ICT80004 Weekly Communication – Week #01

Student Name:

Arun Ragavendhar Arunachalam Palaniyappan

ID: 104837257

Commonwealth Scientific and Industrial Research Organisation (CSIRO)

Organisation:

Industry Supervisor: Dr. Shigang Liu

Date Prepared:

08/08/2025

Internship Week #: 1

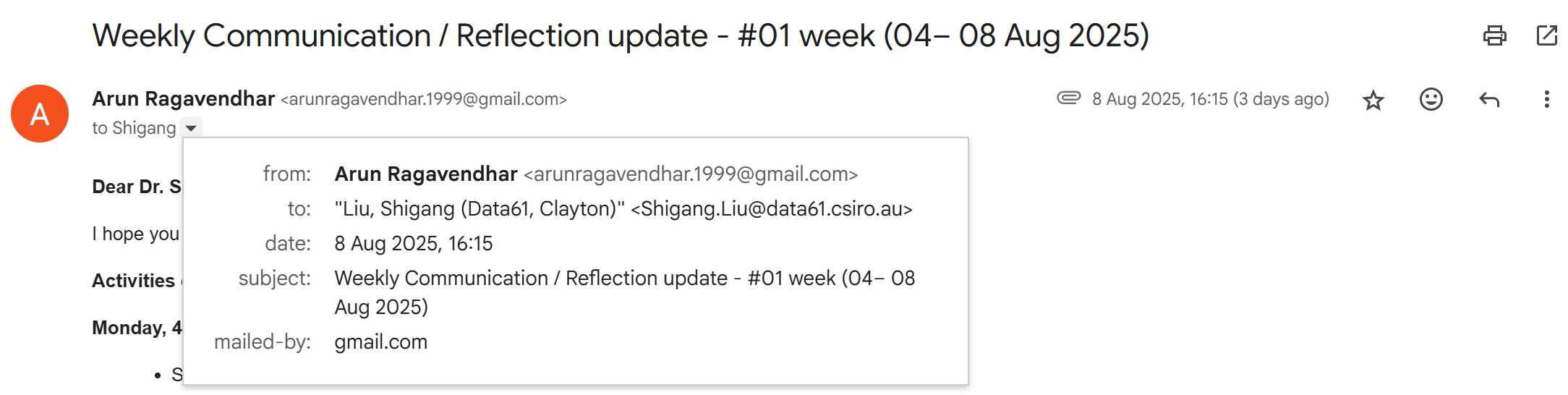
|  |  |  |
| --- | --- | --- |
| Day | Date | Task(s) completed |
| 1 | Monday, 4 Aug  2025 - 4 hours | * Studied and analysed two assigned research papers: Large Language Models and Simple, Stupid Bugs (SStuBs) * Chain-of-Thought Prompting Elicits Reasoning in Large Language Models |
| 1 | Monday, 4 Aug  2025 - 4 hours | * Prepared a PowerPoint presentation summarising key learnings from these papers, including examples of vulnerabilities and defence strategies. |
| 2 | Tuesday, 5 Aug  2025 - 4 hours | * Presented PPT via MS Teams and discussed feedback. Received five research papers for review on prompt injection and defence methods, including works on indirect attacks, LLM exploitation, structured queries, and vulnerability analysis. |
| 2 | Tuesday, 5 Aug  2025 - 4 hours | * Began reading and analysing the first of these papers (Benchmarking and Defending Against Indirect Prompt Injection Attacks on LLMs), noting methodology, results, and relevance to our project. |

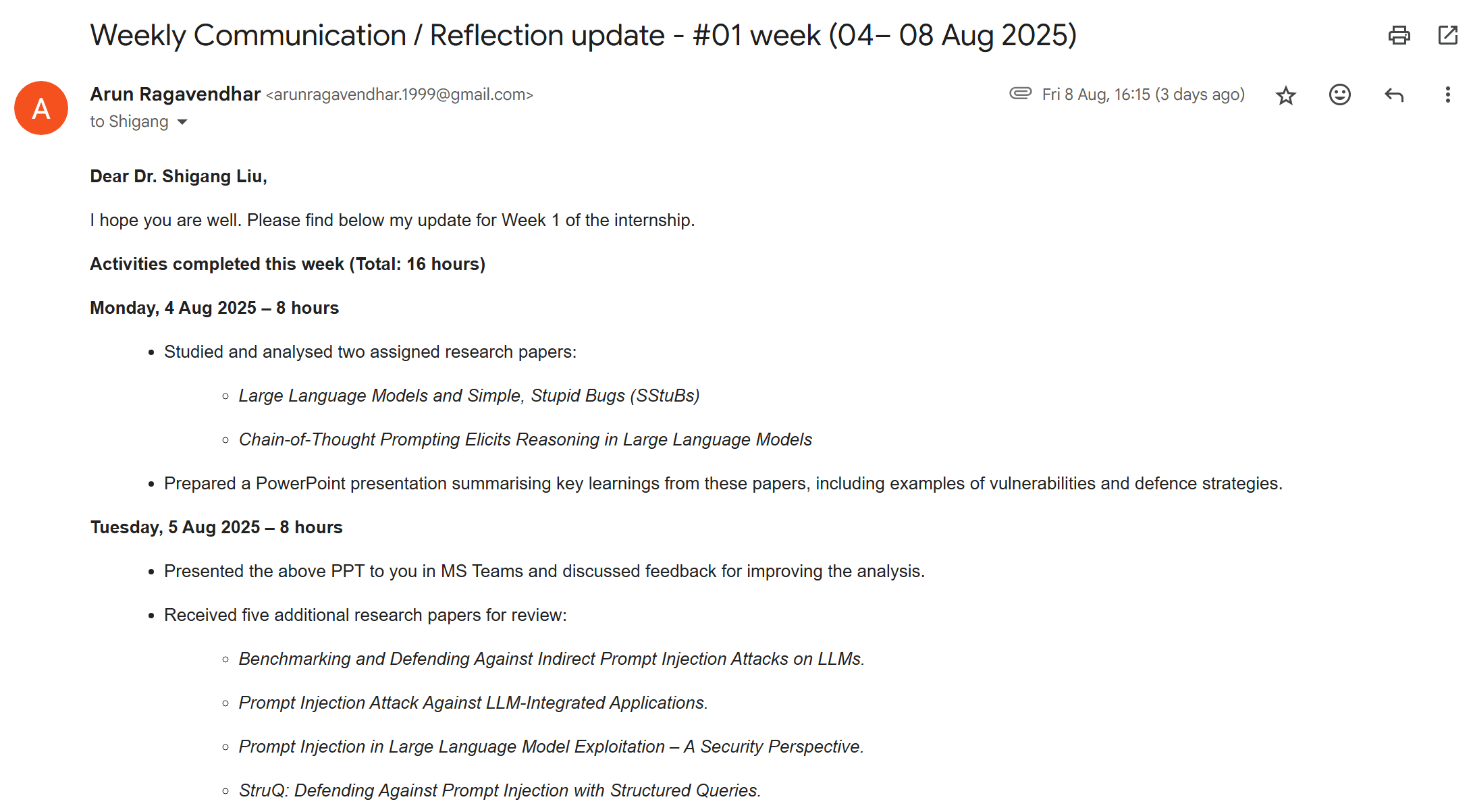
Total hours completed for the week: 16

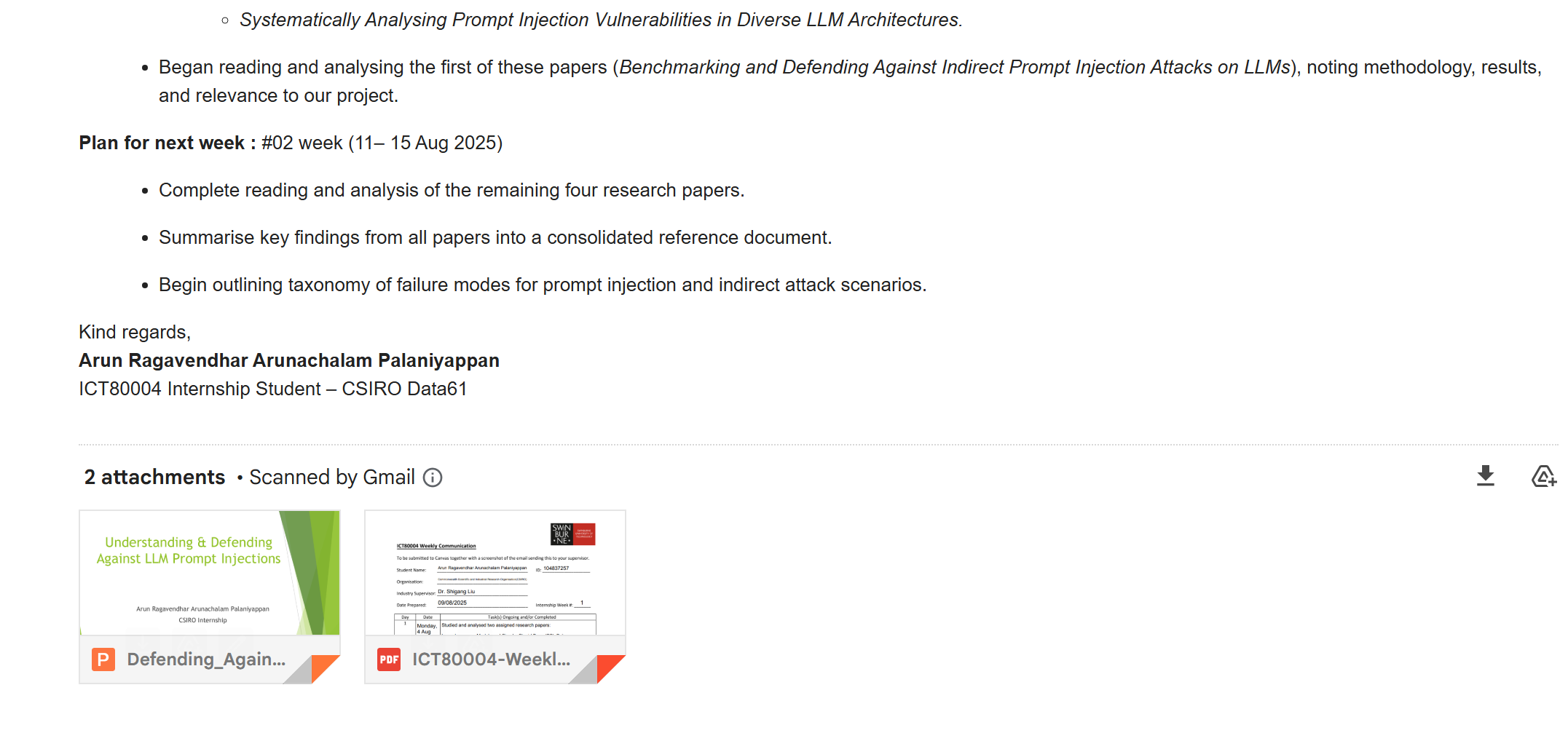
**Plans for next week: #02 week (11– 15 Aug 2025)**

* Complete reading and analysis of the remaining four research papers.
* Summarise key findings from all papers into a consolidated reference document.
* Begin outlining taxonomy of failure modes for prompt injection and indirect attack scenarios.

**Screenshot of Timely EMAIL communication update to the Supervisor at the end of week #01**

****

****

****