Adrian de Wynter

adrian.dewynter@york.ac.uk | LinkedIn | Homepage | Google Scholar

My main research interest is machine cognition (language, understanding, reasoning, etc.) and measurement (especially post-AI). These two are deeply related, since scientific research on AI's cognitive capabilities require carefully-designed experiments and assumptions to draw trustworthy conclusions. My applied work in this area ranges from reasoning (e.g., chain-of-thought verification) to safety (e.g., misalignment). I also study the impact of LLMs and AI in society.

Education

BSc in Engineering, BSc in Physics (double major) 2011-2015
Massachusetts Institute of Technology

MSc in Computer Science 2021
University of Texas at Austin

PhD in Computer Science (ABD; 2022-June 2027 exp.)
University of York
Thesis: Understanding Arguments in Large Language Models
Advisor: Tangming Yuan
Probes the relationship between cognitive capabilities of LLMs; namely, understanding in an informal-logic context (how do you define it? How do you measure it?) and how this works internally.

Employment

Principal Applied Scientist 2022-2026
Microsoft Experiences + Devices
Science lead for the design and deployment of Office AI features and research. These range from Copilot-wide language expansion (42 languages!); prompt suggestions and rewrite in Word; and reliable and accurate evaluation harnesses for Word Agent. My primary research focus is cognition (reasoning, understanding) and measurement (meta-science, and automated evaluation). My day-to-day centres on long-and-medium term investments related to trustworthy measurement and evaluations of and (with) LLMs and agentic systems, and being the technical lead of various applied research teams. I also drive university collaborations (currently BC, KAIST, Columbia, INRIA, and IIT-H).

PGR 2022-Present
University of York Department of Computer Science
I'm a researcher (PGR) in a part-time capacity. My research focus is in exploring epistemic questions around the relationship between understanding and reasoning in LLMs, especially (but not limited to) how it relates to dialogical components such as argumentative reasoning and persuasion. I also mentor MSc and BSc students in their theses.

Applied Scientist 2017-2021
Amazon AWS, Alexa
I worked at Amazon Rekognition (helped launch the video version), AWS (forecasting models for call centres), and Alexa (focused on highly efficient models).

Affiliations

Thinking About Thinking 2026 - present
Fellow

Physical AI Safety Consortium 2026 - present
Researcher

Awards

My work has been covered by venues such as IBM, PC Magazine, Tom's Hardware, 404 Media, and the Turing Post. See my website for links to these.

The Turing Post: 23 Research Papers That Hint Where AI Is Heading 2025
Selected 'Is In-Context Learning Learning?' for its 2025 research roundup.

Microsoft CELA Open Data Award 2025
For the MCFM corpus, a dataset for misgendering detection in 42 languages and dialects.

Microsoft CELA Open Data Award 2023
For the CLANDESTINO corpus, a dataset for localized Spanish toxic-language detection.

Microsoft CELA Open Data Award 2023
For the RTP-LX corpus, a dataset in 38 languages for toxic-language detection.

(Selected) Invited Talks

Assumptionless Machine Cognition 2026
Invited talk at the German Research Centre for Artificial Intelligence (DFKI; 2026). Shortened abstract: I will present a research programme for studying machine cognition without dismissing these behaviours as autocomplete, nor anthropomorphising them. We'll discuss evidence of genuine-but-shallow learning and comprehension capabilities and the meta-scientific crisis this creates; and how they relate to debates on emergence, fairness, and the impact of deploying systems we barely understand.

Keynote: Some Issues in Machine Cognition 2026
BoF at ACL 2026: Social Agents in the Wild: Reasoning, Simulation and Evaluation. Cognition in AI systems is tricky: we do not know how to measure it, and the current claims rely too much in anthropomorphism. In this talk I will describe hurdles, solutions, and emerging results within the field.

The Document Manipulation Benchmark: Challenges in Agentic Document Manipulation 2026
LREC Industry Day: we pose that agentic complexity requires both fine- and coarse-grained measurements to better quantify efficiency and technological leaps.

Will GPT-4 (and 5!) Run DOOM? 2026
ACM I3D SIGGRAPH Symposium on Computer Graphics: measuring cognitive capabilities (spatial orientation, object permanence, etc) from the lens of videogames.

Evaluation of LLMs: Challenges and Directions 2025
Invited talk at MBZUAI's Research Showcase. We pose that in order to perform accurate scientific measurement with LLMs it is needed to lift some pre-existing assumptions on what it means to measure things in science.

The Curse of the Biased Researcher: Common Pitfalls in LLM Evaluation 2023
Microsoft Machine Learning and Data Science (MLADS) Conference: I point out that because LLMs sound reasonable, that does not mean that their output is.

On the Opportunities and Dangers of LLM-Based Evaluation 2023
Microsoft Machine Learning and Data Science (MLADS) Conference: I pose that LLMs are unable to reason fully over their input, and thus the pre-supposed assumption that an LLM will be an appropriate evaluator is false. This eventually became my paper 'Is In-Context Learning Learning?'.

(Selected) Publications

Is In-Context Learning Learning? ICLR 2026
Experimentally determines whether the in-context learning (ICL) mechanism constitutes learning in the sense of Valiant (1984). It finds that yes, ICL can learn from unseen distributions–even independently of natural language–but it is brittle to multiple conditions, such as in versus out-of distribution inputs.

The Thin Line Between Comprehension and Persuasion in LLMs ACL 2026
It measures whether LLMs can understand arguments. This is done through both language- and imitation-game-inspired measurements. It finds that the measurement radically changes the results: one indicates lack of comprehension, and the other shows signs of understanding. It concludes by posing the question as to whether it matters that LLMs cannot fully show comprehension of the context.

Will GPT-4 Run *DOOM*? IEEE ToG 2024
The paper–one of the earliest of its kind–shows that videogames are excellent evaluation harnesses for LLMs due to the dynamicity of their environment. This work, and its 2026 invited talk at I3D SIGGRAPH, also finds that LLMs like GPT-5 and Phi-4 have lacking spatial reasoning and long-term planning capabilities, but aren't entirely devoid of these.

Service

Organiser
Machine cognition meetup at EMNLP 2026–a very informal BoF-like setup with like-minded researchers.

Journal Reviewer
Nature Communications (2025-); Nature Machine Intelligence (2025-); Nature AI & Ethics (2025-); ACM Transactions on Games (2024-); ACM Transactions on Intelligent Systems and Technology (2025)

Conference Reviewer / Program Committee
ARR (2024-), ICLR (2024-), NeurIPS (2024-), ICML (2024-), AAAI (2024-).

Publications

[1] **A. de Wynter**, "An approximation algorithm for optimal subarchitecture extraction," vol. abs/2010.08512, 2020. [Online]. Available: https://arxiv.org/abs/2010.08512

[2] ——, "Turing completeness and Sid Meier's *Civilization*," *IEEE Transactions on Games*, 2022. [Online]. Available: https://doi.org/10.1109/TG.2022.3166874

[3] ——, "An algorithm for learning smaller representations of models with scarce data," *Information Geometry*, 2024. [Online]. Available: https://doi.org/10.1007/s41884-024-00153-0

[4] ——, "Will GPT-4 Run DOOM?" *IEEE Transactions on Games*, 2024. [Online]. Available: https://doi.org/10.1109/TG.2024.3497601

[5] ——, "Algorithmically establishing trust in evaluators," 2025. [Online]. Available: https://arxiv.org/abs/2506.03083

[6] ——, "Awes, laws, and flaws from today's LLM research," in *Findings of the Association for Computational Linguistics: ACL 2025*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds. Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 12 834–12 854. [Online]. Available: https://aclanthology.org/2025.findings-acl.664/

[7] ——, "If Eleanor Rigby had met ChatGPT: A study on loneliness in a post-LLM world," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds. Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 19 898–19 913. [Online]. Available: https://aclanthology.org/2025.acl-long.976/

[8] ——, "Cross-pollination and crop rotation: A field guide for agentic systems research," 2026, preprint available upon request.

[9] ——, "If LLMs have human-like attributes, then so does *Age of Empires II*," 2026. [Online]. Available: https://arxiv.org/abs/2605.31514

[10] ——, "Is in-context learning learning?" in *The Fourteenth International Conference on Learning Representations*, 2026. [Online]. Available: https://openreview.net/forum?id=bISoBbYjjd

[11] ——, "Population physics, population problems: Safety and emergence in LLM societies," 2026. [Online]. Available: https://arxiv.org/abs/2609.33871

[12] ——, "What was said, not what was 'thought': Type-6 logic for CoT verification," 2026. [Online]. Available: https://arxiv.org/abs/2609.38420

[13] **A. de Wynter** and D. J. Perry, "Optimal subarchitecture extraction for BERT," vol. abs/2010.10499, 2020. [Online]. Available: https://arxiv.org/abs/2010.10499

[14] **A. de Wynter**, X. Wang, Q. Gu, and S.-Q. Chen, "On meta-prompting," vol. abs/2312.06562, 2023. [Online]. Available: https://arxiv.org/abs/2312.06562

[15] **A. de Wynter**, X. Wang, A. Sokolov, Q. Gu, and S.-Q. Chen, "An evaluation on large language model outputs: Discourse and memorization," *Natural Language Processing Journal*, vol. 4, p. 100024, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2949719123000213

[16] **A. de Wynter**, I. Watts, T. Wongsangaroonsri, M. Zhang, N. Farra, N. E. Altıntoprak, L. Baur, S. Claudet, P. Gajdušek, Q. Gu, A. Kaminska, T. Kaminski, R. Kuo, A. Kyuba, J. Lee, K. Mathur, P. Merok, I. Milovanović, N. Paananen, V.-M. Paananen, A. Pavlenko, B. P. Vidal, L. I. Strika, Y. Tsao, D. Turcato, O. Vakhno, J. Velcsov, A. Vickers, S. F. Visser, H. Widarmanto, A. Zaikin, and S.-Q. Chen, "RTP-LX: Can LLMs evaluate toxicity in multilingual scenarios?" *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 27, pp. 27 940–27 950, Apr. 2025. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/35011

[17] **A. de Wynter** and T. Yuan, "'I'd Like to Have an Argument, Please": Argumentative Reasoning in Large Language Models," in *Proceedings of COMMA 2024*, C. Reed, M. Thimm, and T. Rienstra, Eds. Frontiers in Artificial Intelligence and Applications, 2024, pp. 73–84. [Online]. Available: https://ebooks.iospress.nl/doi/10.3233/FAIA240311

[18] ——, "The thin line between comprehension and persuasion in LLMs," in *Findings of the Association for Computational Linguistics: ACL 2026*, M. Liakata, V. P. Moreira, J. Zhang, and D. Jurgens, Eds. San Diego, California, United States: Association for Computational Linguistics, Jul. 2026, pp. 6603–6631. [Online]. Available: https://aclanthology.org/2026.findings-acl.329/

[19] R. Hada, V. Gumma, **A. de Wynter**, H. Diddee, M. Ahmed, M. Choudhury, K. Bali, and S. Sitaram, "Are large language model-based evaluators the solution to scaling up multilingual evaluation?" in *Findings of the Association for Computational Linguistics: EACL 2024*, Y. Graham and M. Purver, Eds. St. Julian's, Malta: Association for Computational Linguistics, Mar. 2024, pp. 1051–1070. [Online]. Available: https://aclanthology.org/2024.findings-eacl.71

[20] A. Jangra and **A. de Wynter**, "Fantastic metrics and how to find them: Meta-evaluating metrics for multi-lingual style-personalized generation," 2026. [Online]. Available: https://arxiv.org/abs/2610.19166

[21] A. Jangra, B. Sarrafzadeh, S. Cucerzan, **A. de Wynter***, and S. K. Jauhar*, "Evaluating style-personalized text generation: Challenges and directions," in *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*, H.-Y. Lee, S. Dev, J. Li, and D. Elliott, Eds. Budapest, Hungary: Association for Computational Linguistics, Oct. 2026, pp. 3854–3878. [Online]. Available: https://aclanthology.org/2026.emnlp-main.168/

[22] D. Kim, **A. de Wynter**, H. Chen, H. Kim, and H. Vikalo, "Foundation-preserving adaptation via generalized Rayleigh-quotient optimization," 2026. [Online]. Available: https://arxiv.org/abs/2606.00132

[23] W. Li and **A. de Wynter**, "The Hrunting of AI: Where and how to improve English dialectal fairness," 2026. [Online]. Available: https://arxiv.org/abs/2603.15187

[24] W. Li, A. Zhao, **A. de Wynter**, S.-Q. Chen, P. Karimov, and J. K. Hartshorne, "Does using LLMs in daily life help or hinder learning a second language?" *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 47. [Online]. Available: https://escholarship.org/uc/item/2jh361fg

[25] F. Lin, S. Mao, E. La Malfa, V. Hofmann, **A. de Wynter**, X. Wang, S.-Q. Chen, M. J. Wooldridge, J. B. Pierrehumbert, and F. Wei, "Assessing dialect fairness and robustness of large language models in reasoning tasks," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds. Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 6317–6342. [Online]. Available: https://aclanthology.org/2025.acl-long.317/

[26] S. Sitaram, **A. de Wynter**, I. McCrum, Q. Gu, and S.-Q. Chen, "A multilingual, culture-first approach to addressing misgendering in LLM applications," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, Eds. Suzhou, China: Association for Computational Linguistics, Nov. 2025, pp. 31 147–31 171. [Online]. Available: https://aclanthology.org/2025.emnlp-main.1587/

[27] Y. Zhang, S. Mao, T. Ge, X. Wang, **A. de Wynter**, Y. Xia, W. Wu, T. Song, M. Lan, and F. Wei, "LLM as a mastermind: A survey of strategic reasoning with large language models," *COLM*, 2024. [Online]. Available: https://arxiv.org/abs/2404.01230

Patents

[Pat1] S.-Q. Chen, **A. de Wynter**, M. Zhang, D. Du, Z. Li, A. B. Figueiredo F de Souza, A. Shepherd, C. Mayer, A. Sefid, T. Ahmed, and L. Strika, "An intelligent writing assistant for neurodiverse users," patentus 5 861 843, filed.

[Pat2] S.-Q. Chen, X. Wang, **A. de Wynter**, X. He, Q. Gu, M. Johnson, and U. Bose, "Content assistance processes for foundation model integrations," patentus 413 669-US01, filed.

[Pat3] D. Doyle, **A. de Wynter**, S.-Q. Chen, O. Gauthier, S. Shi, A. Shastri, D. Mac Mathuna, S. Mohamed, and I. McCrum, "Adaptive infrastructure for improved content creation with agentic systems," patentus 505 171-US01, filed.

[Pat4] A. Vickers, **A. de Wynter**, and E. Cadoni, "Creating textual output with the writing style of a specific user using generative artificial intelligence," patentus 5 861 843, filed.