Research Diary

PhD Research Journal

Author: **Annabel Jakob**

Start Date: 10 February 2026

Field: **Computer Science**

Contents

1 Resources

- DPhil Progression Information and Resources

2 04 February 2026

2.1 Research Plan

Today's main tasks:

- Meeting with Seth and Gunes
- Familiarise with DPhil milestones and progression requirements
- Reading relevant literature suggested during supervisor meeting

2.2 Content Details

Meeting Summary

Meeting Notes:

- DPhil Milestones
  - Term 4, Week 0: Transfer status
  - 8th/9th Term: Confirmation of DPhil status
- Finding a research question
  - Bayesian Transformers
    - The Bayesian Geometry of Transformer Attention
    - Transformers Can Do Bayesian Inference
  - AI-assisted proofs
    - Existing tools: LEAN Proof Checker and Xena Project, Autodiff
    - Possible steps:
      1. Compile dataset of theorems presented in previous conference papers (e.g. NeurIPS)
      2. Verify the theorems and proofs in the dataset
      3. Goal: Can we produce *new* theorems and proofs?
    - Other resources:
      - `https://en.wikipedia.org/wiki/Kevin_Buzzard`
      - `https://terrytao.wordpress.com/2025/12/08/the-story-of-erdos-problem-126/`
      - Aristotle: IMO-level Automated Theorem Proving
- Mechanistic interpretability

  - Potentially connected to encrypted backdoors
  - Talk to Marek and Junayed
- Statistical Machine Learning
  - DeepRV: Accelerating spatiotemporal inference with pre-trained neural priors
- Random Number Generators
  - Can a backdoor be hidden in a RNG/ induced through an RNG? I.e. malicious signal induced via carefully chosen "random" numbers?
  - Planting Undetectable Backdoors in Machine Learning Models
  - Oblivious Defense in ML Models: Backdoor Removal without Detection

3 09 February 2026

Today's main tasks:

- Read Bayesian Transformers papers:
  - The Bayesian Geometry of Transformer Attention
  - Transformers Can Do Bayesian Inference

Paper Reading

Paper title: Transformers Can Do Bayesian Inference [1]
Authors: Mueller et al.
Summary:
Bayesian methods are usually slow or mathematically intractable for large datasets. Deep learning is fast but often bad at uncertainty and priors. As a solution, the authors present Prior-Data Fitted Networks (PFNs), which learn the mapping of the data to Bayesian posterior prediction.

Notes:

- PFNs: train transformers to approximate Bayesian posterior predictive distirbutions by sampling synthetic distirbutions, masking labels, and learning to predict them. At inference time, a single forward pass approximates Bayesian inference.
- PFNs *approximate* these posteriors
- PFNs work for any prior distribution that can be sampled from

References

[1] Samuel Müller et al. "TRANSFORMERS CAN DO BAYESIAN INFERENCE". In: *ICLR 2022 - 10th International Conference on Learning Representations*. 2022.