ICT80004 Weekly Communication – Week #03

Student Name: Arun Ragavendhar Arunachalam Palaniyappan ID: 104837257

Organisation: Commonwealth Scientific and Industrial Research Organisation (CSIRO)

Industry Supervisor: Dr. Shigang Liu

Date Prepared: 22/08/2025 Internship Week #: 3

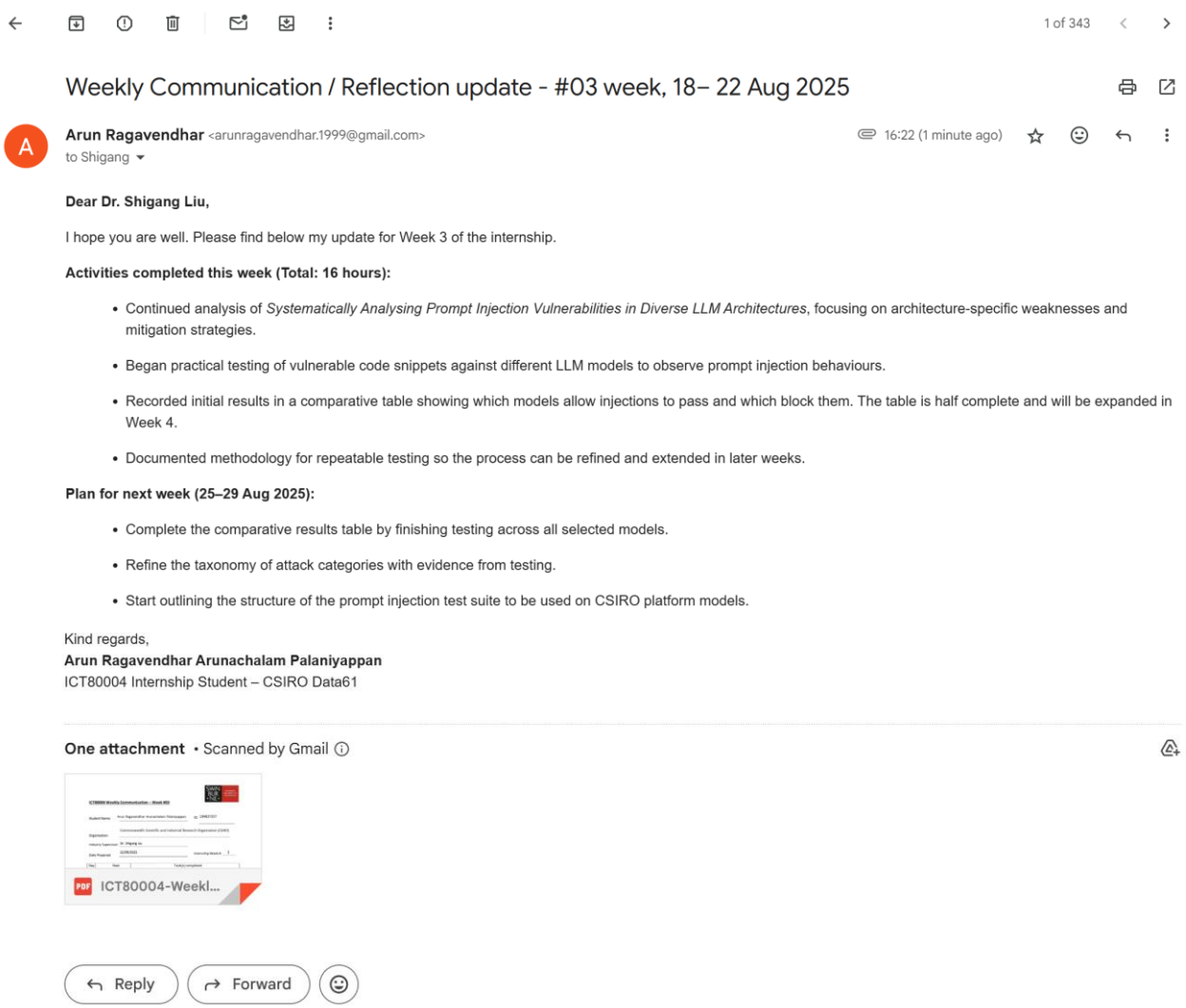
Day	Date	Task(s) completed
1	Monday 18 Aug 2025 8 hours	○ Continued reading and analysis of *Systematically Analysing Prompt Injection Vulnerabilities in Diverse LLM Architectures*. ○ Identified model-specific weaknesses and mitigation strategies. ○ Prepared notes linking this paper's findings to our planned test suite design.
2	Wednesday 20 Aug 2025 8 hours	○ Began practical testing of vulnerable code snippets across different LLM models. ○ Recorded partial results in a comparative table (half complete). ○ Documented testing methodology for reproducibility.

Total hours completed for the week: 16

Plans for next week: #02 week (18– 22 Aug 2025)

- Complete the comparative table of model test results.
- Use results to refine taxonomy of attack categories and defence methods.
- Draft initial structure for a prompt injection test suite to evaluate CSIRO platform models.

Screenshot of Timely EMAIL communication update to the Supervisor at the end of week #02 sent on 22 August 2025 Friday at 4:22 PM (16:22).



1 of 343

Weekly Communication / Reflection update - #03 week, 18– 22 Aug 2025

Arun Ragavendhar <arunragavendhar.1999@gmail.com> 16:22 (1 minute ago)
to Shigang

Dear Dr. Shigang Liu,

I hope you are well. Please find below my update for Week 3 of the internship.

Activities completed this week (Total: 16 hours):

- Continued analysis of *Systematically Analysing Prompt Injection Vulnerabilities in Diverse LLM Architectures*, focusing on architecture-specific weaknesses and mitigation strategies.
- Began practical testing of vulnerable code snippets against different LLM models to observe prompt injection behaviours.
- Recorded initial results in a comparative table showing which models allow injections to pass and which block them. The table is half complete and will be expanded in Week 4.
- Documented methodology for repeatable testing so the process can be refined and extended in later weeks.

Plan for next week (25–29 Aug 2025):

- Complete the comparative results table by finishing testing across all selected models.
- Refine the taxonomy of attack categories with evidence from testing.
- Start outlining the structure of the prompt injection test suite to be used on CSIRO platform models.

Kind regards,
Arun Ragavendhar Arunachalam Palaniyappan
ICT80004 Internship Student – CSIRO Data61

One attachment • Scanned by Gmail

PDF ICT80004-Weekl...

Reply Forward