# 2025 START Program
# CFP Brief

THEME: **03.** **Artificial Intelligence**

SUB-THEME: **3.3.** **Multi-modal life-long memory augmentation system with fast retrieval**

## Context/ Overview

Personal knowledge base has evolved from traditional note-taking or diary to photos, videos, audio recordings etc. These information like our human brains can help us store important memories and moments. We human beings have evolved in the way to leverage the memory tools to help us study new skillsets, monitor our health and in our daily works. Maintaining the outside lifelong memory is of great value to everyone as when we grow old our human brain will gradually and easily forget things. However, as of today, the personal memory database is not well designed for fast retrieval. It takes hours or minutes for you to figure out what you did 1 years ago by checking the albums. A lot of important moments are not recorded. It is still not well organized to maintain all different sources of information organically together.

## Problem Statement

This call for proposal is to develop a personal life-long AI powered memory system aiming at enhancing our memory through technology. The targeting devices for information fetching including smart glasses, watches, rings, and phones etc. With various sensors in our wearable devices such as cameras and microphones in smart glass, GPS/health sensors in smart watches or Ring, information like location, activities, health conditions can be easily read. We are looking for solutions on this problem from two parts.

The first part of CFP is to improve the **context window length** for short-term memory. Traditional LLMs typically operate within a fixed-length context window, which restricts them to considering only a limited amount of preceding text when making predictions. Expanding the context window is a challenging task: the attention mechanism in transformers has a quadratic time complexity relative to the input sequence length; storing key-value (KV) pairs for long contexts demands substantial GPU memory which has become a bottleneck in processing extensive sequences; LLMs/LVMs may struggle to maintain performance over long contexts due to difficulties in capturing and utilizing distant dependencies effectively. There are already a lot of efforts trying to solve these problems such as efficient attention mechanisms, KV caching optimization etc. We are looking for new approaches and innovative ideas to further improve in this area.

The second part of this CFP is on how to improve the long-term memory including how to get useful context information from sensors on the devices we mentioned above, how to organize the information in some data base or knowledge graph with temporal and spatial information, how to fast retrieve the information for real-time Q&A with LLM etc. For example, the user could ask smart glass “where did you last see my wallets?”, “what was the name of person I met yesterday?” “Have I already taken the medicine?” etc., the information will be automatically fetched from memory system and immediately provide answers to our customer.

## Objectives & Scope

To solve the problem above, please consider the following selected topics and focus areas. Feel free to come up new ideas to solve the problem.

## Specific Topics & focus areas*

1. Develop efficient way for **life-long multi-modal** information storage system such as **spatial-temporal knowledge graph** or vector database for **personal** data. The solution could enable multi-modality information fetching with reduced storage size.

2. Develop LLM based fast retrieval algorithms to fetch the useful information out based on user query. Target is for **real-time** Q&A interaction from life-long database.

3. Develop **low-power** light weight model for video/audio perception on smart glasses or phones

4. Develop secure solution for personal multi-modal information storage

5. New areas and technology to enable personal memory augmentation system.

6. Efficient and low-complexity LLM fine-tune methods on personal data

7. Develop new attention mechanisms for long context window length

8. Develop new KV cache algorithms with reduce GPU memory resources

9. Architecture and hardware acceleration by fully leveraging the current GPU resources and instructions for long context window

10. New ideas to increase context window length and low complexity of LLM fine-tune.

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

END OF DOCUMENT