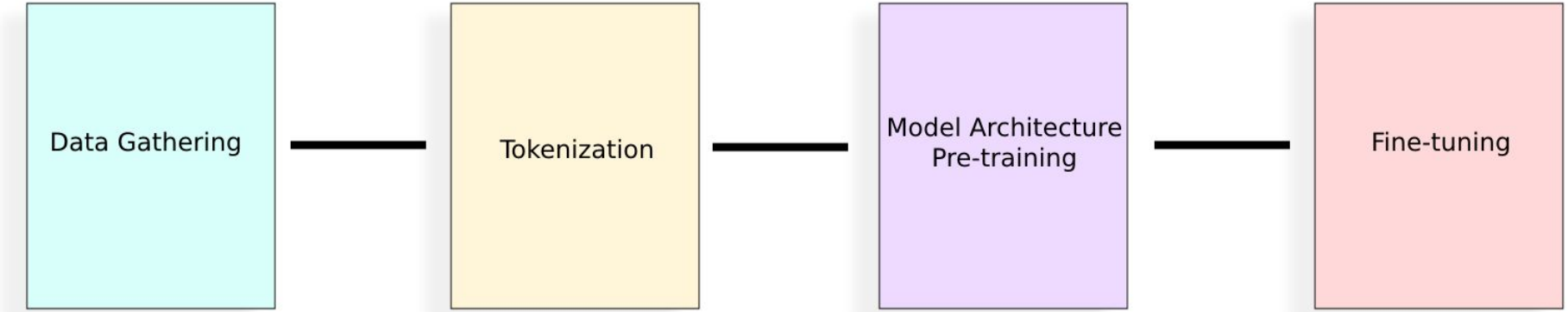
Further Pre-Training on Heart Transplant Research

HUGO TORBIÖRNSSON & GUSTAV ROSENGREN

LUND UNIVERSITY

Training a Large Language Model



LUND UNIVERSITY
LTH
FACULTY OF ENGINEERING

Further Pre-training

Strengthen general language understainding within a domain

Often unmarked

Predict next token

Fine-tuning

Specialize the model for a specific purpose

Labeled data / instructions

Q&A, Classify, Translate ...

LUND UNIVERSITY
LTH
FACULTY OF ENGINEERING

Dataset

- 25 PDFs -- all scientific heart transplant articles

- Parsing and cleanup of irrelevant information → ~25k rows for tokenization
  - Small amount of data in this context

```
IABPs are traditionally inserted via the femoral artery, which limits the patient in terms of Pt, as
ambulation is often institutionally prohibited with the device in the femoral artery position. Not only is
insertion site bleeding a possibility, but the Iabp may become kinked with hip exion leading to
```

LUND UNIVERSITY
LTH FACULTY OF ENGINEERING

Tools

- Lightweight and Optimized Framework

LUNARC

Lund University

- GPU Cluster

- Setup initial pipeline

Method

- Llama-3.1-8B
- LoRA, Low Rank-Adaptation
- Quantization, reduces precision of weights

LUND UNIVERSITY
LTH
FACULTY OF ENGINEERING

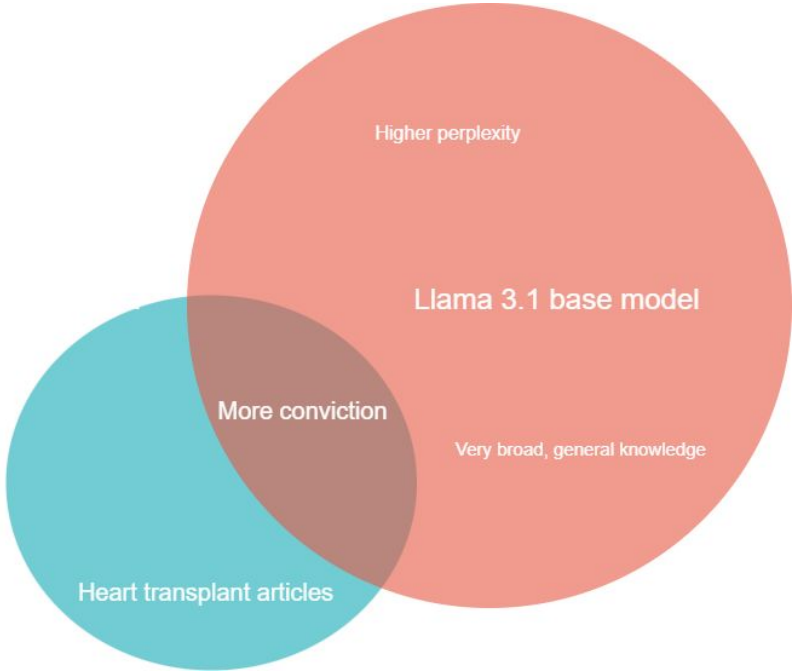
Example

```
prompt = "Heart transplant recipients are prone to"
```

```
Heart transplant recipients are prone to rejection of the new organ.
The body’s immune system recognizes the transplanted organ as foreign and attacks it.
The rejection is typically mild and can be treated with immunosuppressant drugs
But if the attack is severe, it can be fatal.
```

Perplexity

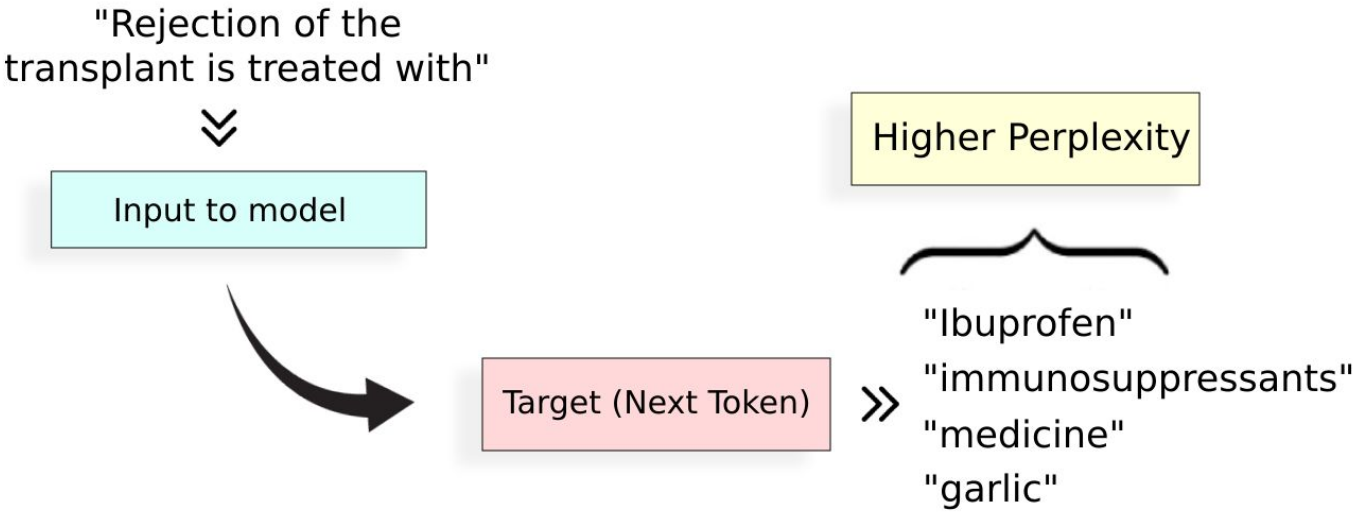
$$PP(W) = \frac{1}{P(w_1, w_2, \ldots, w_N)^{\frac{1}{N}}}$$



Confident about next token -> low perplexity

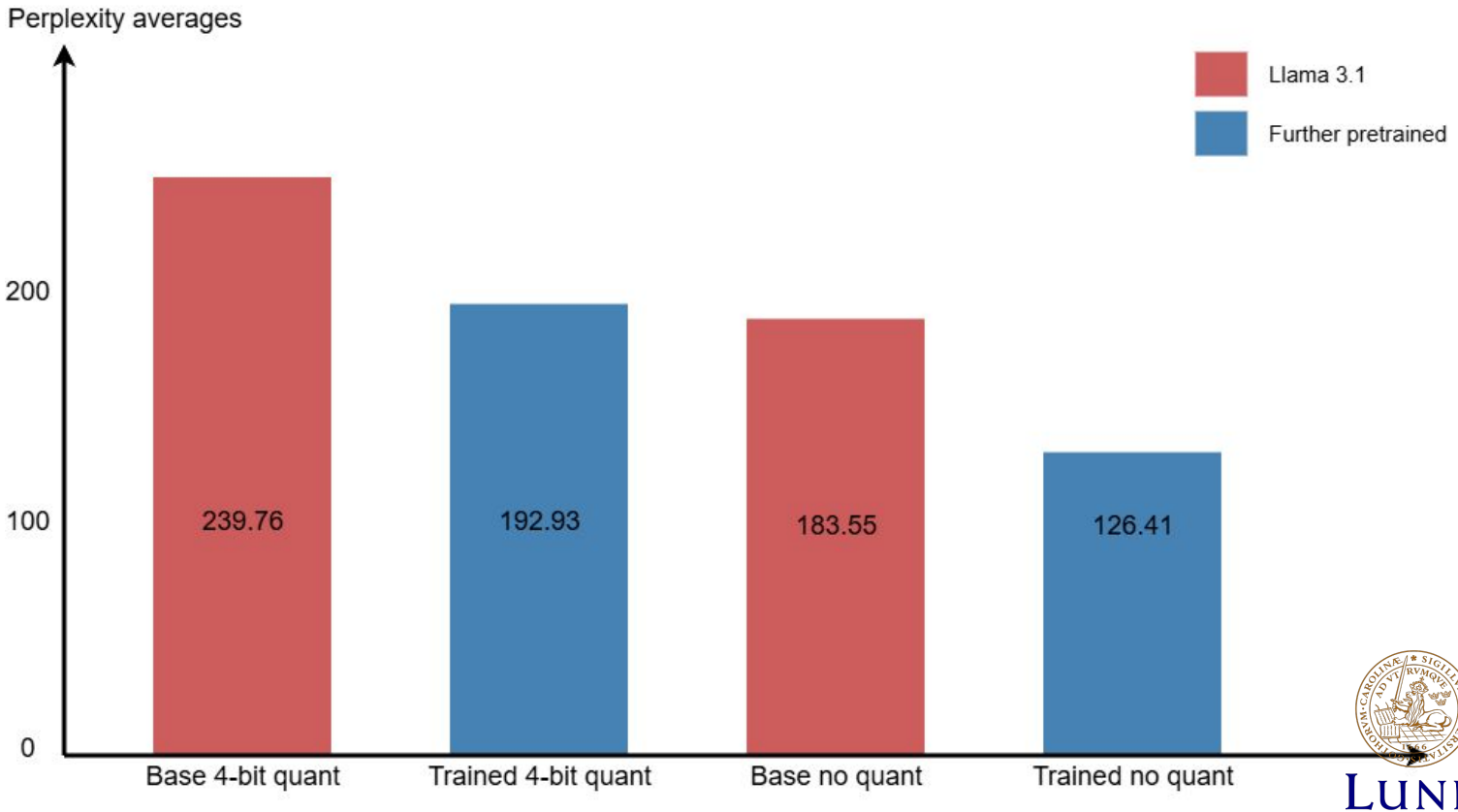
Uncertain about next token -> high perplexity

Perplexity

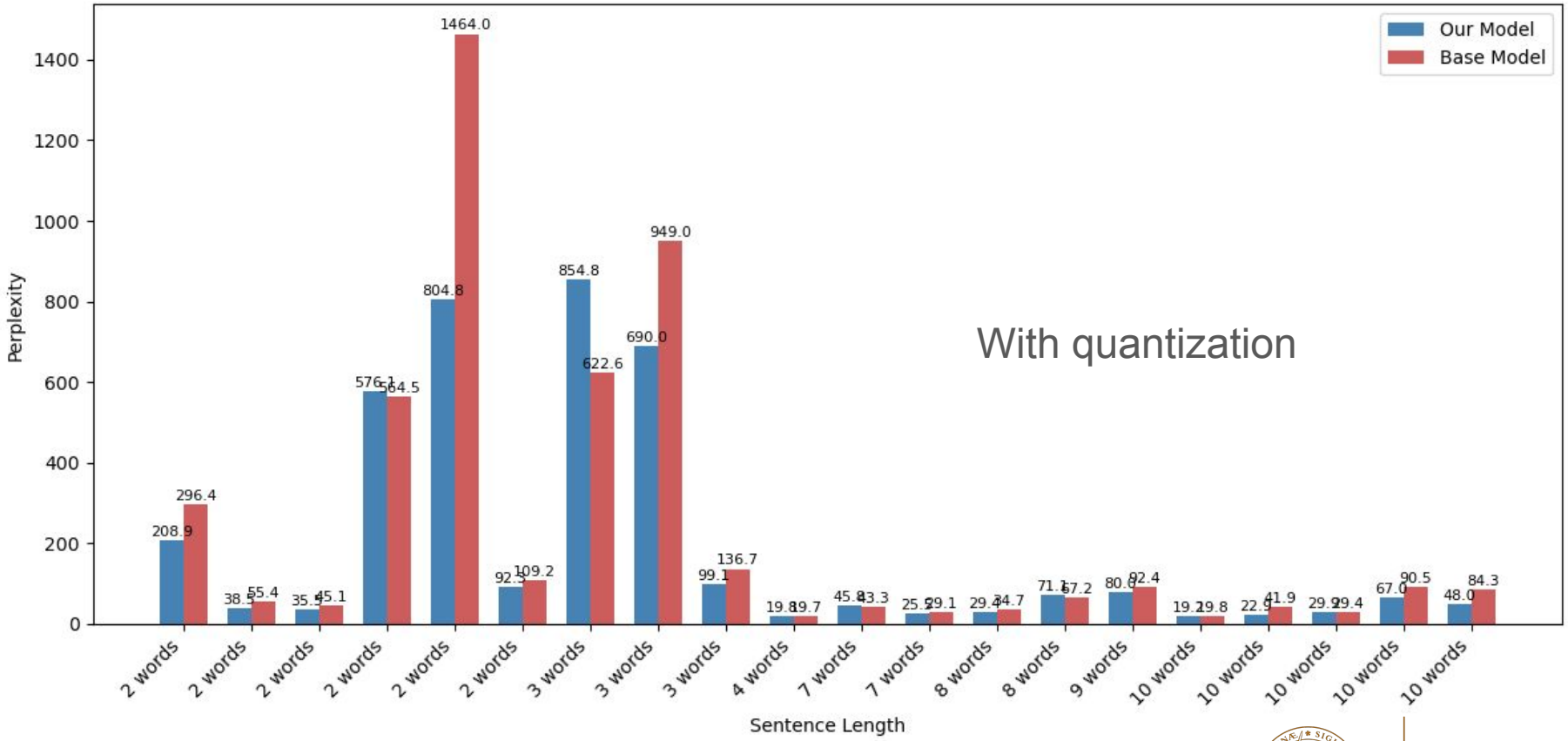


LUND UNIVERSITY
LTH
FACULTY OF ENGINEERING

Results

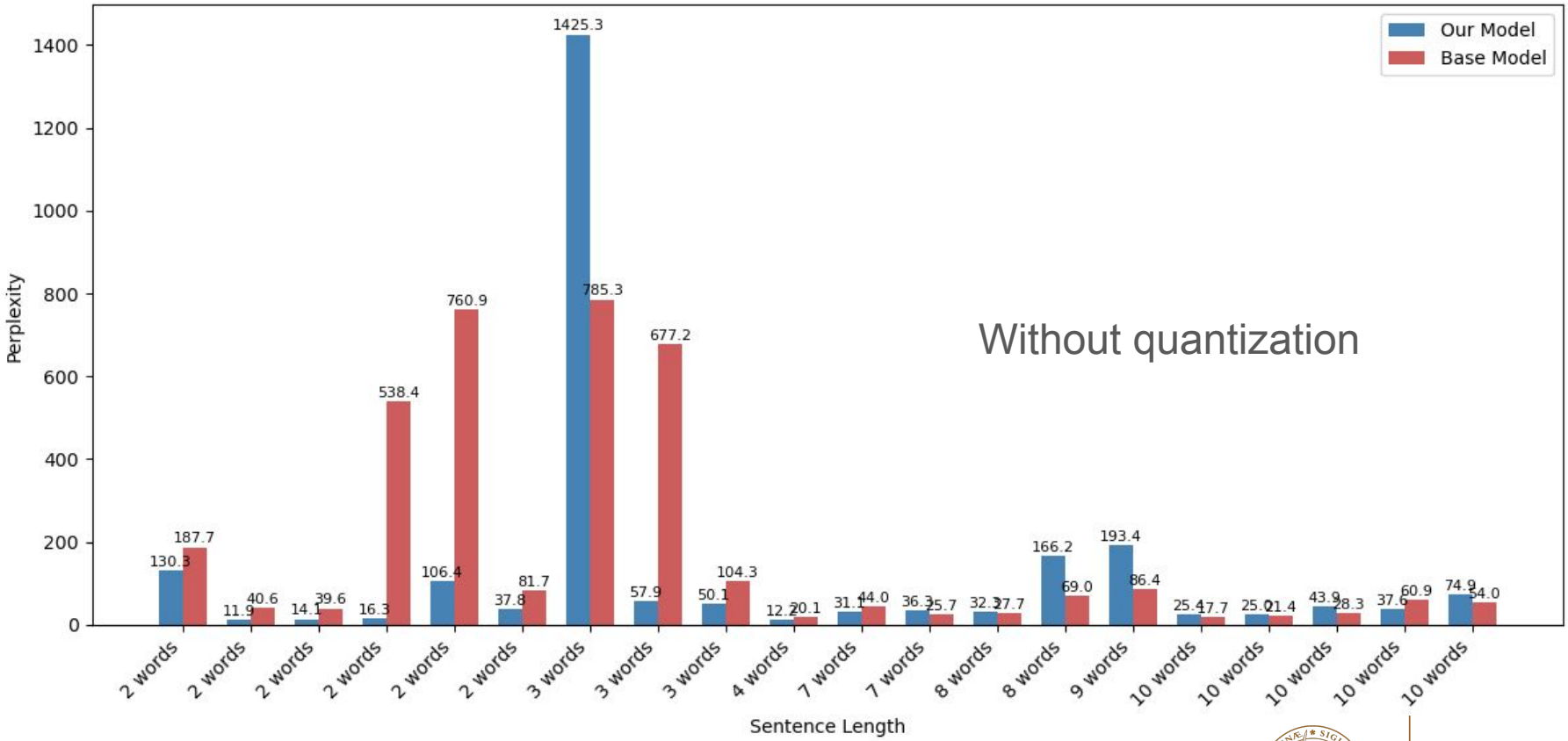


LUND UNIVERSITY
LTH
FACULTY OF ENGINEERING



LUND UNIVERSITY

LTH
FACULTY OF ENGINEERING



LUND UNIVERSITY

LTH
FACULTY OF ENGINEERING

Conclusions

- Highly possible to make a difference - model has improved
- More data -> better results
- If possible - keep quantization to a minimum

Thank you – questions?