Swarnadeep Saha

Research Scientist, FAIR at Meta, Seattle, USA

https://swarnahub.github.io
https://github.com/swarnaHub
swarnadeep@meta.com

RESEARCH INTERESTS

Large Language Models, Reasoning, Planning, Alignment

RESEARCH AND WORK EXPERIENCE

FAIR at Meta (Alignment Team) — **Seattle, USA**
Research Scientist, Manager: Dr. Jason Weston — *August 2024-Present*

- Fundamental research to improve **Reasoning, Memory, and Alignment** of **Large Language Models**.

FAIR Labs at Meta — **Palo Alto, USA**
Research Intern, Mentors: Dr. Xian Li and Dr. Jason Weston — *May 2023-Dec 2023*

- Fundamental research on improving **Large Language Model Evaluation**.
- Paper accepted to **NAACL 2024**.

FAIR Labs at Meta — **Seattle, USA**
Research Intern, Mentor: Dr. Asli Celikyilmaz — *May 2022-Dec 2022*

- Fundamental research on **Multi-step Reasoning** for **Text generation from semi-structured data**.
- Paper accepted to **Findings of ACL 2023**.

Salesforce AI Research — **Palo Alto, USA**
Research Intern, Mentors: Dr. Nazneen Rajani and Dr. Jesse Vig — *June 2021-August 2021*

- Fundamental research on connecting **Interpretability** to **Sample hardness**.
- Paper accepted as oral to **EMNLP 2022**.

IBM Research — **Bangalore, India**
Research Engineer, Manager: Dr. Shantanu Godbole — *July 2017 - June 2019*

- Developed large scale **Machine Learning** solutions for **Intelligent Tutoring Systems (Watson Tutor)**, notably in the areas of **Automatic Short Answer Grading** and **Text Segmentation**.
- **Lab-wide Research Appreciation** award and twice **Manager's Choice** award.

Adobe Systems — **Noida, India**
Member of Technical Staff, Manager: Rajeev Sharma — *June 2014 - July 2015*

- Worked as a full-stack software developer in the **Acrobat Reader Team** of Adobe.

EDUCATION

UNC Chapel Hill — **North Carolina, USA**
Ph.D. in Computer Science, Advisor: Prof. Mohit Bansal — *2019 - 2024*

Thesis: *Multi-Step Reasoning over Natural Language*
Google PhD Fellowship in NLP for 2023 and 2024

Indian Institute of Technology, Delhi **Delhi, India**
M.Tech. in Computer Science, Advisor: Prof. Mausam *2015 - 2017*
Thesis: *Open Information Extraction from Numerical and Conjunctive Sentences*
Best M.Tech Thesis Award in Computer Science

Jadavpur University **Kolkata, India**
B.E. in Computer Science, GPA: 8.72/10.0 *2010 - 2014*

PUBLICATIONS

1. **Learning to Plan & Reason for Evaluation with Thinking-LLM-as-a-Judge**
Swarnadeep Saha, Xian Li, Marjan Ghazvininejad, Jason Weston, Tianlu Wang
Under Review at ICML 2025[pdf]

2. **System-1.x: Learning to Balance Fast and Slow Planning with Language Models**
Swarnadeep Saha, Archiki Prasad, Justin Chih-Yao Chen, Peter Hase, Elias Stengel-Eskin, Mohit Bansal
ICLR 2025[pdf]

3. **MAgICoRe: Multi-Agent, Iterative, Coarse-to-Fine Refinement for Reasoning**
Justin Chih-Yao Chen, Archiki Prasad, **Swarnadeep Saha**, Elias Stengel-Eskin, Mohit Bansal
Under Review[pdf]

4. **MAGDi: Structured Distillation of Multi-Agent Interaction Graphs Improves Reasoning in Smaller Language Models**
Justin Chih-Yao Chen*, **Swarnadeep Saha***, Elias Stengel-Eskin, Mohit Bansal
ICML 2024[pdf]

5. **Branch-Solve-Merge Improves Large Language Model Evaluation and Generation**
Swarnadeep Saha, Omer Levy, Asli Celikyilmaz, Mohit Bansal, Jason Weston, and Xian Li
NAACL 2024[pdf]

6. **ReConcile: Round-Table Conference Improves Reasoning via Consensus among Diverse LLMs**
Justin Chih-Yao Chen, **Swarnadeep Saha**, and Mohit Bansal
ACL 2024[pdf]

7. **Can Language Models Teach Weaker Agents? Teacher Explanations Improve Students via Personalization**
Swarnadeep Saha, Peter Hase, and Mohit Bansal
NeurIPS 2023[pdf]

8. **ReCEval: Evaluating Reasoning Chains via Correctness and Informativeness**
Archiki Prasad, **Swarnadeep Saha**, Xiang Zhou, and Mohit Bansal
EMNLP 2023[pdf]

9. **MURMUR: Modular Multi-Step Reasoning for Semi-Structured Data-to-Text Generation**
Swarnadeep Saha, Xinyan Velocity Yu, Mohit Bansal, Ramakanth Pasunuru, and Asli Celikyilmaz
ACL Findings 2023[pdf]

10. **Summarization Programs: Interpretable Abstractive Summarization with Neural Modular Trees**
Swarnadeep Saha, Shiyue Zhang, Peter Hase, and Mohit Bansal
ICLR 2023[pdf]

11. **Are Hard Examples also Harder to Explain? A Study with Human and Model-Generated Explanations**
    Swarnadeep Saha, Peter Hase, Nazneen Rajani, and Mohit Bansal
    EMNLP 2022[pdf]

12. **Explanation Graph Generation via Pre-trained Language Models: An Empirical Study with Contrastive Learning**
    Swarnadeep Saha, Prateek Yadav, and Mohit Bansal
    ACL 2022[pdf]

13. **ExplaGraphs: An Explanation Graph Generation Task for Structured Commonsense Reasoning**
    Swarnadeep Saha, Prateek Yadav, Lisa Bauer, and Mohit Bansal
    EMNLP 2021[pdf]

14. **multiPRover: Generating a Set of Proofs for Improved Interpretability in Rule Reasoning**
    Swarnadeep Saha, Prateek Yadav, and Mohit Bansal
    NAACL 2021[pdf]

15. **PRover: Proof Generation for Interpretable Reasoning over Rules**
    Swarnadeep Saha, Sayan Ghosh, Shashank Srivastava, and Mohit Bansal
    EMNLP 2020[pdf]

16. **ConjNLI: Natural Language Inference over Conjunctive Sentences**
    Swarnadeep Saha, Yixin Nie, and Mohit Bansal
    EMNLP 2020[pdf]

17. **Pre-Training BERT on Domain Resources for Short Answer Grading**
    Chul Sung, Tejas Dhamecha, **Swarnadeep Saha**, Tengfei Ma, Vinay Reddy, and Rishi Arora
    EMNLP 2019[pdf]

18. **Aligning Learning Objectives to Learning Resources: A Lexico-Semantic Spatial Approach**
    Swarnadeep Saha, Malolan Chetlur, Tejas I. Dhamecha, Shantanu Godbole and others
    IJCAI 2019[pdf]

19. **Creating Scoring Rubric from Representative Student Answers for Improved Short Answer Grading**
    Smit Marvaniya, **Swarnadeep Saha**, Tejas I. Dhamecha, Peter Foltz, Renuka Sindhgatta and Bikram Sengupta
    CIKM 2018[pdf]

20. **Joint Multi-Domain Learning for Automatic Short Answer Grading**
    Swarnadeep Saha, Tejas I. Dhamecha, Smit Marvaniya, Peter Foltz, Renuka Sindhgatta and Bikram Sengupta
    arXiv 1902.09183[pdf]

21. **Open Information Extraction from Conjunctive Sentences**
    Swarnadeep Saha and Mausam
    COLING 2018[pdf]

22. **Balancing Human Efforts and Performance of Student Response Analyzer in Dialog-based Tutors**
    Tejas I. Dhamecha, Smit Marvaniya, **Swarnadeep Saha**, Renuka Sindhgatta and Bikram Sengupta
    AIED 2018[pdf]

23. **Sentence Level or Token Level Features for Automatic Short Answer Grading?: Use Both**
Swarnadeep Saha, Tejas I. Dhamecha, Smit Marvaniya, Renuka Sindhgatta and Bikram Sengupta
AIED 2018[pdf]

24. **Bootstrapping for Numerical Open IE**
Swarnadeep Saha, Harinder Pal and Mausam
ACL 2017[pdf]

ACHIEVEMENTS AND AWARDS

- **Google PhD Fellowship** (one of eight students worldwide) in NLP with full funding for 2 years.
- **Munroe and Rebecca Cobey Fellowship** at UNC Chapel Hill.
- **Best M.Tech. Thesis** in Computer Science at IIT Delhi.
- Lab-wide **Research Appreciation Award** at IBM Research.
- Twice **Manager's Choice Award** at IBM Research.
- **All India Rank of 142** in Graduate Aptitude Test in Engineering (GATE), 2014.
- **State Rank of 96** in West Bengal Joint Entrance Examination (WBJEE), 2010.

PROFESSIONAL SERVICE

- **Area Chair:** EMNLP 2024
- **Conference Reviewer:** NAACL 2024, ACL 2023, EMNLP 2023, NeurIPS 2023, ARR 2022, EMNLP 2022, ARR 2021, EMNLP 2021, NAACL 2021, AAAI 2020, AIED 2019, NAACL 2019, EMNLP 2018.
- **Journal Reviewer:** AI Journal (AIJ), Computational Linguistics (CL).

REFERENCES

- **Dr. Jason Weston**, Senior Director/Research Scientist, Fundamental AI Research (FAIR), Meta.
- **Dr. Mohit Bansal**, John R. & Louise S. Parker Associate Professor of CS, UNC Chapel Hill.
- **Dr. Mausam**, Professor, Jai Gupta Chair of CSE and Founding Head of School of AI, IIT Delhi and Affiliate Professor of CS, University of Washington, Seattle.