

ICT80004 Weekly Communication – Week #02

Student Name: Arun Ragavendhar Arunachalam Palaniyappan ID: 104837257

Organisation: Commonwealth Scientific and Industrial Research Organisation (CSIRO)

Industry Supervisor: Dr. Shigang Liu

Date Prepared: 15/08/2025 Internship Week #: 2

Day	Date	Task(s) completed
1	Monday 11 Aug 2025 8 hours	<ul style="list-style-type: none"> Completed reading and detailed analysis of Benchmarking and Defending Against Indirect Prompt Injection Attacks on Large Language Models (BIPIA). <ul style="list-style-type: none"> Noted attack types (task-irrelevant, task-relevant, targeted), effect of injection position, and differences between code/text tasks. Reviewed black-box and white-box defence methods, including boundary awareness, explicit reminders, data tagging, and adversarial training. Began reviewing Prompt Injection Attack Against LLM-Integrated Applications (HOUYI), focusing on its black-box attack framework, payload components, and workflow steps.
2	Wednesday 15 Aug 2025 8 hours	<ul style="list-style-type: none"> Finished analysis of Prompt Injection Attack Against LLM-Integrated Applications, including context inference, payload generation, and dynamic feedback loop. Reviewed Prompt Injection in Large Language Model Exploitation – A Security Perspective, noting input/output filtering steps, attack simulation using probes, and continuous testing needs. Read and summarised StruQ: Defending Against Prompt Injection with Structured Queries, focusing on separate control/data channels, structured instruction tuning, and performance against manual and optimisation-based attacks. Started Systematically Analysing Prompt Injection Vulnerabilities in Diverse LLM Architectures, recording architecture-specific vulnerabilities and suggested mitigations such as input sanitisation, consensus checking, and targeted fine-tuning.

Total hours completed for the week: 16

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

- Complete analysis of *Systematically Analysing Prompt Injection Vulnerabilities in Diverse LLM Architectures*.
- Consolidate findings from all reviewed papers into a comparative table linking attack types, vulnerabilities, and defence approaches.
- Begin drafting a taxonomy of attack categories and matching defence methods, with a focus on how these will guide the design of our final deliverable — a prompt injection test suite for evaluating CSIRO platform models.
- Outline initial test cases for the suite to assess whether prompt injections can bypass or be blocked by current defences.

Screenshot of Timely EMAIL communication update to the Supervisor at the end of week #02 sent on 15 August 2025 at 3:46 PM (15:46).

Weekly Communication / Reflection update - #02 week, 11– 15 Aug 2025


**Arun Ragavendhar** <arunragavendhar.1999@gmail.com>
to Shigang

Dear Dr. S
I hope you
Activities
Monday, 1

from: Arun Ragavendhar <arunragavendhar.1999@gmail.com>
to: "Liu, Shigang (Data61, Clayton)" <Shigang.Liu@data61.csiro.au>
date: 15 Aug 2025, 15:46
subject: Weekly Communication / Reflection update - #02 week, 11– 15 Aug 2025
mailed-by: gmail.com

pt Injection Attacks

Weekly Communication / Reflection update - #02 week, 11– 15 Aug 2025

**Arun Ragavendhar** <arunragavendhar.1999@gmail.com>
to Shigang

15:46 (12 minutes ago) ☆ ☺ ↶ ⋮

Dear Dr. Shigang Liu,
I hope you are well. Please find below my update for Week 2 of the internship.
Activities completed this week (Total: 16 hours)
Monday, 11 Aug 2025 – 8 hours

- Completed reading and detailed analysis of *Benchmarking and Defending Against Indirect Prompt Injection Attacks on Large Language Models (BIPiA)*.
 - Noted attack types (task-irrelevant, task-relevant, targeted), effect of injection position, and differences between code/text tasks.
 - Reviewed black-box and white-box defence methods, including boundary awareness, explicit reminders, data tagging, and adversarial training.
- Began reviewing *Prompt Injection Attack Against LLM-Integrated Applications (HOUYI)*, focusing on its black-box attack framework, payload components, and workflow steps.

Wednesday, 13 Aug 2025 – 8 hours

- Finished analysis of *Prompt Injection Attack Against LLM-Integrated Applications*, including context inference, payload generation, and dynamic feedback loop.
- Reviewed *Prompt Injection in Large Language Model Exploitation – A Security Perspective*, noting input/output filtering steps, attack simulation using probes, and continuous testing needs.
- Read and summarised *StruQ: Defending Against Prompt Injection with Structured Queries*, focusing on separate control/data channels, structured instruction tuning, and performance against manual and optimisation-based attacks.
- Started *Systematically Analysing Prompt Injection Vulnerabilities in Diverse LLM Architectures*, recording architecture-specific vulnerabilities and suggested mitigations such as input sanitisation, consensus checking, and targeted fine-tuning.

Total hours completed: 16

Plans for next week #03 (18–22 Aug 2025):

- Complete analysis of *Systematically Analysing Prompt Injection Vulnerabilities in Diverse LLM Architectures*.
- Consolidate findings from all reviewed papers into a comparative table linking attack types, vulnerabilities, and defence approaches.
- Begin drafting a taxonomy of attack categories and matching defence methods, with a focus on how these will guide the design of our final deliverable — a prompt injection test suite for evaluating CSIRO platform models.
- Outline initial test cases for the suite to assess whether prompt injections can bypass or be blocked by current defences.

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

One attachment • Scanned by Gmail

