Personal Background

I am a PhD researcher in Computer Science at the University of Copenhagen working on large language model evaluation and responsible AI. My academic background in linguistics and cognitive science, combined with industry experience in applied NLP and generative AI, has shaped a research agenda centered on understanding, evaluating, and responsibly communicating the capabilities of language technologies.

Research Statement

My research is motivated by a longstanding interest in human cognition and language. My training in linguistics and cognitive science led me to questions about how humans represent meaning, process language, and interpret the world through communication. Over time, this interest expanded into machine cognition: if language models are increasingly used as proxies for humans, what kinds of capabilities do we actually want them to possess, and how should we evaluate those capabilities in a scientifically meaningful and socially responsible way?

My work focuses primarily on evaluation. I am interested in the foundational question of what it means for an AI system to perform well, and whether current evaluation practices, tools, and methodological assumptions are adequate for answering that question. This involves not only measuring model performance, but also examining the conceptual choices behind evaluation practices. I see evaluation as both a technical and epistemic problem: it requires us to define which capacities matter, determine how they can be measured, and communicate results in ways that are interpretable and useful across research, industry, and governance settings.

A first strand of my research concerns the evaluation of cognitive aspects of language processing. I am interested in whether and in what ways machines exhibit properties that resemble human language understanding, reasoning, or representation. Rather than taking “human-like” behavior at face value, I aim to investigate it carefully: which aspects of cognition are meaningfully comparable, where the analogies break down, and what such comparisons reveal about both human and machine intelligence.

A second strand of my work addresses methodological questions in AI evaluation. I am interested in how evaluation datasets are designed, how tasks operationalize complex constructs, and how conclusions about model ability are drawn from benchmark results. My goal is to contribute to more rigorous and transparent evaluation frameworks that move beyond narrow performance metrics and instead capture the broader validity, limitations, and intended use of an assessment. In this regard, I see evaluation as an essential scientific infrastructure for AI: without careful methodology, our claims about intelligence, reasoning, or usefulness remain underspecified.

A third strand of my research engages with the regulatory and communicative dimensions of evaluation. As AI systems are deployed to wider publics and increasingly shape decisions across domains, it becomes important not only to evaluate systems well, but also to standardize how evaluation is reported and communicated. I am particularly interested in frameworks that make evaluation results more transparent, comparable, and accountable for

different stakeholders, including researchers, developers, policymakers, and end users. My work on standardized reporting emerges from this concern that evaluation should support informed governance and responsible adoption, not merely internal model comparison.

Across these directions, my broader aim is to build evaluation methods that are technically rigorous, conceptually grounded, and socially responsive. I want to contribute to a research culture in which AI systems are not only optimized for performance, but also examined in terms of what their capabilities mean, how they are measured, and how those measurements are communicated. By combining insights from linguistics, cognitive science, and computer science, I hope to advance a more human-centered and transparent science of AI evaluation.