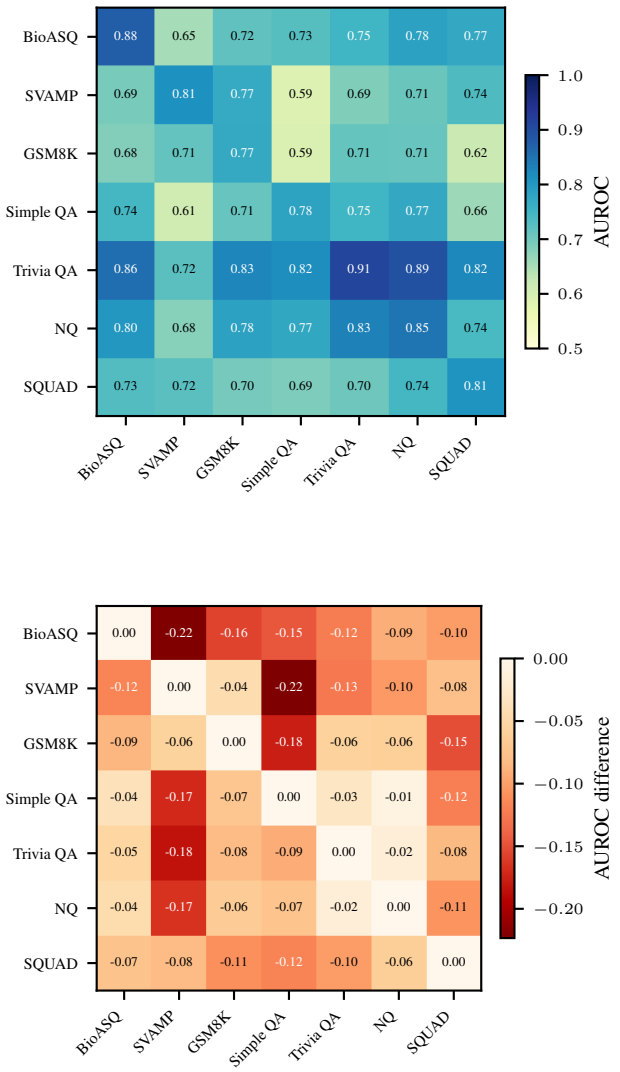


Figure 30. AUROC of probes trained on different tasks on the stop token of the output on last layer. Rows correspond to evaluation tasks while columns correspond to training tasks. The second plot represents the difference between the probe trained on this task and probes trained on the other datasets. Results are averaged over 5 runs.

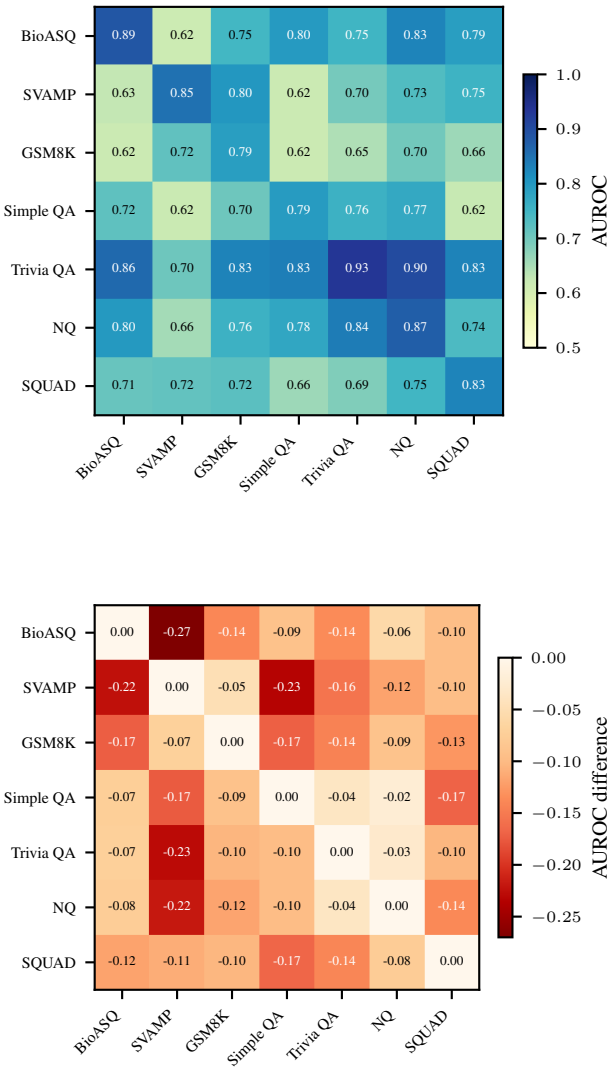


Figure 31. AUROC of probes trained on different tasks on the token before the stop token of the output on last layer. Rows correspond to evaluation tasks while columns correspond to training tasks. The second plot represents the difference between the probe trained on this task and probes trained on the other datasets. Results are averaged over 5 runs.

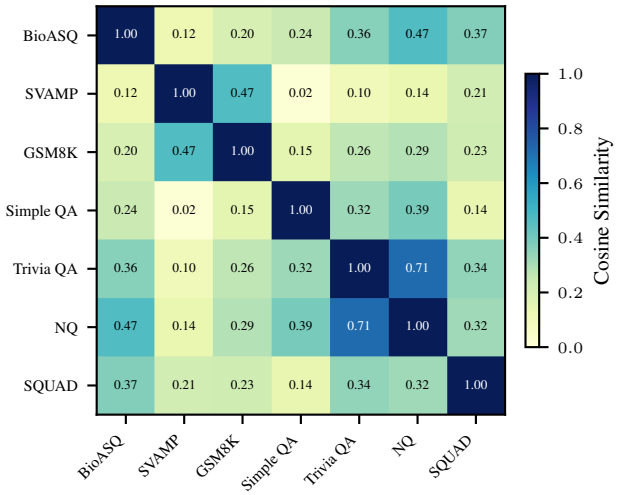


Figure 32. Cosine similarity between probes trained on different datasets using L2 regularization. Results are averaged over 5 runs.

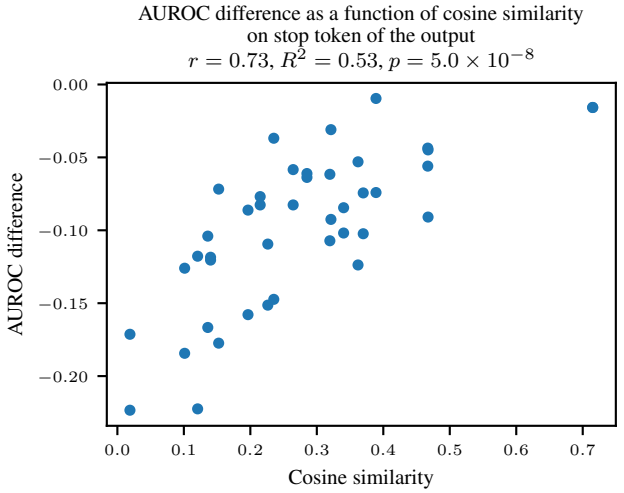


Figure 34. AUROC difference to probe trained on the right dataset as a function of cosine similarity between probes. Results are averaged over 5 runs.

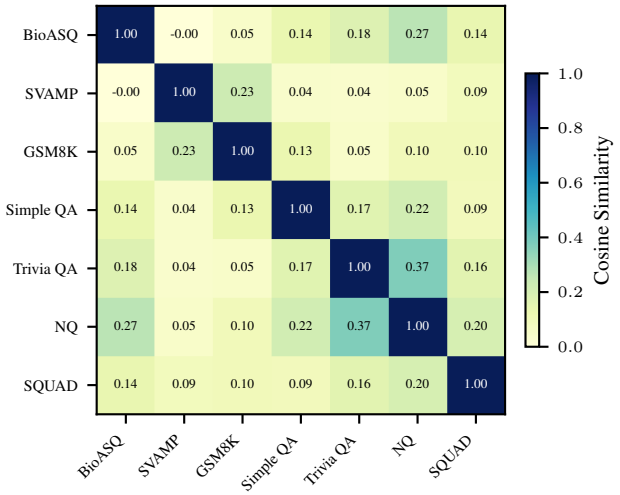


Figure 33. Cosine similarity between probes trained on different datasets using L2 regularization. Results are averaged over 5 runs.

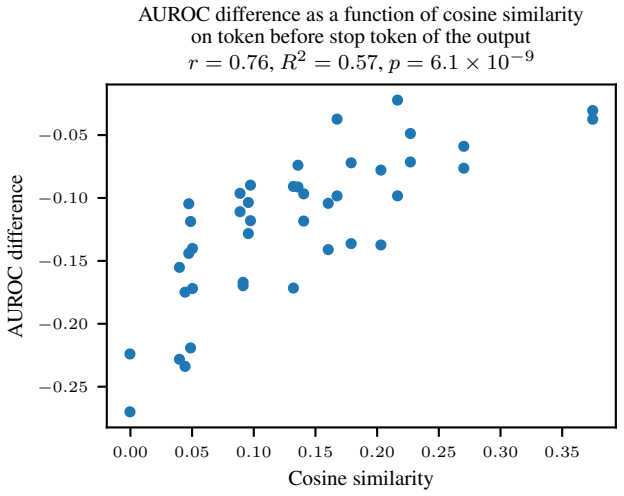


Figure 35. AUROC difference to probe trained on the right dataset as a function of cosine similarity between probes. Results are averaged over 5 runs.

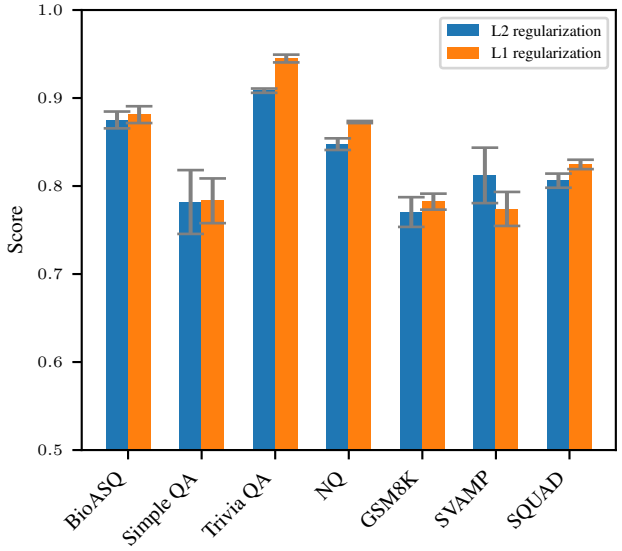


Figure 36. AUROC of linear probes at the stop token of the output, trained with either L1 or L2 regularization.

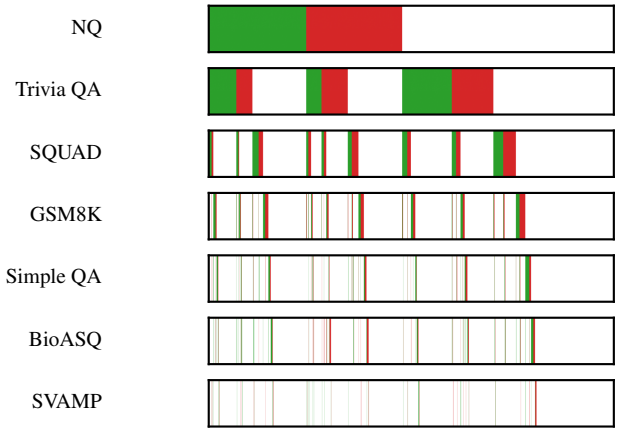


Figure 38. Signed support of sparse probes trained on different datasets at the stop token of the output, using L1 regularization at layer 21. Each row represents one probe trained on the corresponding dataset (y-axis labels). The x-axis shows the 3584 dimensions of the hidden state vector. Green indicates positive coefficients, red indicates negative coefficients, and white indicates zero coefficients (sparsity). Dimensions are sorted by sparsity level across all datasets, with the least sparse dimensions on the left and the most sparse on the right.

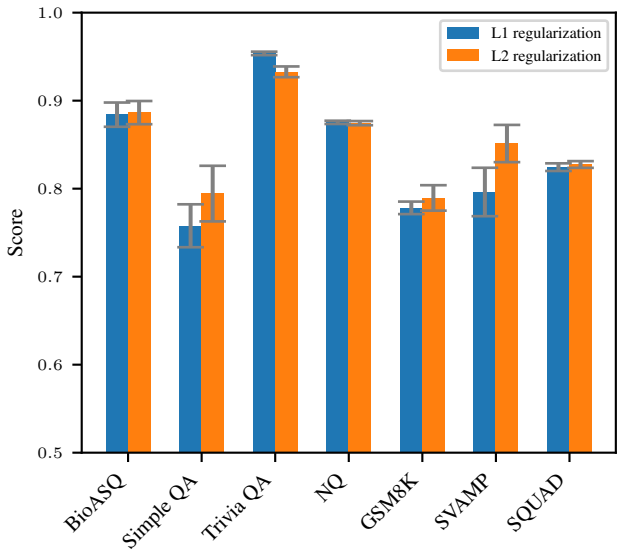


Figure 37. AUROC of linear probes at the token before the stop token of the output, trained with either L1 or L2 regularization.

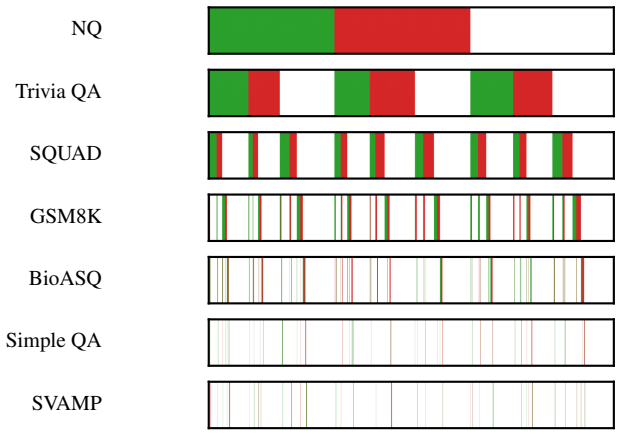


Figure 39. Signed support of sparse probes trained on different datasets at the token before the stop token of the output, using L1 regularization at layer 21. Each row represents one probe trained on the corresponding dataset (y-axis labels). The x-axis shows the 3584 dimensions of the hidden state vector. Green indicates positive coefficients, red indicates negative coefficients, and white indicates zero coefficients (sparsity). Dimensions are sorted by sparsity level across all datasets, with the least sparse dimensions on the left and the most sparse on the right.

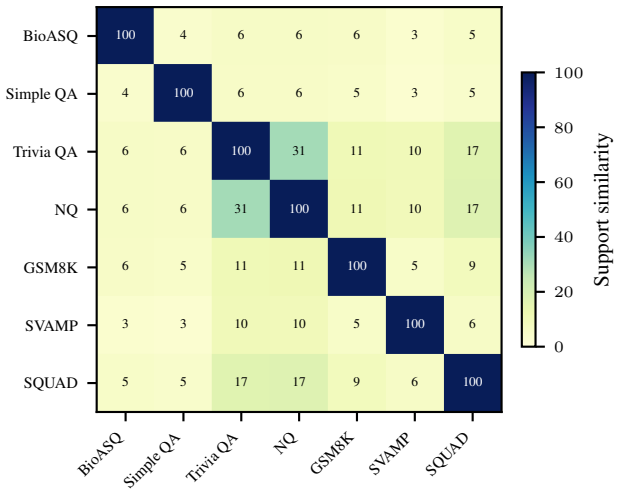


Figure 40. Support overlap between sparse probes trained on different datasets. Darker colors indicate higher overlap percentages. Task pairs with $> 30\%$ overlap (TriviaQA, NQ, SimpleQA) correspond to successful cross-task generalization, while most pairs show $< 15\%$ overlap, explaining generalization failure.

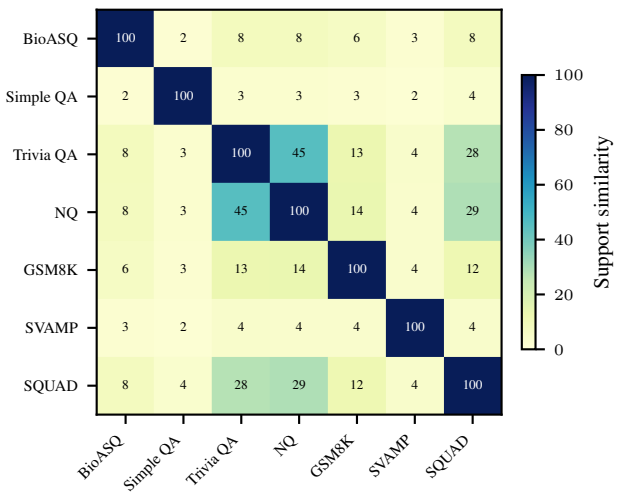


Figure 41. Support overlap between sparse probes trained on different datasets. Darker colors indicate higher overlap percentages. Task pairs with $> 30\%$ overlap (TriviaQA, NQ, SimpleQA) correspond to successful cross-task generalization, while most pairs show $< 15\%$ overlap, explaining generalization failure.

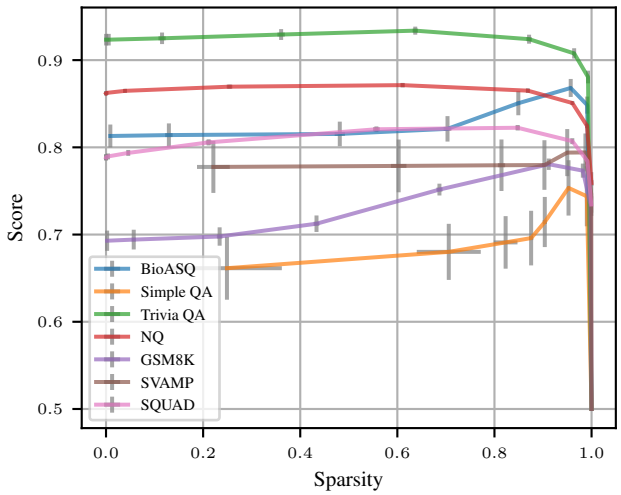


Figure 42. AUROC of probes trained using L1 regularisation as a function of the sparsity level on the stop token of the output.

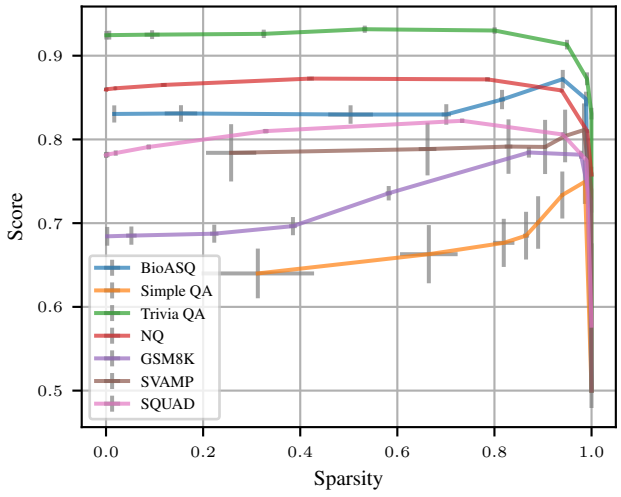


Figure 43. AUROC of probes trained using L1 regularisation as a function of the sparsity level on the token before the stop token of the output.

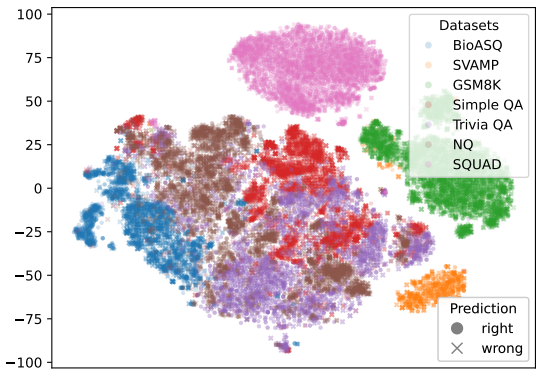


Figure 44. t-SNE plots of the hidden space at layer 21 at the stop token of the output.

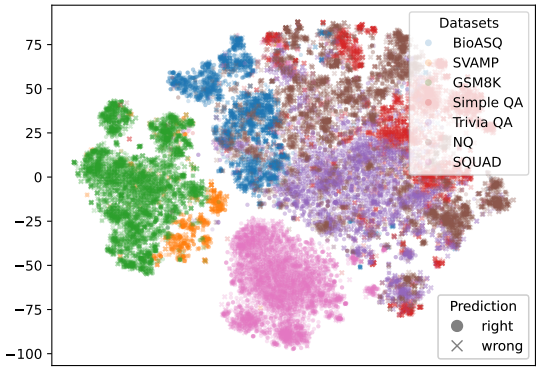


Figure 45. t-SNE plots of the hidden space at layer 21 at the token before stop token of the output.

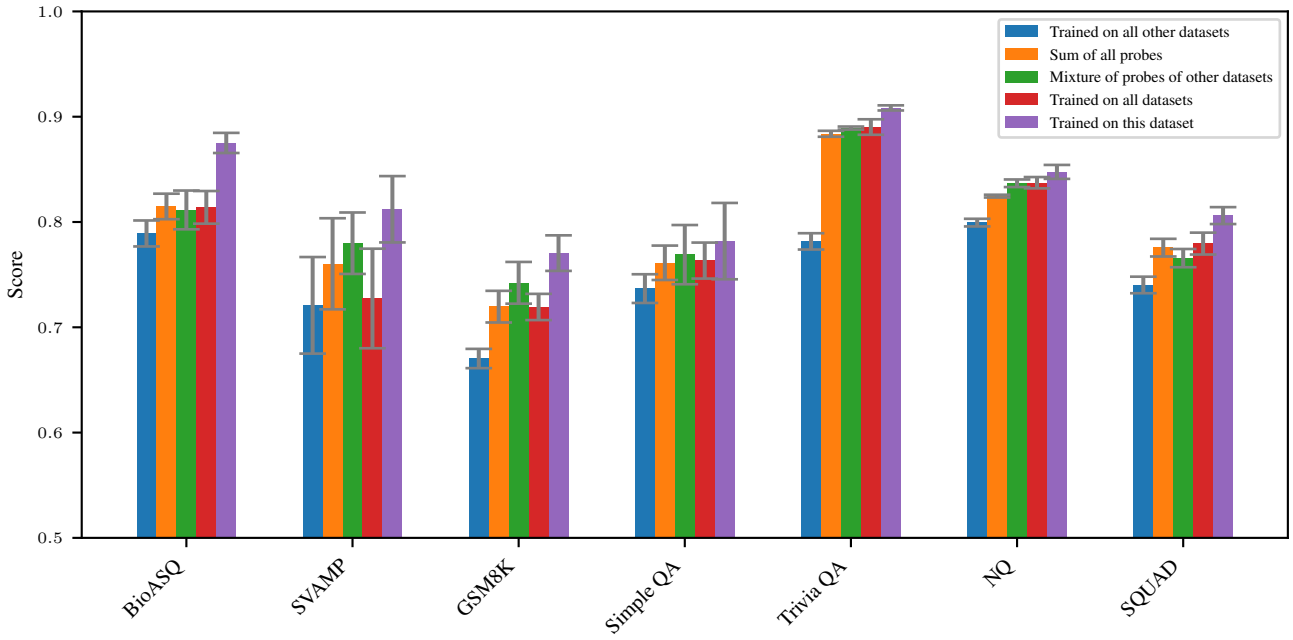


Figure 46. AUROC of linear probes at the stop token of the output in the multi-task setting, using L2 regularization.

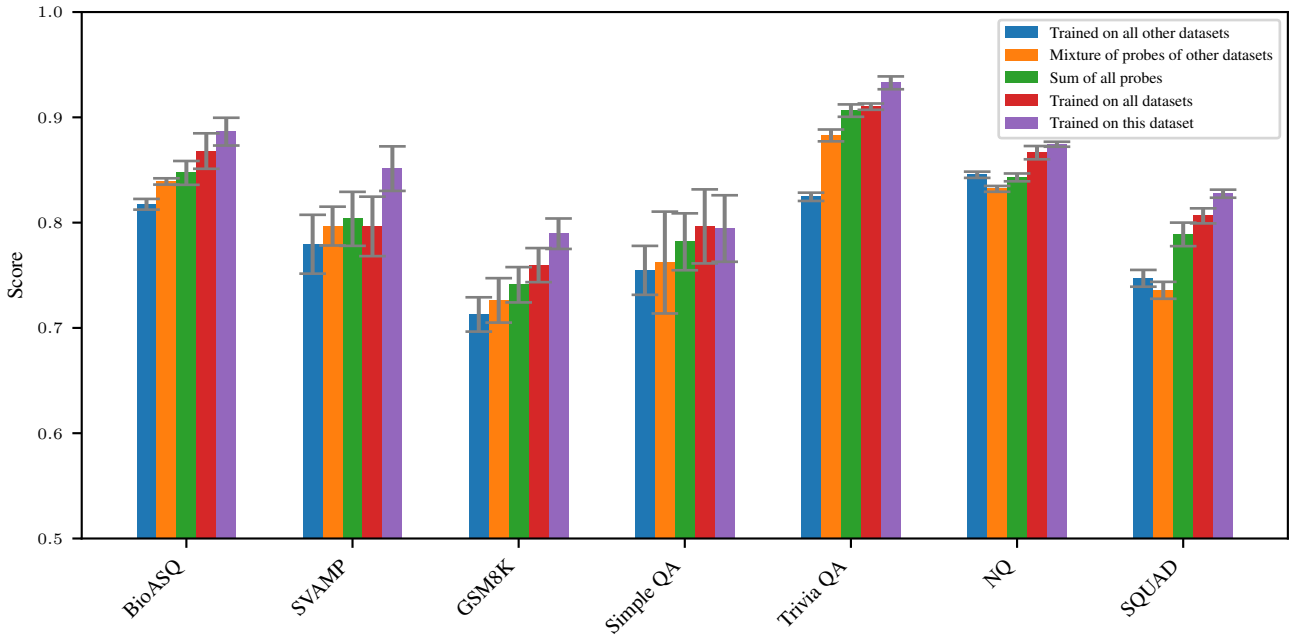


Figure 47. AUROC of linear probes at the token before the stop token of the output in the multi-task setting, using L2 regularization.

B. Llama 3.1 8B Instruct

B.1. Layer 32

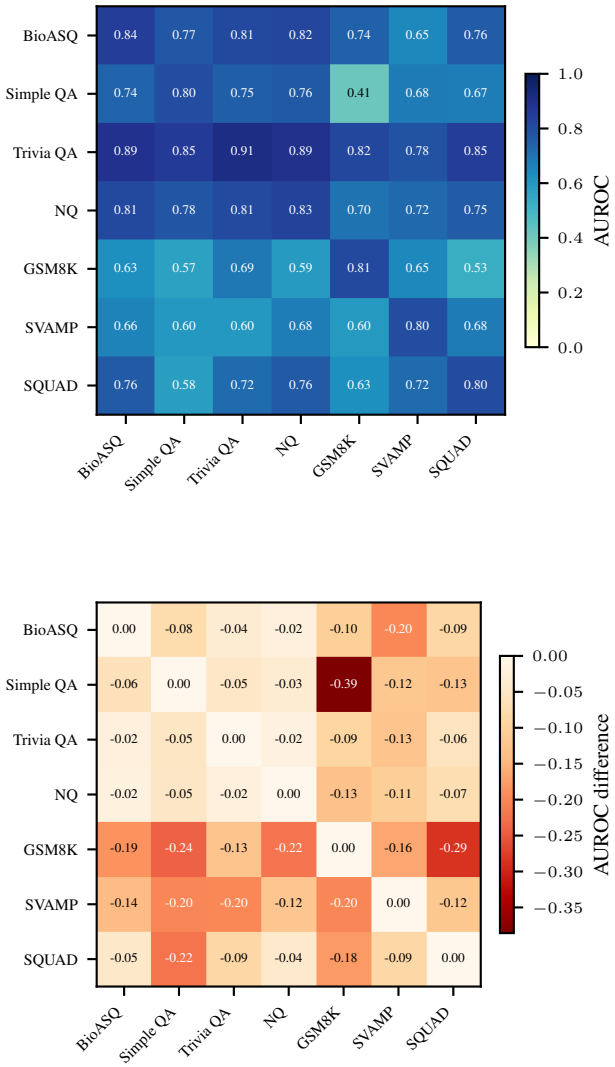


Figure 48. AUROC of probes trained on different tasks on the stop token of the output on last layer. Rows correspond to evaluation tasks while columns correspond to training tasks. The second plot represents the difference between the probe trained on this task and probes trained on the other datasets. Results are averaged over 5 runs.

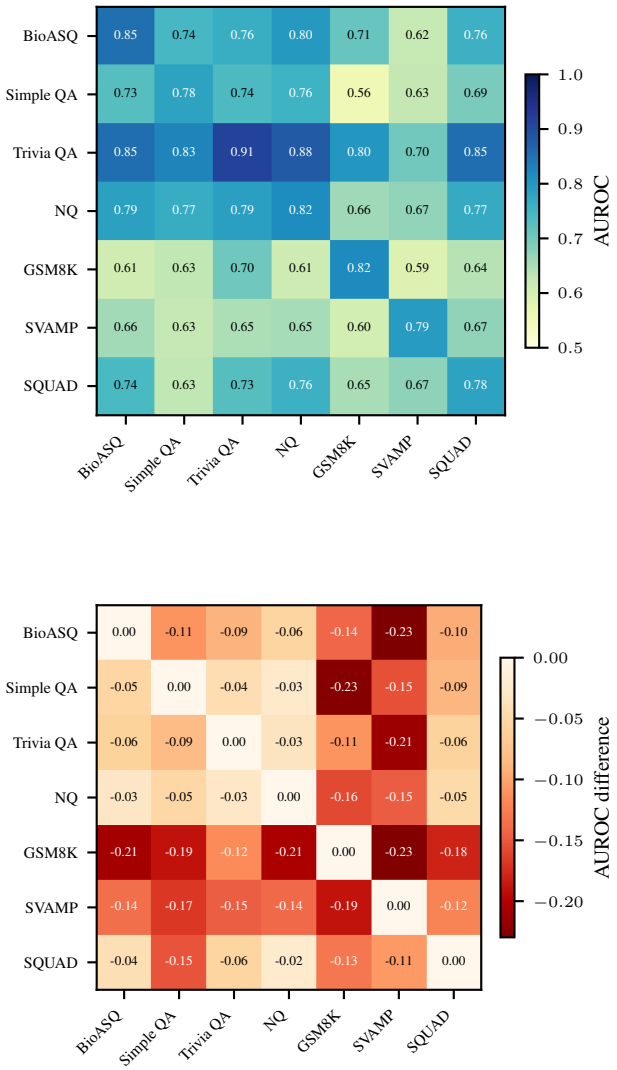


Figure 49. AUROC of probes trained on different tasks on the token before the stop token of the output on last layer. Rows correspond to evaluation tasks while columns correspond to training tasks. The second plot represents the difference between the probe trained on this task and probes trained on the other datasets. Results are averaged over 5 runs.

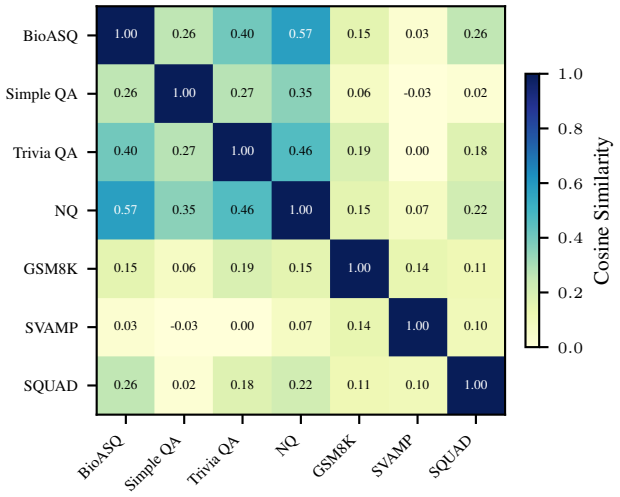


Figure 50. Cosine similarity between probes trained on different datasets using L2 regularization. Results are averaged over 5 runs.

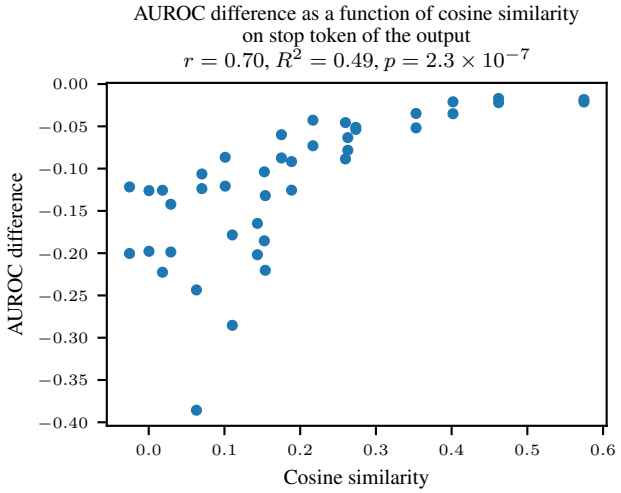


Figure 52. AUROC difference to probe trained on the right dataset as a function of cosine similarity between probes. Results are averaged over 5 runs.

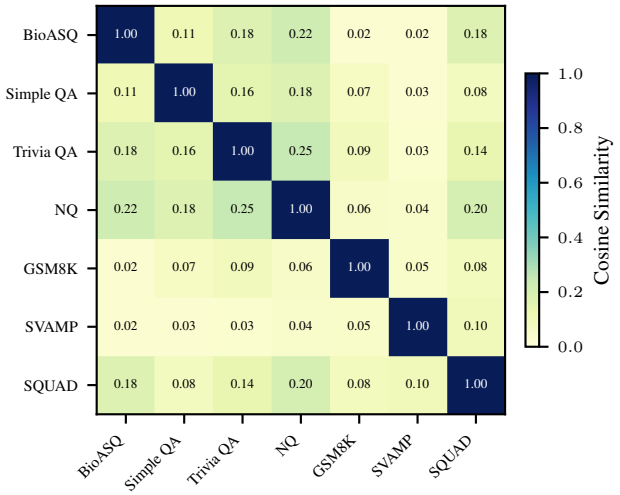


Figure 51. Cosine similarity between probes trained on different datasets using L2 regularization. Results are averaged over 5 runs.

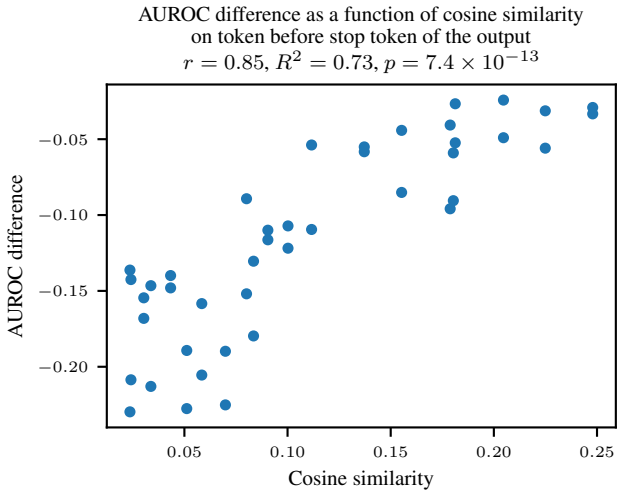


Figure 53. AUROC difference to probe trained on the right dataset as a function of cosine similarity between probes. Results are averaged over 5 runs.

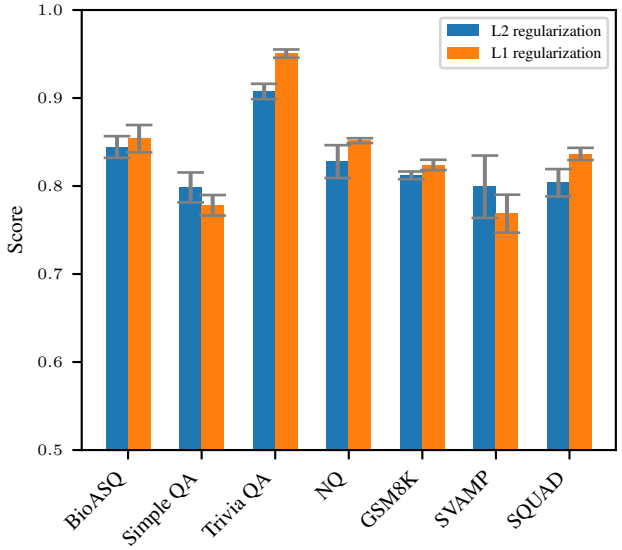


Figure 54. AUROC of linear probes at the stop token of the output, trained with either L1 or L2 regularization.

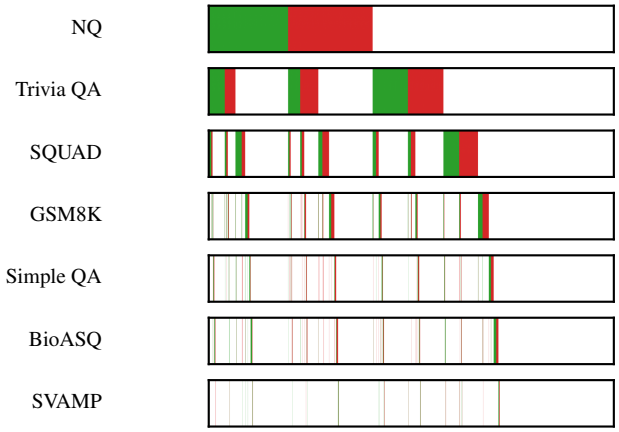


Figure 56. Signed support of sparse probes trained on different datasets at the stop token of the output, using L1 regularization at layer 32. Each row represents one probe trained on the corresponding dataset (y-axis labels). The x-axis shows the 3584 dimensions of the hidden state vector. Green indicates positive coefficients, red indicates negative coefficients, and white indicates zero coefficients (sparsity). Dimensions are sorted by sparsity level across all datasets, with the least sparse dimensions on the left and the most sparse on the right.

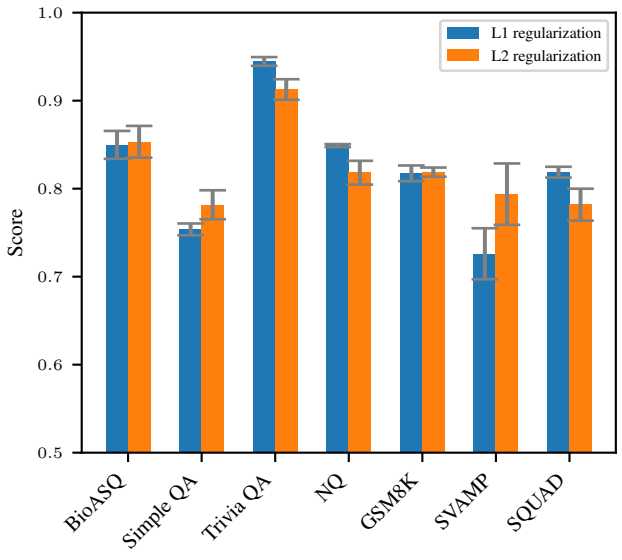


Figure 55. AUROC of linear probes at the token before the stop token of the output, trained with either L1 or L2 regularization.

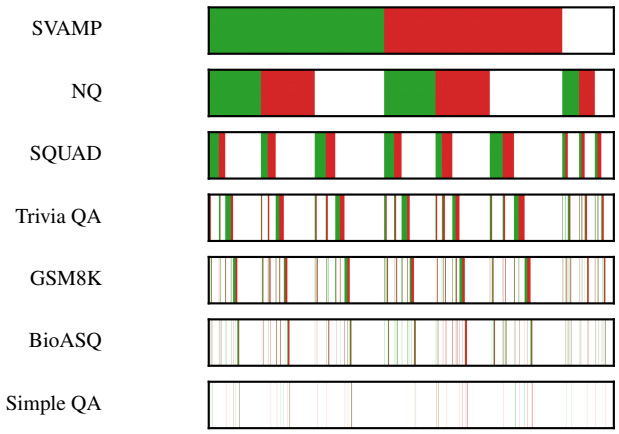


Figure 57. Signed support of sparse probes trained on different datasets at the token before the stop token of the output, using L1 regularization at layer 32. Each row represents one probe trained on the corresponding dataset (y-axis labels). The x-axis shows the 3584 dimensions of the hidden state vector. Green indicates positive coefficients, red indicates negative coefficients, and white indicates zero coefficients (sparsity). Dimensions are sorted by sparsity level across all datasets, with the least sparse dimensions on the left and the most sparse on the right.

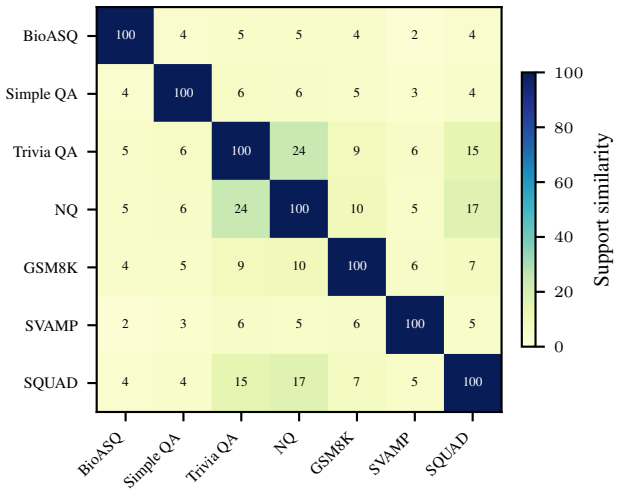


Figure 58. Support overlap between sparse probes trained on different datasets. Darker colors indicate higher overlap percentages. Task pairs with $> 30\%$ overlap (TriviaQA, NQ, SimpleQA) correspond to successful cross-task generalization, while most pairs show $< 15\%$ overlap, explaining generalization failure.

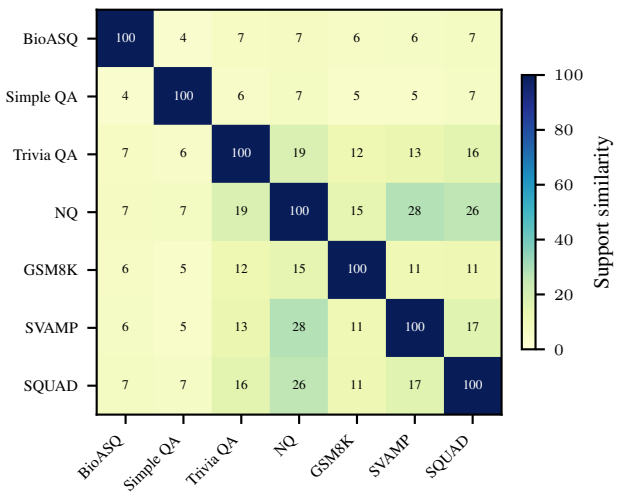


Figure 59. Support overlap between sparse probes trained on different datasets. Darker colors indicate higher overlap percentages. Task pairs with $> 30\%$ overlap (TriviaQA, NQ, SimpleQA) correspond to successful cross-task generalization, while most pairs show $< 15\%$ overlap, explaining generalization failure.

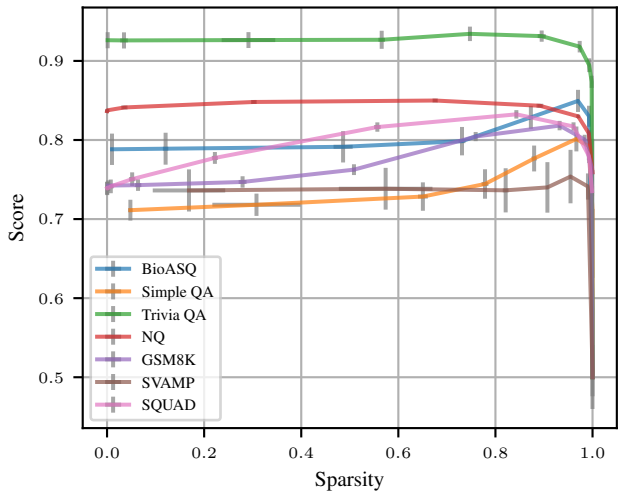


Figure 60. AUROC of probes trained using L1 regularisation as a function of the sparsity level on the stop token of the output.

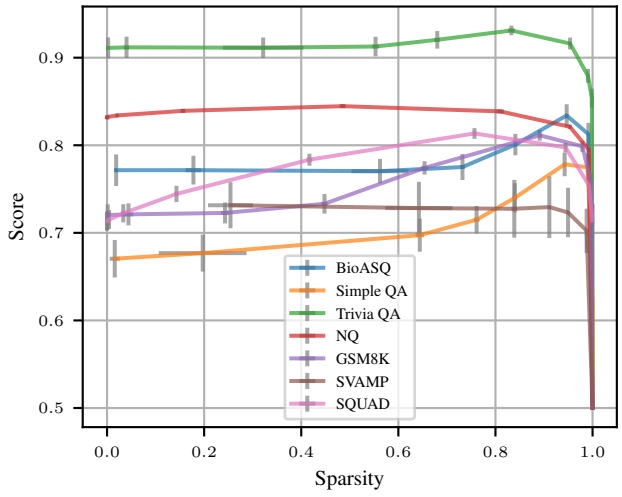


Figure 61. AUROC of probes trained using L1 regularisation as a function of the sparsity level on the token before the stop token of the output.

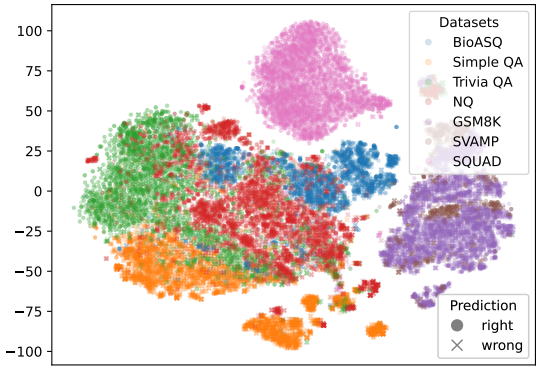


Figure 62. t-SNE plots of the hidden space at layer 32 at the stop token of the output.

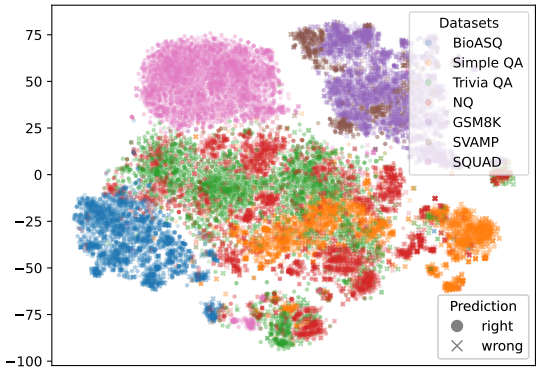


Figure 63. t-SNE plots of the hidden space at layer 32 at the token before stop token of the output.

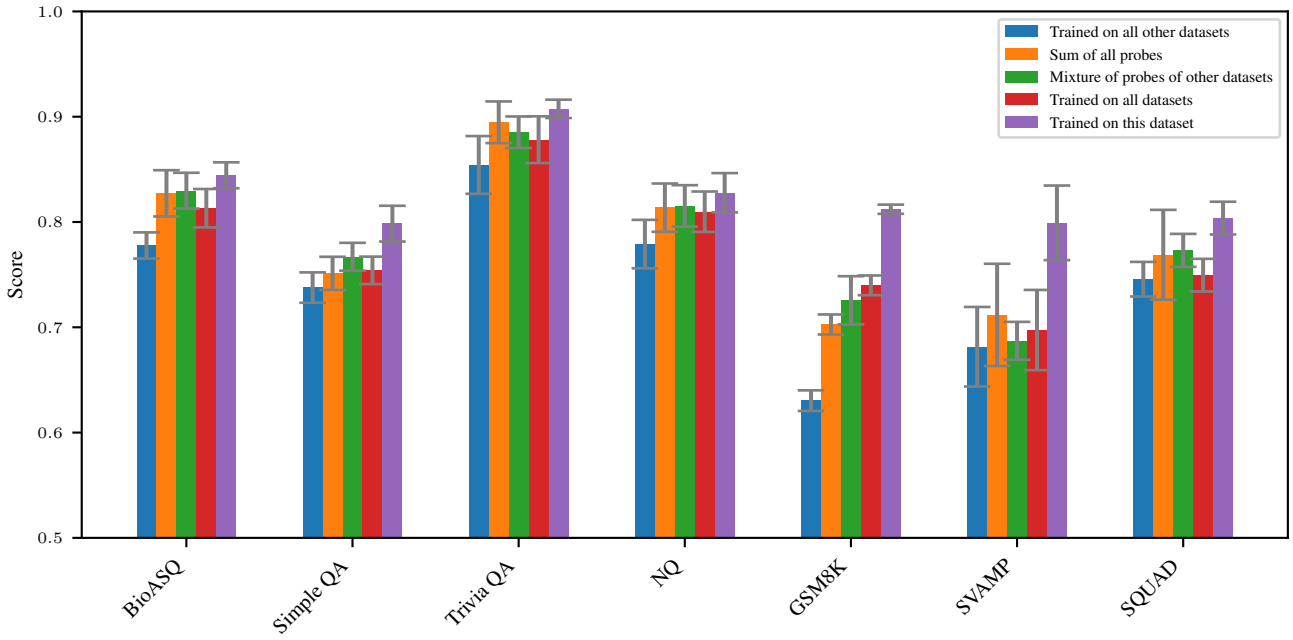


Figure 64. AUROC of linear probes at the stop token of the output in the multi-task setting, using L2 regularization.

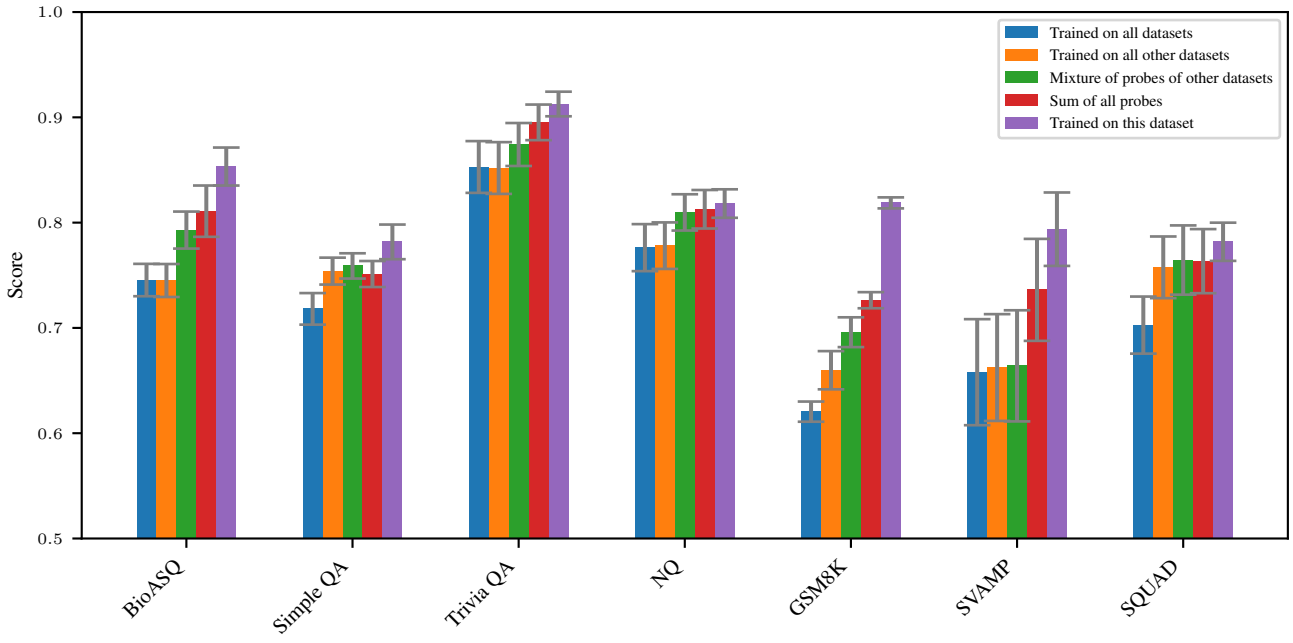


Figure 65. AUROC of linear probes at the token before the stop token of the output in the multi-task setting, using L2 regularization.

B.2. Layer 24

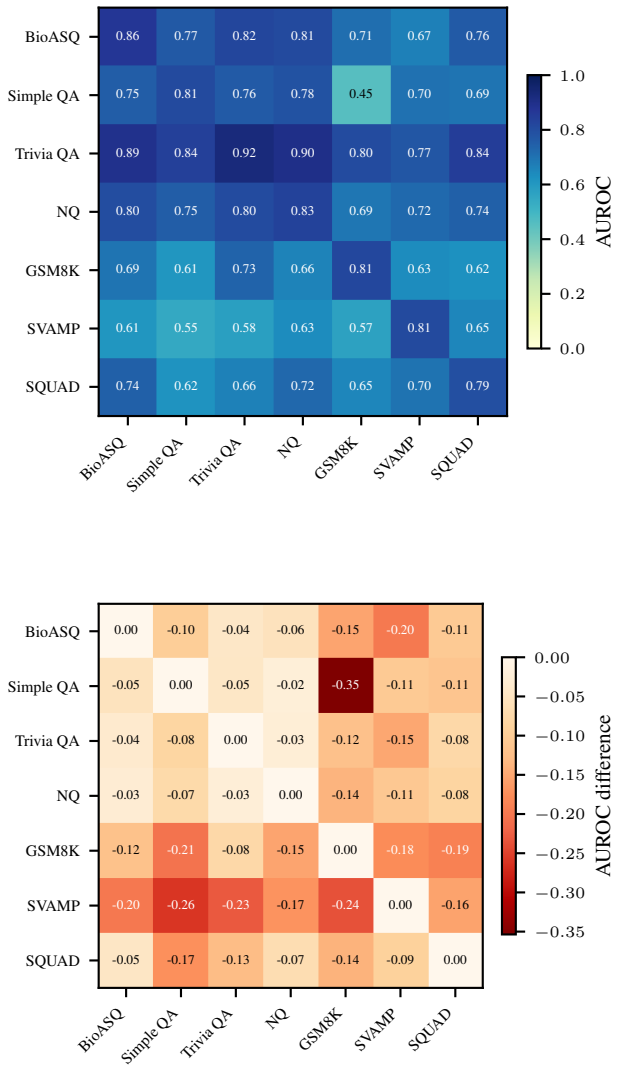


Figure 66. AUROC of probes trained on different tasks on the stop token of the output on last layer. Rows correspond to evaluation tasks while columns correspond to training tasks. The second plot represents the difference between the probe trained on this task and probes trained on the other datasets. Results are averaged over 5 runs.

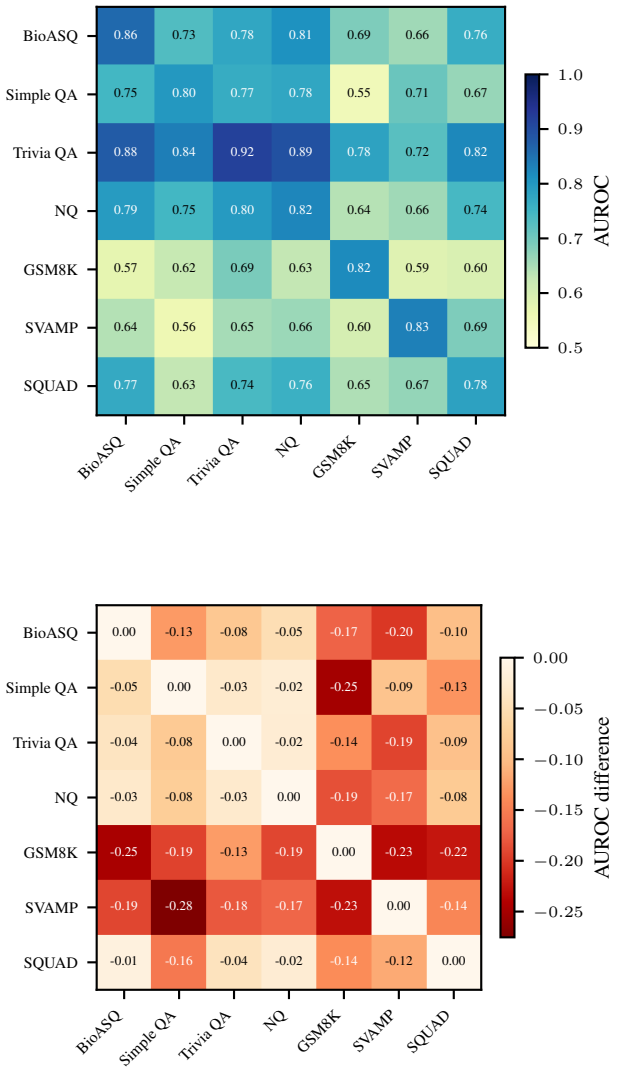


Figure 67. AUROC of probes trained on different tasks on the token before the stop token of the output on last layer. Rows correspond to evaluation tasks while columns correspond to training tasks. The second plot represents the difference between the probe trained on this task and probes trained on the other datasets. Results are averaged over 5 runs.

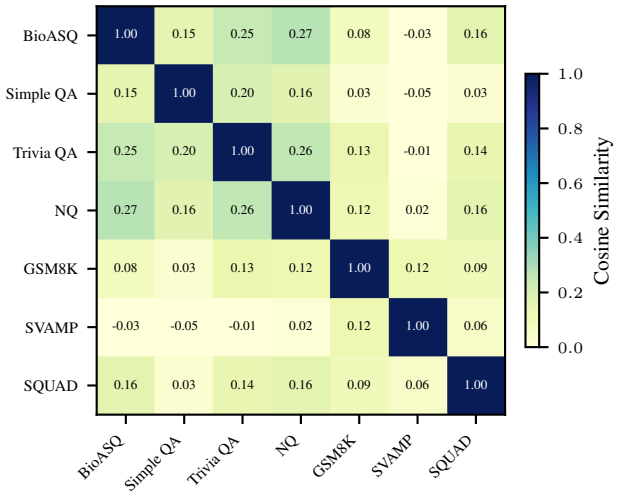


Figure 68. Cosine similarity between probes trained on different datasets using L2 regularization. Results are averaged over 5 runs.

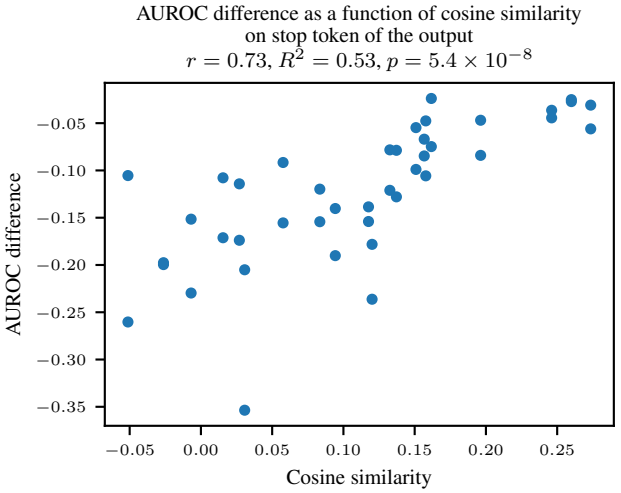


Figure 70. AUROC difference to probe trained on the right dataset as a function of cosine similarity between probes. Results are averaged over 5 runs.

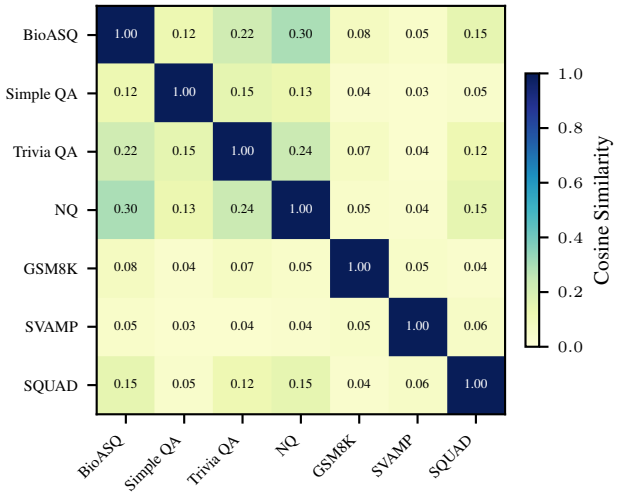


Figure 69. Cosine similarity between probes trained on different datasets using L2 regularization. Results are averaged over 5 runs.

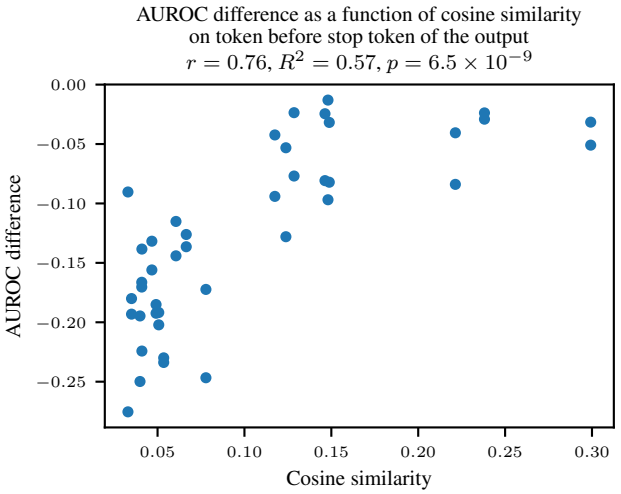


Figure 71. AUROC difference to probe trained on the right dataset as a function of cosine similarity between probes. Results are averaged over 5 runs.

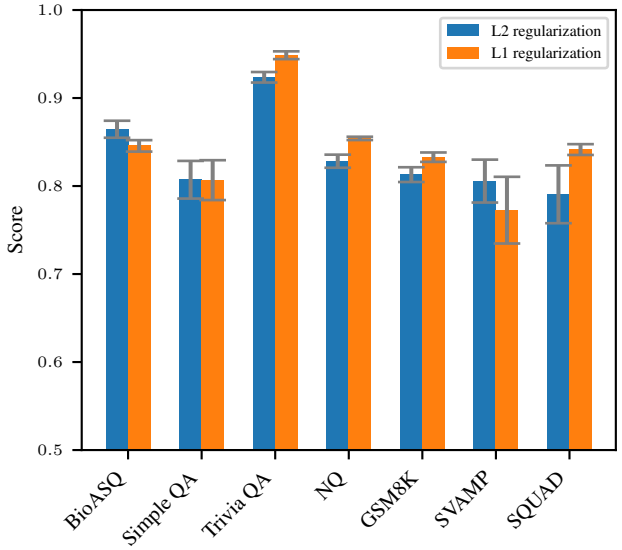


Figure 72. AUROC of linear probes at the stop token of the output, trained with either L1 or L2 regularization.

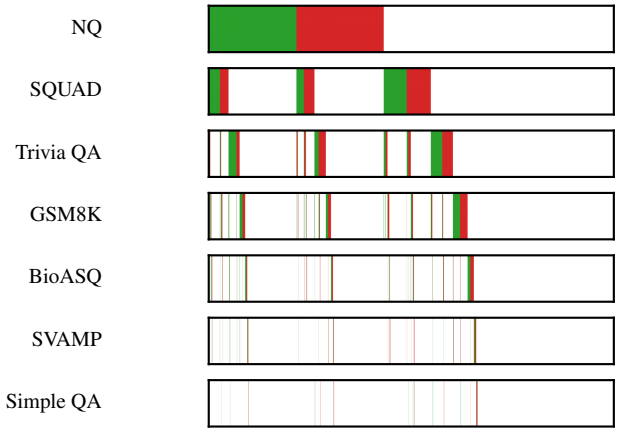


Figure 74. Signed support of sparse probes trained on different datasets at the stop token of the output, using L1 regularization at layer 24. Each row represents one probe trained on the corresponding dataset (y-axis labels). The x-axis shows the 3584 dimensions of the hidden state vector. Green indicates positive coefficients, red indicates negative coefficients, and white indicates zero coefficients (sparsity). Dimensions are sorted by sparsity level across all datasets, with the least sparse dimensions on the left and the most sparse on the right.

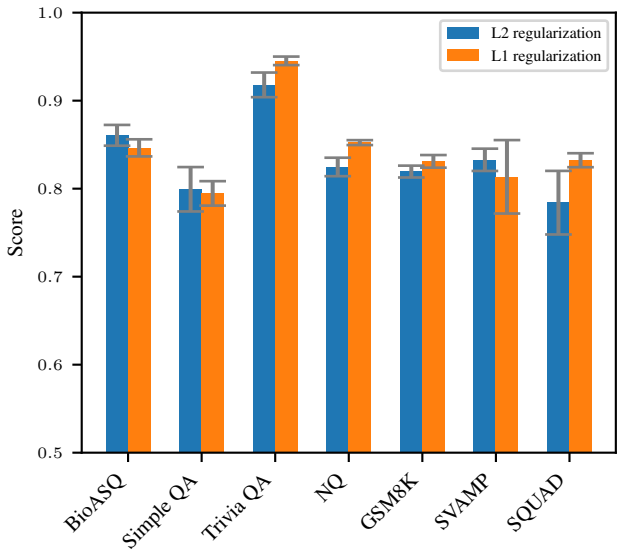


Figure 73. AUROC of linear probes at the token before the stop token of the output, trained with either L1 or L2 regularization.

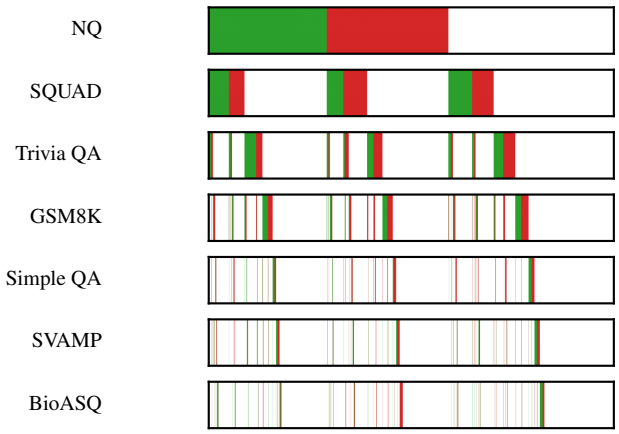


Figure 75. Signed support of sparse probes trained on different datasets at the token before the stop token of the output, using L1 regularization at layer 24. Each row represents one probe trained on the corresponding dataset (y-axis labels). The x-axis shows the 3584 dimensions of the hidden state vector. Green indicates positive coefficients, red indicates negative coefficients, and white indicates zero coefficients (sparsity). Dimensions are sorted by sparsity level across all datasets, with the least sparse dimensions on the left and the most sparse on the right.

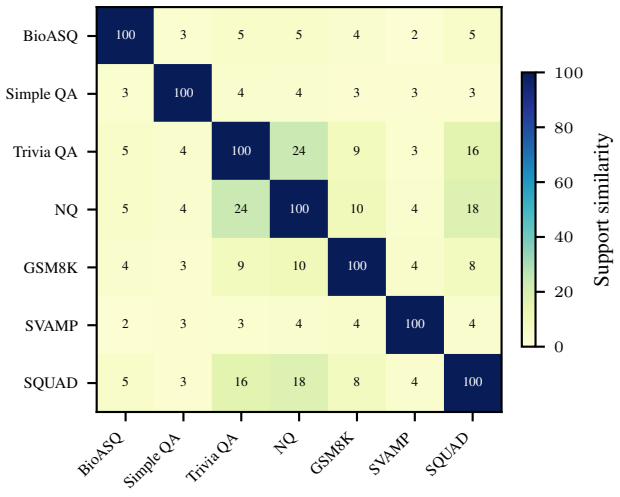


Figure 76. Support overlap between sparse probes trained on different datasets. Darker colors indicate higher overlap percentages. Task pairs with $> 30\%$ overlap (TriviaQA, NQ, SimpleQA) correspond to successful cross-task generalization, while most pairs show $< 15\%$ overlap, explaining generalization failure.

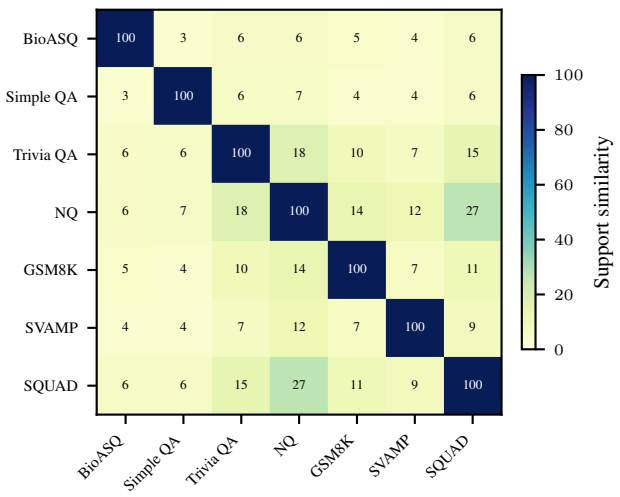


Figure 77. Support overlap between sparse probes trained on different datasets. Darker colors indicate higher overlap percentages. Task pairs with $> 30\%$ overlap (TriviaQA, NQ, SimpleQA) correspond to successful cross-task generalization, while most pairs show $< 15\%$ overlap, explaining generalization failure.

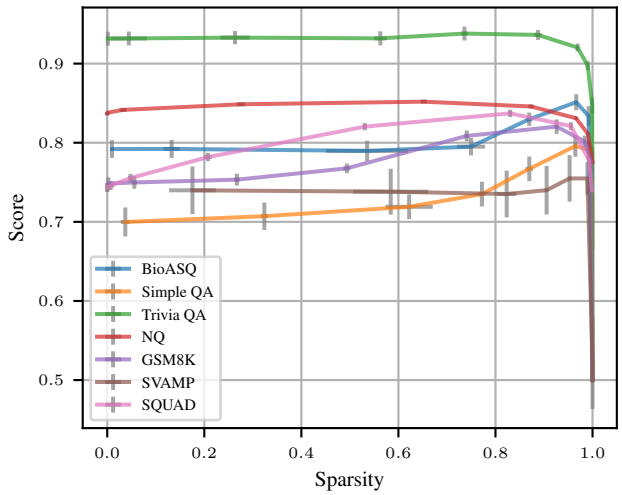


Figure 78. AUROC of probes trained using L1 regularisation as a function of the sparsity level on the stop token of the output.

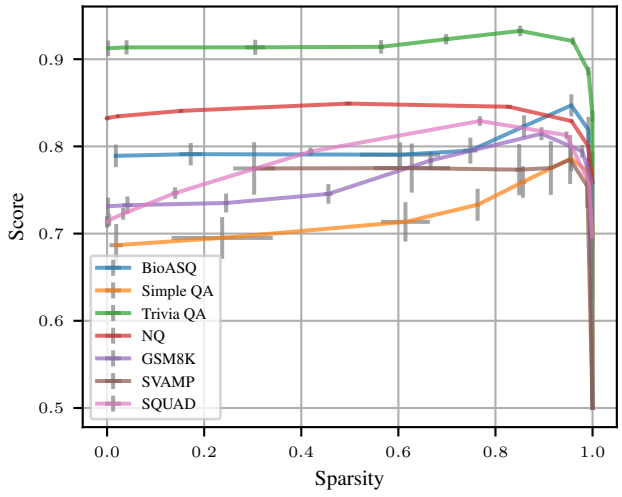


Figure 79. AUROC of probes trained using L1 regularisation as a function of the sparsity level on the token before the stop token of the output.

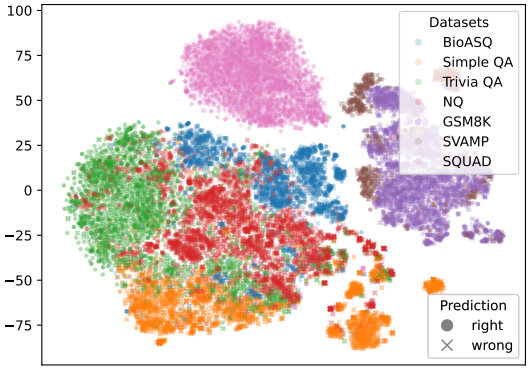


Figure 80. t-SNE plots of the hidden space at layer 24 at the stop token of the output.

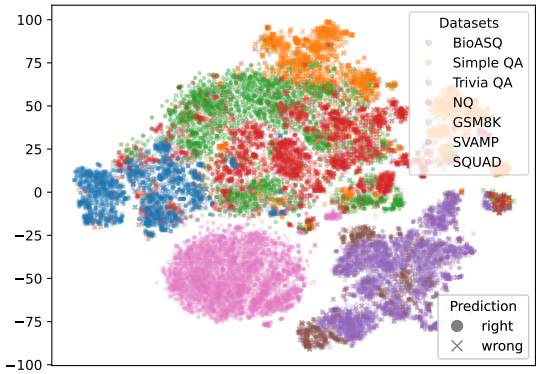


Figure 81. t-SNE plots of the hidden space at layer 24 at the token before stop token of the output.

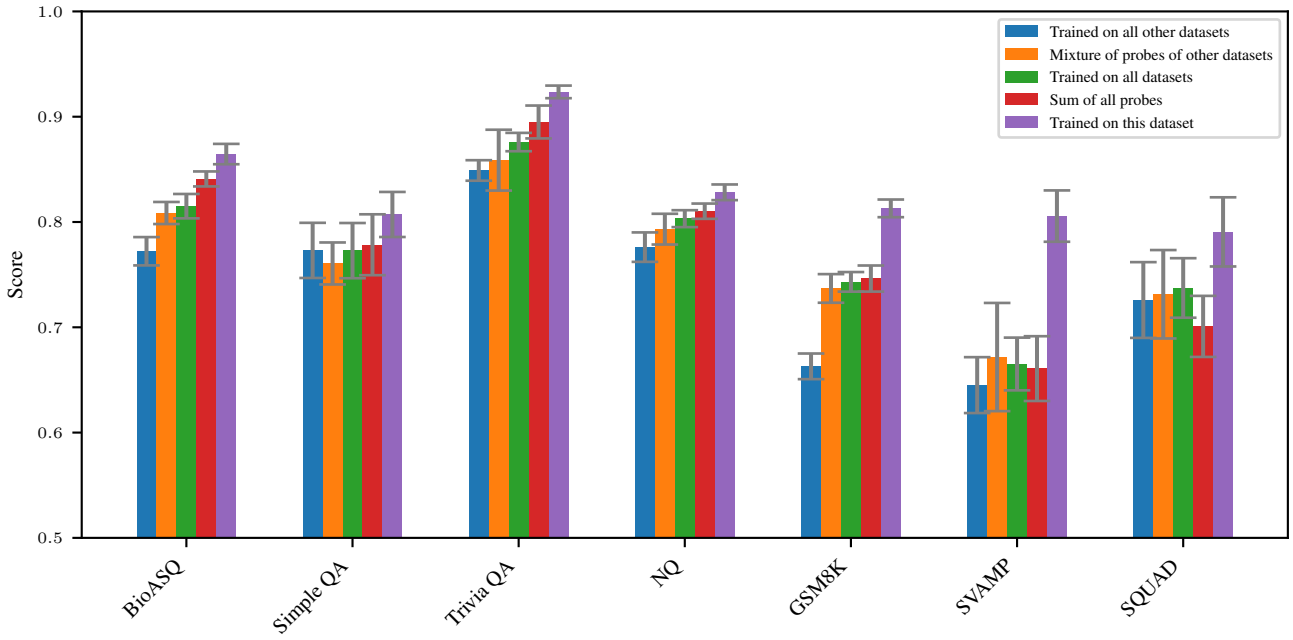


Figure 82. AUROC of linear probes at the stop token of the output in the multi-task setting, using L2 regularization.

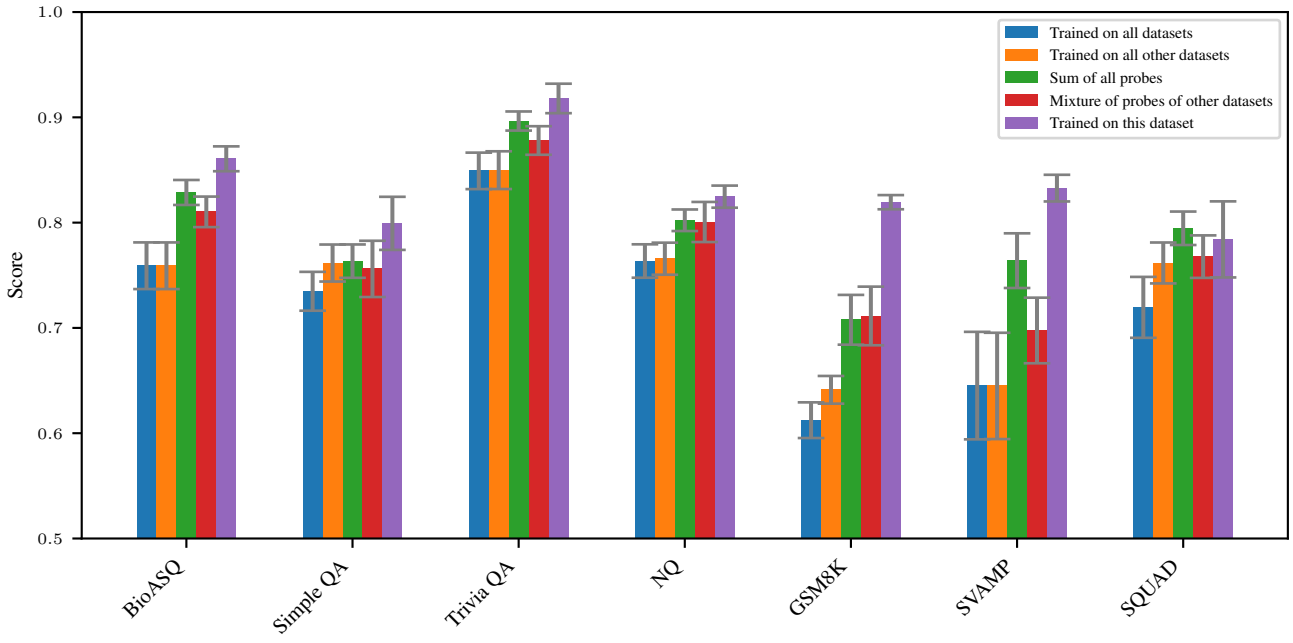


Figure 83. AUROC of linear probes at the token before the stop token of the output in the multi-task setting, using L2 regularization.

C. Phi-4 Mini Instruct

C.1. Layer 32

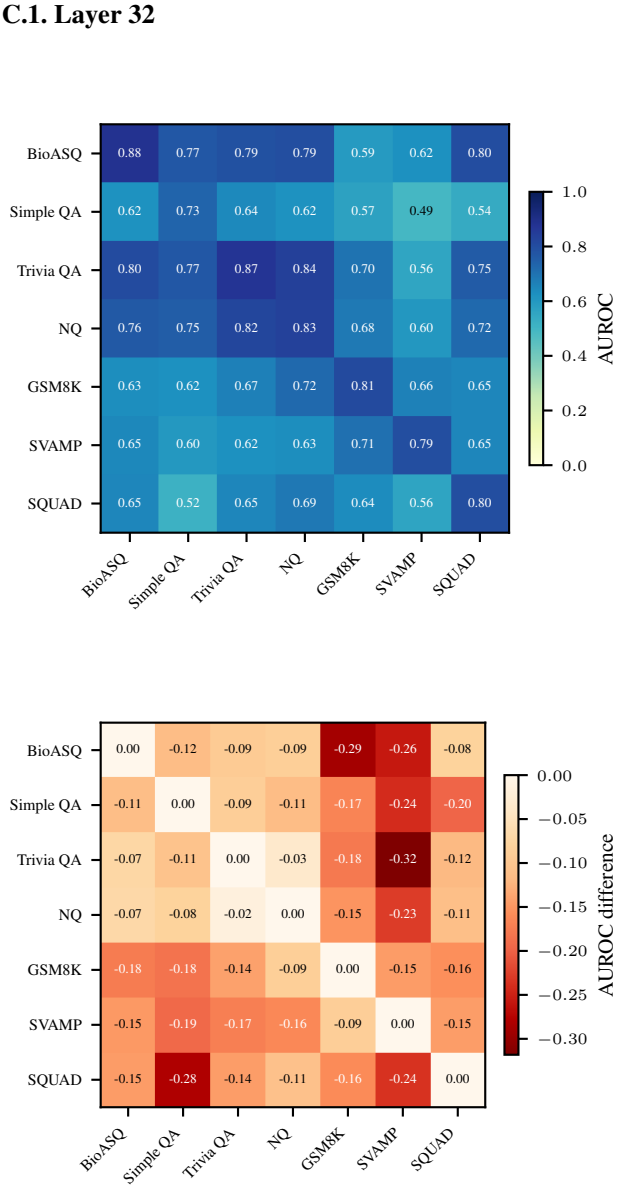


Figure 84. AUROC of probes trained on different tasks on the stop token of the output on last layer. Rows correspond to evaluation tasks while columns correspond to training tasks. The second plot represents the difference between the probe trained on this task and probes trained on the other datasets. Results are averaged over 5 runs.

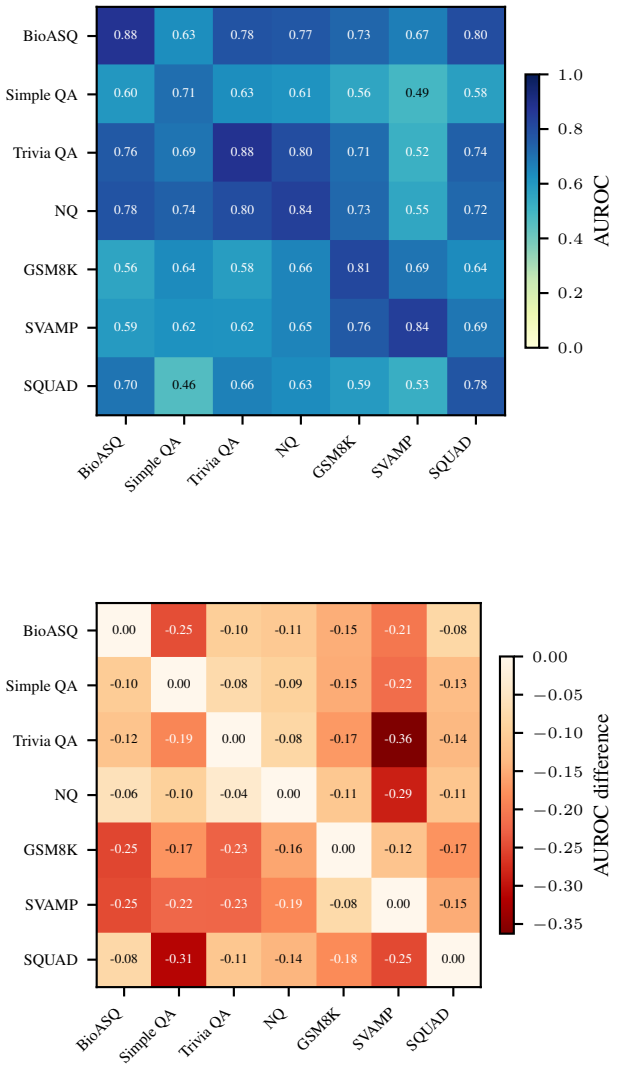


Figure 85. AUROC of probes trained on different tasks on the token before the stop token of the output on last layer. Rows correspond to evaluation tasks while columns correspond to training tasks. The second plot represents the difference between the probe trained on this task and probes trained on the other datasets. Results are averaged over 5 runs.

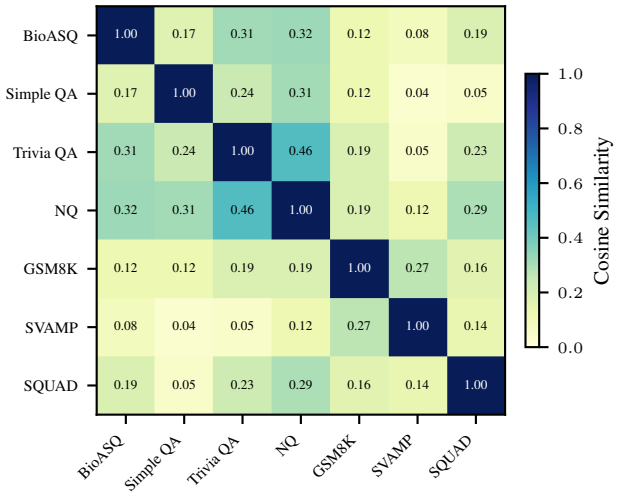


Figure 86. Cosine similarity between probes trained on different datasets using L2 regularization. Results are averaged over 5 runs.

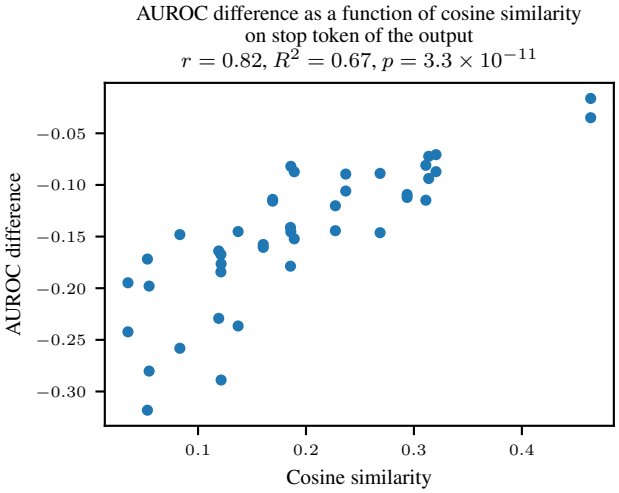


Figure 88. AUROC difference to probe trained on the right dataset as a function of cosine similarity between probes. Results are averaged over 5 runs.

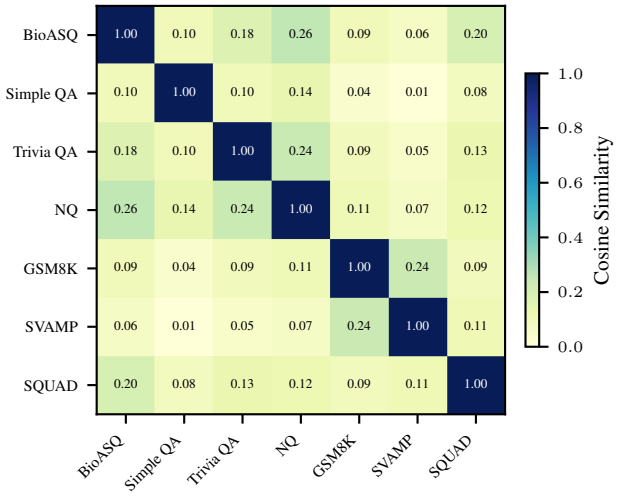


Figure 87. Cosine similarity between probes trained on different datasets using L2 regularization. Results are averaged over 5 runs.

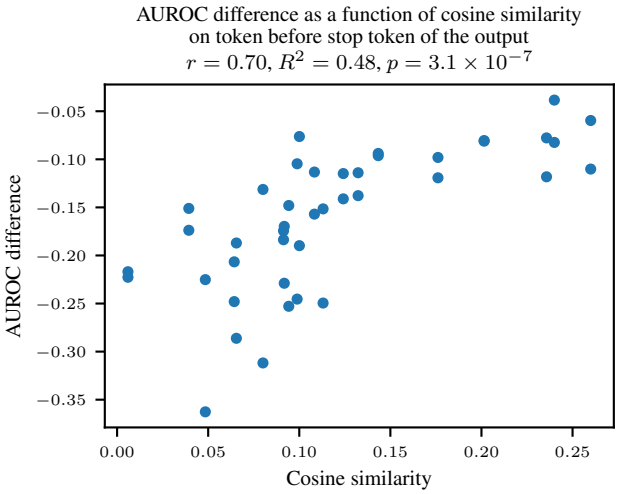


Figure 89. AUROC difference to probe trained on the right dataset as a function of cosine similarity between probes. Results are averaged over 5 runs.

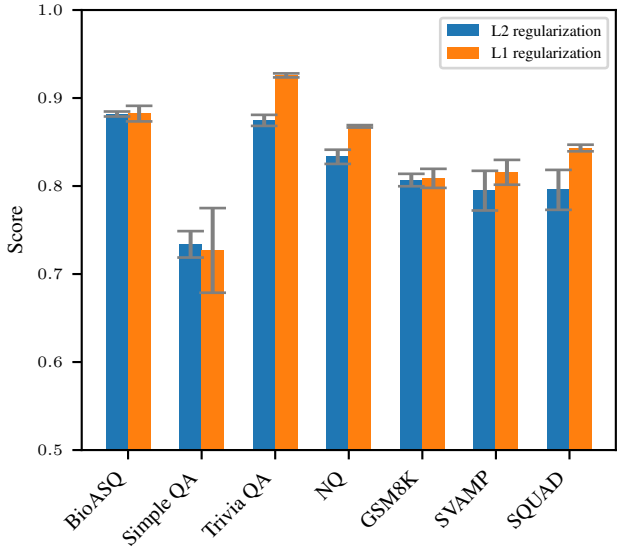


Figure 90. AUROC of linear probes at the stop token of the output, trained with either L1 or L2 regularization.

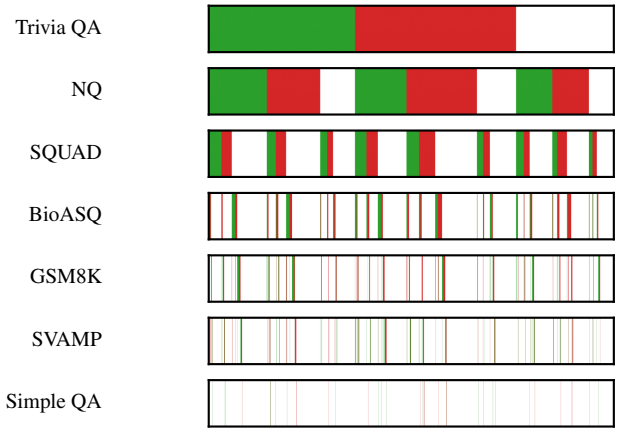


Figure 92. Signed support of sparse probes trained on different datasets at the stop token of the output, using L1 regularization at layer 32. Each row represents one probe trained on the corresponding dataset (y-axis labels). The x-axis shows the 3584 dimensions of the hidden state vector. Green indicates positive coefficients, red indicates negative coefficients, and white indicates zero coefficients (sparsity). Dimensions are sorted by sparsity level across all datasets, with the least sparse dimensions on the left and the most sparse on the right.

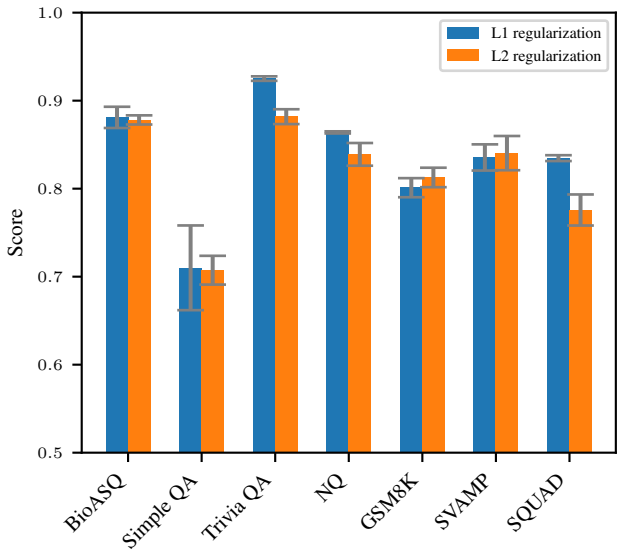


Figure 91. AUROC of linear probes at the token before the stop token of the output, trained with either L1 or L2 regularization.

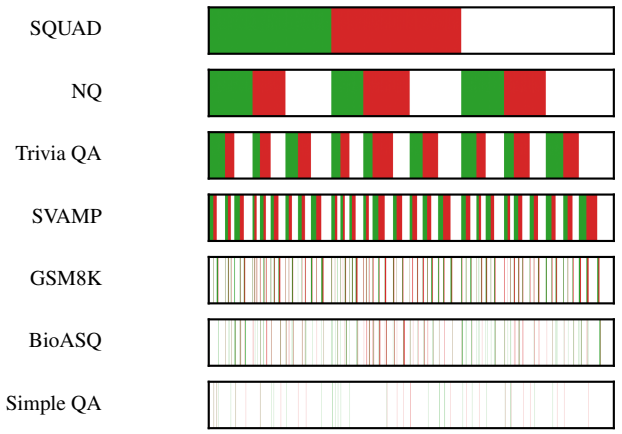


Figure 93. Signed support of sparse probes trained on different datasets at the token before the stop token of the output, using L1 regularization at layer 32. Each row represents one probe trained on the corresponding dataset (y-axis labels). The x-axis shows the 3584 dimensions of the hidden state vector. Green indicates positive coefficients, red indicates negative coefficients, and white indicates zero coefficients (sparsity). Dimensions are sorted by sparsity level across all datasets, with the least sparse dimensions on the left and the most sparse on the right.

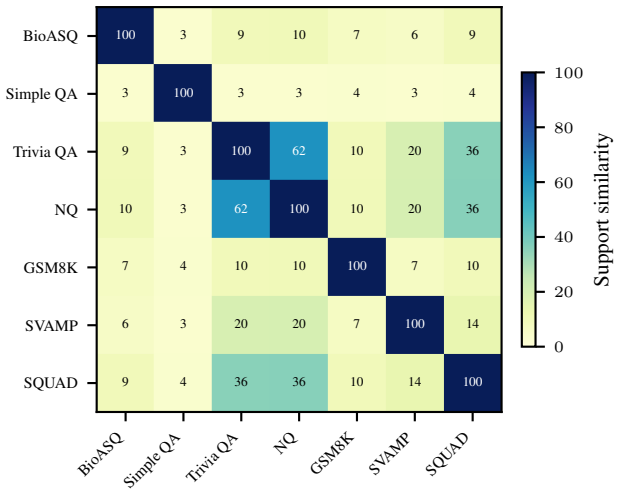


Figure 94. Support overlap between sparse probes trained on different datasets. Darker colors indicate higher overlap percentages. Task pairs with $> 30\%$ overlap (TriviaQA, NQ, SimpleQA) correspond to successful cross-task generalization, while most pairs show $< 15\%$ overlap, explaining generalization failure.

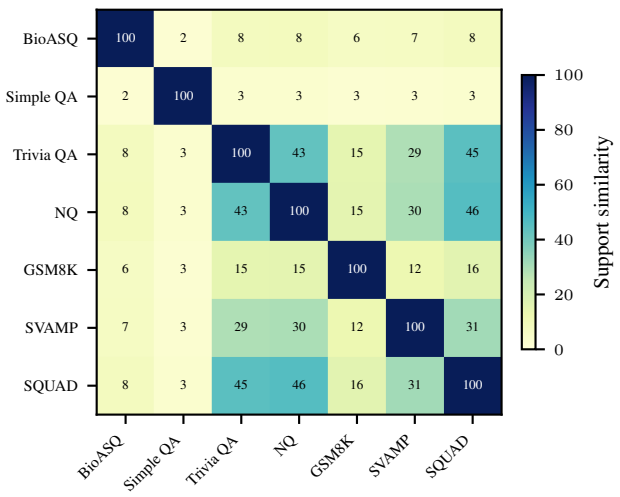


Figure 95. Support overlap between sparse probes trained on different datasets. Darker colors indicate higher overlap percentages. Task pairs with $> 30\%$ overlap (TriviaQA, NQ, SimpleQA) correspond to successful cross-task generalization, while most pairs show $< 15\%$ overlap, explaining generalization failure.

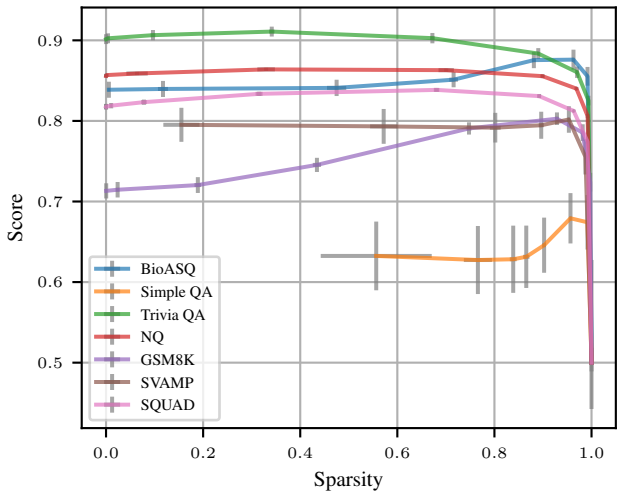


Figure 96. AUROC of probes trained using L1 regularisation as a function of the sparsity level on the stop token of the output.

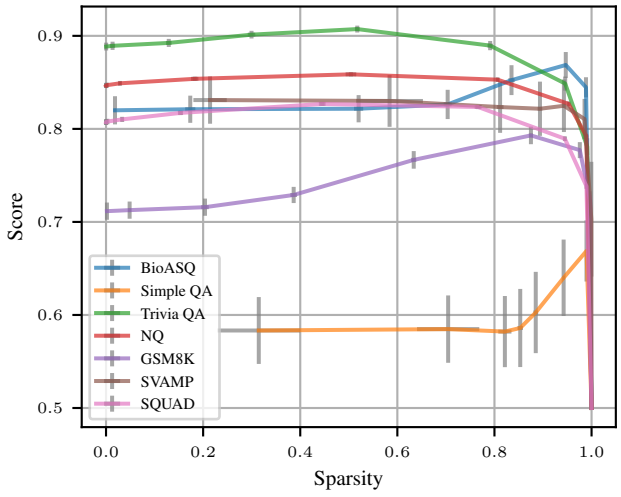


Figure 97. AUROC of probes trained using L1 regularisation as a function of the sparsity level on the token before the stop token of the output.

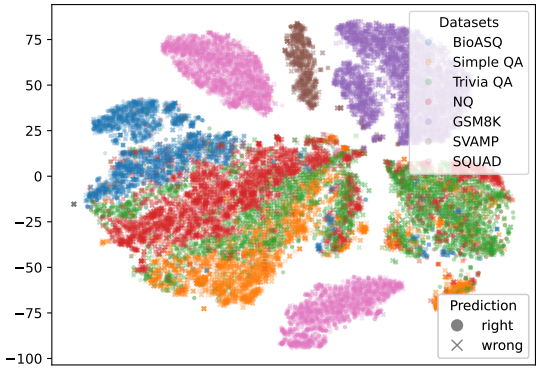


Figure 98. t-SNE plots of the hidden space at layer 32 at the stop token of the output.

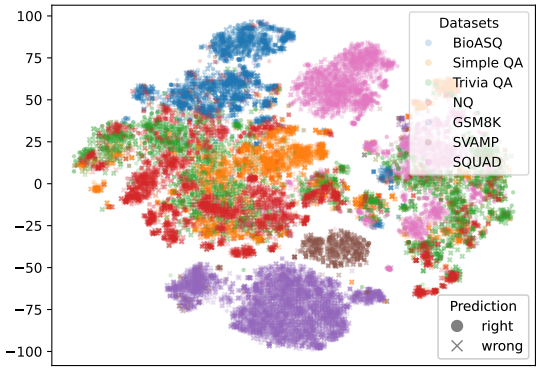


Figure 99. t-SNE plots of the hidden space at layer 32 at the token before stop token of the output.

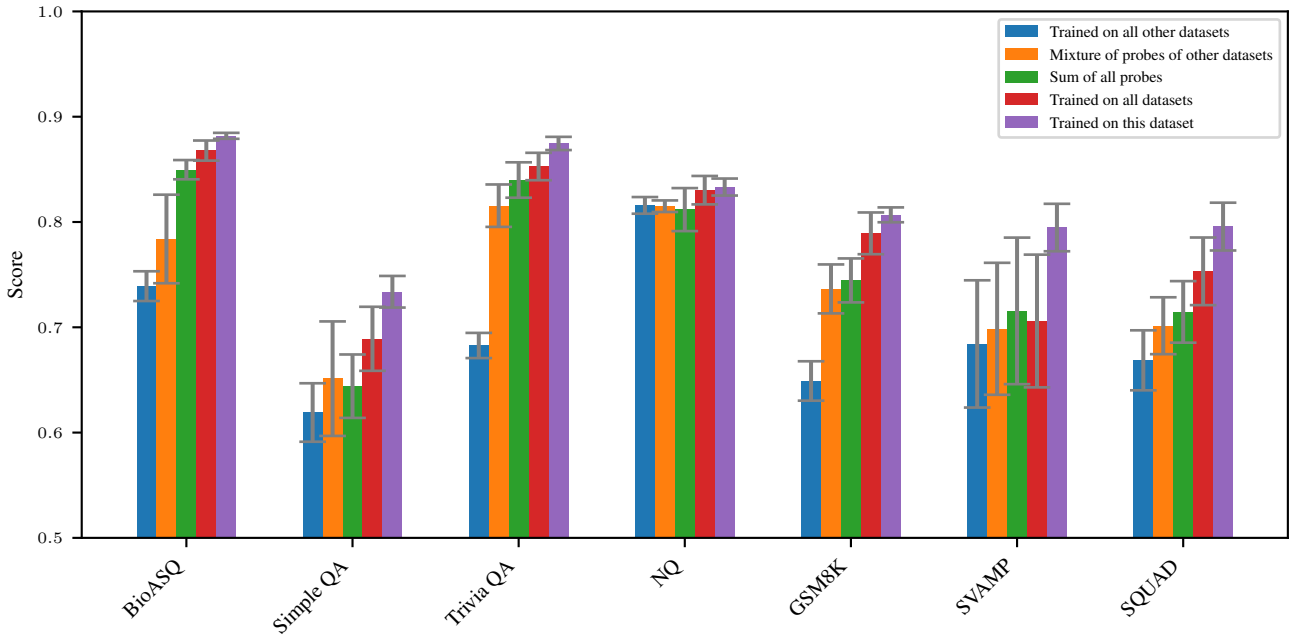


Figure 100. AUROC of linear probes at the stop token of the output in the multi-task setting, using L2 regularization.

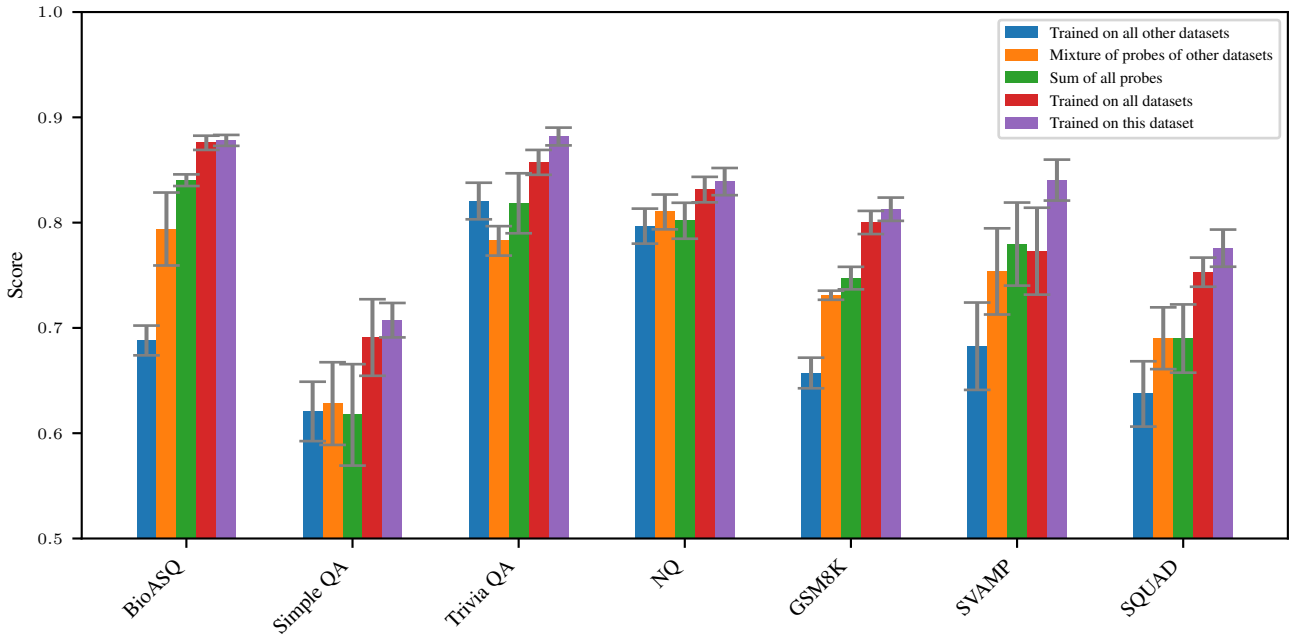


Figure 101. AUROC of linear probes at the token before the stop token of the output in the multi-task setting, using L2 regularization.

C.2. Layer 24

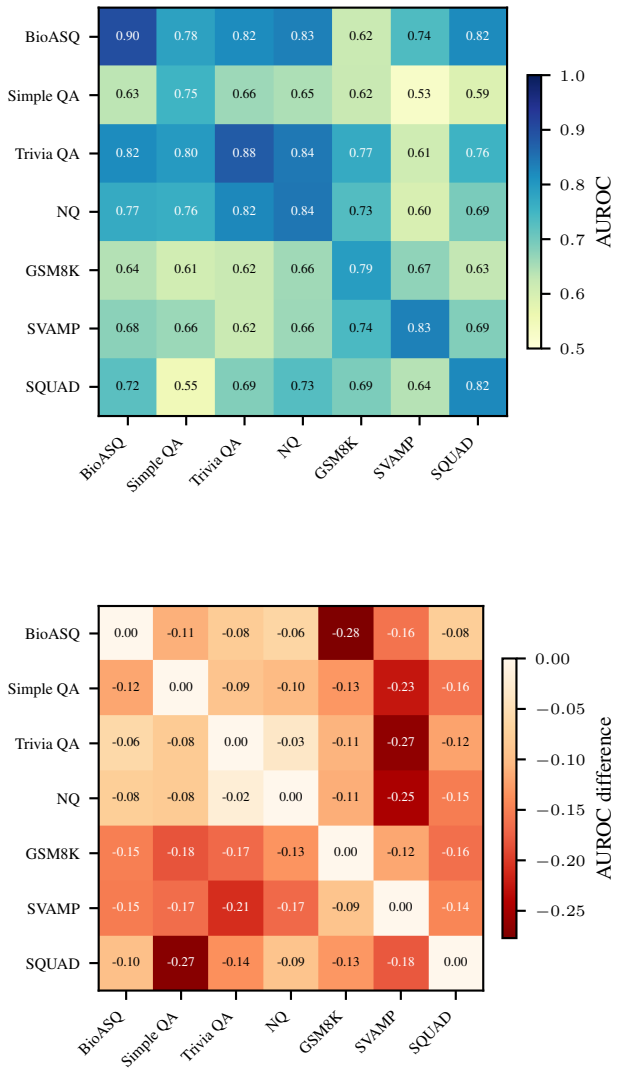


Figure 102. AUROC of probes trained on different tasks on the stop token of the output on last layer. Rows correspond to evaluation tasks while columns correspond to training tasks. The second plot represents the difference between the probe trained on this task and probes trained on the other datasets. Results are averaged over 5 runs.

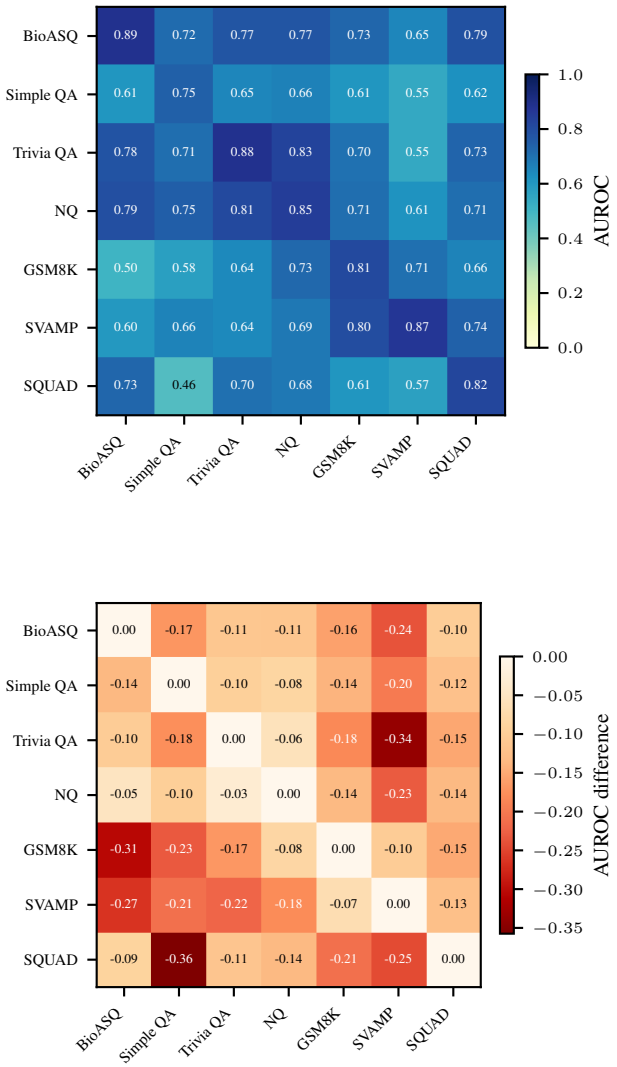


Figure 103. AUROC of probes trained on different tasks on the token before the stop token of the output on last layer. Rows correspond to evaluation tasks while columns correspond to training tasks. The second plot represents the difference between the probe trained on this task and probes trained on the other datasets. Results are averaged over 5 runs.

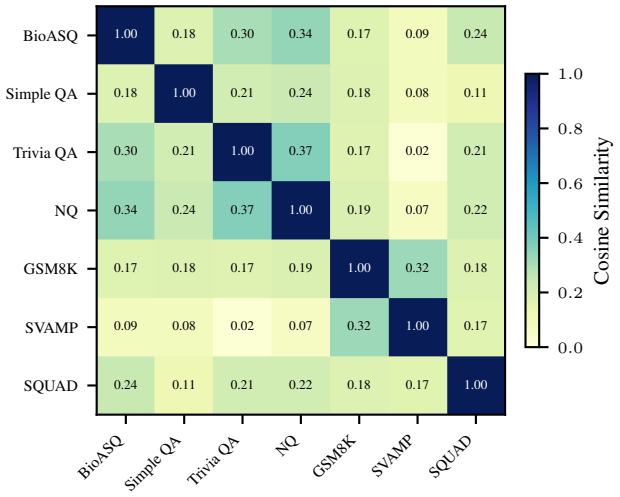


Figure 104. Cosine similarity between probes trained on different datasets using L2 regularization. Results are averaged over 5 runs.

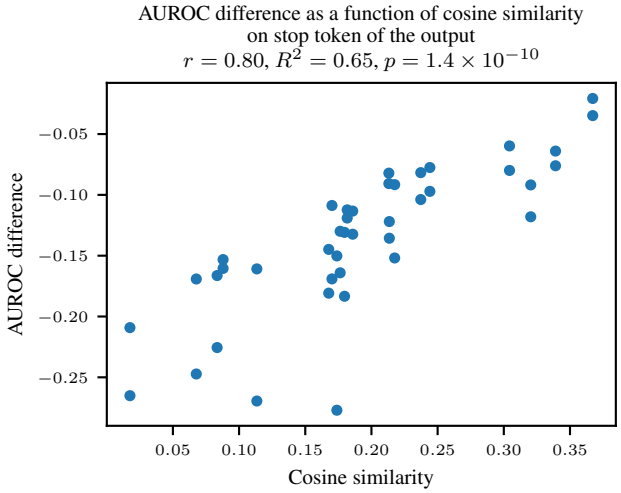


Figure 106. AUROC difference to probe trained on the right dataset as a function of cosine similarity between probes. Results are averaged over 5 runs.

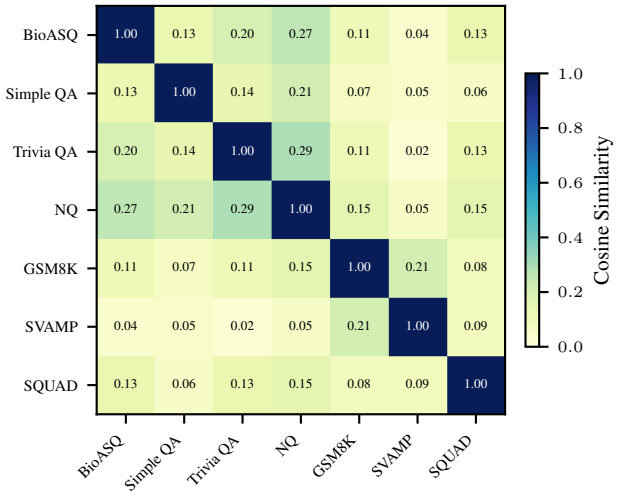


Figure 105. Cosine similarity between probes trained on different datasets using L2 regularization. Results are averaged over 5 runs.

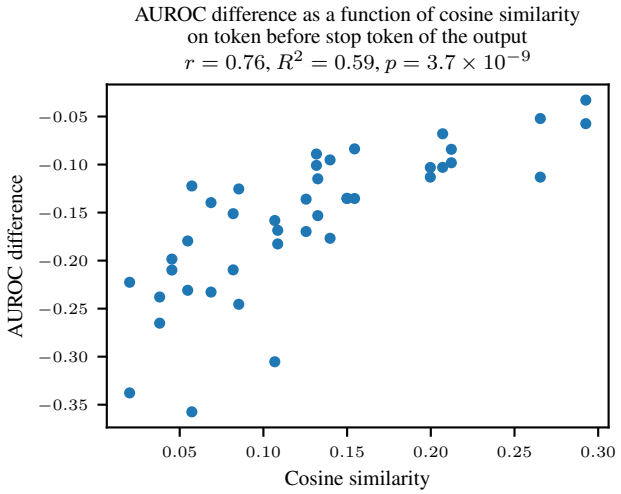


Figure 107. AUROC difference to probe trained on the right dataset as a function of cosine similarity between probes. Results are averaged over 5 runs.

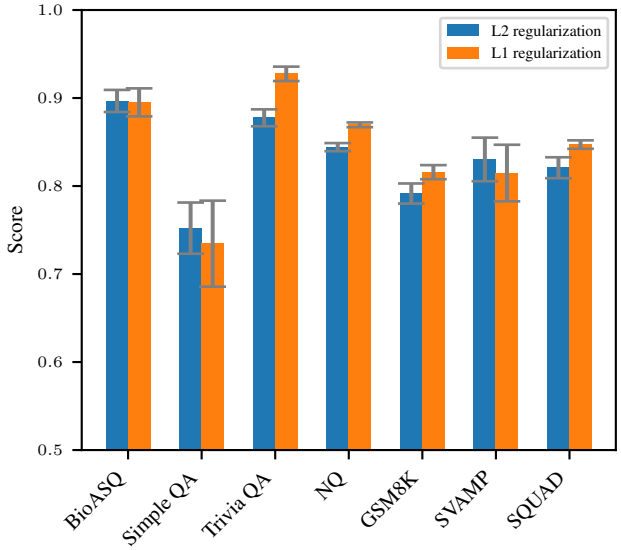


Figure 108. AUROC of linear probes at the stop token of the output, trained with either L1 or L2 regularization.

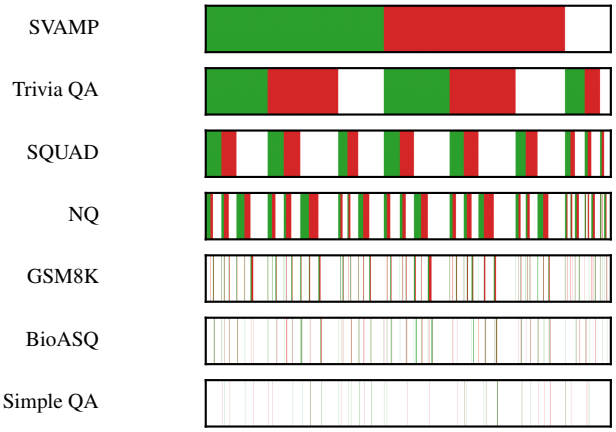


Figure 110. Signed support of sparse probes trained on different datasets at the stop token of the output, using L1 regularization at layer 24. Each row represents one probe trained on the corresponding dataset (y-axis labels). The x-axis shows the 3584 dimensions of the hidden state vector. Green indicates positive coefficients, red indicates negative coefficients, and white indicates zero coefficients (sparsity). Dimensions are sorted by sparsity level across all datasets, with the least sparse dimensions on the left and the most sparse on the right.

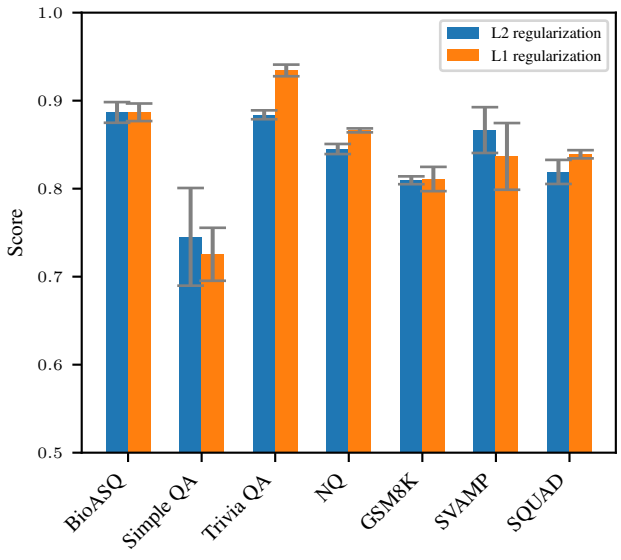


Figure 109. AUROC of linear probes at the token before the stop token of the output, trained with either L1 or L2 regularization.

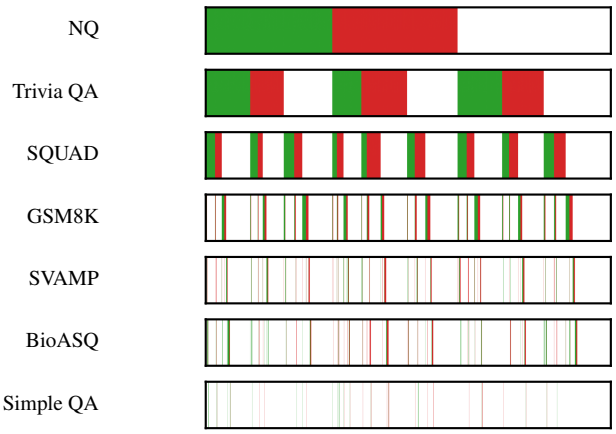


Figure 111. Signed support of sparse probes trained on different datasets at the token before the stop token of the output, using L1 regularization at layer 24. Each row represents one probe trained on the corresponding dataset (y-axis labels). The x-axis shows the 3584 dimensions of the hidden state vector. Green indicates positive coefficients, red indicates negative coefficients, and white indicates zero coefficients (sparsity). Dimensions are sorted by sparsity level across all datasets, with the least sparse dimensions on the left and the most sparse on the right.

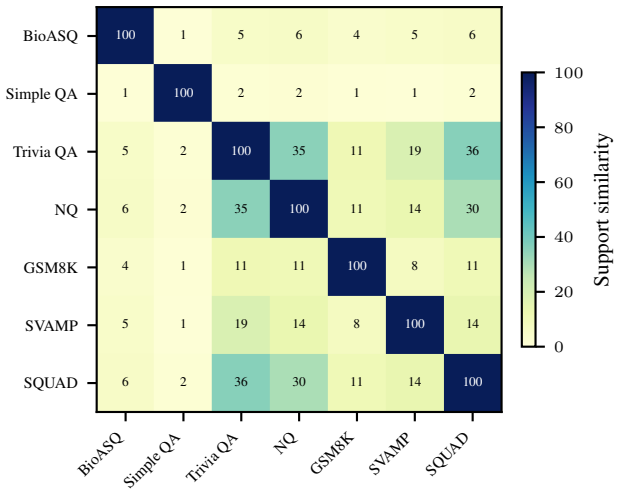


Figure 112. Support overlap between sparse probes trained on different datasets. Darker colors indicate higher overlap percentages. Task pairs with $> 30\%$ overlap (TriviaQA, NQ, SimpleQA) correspond to successful cross-task generalization, while most pairs show $< 15\%$ overlap, explaining generalization failure.

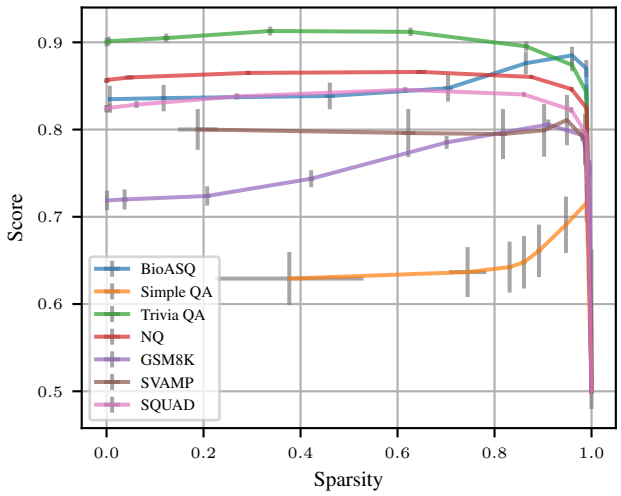


Figure 114. AUROC of probes trained using L1 regularisation as a function of the sparsity level on the stop token of the output.

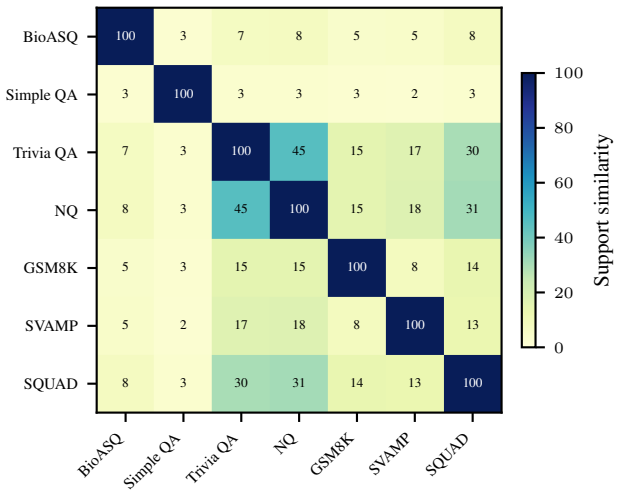


Figure 113. Support overlap between sparse probes trained on different datasets. Darker colors indicate higher overlap percentages. Task pairs with $> 30\%$ overlap (TriviaQA, NQ, SimpleQA) correspond to successful cross-task generalization, while most pairs show $< 15\%$ overlap, explaining generalization failure.

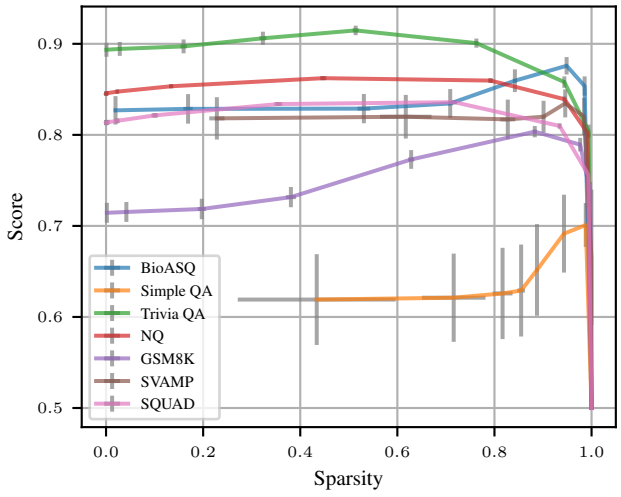


Figure 115. AUROC of probes trained using L1 regularisation as a function of the sparsity level on the token before the stop token of the output.

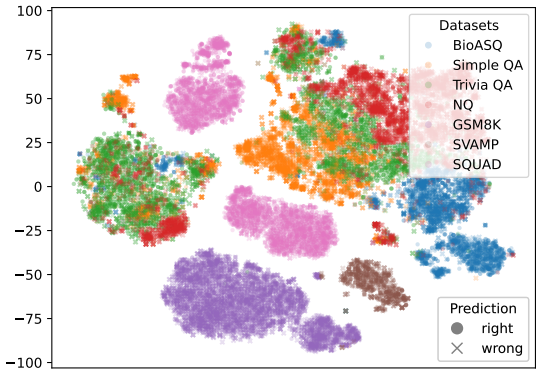


Figure 116. t-SNE plots of the hidden space at layer 24 at the stop token of the output.

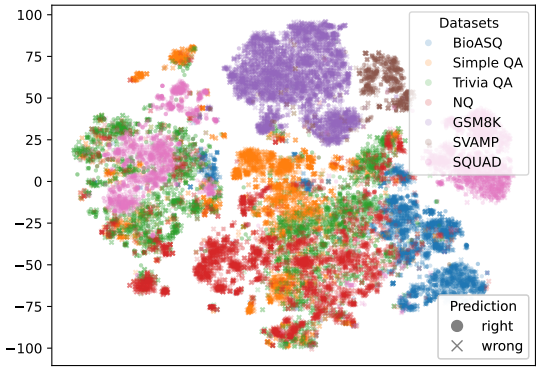


Figure 117. t-SNE plots of the hidden space at layer 24 at the token before stop token of the output.

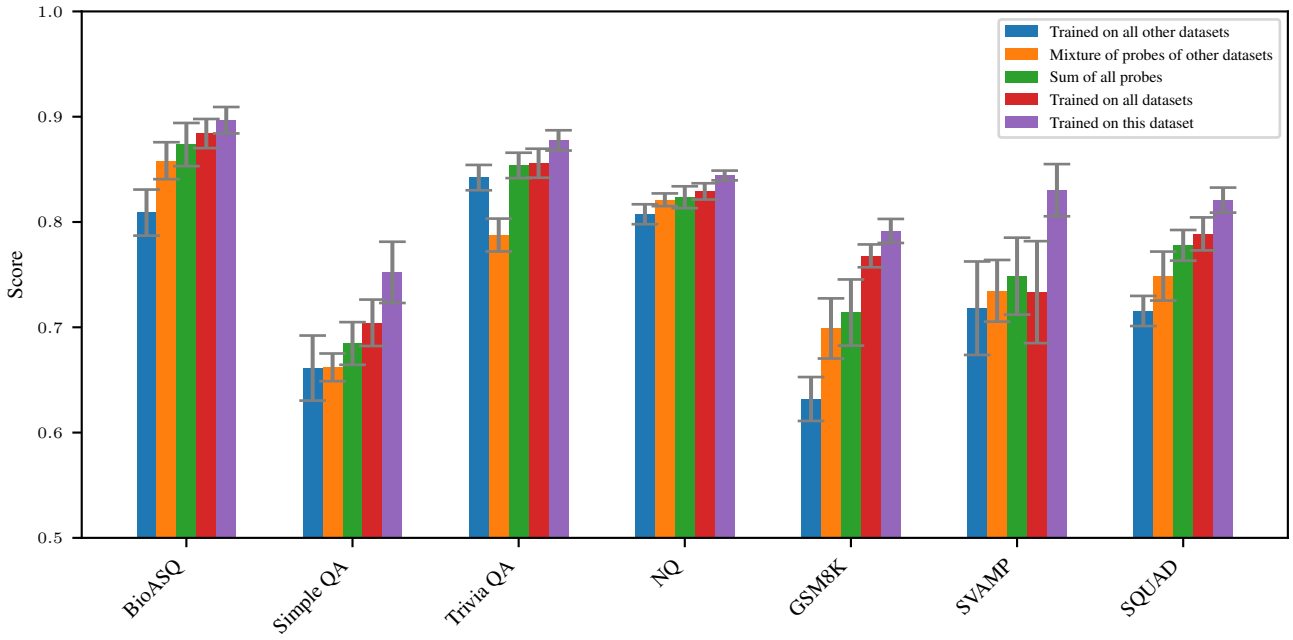


Figure 118. AUROC of linear probes at the stop token of the output in the multi-task setting, using L2 regularization.

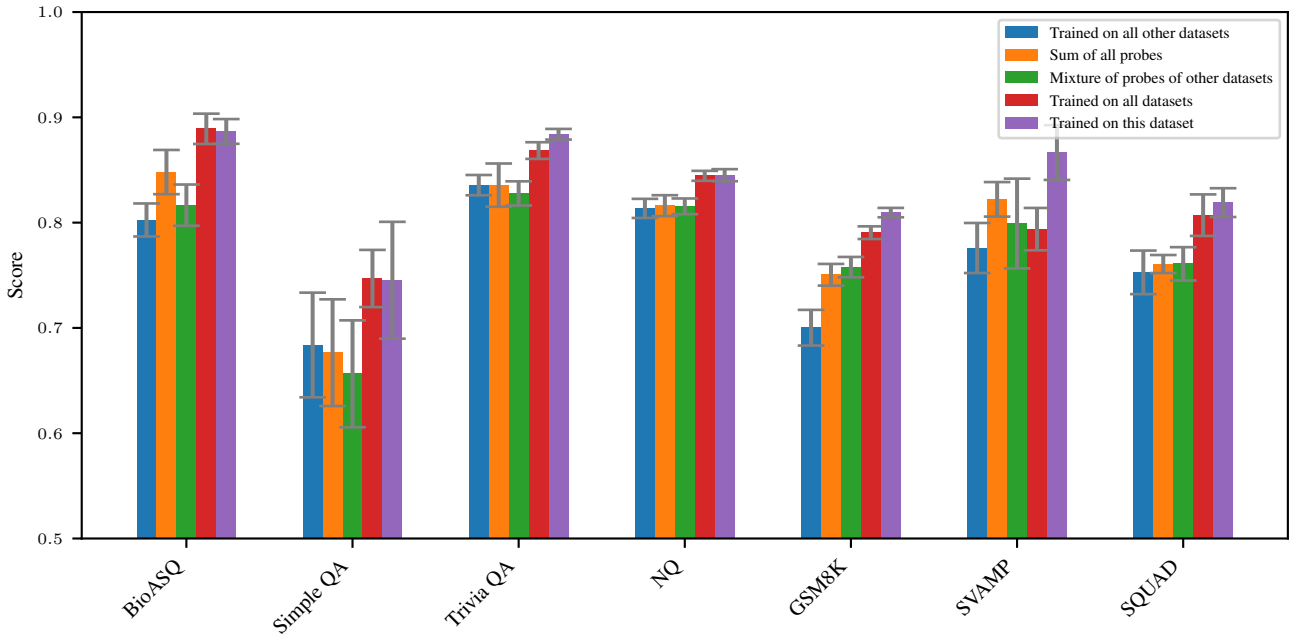


Figure 119. AUROC of linear probes at the token before the stop token of the output in the multi-task setting, using L2 regularization.