# 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 ~Sept. 2026)

**Erasmus Mundus Language and Communication Technologies (LCT) dual-degree master's program**[1]
**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
Thesis: *Investigating Large Language Models' Representations Of Plurality Through Probing Interventions*

**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:

**Michael Hanna,** Ollie Liu, and Alexandre Variengien. 2023. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. To appear 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. To appear 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. To appear in 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. **(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*

[1] 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.
*Equal contribution

*Computational Linguistics*, pages 835–848, 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*, pages 5597–5612, 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:

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

- Learned mechanistic interpretability techniques as part of the REMIX program.
- Used causal interventions to study low-level mechanisms underlying GPT-2’s behavior on linguistic tasks.

**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.
- Conducted empirical experiments comparing BERT’s hypernym discovery performance to existing systems’.

**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

**Teaching Assistant (TA)**, Institute for Logic, Language, and Computation, University of Amsterdam

- **Higher Cognitive Functions**
  - Crafted and assessed written assignments for a master’s-level cognitive science course.
- **Interpretability and Explainability in AI**

  - 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 Skills:

- **Programming and Markup Languages**: Python, C, LaTeX, Elm, Scratch
- **Human Languages**: English (native), Spanish (fluent)
- **Machine Learning Frameworks**: PyTorch, Tensorflow 2.0, scikit-learn

## 6 Scholarships & Honors:

- **European Laboratory for Intelligent Systems (ELLIS) PhD**: a selective PhD meta-program supporting co-supervision and research visits throughout Europe
- **LCT Scholarship**: scholarship for 2 years of funded master's study of computational linguistics
- **Enrico Fermi Scholar:** top 5% of undergraduate major (computer science)
- **Georgiana Simpson Scholar:** top 5% of undergraduate major (linguistics)
- **Phi Beta Kappa**: academic achievement honors fraternity (top ~5% of overall undergraduate class)
- **Summa Cum Laude (Undergraduate)**