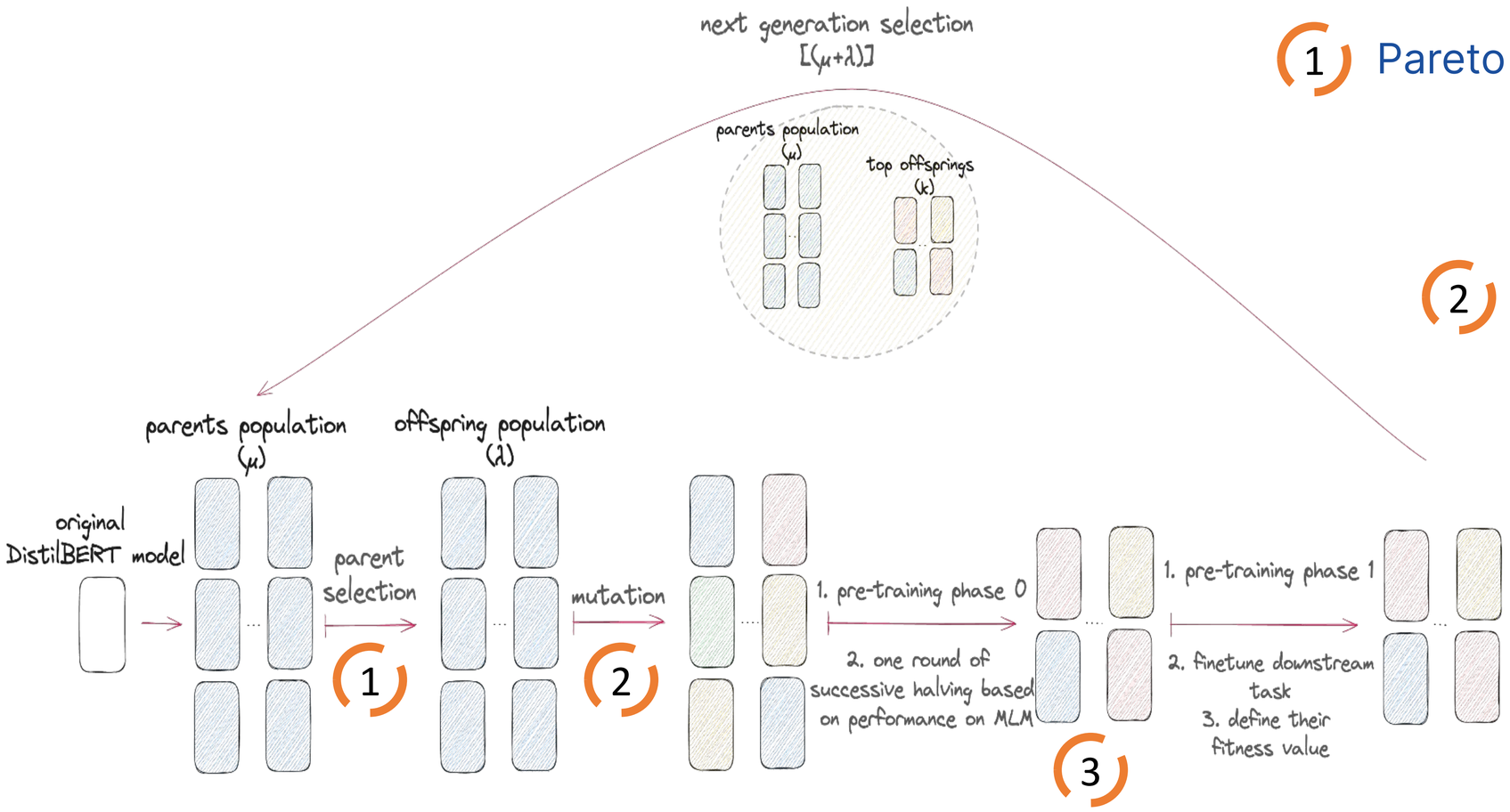
Resource-constrained Neural Architecture Search on Language Models: A case study

Universiteit Leiden
The Netherlands

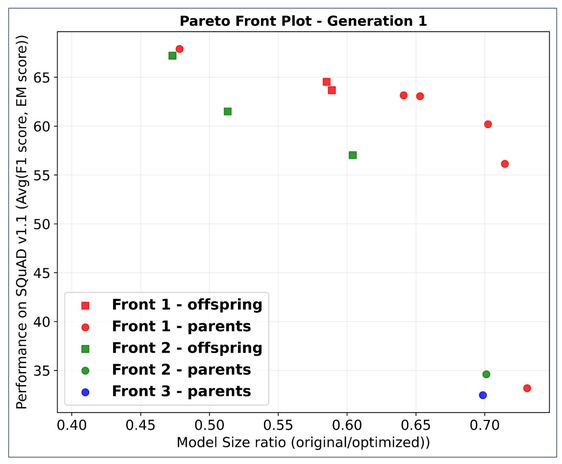
INTRODUCTION

- Complex transformer architecture, limited intuition towards optimization
- **Resource-constrained** Neural Architecture Search (NAS) on the transformer-encoder macro-architecture (case study)

METHOD (PIPELINE)



1 Pareto Fronts

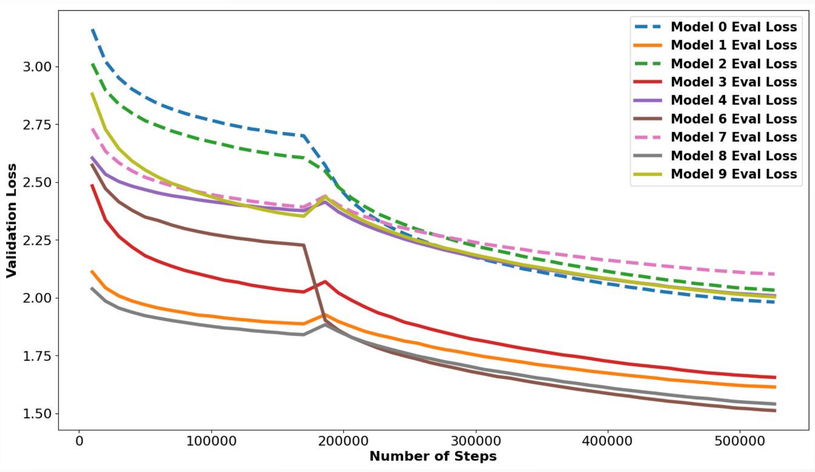


2
- two-level hierarchical search space
  - transformer encoder block (add/remove)
  - Feed-forward Neural Network (FNN) block (add/alter)
  - Multi-head self-attention (alter)

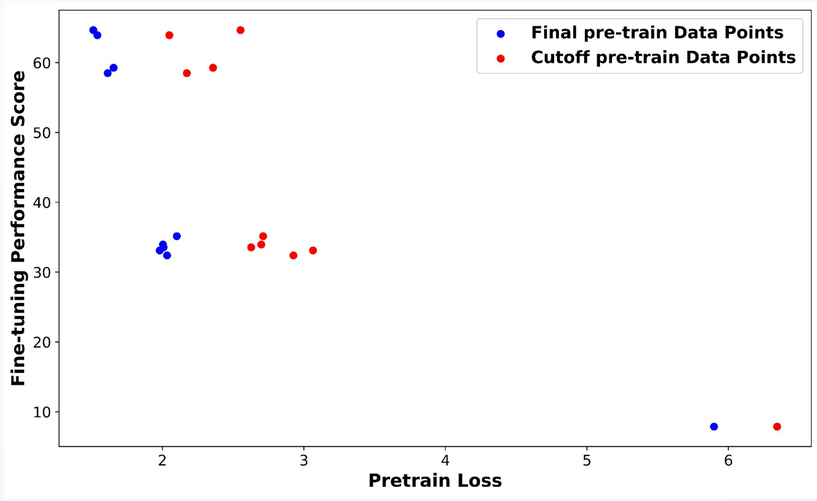
3 two-phase training with model discarding

RESULTS

Multi-fidelity optimization:
minimal crossing



Cutoff criterion:
Strong correlation



DISCUSSION

- Strong Correlation between **pre-training** and **downstream task** performance
- Applicability of Learning Curve analysis and model selection mechanism

Challenges & Future work

- Search space definition
- LLM pre-training cost
  - weight inheritance
  - parameter/data/memory efficient pre-training

Andreas Paraskeva (Leiden University)
João Pedro Reis (Universidade do Porto)
Suzan Verberne (Leiden University)
Jan N. van Rijn (Leiden University)

contact us

full paper