PhD Program between Freie Universität Berlin (FUB) and the China Scholarship Council (CSC)

Open PhD Position at Freie Universität Berlin, offered only to Chinese CSC scholarship candidates 2027

Department/Institute: Department of Education and Psychology

Subject area: Cognitive Neuroscience

Name of Supervisor: Radoslaw Martin Cichy

Number of open PhD positions: 1

Type of the PhD Study: full-time only

Project title: Metacognition in Large Language Models: Quantifying Confidence and Calibration

PhD Project description:

Reliable artificial intelligence requires systems that can assess when their answers are likely to be correct. While modern large language models (LLMs) produce highly fluent text, they lack robust mechanisms to evaluate their own correctness, frequently leading to hallucinations and miscalibrated confidence.

Key goals of this PhD project are:

- Collect and curate a set of benchmark datasets with sufficient variability in LLM accuracy and confidence.
- Implement and evaluate different methods of confidence elicitation in LLMs (e.g. explicit probability judgements versus response consistency)
- Implement and validate a measure of metacognitive efficiency of LLMs.
- Investigate which confidence elicitation methods maximize metacognitive efficiency
- Compare LLM metacognition across models and domains (e.g., general knowledge versus fluid intelligence).
- Assess reliability of LLM confidence generation (e.g., test-retest reliability)

The project will use a computational toolbox developed by one of the supervisors, to robustly estimate metacognitive parameters. For more information, see: https://re-meta.github.io/

Overall, this research will transfer insights from human metacognition research to establish rigorous, much-needed evaluation frameworks for machine confidence. Its findings are highly relevant for the development of artificial systems that communicate uncertainty faithfully and identify when additional verification or human review is needed.

The project is suited for highly motivated applicants with interests in computer science, cognitive science, psychology, and related fields.

Language requirements:

A score of ≥ 6.5 on IELTS, or a score of ≥ 95 on TOEFL iBT (corresponding to ≥ 5 on the new 1-6 scale).

Academic requirements:

The applicant should have a Master degree in one of the following (or related) disciplines: Computer Science, Artificial Intelligence, Cognitive Science, Computational Neuroscience, Biomedical engineering.

The candidate should have a strong interest in artificial intelligence and cognitive science, and excellent coding skills (in particular, Python).

Information of the professor or research group leader (website, awards etc.):

Supervisor
Prof. Radoslaw Cichy

Website of the research group:
https://www.ewi-psy.fu-berlin.de/en/einrichtungen/arbeitsbereiche/neural_dyn_of_vis_cog/index.html

Awards:
2026: BMFTR grant GROWUPAI (Co-PI with Gemma Roig (GU Frankfurt), 1.1M€) - Visual AI that grows up like the human brain for improved efficiency and robustness
2024: ERC Consolidator Grant 101123101 (PI, 2.3M€) A theory and model of the neural
transformations mediating human object perception (TRANSFORM)
2022: Early Career Impact Award from the Federation of Associations in Behavioral and Brain
Sciences (FARBS)
2020: Scout of the Henriette Herz Scouting Program of the Alexander von Humboldt Foundation
2020: Neuroimage Paper of the Year
2019 Fellow of the Research Group “Cognitive Behavior of Humans, Animals and Machines:
Situation Model Perspectives” (Center for Interdisciplinary Research, University Bielefeld)

Key publications:

Gifford AT, Jastrzębowska MA, Singer JJD, Cichy RM (2025) In silico discovery of representational relationships across visual cortex. Nat Human Behav; doi: 10.1038/s41562-025-02252-z

Lu Z*, Doerig A*, Bosch V*, Krahmer B, Kaiser D*, Cichy RM*, Kietzmann TC* (2025). End-to-
end topographic networks as models of cortical map formation and human visual behaviour. Nat
Human Behav; doi: 0.1038/s41562-025-02220-7.

Lahner B, Dwivedi K, Iamshchinina P, Graumann M, Lascelles A, Roig G, Gifford AT, Pan B, Jin S, Murtz NAR, Kay K, Oliva A*, Cichy RM* (2024) BOLD Moments: modeling short visual events
through a video fMRI dataset and metadata. Nat Comm 15: 6241; doi: 10.1038/s41467-024-50310-3.

Graumann M, Ciuffi C, Dwivedi K, Roig G, Cichy RM (2022) The spatiotemporal neural dynamics
of object location representations in the human brain. Nat Human Behav 6: 796–811; doi
10.1038/s41562-022-01302-0.

Cichy RM, Oliva A (2020) A M/EEG-fMRI Fusion Primer: Resolving Human Brain Responses in
Space and Time. Neuron 107(5): 772-781; doi: 10.1016/j.neuron.2020.07.001.

Cichy RM & Kaiser D (2019) Deep neural networks as scientific models. Trends Cogn Sci 23(4): 305-317; doi: 10.1016/j.tics.2019.01.009.

Guggenmos M (2022) Reverse Engineering of Metacognition. eLife 11: e75420; doi: 10.7554/eLife.75420.

Co- supervisor:
Prof. Matthias Guggenmos, Health and Medical University Potsdam
Professor for Computational Cognitive Neuroscience
https://coconeuro.github.io/

Please note:
In a first step, the complete application should be uploaded to the online portal (https://fuberlin.moveon4.de/form/60acfece5d328710e40bdbd5/eng) for evaluation by January 15th, 2027.