**Plan of work (appended to thesis contract)**

Name of student Ino van de Wouw

Degree programme Text Mining student number 2819716

Name of supervisor(s)/ Antske Fokkens

Chair ……................................................................................

Second reader/ Pia Sommerauer

Chair ….. ………….………………………………………………………………….

Working title Minimalist Language Models: A headless training approach on BabyLM Data.

Topic Train a language model with a learning objective defined as Contrastive Weight Tying (CWT), also called a headless language model (Godey et al., 2023), in combination with the data of the BabyLM shared task (Hu et al., 2024).

Aim and relevance The thesis aims to combine two methods for efficient language model training, which have not been combined before. While headless architectures have primarily been studied in large-scale contexts, exploring their behaviour at smaller scales could determine whether this headless approach remains effective with limited parameters and data. Furthermore, the BabyLM challenge includes an interesting direction in LM development that counters the urge to infinitely scale up training datasets. This poses the following research question: How does a headless language model trained on the BabyLM 10-million token dataset compare in performance to a standard prediction-headed model, and under what conditions might it outperform its traditional counterpart? A first hypothesis is that the Headless LM will outperform the traditional counterpart on the 10M token dataset.

Problem definition Train a language model with a learning objective defined by Godey et al. (2023), this learning objective replaces probability predictions for reconstruction of input embeddings, which Godey calls Contrastive Weight Tying (CWT). The language model will be trained on the BabyLM shared task data. The BabyLM shared task goal is to focus language model training on architecture instead of prioritizing ever-growing training datasets. The data is divided into two tracks, a 10 million token and a 100 million token version. In this thesis the 10 million token version will be used.

Data-collection/ research method

The research method is an experiment: the headless language model architecture will be adapted from Godey, after training the language model will be evaluated on the evaluation pipeline proposed by the 2025 BabyLM task (Charpentier et al., 2025).

Provisional organization of chapters

Abstract -> Introduction -> Methodology -> Results -> Discussion -> Conclusion

Timetable

• start date 31/3/2025

• orientation phase 1/2/2025 – 31/3/2025

• execution phase 31/3/2025 – 30/4/2025 parallel: setting up the experiment & writing background and methodology

(submission dates for chapters) 1/5/2025 – 15/5/2025 overflow experiment & evaluation

• completion phase 15/5/2025 – 27/6/2025 writing the report

• planned thesis presentation date 27th or 28th of May

• planned graduation date 27/6/2025

Remarks At the moment of the thesis plan revision the experiment is set-up to be executed.

Provisional book list

Charpentier, L., Choshen, L., Cotterell, R., Gul, M. O., Hu, M., Jumelet, J., Linzen, T., Liu, J., Mueller, A., Ross, C., Shah, R. S., Warstadt, A., Wilcox, E., & Williams, A. (2025). *BabyLM Turns 3: Call for papers for the 2025 BabyLM workshop* (arXiv:2502.10645). arXiv. https://doi.org/10.48550/arXiv.2502.10645

Godey, N., Clergerie, É. de la, & Sagot, B. (2023). *Headless Language Models: Learning without Predicting with Contrastive Weight Tying* (arXiv:2309.08351). arXiv. https://doi.org/10.48550/arXiv.2309.08351

Hu, M. Y., Mueller, A., Ross, C., Williams, A., Linzen, T., Zhuang, C., Cotterell, R., Choshen, L., Warstadt, A., & Wilcox, E. G. (2024). *Findings of the Second BabyLM Challenge: Sample-Efficient Pretraining on Developmentally Plausible Corpora* (arXiv:2412.05149). arXiv. https://doi.org/10.48550/arXiv.2412.05149