Finding multi-dimensional subspaces in LLM efficiently with kernels

Thomas Winninger
Télécom SudParis & ENS Paris-Saclay

Context

Subspaces in LLMs can be multi-dimensional and non-linear. However, if methods exist to extract such multi-dimensional subspaces (with optimisation or SAE), we lack efficient and robust principled methods. Given recent improvements in steering via kernel methods, I tried to apply these methods to search and analyse multi-dimensional subspaces, and it performed remarkably well.

In this exploratory paper, I propose:

- An adaptation of the RFM-AGOP framework to extract multi-dimensional subspaces.
- A probe based initialization to improve stability.
- An analysis of the effect of steering with this method on both reasoning and non-reasoning models over long reasoning traces.

Future work is needed to better understand the relations between subspaces found by different methods. If confirmed, RFM could be a cheap and scalable complement to existing subspace extraction methods in LLMs.

Method

RFM AGOP framework

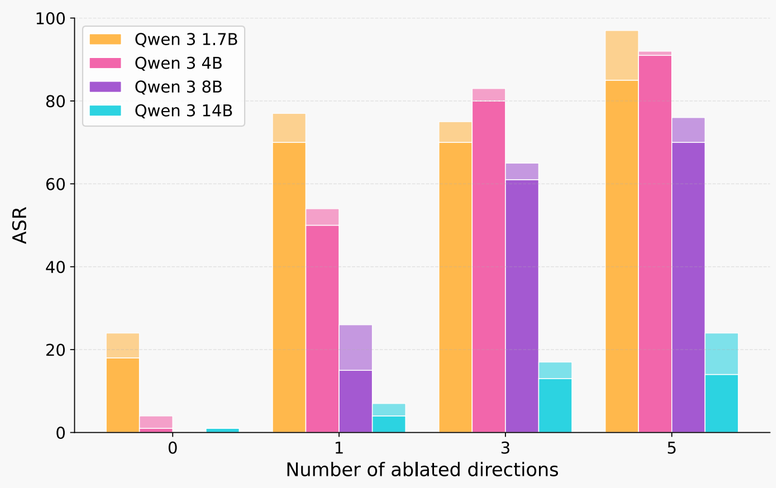
Given a dataset of contrastive pairs (harmful-harmless), we extract activations from the residual stream at each layer and last token position. These activations in $\mathbb{R}^d$ are used to solve a kernel ridge regression with the Mahalanobis Laplace Kernel:

$$K_M(u,v) = \exp\left(-\frac{1}{L}\sqrt{(u-v)^\top M (u-v)}\right)$$

With the coefficient α from the regression, this gives a classifier $f_{\alpha,M}$ between harmful and harmless activations. This classifier depends on M, which is updated via the Average Gradient Outer Product:

$$M_{t+1} = \frac{1}{n}\sum_{i=1}^{n} \nabla_x f_{\alpha_t,M_t}(x_i) \nabla_x f_{\alpha_t,M_t}(x_i)^\top$$

n being the number of pairs. The first eigenvectors of M are the directions of largest change, thus our refusal subspace.



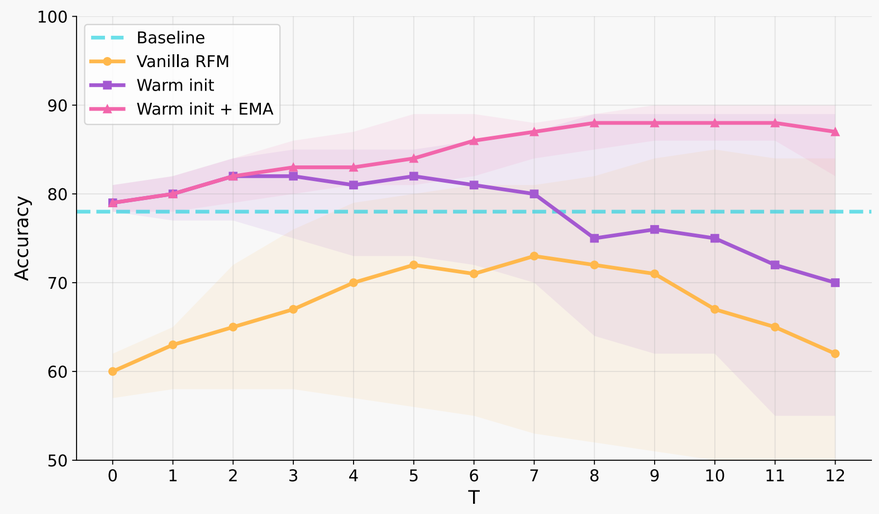
To test the retrieved directions, we ablated each direction and evaluated the Attack Success Rate (ASR) on harmful sentences. While small models are already sensible with one ablated direction, larger models need the ablation of at least three directions.

Probe-Informed initialization

To improve stability and warm up the algorithm, I tried multiple initializations besides the original $M_0 = I$. The best candidate was

$$M_0 = \beta w w^\top + (1-\beta)\Sigma_{X,k}$$

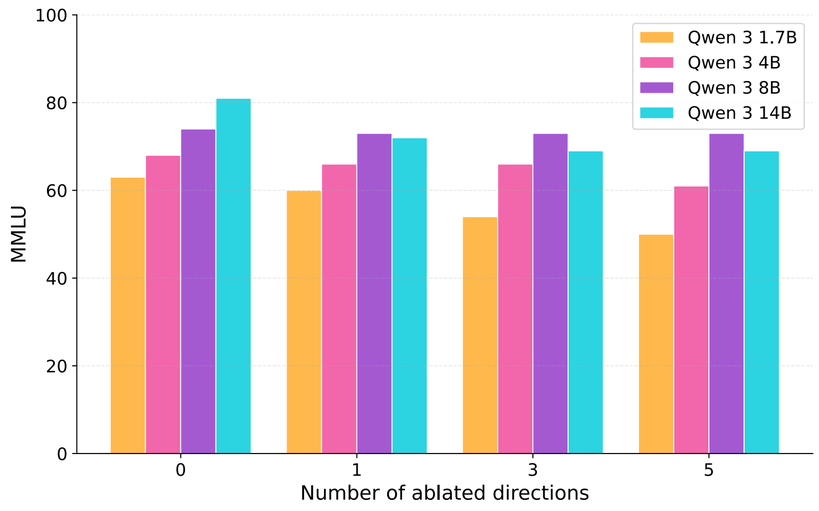
where w is the direction of a linear probe $p(x) = w^\top x + b$ and $\Sigma_{X,k}$ is the rank-k truncated empirical covariance matrix of the activations, with $k = 10$ to help AGOP explore multiple directions.



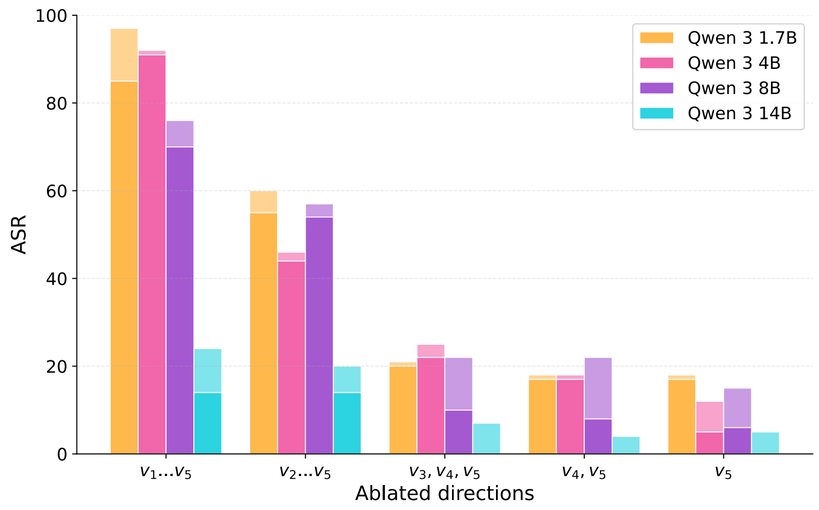
Not using the warm init leads to performance worse than the baseline – the one-dimensional linear probe. Additionaly, using the Exponential Moving Average (EMA) on M seems to improve stability.

Ablations

- Removing the directions found by RFM do not impact the model's performance on MMLU:



- Directions are decreasingly important:



Limitations

The most straightforward limitation is the lack of evaluation on larger models. But the most problematic is measuring refusal, which is hard and subjective, especially on long reasoning traces.

Furthermore, the method relies on the behaviour to be linearly represented in activations, which is not always the case.