2026 START Program
CFP Brief

THEME: **02.** **Agentic, Artificial Intelligence**

SUB-THEME: **2.1.** **Multimodal Reasoning Agents for Complex, Long-Horizon Tasks**

Context/ Overview

Recent foundational models have demonstrated remarkable multimodal understanding describing scenes, answering questions, generating code etc, yet they consistently fail when asked to act on that understanding over multiple steps. An agent can identify a misconfigured network setting in a screenshot but cannot reliably execute the six-step troubleshoot sequence to fix it. This gap between understanding and sustained, reliable action represents one of the open challenges in AI. Closing it requires a multi-modal reasoning agentic system that can integrate vision, audio, text, sensor telemetry, and environmental context not merely to comprehend but to plan, execute, adapt and recover across extended task horizons. These agents will form the foundation for truly autonomous digital companions and physical-world assistants capable of orchestrating complex workflows across applications, devices and environments.

Problem Statement

Despite significant progress in multimodal understanding, current-agent remain predominantly reactive and single-step in nature. They lack long-horizon planning, persistent state-tracking and adaptive reasoning capabilities required to solve complex, real-world tasks. Existing systems struggle to decompose high-level goals into coherent sub-task sequences, recover gracefully from errors mid-execution, or maintain context when transitioning between digital and physical domains. There is also a gap in how agents build and leverage the internal representation of their environment whether to predict future states, anticipate user needs or dynamically adapt their interaction. Moreover, majority of the current approaches lack tight integration between reasoning and action necessary for agents to dynamically revise plans based on new observation, learn and execute from failures, adapt the strategies in real time, leading to system that can reason but fail to act reliably.

Objectives & Scope

This call for proposals seeks to advance in multimodal reasoning agents anchored in a single unifying question: *how can multimodal reasoning be translated into reliable, multi-step action in complex environments*. The goal is to develop agent architectures that can perceive, reason, plan and act on digital/physical domains – handling complex task with robust performance, minimal latency and adaptive behavior.

Specific Topics & focus areas*

1. **Long-Horizon reasoning and Adaptive Planning**
   Developing strategies for hierarchical task decomposition, persistent state tracking and planning over extended time horizons including recovery strategies that enable agents to operate robustly in open-ended environments

2. **Multi-Modal Generative Mobile/UI World Models**
   Grounding diverse set of inputs into actionable representations of the environment, including approaches such as world models for prediction future states, generative UI that dynamically constructs, or adapts interfaces based on context and semantic representation that support downstream reasoning.

3. **Sustained multi-step execution in complex environments**
   Executing extended action sequences that maintain coherence over many steps, manage cascading dependencies between action and adapt to changing conditions

※ The topics are not limited to the above examples and the participants are encouraged to propose other original ideas.

END OF DOCUMENT