Akshat Singh Jaswal

github.com/akshat-sj linkedin.com/in/akshat-sj sja.akshat@gmail.com +91 7975080168

Education

PES University Sep 2022 - Jun 2026
Bachelor of Technology in Computer Science

Experience

Lossfunk *ML Researcher* July 2025 – Present

- Conducted research on world models for long-horizon environment prediction, analyzing failure modes across representation learning, dynamics modeling, and rollout accuracy to improve generalization and sample efficiency.
- Investigated meta-RL formulations to discover variants of standard RL algorithms using evolutionary search, exploring population-based optimization to identify novel policies that outperform fixed-update baselines.
- Extended the evolutionary framework to jointly optimize reward functions, architectural choices, and training hyperparameters, enabling implicit improvements in agent learning dynamics without manual reward shaping.

Publications

It Takes Two: A Dual Stage Approach for Terminology-Aware Translation ACL Anthology
Akshat Singh Jaswal *WMT 2025 (co-located with EMNLP 2025)*

- Proposed DuTerm, a two-stage approach to terminology-aware machine translation combining NMT adaptation with LLM-based post-editing.
- Built a high-quality synthetic terminology corpus via LLM generation and COMET-QE filtering, and adapted NLLB-200 (3.3B) with parameter-efficient fine-tuning.
- Achieved SOTA chrF under proper terminology constraints (71.0 avg) with ∼99% term accuracy on the WMT 2025 shared task, outperforming constraint-based, ICL-driven, and preference-optimized systems.

AWE: Adaptive Agents for Dynamic Web Penetration Testing NDSS
Akshat Singh Jaswal, Ashish Baghel *NDSS LAST-X 2026*

- Designed AWE, a memory-augmented multi-agent framework for autonomous web penetration testing combining vulnerability-specific exploitation pipelines with LLM orchestration.
- Achieved 87% XSS and 66.7% blind SQLi success on the 104-challenge XBOW benchmark, outperforming MAPTA while using 98% fewer tokens.

Projects More on github.com/akshat-sj

FlashAttention | *CUDA, PyTorch* https://github.com/akshat-sj/flashattention

- Implemented kernels of Flash Attention in raw CUDA to accelerate the attention mechanism in transformers.
- Optimized memory access patterns for small workloads, reducing the execution time of the kernel from 15.842ms (PyTorch) to 6.172ms.
- Achieved performance improvements by leveraging techniques like coalesced memory access, loop unrolling, and optimized thread-block configurations, optimal shared memory utilization etc.

Bitmind | *Llama2, AMD XDNA, BM25* https://github.com/akshat-sj/bitmind

- Developed an AI gaming assistant utilizing AMD hardware and NPU acceleration to optimize LLaMA2 through quantization (AWQ), achieving 1.5 tokens/second processing speed and support for various game titles.
- Implemented a custom retrieval system using BM25, featuring advanced tokenization and dynamic passage construction for enhanced contextual accuracy in text generation and document scoring.
- Awarded an AMD mini PC ($1000) through the AMD Pervasive AI Contest for support of on-device fine-tuning and deployment of OPT/LLaMA-based game assistants.

KRQueue | *Python, Kafka, Redis* https://github.com/akshat-sj/KRQueue

- Engineered a distributed task queue system using Kafka and Redis, featuring fault tolerance, task reassignment, exponential backoff retries, and worker failure detection for robust task coordination in a highly distributed setup.
- Worked on adding advanced features like task chaining with self-balancing dependencies, task broadcasting, load balancing, distributed logging, and task grouping for scalable, efficient, and reliable task execution.