

Arkadiy Saakyan

a.saakyan@cs.columbia.edu asaakyan.github.io GitHub Google Scholar

EDUCATION

Columbia University <i>Ph.D. in Computer Science (AI/NLP)</i> Advisor: Prof. Smaranda Muresan Committee: Kathleen McKeown, Julia Hirschberg, Diyi Yang, Chris Callison-Burch	2022 – 2027
Columbia University <i>M.S. in Computer Science (en route to Ph.D.)</i>	2022 – 2023
Columbia University, Columbia College <i>B.A. in Computer Science, Concentration in Mathematics</i>	2017 – 2021

EXPERIENCE

Google DeepMind <i>Student Researcher, New York, NY</i> Cultural values in AI safety. Hosts: Charvi Rastogi, Lora Aroyo.	Jun 2025 – Sep 2025
Google DeepMind <i>Student Researcher, Mountain View, CA</i> Evaluation of AI critic models. Hosts: Najoung Kim, Deepak Ramachandran.	May 2024 – Aug 2024
Amazon Web Services, AI Labs <i>Applied Scientist Intern, New York, NY</i> System prompt following. Hosts: Giovanni Paolini, Yassine Benajiba.	May 2023 – Aug 2023
UBS, Data Analytics Lab <i>Data Scientist, New York, NY</i>	Jul 2021 – Aug 2022

PUBLICATIONS

* Equal contribution.

- 1 **Arkadiy Saakyan**, Charvi Rastogi, Lora Aroyo. *Quantifying the Saliency of Geo-Cultural Values for Pluralistic Safety Alignment*. ICML 2026.
- 2 **Arkadiy Saakyan**, Najoung Kim, Smaranda Muresan, Tuhin Chakrabarty. *Death of the Novel(ty): Beyond n-Gram Novelty as a Metric for Textual Creativity*. ICLR 2026.
- 3 Nikhil Reddy Varimalla, Yunfei Xu, Meng Fan Wang, **Arkadiy Saakyan**, Smaranda Muresan. *VideoNorms: Benchmarking Cultural Awareness of Video Language Models*. Findings of AACL 2027, IC2S2 2026; C3NLP @ ACL 2026.
- 4 **Arkadiy Saakyan**, Shreyas Kulkarni, Tuhin Chakrabarty, Smaranda Muresan. *Understanding Figurative Meaning through Explainable Visual Entailment*. NAACL 2025.
- 5 **Arkadiy Saakyan**, Smaranda Muresan. *ICLEF: In-Context Learning with Expert Feedback for Explainable Style Transfer*. ACL 2024.
- 6 Sky CH-Wang*, **Arkadiy Saakyan***, Oliver Li, Zhou Yu, Smaranda Muresan. *Sociocultural Norm Similarities and Differences via Situational Alignment and Explainable Textual Entailment*. EMNLP 2023.

- 7 Oliver Li, Mallika Subramanian, **Arkadiy Saakyan**, Sky CH-Wang, Smaranda Muresan. *NormDial: A Comparable Bilingual Synthetic Dialog Dataset for Modeling Social Norm Adherence and Violation*. EMNLP 2023.
- 8 Tuhin Chakrabarty, Kanishk Singh, **Arkadiy Saakyan**, Smaranda Muresan. *Learning to Follow Object-Centric Image Editing Instructions Faithfully*. Findings of EMNLP 2023.
- 9 Tuhin Chakrabarty*, **Arkadiy Saakyan***, Olivia Winn*, Artemis Panagopoulou, Yue Yang, Marianna Apidianaki, Smaranda Muresan. *I Spy a Metaphor: Large Language Models and Diffusion Models Co-Create Visual Metaphors*. Findings of ACL 2023.
- 10 Tuhin Chakrabarty, **Arkadiy Saakyan**, Debanjan Ghosh, Smaranda Muresan. *FLUTE: Figurative Language Understanding through Textual Explanations*. EMNLP 2022.
- 11 Tuhin Chakrabarty, **Arkadiy Saakyan**, Smaranda Muresan. *Don't Go Far Off: An Empirical Study on Neural Poetry Translation*. EMNLP 2021.
- 12 **Arkadiy Saakyan**, Tuhin Chakrabarty, Smaranda Muresan. *COVID-Fact: Fact Extraction and Verification of Real-World Claims about COVID-19 Pandemic*. ACL 2021.

TEACHING

LLMs: Foundations and Ethics (BC3997) <i>Teaching Assistant, Columbia University</i> Instructor: Prof. Smaranda Muresan	Spring 2025
Natural Language Processing (BC3997) <i>Teaching Assistant, Columbia University</i> Instructor: Prof. Smaranda Muresan	Fall 2024
Neural Networks and Deep Learning (COMS W4995) <i>Teaching Assistant, Columbia University</i> Instructor: Prof. Richard Zemel	Fall 2023

MENTORSHIP

Sofia Hirao Barnard, B.A.	2026 – present
Cecilia Yang Columbia, M.S.	2026 – present
Nikhil Varimalla Columbia, M.S. · 1 paper	2024 – 2026
Lorraine (Yunfei) Xu Columbia, B.A. · 1 paper	2024 – 2026
Meng Fan Wang Columbia, B.A. · 1 paper	2024 – 2026
Shreyas Kulkarni Columbia, M.S. · 1 paper	2023 – 2024
Kanishk Singh Columbia, M.S. · 1 paper	2023 – 2024
Oliver Li Columbia, B.A. · 2 papers	2022 – 2023
Mallika Subramanian Columbia, M.S. · 1 paper	2022 – 2023

HONORS & AWARDS

Google Cloud Research Credits (\$10,000) For a research project on cultural norm understanding in videos.	Spring 2025
Columbia Data Science Institute Best Student Project For the EMNLP 2021 paper <i>Don't Go Far Off: An Empirical Study on Neural Poetry Translation</i> .	Spring 2022

USC Annenberg Fellowship (declined)	Fall 2022
ACL Diversity & Inclusion Award	Summer 2021
Columbia Undergraduate Scholars Program Summer Research Grant (\$3,500)	Summer 2020
John Jay National Scholarship	2017 – 2021
Financial aid and special programming for outstanding students.	

SERVICE

Reviewer	NeurIPS (2026), COLM (2026), ICML (2026), ICLR (2026), CHI (2026), ACL (2023, 2024), EMNLP (2023, 2024), NAACL (2023), EACL (2023, 2026)
Organizer	Shared Task on Understanding Figurative Language (FigLang 2022) Understanding of Figurative Language through Visual Entailment Shared Task (FigLang 2024)

SKILLS

Research	LLM evaluation, alignment & safety, multimodal (vision–language) models, interpretability, explainability, figurative language, computational social science
Programming	Python, PyTorch, Hugging Face (Transformers, Datasets), SQL, Git, $\LaTeX$
Platforms	Google Cloud, AWS, Slurm / GPU clusters

MEMBERSHIPS

Association for Computational Linguistics (ACL)	2021 – present
Queer in AI	2021 – present