Melanie Sclar

PhD Candidate, University of Washington | msclar@cs.washington.edu |

Education

University of Washington Sept 2021–present
PhD in Computer Science
Advisors: Yejin Choi, Yulia Tsvetkov
Dissertation: *Structure-Guided Approaches for Robust Language Model Reasoning. GPA: 3.98/4.0.*

Universidad de Buenos Aires 2017
Licenciatura en Ciencias de la Computación (6-year integrated Master's and Bachelor's degree)
Dissertation: *Analysis and Prediction of Human Visual Search. GPA: 9.42/10 (equiv. 3.96/4.0).*

Work Experience

Visiting Researcher Sept 2023–Sept 2025
Meta FAIR (Fundamental AI Research), Superintelligence Labs
Developed robust evaluation and synthetic data generation methods for theory of mind in LLMs (ICLR 2025), interpretable personalized preference models (COLM 2025 Oral), and post-training for theory of mind in LLMs.

Research Intern Apr–Aug 2021
Carnegie Mellon University | Advisors: Yonatan Bisk, Graham Neubig
Designed first fully-symmetric multi-agent environment requiring theory of mind (ToM). Showed multi-agent RL models benefit from explicit ToM modeling but still achieve <50% of simple heuristics' performance (ICML'22).

Lead Machine Learning Engineer Nov 2019–Apr 2021
ASAPP
Built context-aware message suggestion system to speed up customer support interactions for Fortune 500 companies. Designed extensible architecture combining neural encoders with tree-based ranking, allowing rapid addition of new customer-vetted responses without model retraining (+15% CTR in the core company product).

Lead Machine Learning Engineer Jan 2018–Oct 2019
BrightSector Algorithms
Led 5-person team researching and deploying NLP algorithms at scale for Mercado Libre (Latin America's largest e-commerce platform). Developed named entity recognition models for product attribute extraction, classification systems for item categorization, and clustering methods to consolidate 200M+ listings.

Software Engineering Intern Jan–Apr 2016, Jan–Apr 2015
Facebook Inc.
(2016) Researched adaptive video stabilization to optimize compute while preserving video quality (2 patents).
(2015) Built recommendation algorithms for Feed content using sentiment analysis and trending topic detection.

Selected Honors & Awards

Mathematics Olympiads

National Champion, Argentine Mathematical Olympiad (2011)
Bronze Medal, Ibero-American Mathematical Olympiad (2011)
Silver Medal, South American Mathematical Olympiad (2010)
Second place, Inter-University Argentine Mathematical Competition (CIMA) (2016)
Bronze Medal, Ibero-American Youth Mathematics Competition (2008)

Computer Science and Machine Learning Olympiads

Latin America Champion, ICPC World Finals (2015) – International Collegiate Programming Competition
Champion, South America South Regional ICPC (2013, 2014); 3rd place (2012)
National Champion, Argentine ICPC Competition (2014); 2nd place (2012); 3rd place (2013)
First (2017) and second place (2018), Universidad de Buenos Aires Computing School ML nationwide competition

Selected Publications **denotes equal contribution*

Robustness, Model Behavior Quantification and Understanding

Sclar, M., Choi, Y., Tsvetkov, Y., Suhr, A. (2024). Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. **ICLR 2024 (500+ cits)**.

Sclar, M.*, Dziri, N.*, Lu, X.*, et al. (2023). Faith and Fate: Limits of Transformers on Compositionality. **NeurIPS 2023, Spotlight (600+ citations)**.

Lu, X., **Sclar, M.**, [...], Choi, Y. (2025). AI as Humanity's Salieri: Quantifying Linguistic Creativity of Language Models via Systematic Attribution of Machine Text against Web Text. **ICLR 2025, Oral**.

Lin, B. Y., Ravichander, A., Lu, X., Dziri, N., **Sclar, M.**, [...], Choi, Y. (2024). The unlocking spell on base llms: Rethinking alignment via in-context learning. ICLR 2024.

Theory of Mind Reasoning, Personalization, Modeling Human Behavior

Sclar, M., Kumar, S., West, P., Suhr, A., Choi, Y., Tsvetkov, Y. (2023). Minding Language Models' (Lack of) Theory of Mind: A Plug-and-Play Multi-Character Belief Tracker. **ACL 2023, Outstanding Paper Award**.

Li, SS., **Sclar, M.**, et al., PrefPalette: Personalized Preference Modeling with Latent Attributes. (2025). **COLM 2025, Oral**.

Sclar, M., Yu, J., Fazel-Zarandi, M., Tsvetkov, Y., Bisk, Y., Choi, Y., Celikyilmaz, A. (2025). Explore Theory of Mind: Program-guided adversarial data generation for theory of mind reasoning. ICLR 2025.

Sclar, M., Neubig, G., Bisk, Y. (2022). Symmetric Machine Theory of Mind. ICML 2022.

Sclar, M.*, Bujia, G.*, Vita, S., Solovey, G., Kamienkowski, J.E. (2020). Modeling human visual search: A combined bayesian searcher and saliency map approach for eye movement guidance in natural scenes. SVRHM NeurIPS Workshop 2020, Oral presentation; **NVIDIA Diversity in AI Best Paper Award**.

Patents

Wolf, W.A., **Sclar, M.**, et al. (2024). Processing clusters with mathematical models for message suggestion. U.S. Patent 11,985,102.
Sclar, M., et al. (2020). Neural network to optimize video stabilization parameters. U.S. Patent 10,582,211.
Sclar, M., et al. (2019). Foreground detection for video stabilization. U.S. Patent 10,506,248.

Selected Service, Outreach & Teaching

Teaching: Head Teaching Assistant, University of Washington (two quarters); Teaching Assistant, Universidad de Buenos Aires (9 semesters; 2014–2017, 2019); Mathematics Olympiad Coach (2012–2016), plus others below.

Organization: Bridging Language, Agent, and World Models for Reasoning and Planning Workshop (NeurIPS 2025); South America Topcoder Open Regionals (2019); 20+ programming and mathematics competitions (2012–2018); Organizer & instructor at competitive programming training camps (2014–2018; Brazil & Argentina)

Outreach: Weekly programming instructor for underrepresented students in partnership with National University of San Martin, Argentina (2016–2018); Invited speaker at FemIT 2020, conference for Latin American women in tech (1500+ attendees); University of Buenos Aires' annual Computer Science Week Invited Speaker (2017–2020)

Jury Member: Argentine Informatics Olympiad (2018–present); Argentine Mathematics Olympiad (2020)

Invited Speaker: Visions of Language Modeling Workshop @ COLM (2025); Oracle ML (2024), CMU (2024), Universidad Torcuato di Tella (2024), among others.

Skills

Programming: Python (advanced); C++, Java, MATLAB, SQL (intermediate); R, Bash, PHP/Hack (basic)
Languages: Spanish (native), English (fluent, TOEFL 119/120), Portuguese (advanced), French (intermediate)