# Temporal Task Diversity & Transformer Training

## Setting:

Transformer **in-context linear regression**, finite and ***changing*** **distribution** of regression tasks.

To generate a sequence:

1. Sample **task** $t$ from time-varying distribution.
2. Sample iid **regressors** $x_1, x_2, ..., x_K$ from $\mathcal{N}(0, I)$.
3. Sample iid **noise** $\varepsilon_1, \varepsilon_2, ..., \varepsilon_K$ from $\mathcal{N}(0, \sigma^2)$.
4. Compute **labels** $y_i = t \cdot x_i + \varepsilon_i$ for all $i$.
5. Sequence is $x_1, y_1, x_2, y_2, ..., x_K, y_K$.

## Prior result:

For **stationary** task distributions, *two regimes:*

- **Low number of tasks:** Models learn prediction strategy specialised to the fixed, finite task set *(memorisation)*.
- **High number of tasks:** Models learn a general prediction strategy that works for training tasks and new tasks *(generalisation)*.

*[Raventós et al., arXiv:2306.15063]*

## Our result:

For **non-stationary** task distributions (e.g., MALA with variable step size):

- **Low number of tasks + slow change:** Models continually specialise to the *moving* task set *(continual memorisation)*.
- **Low number of tasks + fast change:** Models learn a general prediction strategy *(stable generalisation)*.


Afiq Abdillah **Effiezal Aswadi**
Independent, *equal contribution*
Oliver **Britton**
University of Oxford, *equal contribution*
Ross **Baker**
University of Oxford, *equal contribution*
Matthew **Farrugia-Roberts**
University of Oxford, matthew@far.in.net


# Understanding continual learning: ***Continually changing data*** creates an ***inductive bias*** in favour of models that ***generalise over time***.