Michael Hanna

Amsterdam, The Netherlands | m.w.hanna@uva.nl | hannamw.github.io

1 Education:

University of Amsterdam, Amsterdam, The Netherlands
PhD, Computational Linguistics (begun Sept. 2022; expected ~Oct. 2026)

Charles University, Prague, Czech Republic[†] (Sept. 2022)
MS, Computer Science; specialization in computational linguistics; GPA: 1 (excellent) / A, with honors

University of Trento, Trento, Italy[†] (July 2022)
MS, Cognitive Science; specialization in computational linguistics; GPA: 110/110, with honors

University of Chicago, Chicago, IL, USA (June 2020)
BS with Honors, Computer Science, specialization in machine learning; GPA: 3.95
BA with Honors, Linguistics; GPA: 3.96
Honors Thesis: *Measuring the Interpretability of Latent-Space Representations of Sentences from Variational Autoencoders*.

2 Publications:

Curt Tigges, **Michael Hanna**, Qinan Yu, and Stella Biderman. 2024. LLM Circuit Analyses Are Consistent Across Training and Scale. *Advances in Neural Information Processing Systems (to appear).* **(NeurIPS 2024)**

Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. 2024. Have Faith in Faithfulness: Going Beyond Circuit Overlap When Finding Model Mechanisms. *First Conference on Language Modeling (to appear)*. **(COLM 2024)**

Frank Wildenburg, **Michael Hanna**, and Sandro Pezzelle. 2024. Do Pre-Trained Language Models Detect and Understand Semantic Underspecification? Ask the DUST!. *Findings of the Association for Computational Linguistics*. **(ACL Findings 2024)**

Michael Hanna, Ollie Liu, and Alexandre Variengien. 2023. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. In *Advances in Neural Information Processing Systems.* **(NeurIPS 2023)**

Michael Hanna, Yonatan Belinkov, and Sandro Pezzelle. 2023. When Language Models Fall in Love: Animacy Processing in Transformer Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. **(EMNLP 2023)**

Abhijith Chintam, Rahel Beloch, Willem Zuidema, **Michael Hanna***, and Oskar van der Wal*. 2023. Identifying and Adapting Transformer-Components Responsible for Gender Bias in an English Language Model. In

[†]These degrees were part of the Erasmus Mundus LCT dual-degree master's program. I spent the 2020-2021 academic year at Charles University and the 2021-2022 academic year at the University of Trento. My master's thesis, joint between the two, was: *Investigating Large Language Models' Representations Of Plurality Through Probing Interventions*
*Equal contribution

Proceedings of the Sixth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP. **(BlackboxNLP 2023)**

Jaap Jumelet, **Michael Hanna***, Marianne de Heer Kloots*, Anne Langedijk*, Charlotte Pouw*, and Oskar van der Wal*. 2023. ChapGTP, ILLC's Attempt at Raising a BabyLM: Improving Data Efficiency by Automatic Task Formation. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning* **(BabyLM Challenge 2023)**

Michael Hanna, Roberto Zamparelli, and David Mareček. 2023. The Functional Relevance of Probed Information: A Case Study. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Dubrovnik, Croatia. Association for Computational Linguistics. **(EACL 2023)**

Michael Hanna*, Federico Pedeni*, Alessandro Suglia, Alberto Testoni, and Raffaella Bernardi. 2022. ACT-Thor: A Controlled Benchmark for Embodied Action Understanding in Simulated Environments. In *Proceedings of the 29th International Conference on Computational Linguistics*, Gyeongju, Republic of Korea. International Committee on Computational Linguistics. **(COLING 2022)**

Michael Hanna and Ondřej Bojar. 2021. A Fine-Grained Analysis of BERTScore. In *Proceedings of the Sixth Conference on Machine Translation*. Punta Cana, Dominican Republic (Online). Association for Computational Linguistics. **(WMT 2021)**

Michael Hanna and David Mareček. 2021. Analyzing BERT's Knowledge of Hypernymy via Prompting. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. Punta Cana, Dominican Republic. Association for Computational Linguistics. **(BlackboxNLP 2021)**

3 Work Experience:

Graduate Teaching Assistant, University of Amsterdam (June 2023 – present)

- Designed, taught, and assessed a week-long master's-level workshop on mechanistic interpretability, including interactive materials (Jupyter Notebooks).
- Advised student projects in mechanistic interpretability.
- Crafted and assessed written assignments for a master's-level cognitive science course.

Research Resident, Redwood Research (Berkeley, CA) (Jan. 2023 – Feb. 2023)

- Learned mechanistic interpretability techniques as part of the REMIX program.
- Studied low-level mechanisms underlying GPT-2's behavior on a math task. This led to a NeurIPS paper, *How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model*.

Research Intern, Charles University, Institute of Formal and Applied Linguistics (Mar. 2021 – Aug. 2021)

- Used prompting to probe BERT for knowledge of hypernyms of common words, and compared BERT's hypernym discovery performance to existing systems'. This led to a BlackBoxNLP paper, *Analyzing BERT's Knowledge of Hypernymy via Prompting*.

Research Assistant, University of Chicago, Department of Linguistics (Jan. 2020 – Jun. 2020)

- Used unsupervised clustering to test if ELMo embeddings of polysemous words were embedded in distinct clusters in the embedding space; this could allow for unsupervised learning of word senses.
- Used zero-shot probing tasks to explore the relationship between BERT's (masked) language modeling abilities / pre-training and its high performance on down-stream tasks.

Software Engineering Intern, Orbital Insight (Boston, MA) (Summer 2019)
- As part of a transition between geodata providers, used Python / scikit-learn to detect inaccurate geo-datapoints from the new data provider. This reduced by 10x the median error for datapoints.
- Wrote monitors in Python that both tracked and plotted trends in data, and sent alerts when anomalies were detected. Wrote Dockerfiles for easy deployment to Kubernetes.

Student Programmer, University of Chicago STEM Education (Feb. 2018 - June 2018)
- Developed projects in Scratch to teach students (grades K-8) math and CS fundamentals.

4 Teaching Experience

Guest Lecturer
- **Politecnico di Torino** (May 2024, Explainable and Trustworthy AI): *Intro to Mechanistic Interpretability*
- **Technion** (March 2024, NLP): *Introduction to Interpretability*

Teaching Assistant (TA), Institute for Logic, Language, and Computation, University of Amsterdam
- **Higher Cognitive Functions** (2023)
  - Crafted and assessed written assignments for a master's-level cognitive science course.
- **Interpretability and Explainability in AI** (2023, 2024)
  - Designed, taught, and assessed a week-long master's-level workshop on mechanistic interpretability, including interactive materials (Jupyter Notebooks).
  - Advised student projects in mechanistic interpretability.

Board Member, Board Manager (2019), Splash! Chicago (Sept. 2016 – Jun. 2020)
- Led Splash! Chicago, a volunteer student group organizing large (100-student) educational events where high school students can learn from university students. Taught linguistics classes for Splash! Chicago.

Grader, University of Chicago, Department of Computer Science (Fall 2018 – Summer 2020)
- Graded student projects, provided feedback regarding errors and areas to improve. Courses graded include Intro to CS, Intro to Comp. Systems, Comp. Architecture, Time Series Analysis and Stochastic Processes.

5 Invited Talks:

University of Copenhagen CoAStaL NLP (July 2024): *Introduction to circuits*
UT Austin Computational Linguistics (November 2023): *A circuit for greater-than in GPT-2*
DeepMind Language Model Interpretability Team (November 2023): *A circuit for greater-than in GPT-2*
Technion NLP Laboratory (March 2023): *Mechanistic interpretability: circuits and circuit-finding*
University of Amsterdam (ILLC) NLPitch (October 2022): *The functional relevance of probed information*

6 Academic Service

Tutorial on *Transformer-Specific Interpretability*. Hossein Mohebbi, Jaap Jumelet, **Michael Hanna**, Afra Alishahi, and Jelle Zuidema. See the tutorial proposal, materials, and recording. (**EACL 2024**)

Reviewing
- ACL ARR: February, April, June, August 2024
- ICML2024 Workshops: Mechanistic Interpretability Workshop, LLMs and Cognition

- NeurIPS2024 Workshops: Interpretable AI, Behavioral ML
- BlackBoxNLP 2023, 2024
- CLiC-It 2024

7 SCHOLARSHIPS & HONORS:

OpenAI Superalignment Fellowship (2024): fellowship providing a stipend and research funding for work on mechanistic interpretability of large language models
Alvise Comel Master's Thesis Prize (2023): prize for the top 2 master's theses on AI and cognitive neuroscience at the University of Trento's Center for Mind and Brain Sciences
European Laboratory for Intelligent Systems (ELLIS) PhD (2022-present): a selective PhD meta-program supporting co-supervision and research visits throughout Europe
LCT Scholarship (2020-2022): scholarship funding 2 years of master's study of computational linguistics
Enrico Fermi Scholar (2020): top 5% of undergraduate major (computer science)
Georgiana Simpson Scholar (2020): top 5% of undergraduate major (linguistics)
Summa Cum Laude (2020)
Phi Beta Kappa (2019-2020): academic achievement honors fraternity (top ~5% of undergraduate class)