

# Shivam Shandilya

## Research Fellow, Microsoft Research

@ email - shivam.stpaulsdarjeeling@gmail.com    Github    Google Scholar




## Education

Aug 2019 May 2023	<b>Birla Institute of Technology, Mesra</b> Bachelor of Technology in Electrical and Electronics Engineering	GPA: 8.35/10
----------------------	---	--------------

## Experience

Jul 2023 Present	<b>Microsoft Research</b> <i>Research Fellow / Advisor: <a href="#">Dr. Menglin Xia</a></i> > <b>Prompt Compression for black-box LLMs:</b> Developed a low- latency framework to efficiently compress unstructured contexts; research currently under review for publication. > <b>Self-Assessing LLM framework with Context-Aware Criteria (SALC):</b> Designed and implemented a framework that dynamically integrates instance-specific knowledge to enhance long-form response evaluation; research currently under review for publication and patent. > <b>Context Optimization for Agent-Based Frameworks:</b> Currently exploring methods to optimize long contexts within agent-based systems and enhance end-to-end efficiency in multi-agent interactions.	<b>Bangalore, India</b>
Jan 2023 Jul 2023	<b>Microsoft Research</b> <i>Research Intern / Advisor: <a href="#">Dr. Jorge Tavares</a></i> > <b>NAS framework for efficient architecture search:</b> Developed a NAS framework that can efficiently derive lightweight task-specific heads for frozen pre-trained LMs that can be bootstrapped for a new task. Work under patent submission. > <b>Efficient model design:</b> Worked on creating smaller proxy models that maintain performance quality while being easier to deploy and fine-tune, thus reducing costs.	<b>Bangalore, India</b>
May 2022 Sept 2022	<b>Google Summer of Code</b> <i>Student contributor under Python Software Foundation / Mentor: <a href="#">Mariano Reingart</a></i> > <b>Contributed to PyZombis project:</b> PyZombis is a community course to teach Python to the Spanish-speaking community, based on a Brazilian MOOC. Worked on improving the course, with an interactive environment for the students, that allows them to visualize and try active code on the site as they learn. Contribution details: <a href="#">GSOC 2022</a>	<b>Remote</b>
May 2022 Jul 2022	<b>CoEAMT IIT Kharagpur</b> <i>Research Intern / Advisor: <a href="#">Dr. Surjya Kanta Pal</a></i> > <b>Object detection and tracking using Deep Learning methods:</b> My work as a research intern was to find a low-cost yet effective solution for object detection and collision free handling in industrial robots. Implemented a YOLOR (You Only Learn One Representation) based method that can quickly generalize to new and unseen objects.	<b>Kharagpur, India</b>

## Conference Publications

- [C.3] TACO-RL: Task Aware Prompt Compression Optimization with Reinforcement Learning   
Shivam Shandilya, Menglin Xia, Supriyo Ghosh, Huiqiang Jiang, Jue Zhang, Qianhui Wu, Victor Rühle  
[Under Review]
- [C.2] Unveiling Context-Aware Criteria in Self-Assessing LLMs   
Taneesh Gupta, Shivam Shandilya, Xuchao Zhang, Supriyo Ghosh, Chetan Bansal, Huaxiu Yao, Saravan Rajmohan  
[Under Review]
- [C.1] Streetwise Agents: Empowering Offline RL Policies to Outsmart Exogenous Stochastic Disturbances in RTC   
Aditya Soni, Mayukh Das, Anjaly Parayil, Supriyo Ghosh, Shivam Shandilya, Ching-An Cheng, Vishak Gopal, Sami Khairy, Gabriel Mittag, Yasaman Hosseinkashi, Chetan Bansal  
[Under Review]

## Selected Research Projects

---

### Task Aware Prompt Compression

Advisor: [Dr. Menglin Xia](#)

- › **TACO-RL:** A RL based framework that utilizes task-specific reward signals to fine-tune encoder based compressors. These encoder models need to be trained using labels only once, after which they can be adapted to other tasks in similar domains by using the target LLM's outputs as reward signals.
- › **Performance:** Improved task performance by at least **8% over SOTA** prompt compression baselines with better performance at harsher compression ratios.
- › **Efficiency Improvement:** Lightweight **encoder based compressors** allow for **low-latency compression** of prompts leading to cost effective end-to-end evaluation. [[Paper](#)]

### Framework for Self-Assessing LLMs

Advisors: [Dr. Xuchao Zhang](#), [Prof. Huaxiu Yao](#)

- › **SALC Framework:** LLM as a Judge like framework for Self-Assessing LLMs with Context-Aware Criteria, enabling dynamic, instance-specific evaluation criteria for adaptable, accurate assessments across diverse tasks.
- › **Performance:** Achieved an average of **4.8% improvement** over baseline evaluation frameworks and up to **12% increase** in LC WinRate on AlpacaEval-2 for DPO.
- › **Efficiency:** Fine-tuned smaller models via knowledge distillation, matching or surpassing larger models' performance with significantly lower computational costs. [[Paper](#)]

## Selected Honors and Awards

---

- › **Smart India Hackathon Finalist:** Selected for the final round of Smart India Hackathon (SIH) for a project under Indian Space Research Organization (ISRO).
- › **Placed 25th** in IBM India Qiskit Challenge, a Quantum Computing Hackathon held by IBM India. [Challenge Link](#)
- › Placed among the **top 10** teams of Eyantra Robotics Competitions in 2020 and 2021. [E-yantra](#)