Research Statement

My proposed research for the PhD programme at the University of Copenhagen will investigate explainable AI at the intersection of computational language models and human linguistic cognition. The primary goal is to develop a better understanding of how language models represent linguistic knowledge, how this differs from human cognition, and whether insights from human language learning can help make AI systems more transparent and trustworthy.

The central research question is: **Can aligning language model learning with human cognitive processes enhance the explainability and reliability of NLP systems?** This question leads to two core hypotheses: (1) that current language models represent language in ways that diverge significantly from human mental representations, and (2) that incorporating cognitively-informed constraints or insights can reduce model biases and hallucinations while improving interpretability.

The key research gap lies in the limited integration between cognitive science and explainable AI for language models. While psycholinguistics offers deep insights into human language processing, these are rarely applied to the evaluation or development of modern NLP systems. I aim to bridge this gap by systematically comparing model internals with cognitive data, seeking ways to explain *why* models make certain predictions, not just *what* they predict.

Research Plan and PhD Activities:
Year 1 – Foundation and Exploration
I will start the programme by completing relevant PhD-level courses and defining the research scope. I will also conduct a thorough review of literature in explainable AI and psycholinguistics, identifying benchmarks for cognitive alignment. Early experiments will compare language model outputs and internal states to human behavioral data using psycholinguistic tasks and probing techniques. I will also assist in teaching and begin sharing my work in academic settings.

Year 2 – Comparative Studies and Model Analysis
The second year will focus on empirical studies. I will design experiments that analyze how language models encode syntax, semantics, and pragmatic cues, and compare these with human processing data. Special attention will be given to hallucinations and bias in models, examining whether these mirror or diverge from cognitive errors in humans. During this phase, I plan to submit at least one paper to a high-impact NLP or AI venue. I will also present findings at workshops and recieve feedback from the research community.

Year 3 – Synthesis, Dissemination, and Collaboration
In the final year, I will synthesize results to propose interpretable modeling strategies or diagnostic frameworks inspired by human cognition. I plan to collaborate with an international institution focused on psycholinguistics or model interpretability, fostering interdisciplinary perspectives. I will continue contributing to teaching, finalise multiple publications, and prepare and defend the PhD thesis. This work will not only advance scientific understanding but could lead to more explainable and socially responsible AI systems.

By grounding explainability in cognitive theory, this research could help build AI tools that are more understandable, safer, and better aligned with human values.