The Workshop on (Large and Small) Language Models on Agentic AI: Efficiency, Modularity, and Deployment (LSLMoAA'25)

1. Organizers

- Prof. Ah-Hwee TAN, Singapore Management University, Singapore
- Dr. Hongzhi Kuai, Chongqing University of Posts and Telecommunications,
  China
- Prof. Xiaohui Tao, University of Southern Queensland, Australia

2. Motivation and Goals

Agentic AI systems are increasingly pervasive—from autonomous assistants to task-oriented agents—yet most rely on large, generalist LMs that incur high latency, cost, and infrastructure demands. Recent work argues that small language models (SLMs) (≈<10 B parameters) offer sufficient task-specific capabilities, lower inference latency, and greater deployment flexibility for agentic subtasks . This half-day workshop seeks to:

1. Critically assess the technical, economic, and environmental trade-offs of SLM-first vs. LLM-centric agent architectures.
2. Showcase state-of-the-art SLMs and conversion techniques for migrating existing LLM-based agents to SLM-powered modules.
3. Foster collaboration between researchers and practitioners on benchmarks, tooling, and best practices for SLM deployment in agentic settings.

3. Scope and Topics

We invite contributions on topics including, but not limited to:

- LSLM Capabilities & Benchmarks: empirical evaluations of LSLMs for commonsense reasoning, tool-calling, and instruction following in agents
- Agentic Architectures: designs for heterogeneous agent systems that route tasks between SLM and LLM modules.
- Conversion Algorithms: pipelines for logging, clustering, fine-tuning, and deploying LSLM specialists in place of LLM calls.
- Economics & Sustainability: cost-model analyses, energy-efficiency studies, and environmental impact assessments of SLM-first deployments.
- On-Device & Edge Inference: frameworks and case studies demonstrating offline or on-device LSLM inference (e.g., consumer-grade GPUs, mobile).

- Tooling & Infrastructure: infrastructure for logging agentic interactions, automated data curation, and PEFT (LoRA/QLoRA) workflows.
- Agentic Robotics & Physical-World Integration: LSLM-powered control loops for drones, robots, and IoT actuators; Safety and real-time constraints in tangible agentic systems
- Digital Twins & Simulation Environments: Creating high-fidelity simulators for training and stress-testing LSLM agents; Transfer learning from virtual to physical deployments
- Collective Intelligence & Multi-Agent Coordination: Swarm-style cooperation among lightweight language agents; Consensus algorithms and negotiation protocols
- Resilience & Fault Tolerance: Designing graceful degradation when SLM modules fail or degrade; Self-healing pipelines and automated rollback strategies
- Cloud-Edge Continuum & Hybrid Orchestration: Dynamic offloading strategies between edge-hosted SLMs and cloud LLMs; Latency-aware scheduling and load balancing
- Ethical & Security Considerations: privacy implications of decentralized SLM inference, bias and democratization, and secure deployment.

4. Format and Schedule

Duration: Full-day (6 hours) co-located with WI-IAT'25 with the following structure:

- Opening (10 min): summarize outcomes and establish a community roadmap
- Keynote: "The Case for LSLM-First Agentic AI (tentative)" (60 min)
- 10 minutes break
- Paper Presentation Session I (20 min presentation + Q&A)
- 10 minutes break
- Paper Presentation Session II (3 x 20 min presentation + Q&A)
- 10 minutes break
- Paper Presentation Session III (3 x 20 min presentation + Q&A)
- 10 minutes break
- Panel Discussion (60 min): "Barriers and Opportunities for LSLM Development and Adoption"
- Closing & Roadmap (10 min): summarize outcomes and establish a community roadmap

5. Expected Audience Researchers and practitioners in:

- Natural Language Processing & Language Modeling
- Agent Architectures & Multi-agent Systems
- AI Infrastructure & Edge Computing

- Sustainable and Responsible AI

6. Submission and Review

Call for Papers: 6-pages
Important Dates:

- Abstract submission deadline: September 15, 2025
- Notification of acceptance: October 6, 2025
- Workshop Day: November 15, 2025

Submissions will be reviewed by the organizing committee for technical quality, relevance to SLM-agentic integration, and potential impact.

7. Outcomes and Follow-up

- Publication of WI-IAT'25 workshop proceedings
- Establishment of an online repository for SLM-agentic benchmarks and conversion scripts
- Formation of a working group to define standard metrics and best practices

By bringing together leading researchers and practitioners, this workshop will chart the path toward more efficient, modular, and sustainable agentic AI using small language models. We look forward to your participation!