

Adrian de Wynter

adewynter@microsoft.com | Homepage | Google Scholar

My main research interests is cognition in machines (language, understanding, reasoning, etc.), and measurement in science (especially post-AI). These two are deeply related, since scientific research on cognitive capabilities require carefully-designed experiments and assumptions to draw trustworthy conclusions. I also do research on computational social science; namely, the impact of LLMs and AI in society.

Education

BSc in Engineering, BSc in Physics

Massachusetts Institute of Technology

MSc in Computer Science

University of Texas at Austin

PhD in Computer Science

(ABD; Oct 2026 exp.)

University of York

Thesis: Understanding and Persuasion 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-Present

Microsoft Word

I am a principal scientist in Copilot. I lead, design, and deploy Word- and Office- AI features and research. My focus at the moment 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. I also study computational social science topics such as toxicity, overreliance in LLMs, and multilingualism. 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.

Applied Scientist

2017-2021

Amazon AWS, Alexa

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

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.

Thinking About Thinking Fellow

2026

For my work on LLM cognition, such as understanding and learning.

The Turing Post: 23 Research Papers That Hint Where AI Is Heading

2025

For my paper 'Is In-Context Learning Learning?' (not an award, still cool).

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 Technische Universität Berlin (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. And yes, we will cover the *Age of Empires II* paper.

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. I know it's commonsense now, but it wasn't in 2023.

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?'.

Service

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-). Everyone does these though.

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] —, “A field guide for agentic systems research: Open problems and next steps,” 2026. [Online]. Available: Forthcoming
- [9] —, “A field guide for agentic systems research: Open problems and next steps,” 2026, forthcoming.
- [10] —, “If LLMs have human-like attributes, then so does *Age of Empires II*,” 2026. [Online]. Available: <https://arxiv.org/abs/2605.31514>
- [11] —, “Is in-context learning learning?” in *The Fourteenth International Conference on Learning Representations*, 2026. [Online]. Available: <https://openreview.net/forum?id=bIS0BbYjjd>
- [12] **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>
- [13] **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>
- [14] **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>
- [15] **A. de Wynter**, I. Watts, T. Wongsangaroonsri, M. Zhang, N. Farra, N. E. Altintoprak, 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>
- [16] **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>
- [17] —, “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/>
- [18] 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>
- [19] A. Jangra, B. Sarrafzadeh, S. Cucerzan, **A. de Wynter***, and S. K. Jauhar*, “Evaluating style-personalized text generation: Challenges and directions,” in *The Fifth Workshop on Natural Language Generation, Evaluation, and Metrics (GEM) Workshop (ACL 2026)*, 2026. [Online]. Available: <https://arxiv.org/abs/2508.06374>
- [20] 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>
- [21] 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>

- [22] 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>
- [23] 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/>
- [24] 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/>
- [25] 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.