Nordisk Familjebok as a Resource for Generating Grammatically Correct Swedish Text

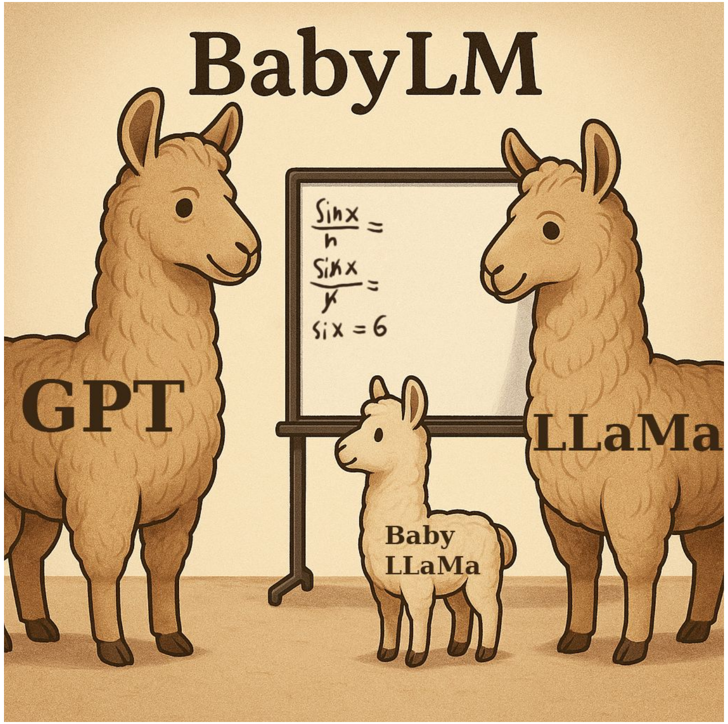
Christopher Meltzer & Morteza Rezaei

Lund University

How well can *Nordisk familjebok* generate grammatically correct Swedish text?

Background

- Inspired by the BabyLM challenge
  - Limited-size dataset (10M or 100M)
  - Best possible result
- The purpose with the BabyLM challenge:
  - Data efficiency
  - Bridge the gap between human and AI
  - Address low-resource NLP challenges
  - And more
- Based on a project from last year:
  - *Small Models, Big Impact: The BabyLM Challenge* by Georgios Chavales and Edoardo Vaira
- In our case:
  We focus more on the chosen dataset



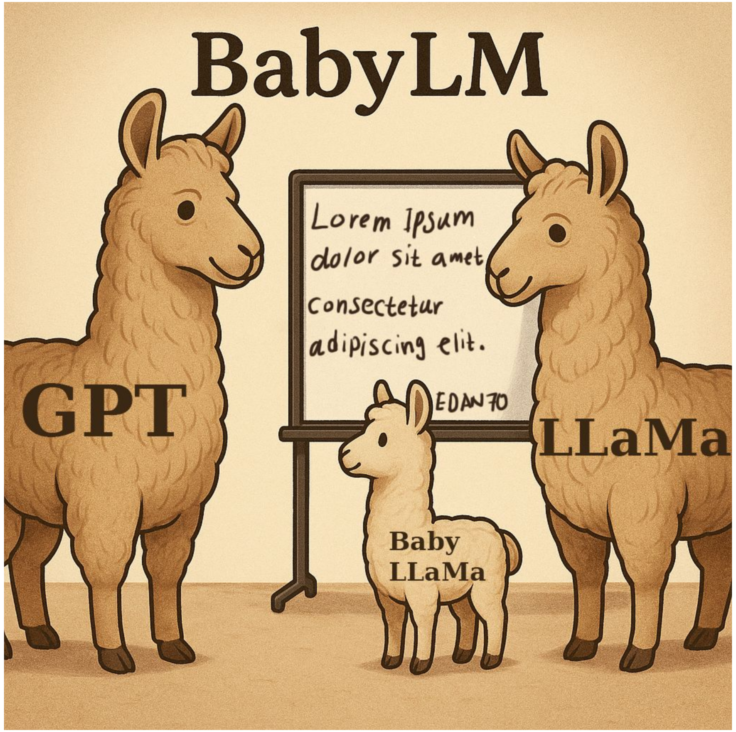
Dataset

- 2nd edition of *Nordisk Familjebok*
- Project Runeberg
- Train and test datasets
  - 19.2 + 4.6 M words (134 + 33 MB)
- Swedish Wikipedia
  - Abstracts of articles
  - Match the size

WIKIPEDIA
Den fria encyklopedin

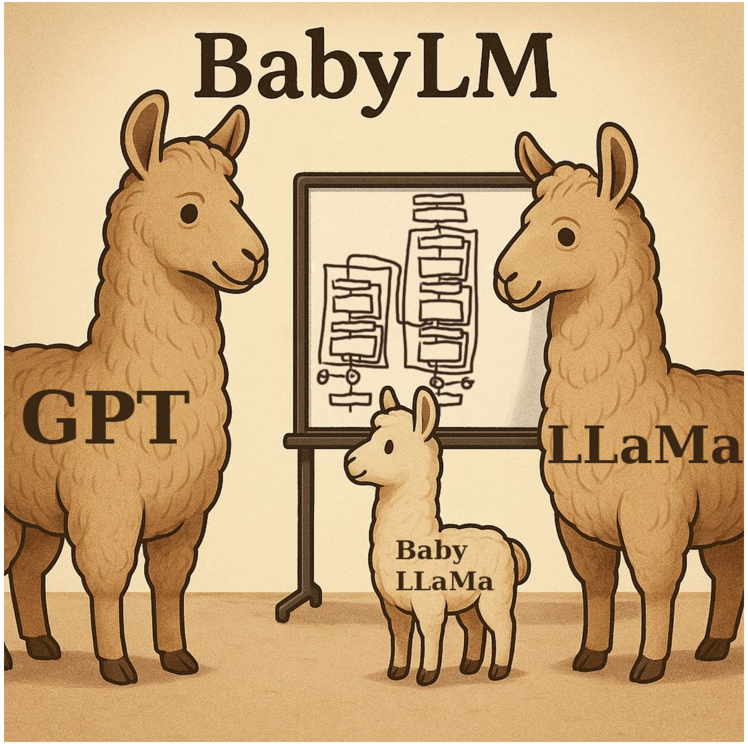
Tokenization

- Wordpiece
- Sequence length of 128
- Vocabulary size of 16 000
- Special tokens:
  - <s>
  - </s>
  - <pad>
  - [UNK]
  - [MASK]

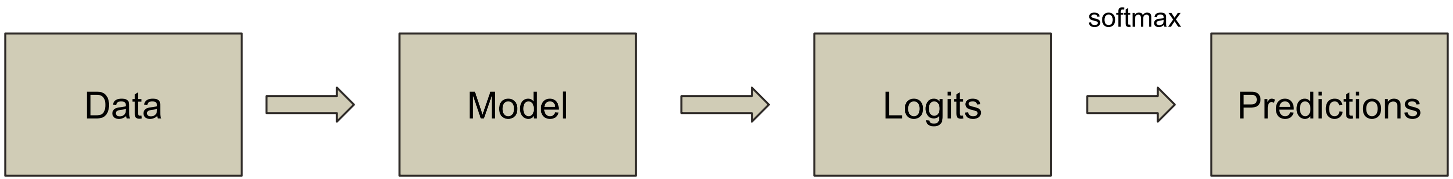


Knowledge Distillation

- What is knowledge distillation?
    - Large model (teacher)
    - Smaller model (student)
    - Student to mimic the teacher
- How is it performed?
    - Training the teacher model(s)
    - Training the student model
        - By combining 2 losses
            - $L = \alpha L_{CE} + (1 - \alpha) L_{KL}$
            - CE: Cross Entropy Loss
            - KL: Kullback–Leibler divergence



Training Teacher Models



Data

Model

Logits

softmax

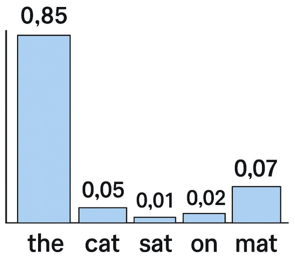
Predictions

Context: “The cat sat on”

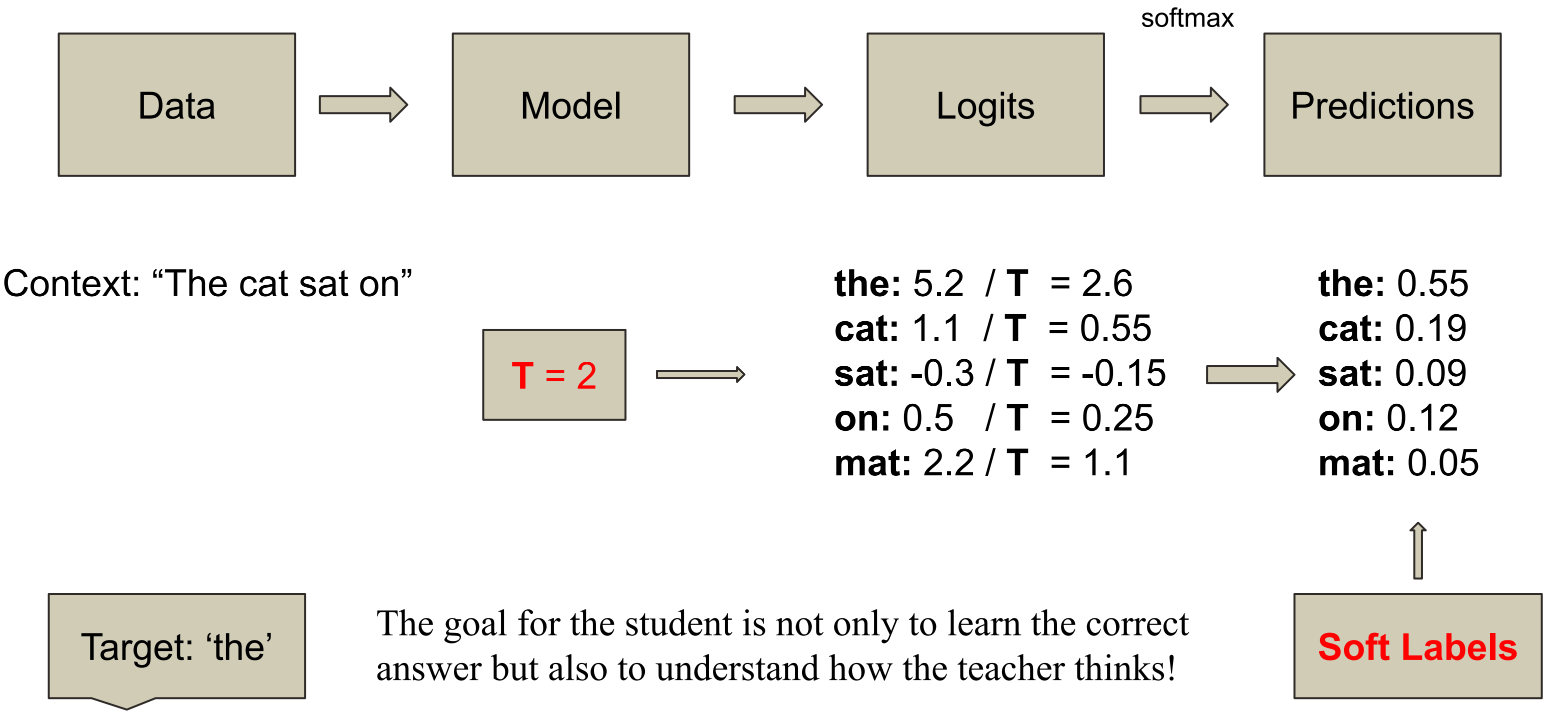
the: 5.2
cat: 1.1
sat: -0.3
on: 0.5
mat: 2.2

the: 0.85
cat: 0.05
sat: 0.01
on: 0.02
mat: 0.07

Target: ‘the’



Temperature and Soft Labels



Data → Model → Logits → Predictions (softmax)

Context: “The cat sat on”

T = 2

the: 5.2 / **T** = 2.6
cat: 1.1 / **T** = 0.55
sat: -0.3 / **T** = -0.15
on: 0.5 / **T** = 0.25
mat: 2.2 / **T** = 1.1

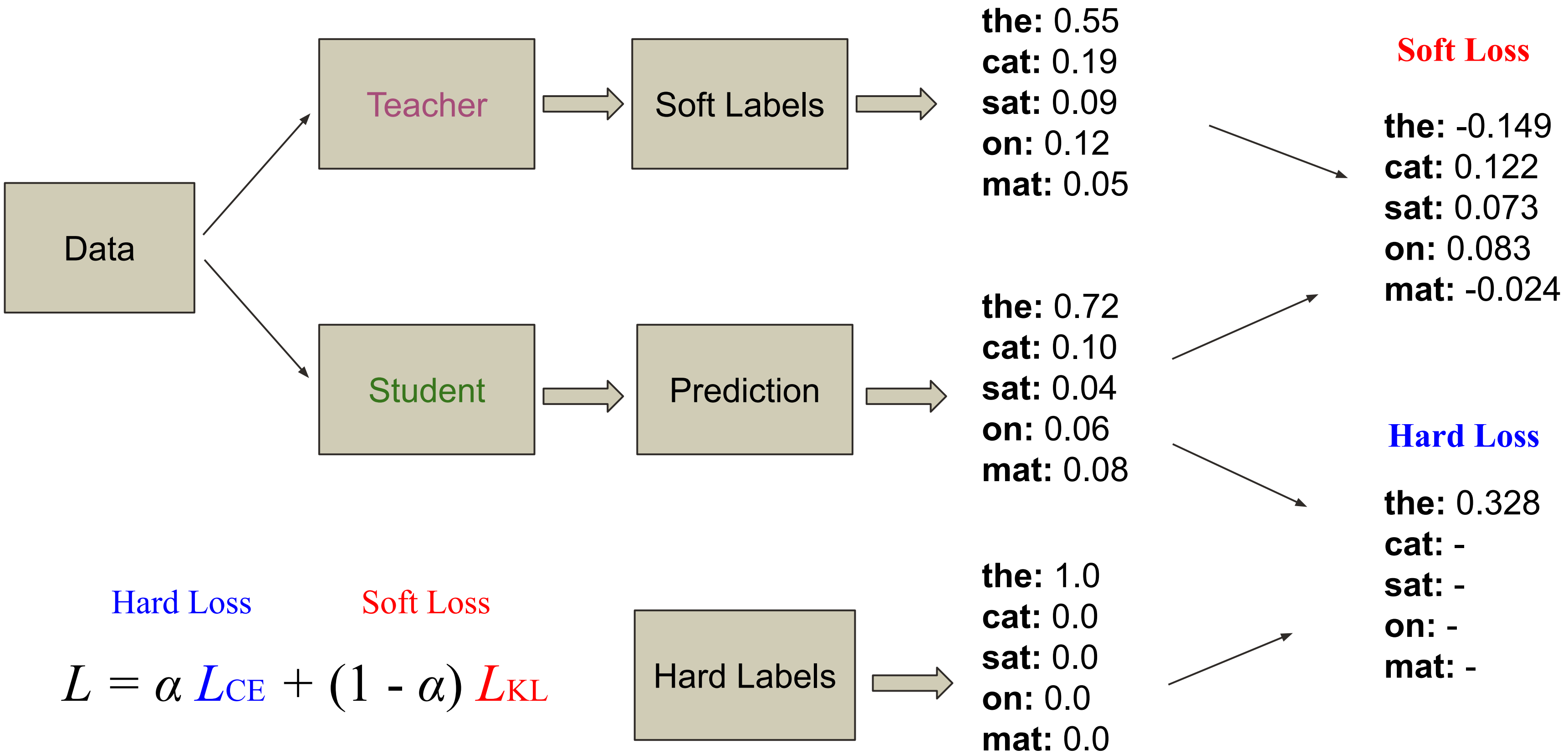
the: 0.55
cat: 0.19
sat: 0.09
on: 0.12
mat: 0.05

Soft Labels

Target: ‘the’

The goal for the student is not only to learn the correct answer but also to understand how the teacher thinks!

Training Student Model



Data

Teacher

Soft Labels

the: 0.55
cat: 0.19
sat: 0.09
on: 0.12
mat: 0.05

Student

Prediction

the: 0.72
cat: 0.10
sat: 0.04
on: 0.06
mat: 0.08

Hard Labels

the: 1.0
cat: 0.0
sat: 0.0
on: 0.0
mat: 0.0

Soft Loss

the: -0.149
cat: 0.122
sat: 0.073
on: 0.083
mat: -0.024

Hard Loss

the: 0.328
cat: -
sat: -
on: -
mat: -

$$L = \alpha \overset{\text{Hard Loss}}{L_{CE}} + (1 - \alpha) \overset{\text{Soft Loss}}{L_{KL}}$$

Our Implementation

Model	Model Size	Training Epochs
GPT2	774M	4
LLaMa	360M	4
RoBERTa	355M	4
OPT	350M	4
LLaMa as student	58M	6

- ➔ α = *0.5*
- ➔ T = *2*
- ➔ ***Soft Labels:*** *average of the teachers*

Evaluation

- DaLAJ-GED-Superlim 2.0 dataset
- Both correct and incorrect sentences
- Error span and corrections
- Only utilize incorrect sentences
- Compare perplexity
- Measure accuracy

Ytterligare en fördel är att de får mig att känna mig som hemma eftersom de är ***mitt*** *familj nu.*

Ytterligare en fördel är att de får mig att känna mig som hemma eftersom de är ***min*** *familj nu.*

Results

Accuracy on Superlim

Model	Nordisk familjebok	Wikipedia
GPT	72.96	78.76
LLaMa	**73.56**	78.53
OPT	72.06	-
RoBERTa	53.02	-
Student of GPT + LLaMa	72.77	**79.09**
Student of GPT + OPT	72.63	-
Student of GPT + RoBERTa	71.28	-
Student of GPT + LLaMa + OPT	72.67	-

Results

Generation example

Nordisk familjebok

<s> < b > Lunds tekniska högskola </ b > i Stockholm bildades i början af 1860 - talet af < sp > G. A. Lundgren </ sp > (se denne) och har till uppgift att utbilda och underhålla ingenjörerna i de svenska och finska industrierna. Detta institut fick sitt första organisationsverk, som f. n. (1908) omfattar äfven ingenjörs - och maskiningenjörer, vid hvilka ingenjörer, ingenjörer, ingenjörer, ingenjörer, ingenjörer, ingenjörer och ingenjörer äro anställda. Sedan 1876 har det varit förenadt med ett institut för elektroteknik och elektroteknik, som omfattar tekniskt och

Wikipedia

<s> Lunds tekniska högskola (RAU) är en svenskspråkig och internationell institution i Lund som grundades 1962 av Fredrik von Heidelberg och Janne Wijnberg. Det är en av de äldsta institutionerna i Sverige och har cirka 1 300 studenter. Akademin har en filial i Lund, som är en av Sveriges största studentnationer och är en av de största och mest inflytelserika i Sverige. </s>

Conclusion

- ~72 % in comparison to ~79 %
- Wikipedia better
    - More similar to Superlim
    - More modern language
    - Fewer abbreviations
- Models perform approximately the same (except RoBERTa)

Questions?