SEYED ARSHAN DALILI

sbd5760@psu.edu arshandalili.github.io arshandalili +1 814 280 3343 State College, PA

Education

- **Pennsylvania State University** — Jan 2025 – Present
  Ph.D., Computer Science and Engineering | GPA: 3.5/4.0 — *State College, PA*
  Advisor: Prof. Mehrdad Mahdavi | Co-advisor: Prof. Wenpeng Yin
- **Sharif University of Technology** — Sep 2019 – Jul 2023
  B.Sc., Computer Engineering | GPA: 3.9/4.0 — *Tehran, Iran*

Publications & Preprints

- **Subspace-Aware Sparse Autoencoders for Mechanistic Interpretability (SASA)**
  S. A. Dalili*, M. Mahdavi.* Submitted to ICML 2026.
- **Model Merging via Multi-Teacher Knowledge Distillation**
  S. A. Dalili*, M. Mahdavi.* Submitted to ICML 2026. arXiv:2512.21288.
- **Imaginations of WALL-E: Reconstructing Experiences with an Imagination-Inspired Module**
  Z. S. Taghavi, S. Gooran, ***S. A. Dalili****, H. Amirzadeh, M. J. Nematbakhsh, H. Sameti.* arXiv:2308.10354.
- **SUT at SemEval-2023 Task 1: Prompt Generation for Visual Word Sense Disambiguation**
  O. Ghahroodi, ***S. A. Dalili****, S. Mesforoush, E. Asgari.* SemEval 2023 (ACL Anthology).
- **AIMA at SemEval-2024 Task 3: Simple Yet Powerful Emotion Cause Pair Analysis**
  A. Ghahramani Kure, M. Dehghani, M. M. Abootorabi, N. Ghazizadeh, ***S. A. Dalili****, E. Asgari.* SemEval 2024 (ACL Anthology).
- **AIMA at SemEval-2024 Task 10: History-Based Emotion Recognition in Hindi-English Code-Mixed Conversations**
  M. M. Abootorabi, N. Ghazizadeh, ***S. A. Dalili****, A. Ghahramani Kure, M. Dehghani, E. Asgari.* SemEval 2024 (ACL Anthology).

Research Experience

- **Graduate Research Assistant** — Jan 2025 – Present
  Advisor: Prof. Mehrdad Mahdavi — *Pennsylvania State University*
  - **Mechanistic interpretability & representation learning:** decomposing learned representations of LLMs (e.g., via sparse autoencoders) to interpret how they work.
  - **Federated learning & model merging:** personalization for heterogeneous clients; data-efficient merging of fine-tuned experts via distillation.
  - **Hallucination mitigation:** reducing factual errors through steering-based interventions at inference time.
  - **Cancer multi-omics:** discovering regulatory roles of gene expression in cancer via multi-omics integration (joint with Prof. Jeffrey Maurice Peters).
  - **Alzheimer's biomarkers:** extracting biomarkers and causal relationships using foundation models (joint with Prof. Wenpeng Yin).
- **Research Assistant** — Aug 2024 – Nov 2024
  Supervisor: Prof. Iryna Gurevych — *UKP Lab, TU Darmstadt*
  – Implemented Fisher-weighted averaging and uncertainty-based gradient matching; contributed code to MergeKit.

Teaching

- **Teaching Assistant (Penn State)** CMPSC 448: Machine Learning and Algorithmic AI, Penn State (Spring 2025). Instructor: Prof. Mehrdad Mahdavi.
- **Teaching Assistant (Sharif University of Technology)** 2020–2023: Artificial Intelligence (×3), Modern Information Retrieval, Computer Architecture, Advanced Programming, Circuits.

Work Experience

- **NLP Intern, ASR Gooyesh Pardaz** — May 2023 – Jul 2023
  Semantics & multimodal NLP; built large-scale IR models for Persian — *Tehran, Iran*
- **Tutor, High School Mathematics and Physics** — Jun 2020 – Feb 2021
  Private tutoring — *Sari, Iran*

Academic Service

- **Reviewer** KDD 2025 Workshop: SciSocLLM.
- **Committee Member** Mid-Atlantic Student Colloquium on Speech, Language and Learning (MASC-SLL 2025).

Skills

Python, PyTorch, Transformers, NumPy, Pandas, NetworkX, scikit-learn; CUDA, SQL, Django, Bash scripting; Git, Docker; LaTeX; Linux.