



รายงานวิชา 2100301  
การฝึกงานวิศวกรรม (ENGINEERING PRACTICE)

จัดทำโดย

นายโชติพิสิฐ อุดลสีหวัฒน์

รหัสประจำตัว 6531313221

ภาควิชาวิศวกรรมคอมพิวเตอร์

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

หน่วยงานที่ฝึกงาน

Japan Advanced Institute of Science and Technology

ช่วงระยะเวลาการฝึกงาน ตั้งแต่วันที่ 2 มิถุนายน พ.ศ. 2568

ถึงวันที่ 31 กรกฎาคม พ.ศ. 2568

รวมระยะเวลาการฝึกงาน 8 สัปดาห์ / 40 วันทำการ / 320 ชั่วโมง

คณะวิศวกรรมศาสตร์

จุฬาลงกรณ์มหาวิทยาลัย

## คำนำ

รายงานฉบับนี้เป็นส่วนหนึ่งของรายวิชา 2100301 การฝึกงานวิศวกรรม (ENGINEERING PRACTICE) จัดทำขึ้นเพื่อบันทึกประสบการณ์และสรุปผลการฝึกงานของผู้จัดทำ ณ Japan Advanced Institute of Science and Technology (JAIST) ประเทศญี่ปุ่น ระหว่างวันที่ 2 มิถุนายน พ.ศ. 2568 ถึงวันที่ 31 กรกฎาคม พ.ศ. 2568 รวมระยะเวลา 8 สัปดาห์

การฝึกงานครั้งนี้มุ่งเน้นการทำวิจัยในหัวข้อ AI Safety และ Chain of Thought Faithfulness โดยผู้จัดทำได้มีโอกาสศึกษางานวิจัยที่เกี่ยวข้อง เข้าร่วมกิจกรรมวิชาการของหน่วยงาน เช่น Reading Group และ Joint Meeting รวมถึงการทดลองและวิเคราะห์ข้อมูลเชิงลึกกับนักวิจัยในสถาบัน ซึ่งช่วยเสริมสร้างทักษะด้านการวิจัย การสื่อสารภาษาอังกฤษ และการทำงานในสภาพแวดล้อมนานาชาติ

รายงานนี้ประกอบไปด้วย 8 บท ได้แก่ 1) วัตถุประสงค์ของการฝึกงาน 2) ช่วงเวลาในการฝึกงานและ 3) สภาพการทำงานในระหว่างฝึกงาน 4) ผู้ควบคุมดูแลการฝึกงาน 5) รายละเอียดของหน่วยงานที่ไปฝึก 6) รายละเอียดของงานที่ทำ 7) ประโยชน์ที่ได้รับ 8) ปัญหา อุปสรรค และข้อเสนอแนะ 9) ภาคผนวก

ผู้จัดทำหวังว่ารายงานนี้จะให้ข้อมูลและรายละเอียดที่มีประโยชน์กับผู้อ่าน

นายโชติพิสิฐ อดุลสีหวัฒน์

ผู้จัดทำ

## สารบัญ

คำนำ .....	2
วัตถุประสงค์ของการฝึกงาน .....	4
ช่วงเวลาในการฝึกงานและสภาพการทำงานในระหว่างฝึกงาน .....	4
ผู้ควบคุมดูแลการฝึกงาน .....	5
รายละเอียดของหน่วยงานที่ไปฝึก .....	5
รายละเอียดของงานที่ทำ .....	8
ประโยชน์ที่ได้รับ .....	10
ปัญหา อุปสรรค และข้อเสนอแนะ .....	10
ภาคผนวก .....	11

### วัตถุประสงค์ของการฝึกงาน

1. ทำงานวิจัยเกี่ยวกับ Interpretability และ Chains of Thought ของ Large Language Model
2. ศึกษางานวิจัยอื่นๆ ที่เกี่ยวข้อง
3. ศึกษาวิธีการอ่านงานวิจัย
4. จัดทำรายงานวิชาการ
5. พัฒนาทักษะการใช้ภาษาอังกฤษ
6. เรียนรู้การใช้ชีวิตในสังคมต่างประเทศ

### ช่วงเวลาในการฝึกงานและสภาพการทำงานในระหว่างฝึกงาน

#### 1. ช่วงเวลาในการฝึกงาน

วันจันทร์ที่ 2 มิถุนายน พ.ศ. 2568 ถึงวันพฤหัสบดีที่ 31 กรกฎาคม พ.ศ. 2568

#### 2. เวลาทำงาน

วันจันทร์ถึงวันศุกร์ เวลา 10:00 น. – 18:00 น. โดยมีเวลาพักได้แก่ 12:00 น. – 13:00 น.

#### 3. สถานที่ทำงาน

เลือกทำงานที่บ้านหรืออาคารเรียนก็ได้ (Hybrid)

#### 4. สภาพการทำงาน

แต่งกายสุภาพ นั่งทำงานในห้องทำงาน นำเสนอความคืบหน้าของงานทุกวันพุธ

## ผู้ควบคุมดูแลการฝึกงาน

### 1. Shirai Kiyaoki

ตำแหน่ง: Associate Professor

หน้าที่: Supervisor, ผู้ดูแลและประเมินการฝึกงาน



### 2. Natthawut Kertkeidkachorn

ตำแหน่ง: Assistant Professor

หน้าที่: Supervisor



## รายละเอียดของหน่วยงานที่ไปฝึก

### 1. ที่ตั้ง



Japan Advanced Institute of Science and Technology, 1 Chome-1 Asahidai, Nomi, Ishikawa 923-12112.

### 2. ประวัติโดยย่อ

JAIST was founded in October 1990 as the first independent national graduate school, to carry out graduate education based on research at the highest level in advanced science and technology. JAIST aims at establishing an ideal model of graduate education for Japan. JAIST was incorporated as a National University Corporation in April 2004.

In our admission decisions we place the most significant weight on the motivation of the student as demonstrated in the personal interview. JAIST admits highly motivated students, including advanced undergraduate students (who have completed at least three years of undergraduate study), professionals, and international students, regardless of undergraduate specialization.

### **3. ขอบเขตของหน่วยงาน**

#### Mission of JAIST

JAIST endeavors to foster leaders capable of contributing to the making of a future world by creation of science and technology through its most advanced education and research in an ideal academic environment.

#### Vision of JAIST

Japan Advanced Institute of Science and Technology (JAIST) aims to become a world's top research university for innovation creation. While advancing the sophistication and excellence of its original researches, JAIST opens up the future of science and technology and contributes to sustainable development through new co-creation based on global-scale collaborations with domestic and overseas universities, research institutes and industry.

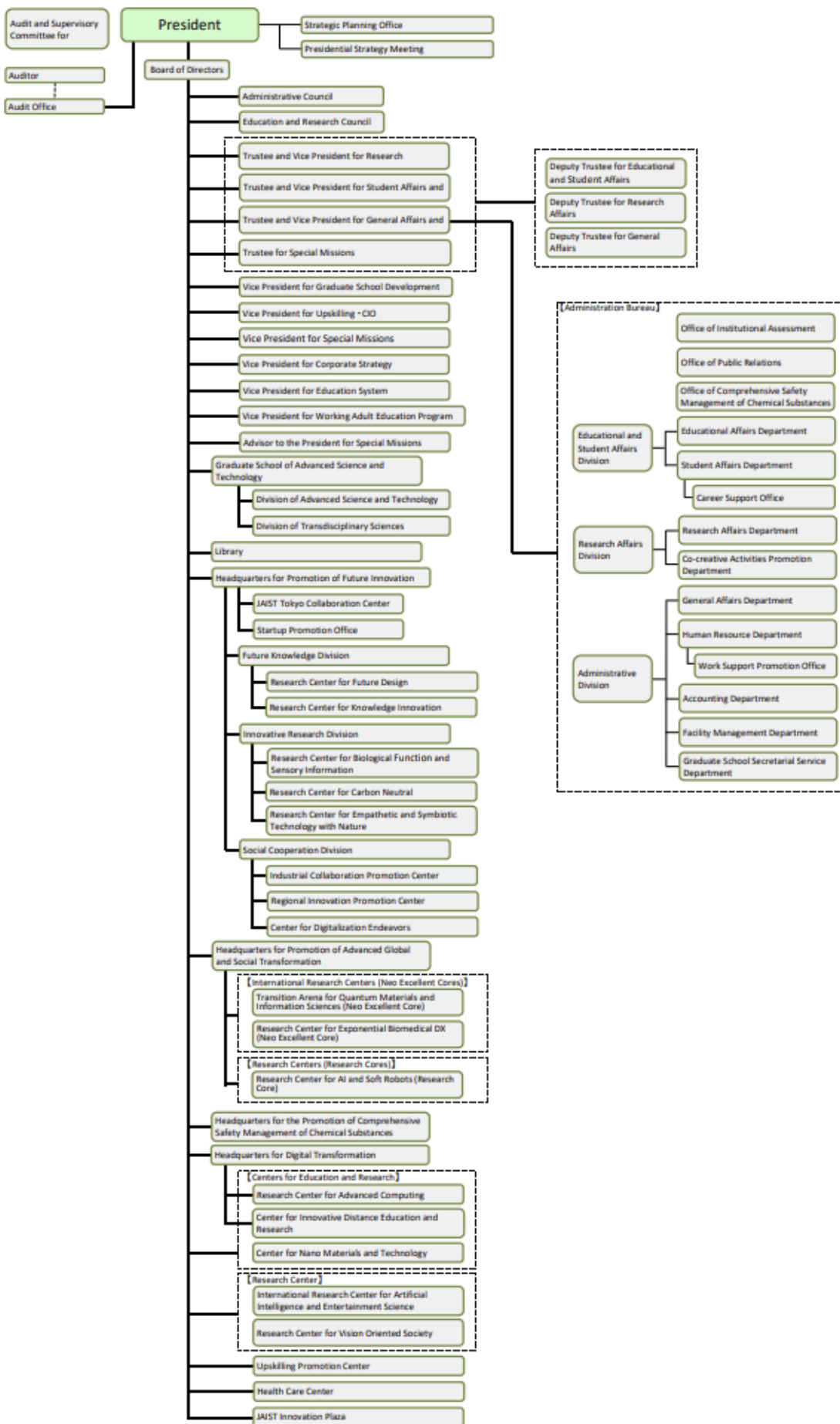
#### Goals of JAIST

JAIST develops leaders in society or industry who hold credible expertise in frontier science and technology, broad perspectives, high level of autonomy and communication ability, through its systematic advanced graduate education.

JAIST, to contribute to societies with research outcomes, creates a center of excellence for advancement of research for solving problems of our world and society and developing new fields through a variety of basic research.

JAIST fosters active global human resources by promoting faculty and student exchanges with leading institutes overseas and globalizing its education and research.

### **4. ระบบบริหารของหน่วยงาน**



## รายละเอียดของงานที่ทำ

### 1. ลักษณะของงานที่ทำ

ทำการศึกษาวิจัยในหัวข้อ AI Safety และ Chain-of-Thought (CoT) Faithfulness โดยเน้นการอ่านบทความวิจัย การเข้าร่วม Reading Group และ Joint Meeting กับอาจารย์ที่ปรึกษาและนักวิจัย ตลอดจนการทดลองประเมิน Reasoning Faithfulness และ Randomness ใน Large Language Models

งานที่ทำครอบคลุมการอ่านและสรุปงานวิจัยด้าน AI Safety เช่น Jailbreaks, Alignment Faking, Hidden/Unfaithful CoT, Reward Hacking, Mechanistic Interpretability การทดลองด้วย Reasoning Games และ Word/Number Randomization การใช้ Neuronpedia Tools เช่น Sparse Autoencoder, Attribution Graph, Circuit Tracing เพื่อวิเคราะห์ Randomness

เป้าหมายหลักคือเพื่อทำความเข้าใจกลไกที่ทำให้ LLM แสดง Unfaithful Reasoning และเพื่อหาแนวทางการตรวจจับและลดความเสี่ยงเหล่านี้

### 2. ตารางเวลาที่ใช้ในแต่ละขั้นตอนการฝึกงาน

ช่วง	ช่วงวันที่	รายละเอียดงาน
1	2 มิ.ย. 67 – 13 มิ.ย. 67	Onboarding Review NLP background Campus Tour อ่านงานวิจัย AI Safety
2	16 มิ.ย. 67 – 27 มิ.ย. 67	อ่านงานวิจัยเพิ่มเติม (Mechanistic Interpretability, RLs on Alignment, Faithfulness, Reward Hacking, Sleeper Agent, Steering Vectors) ประเมิน CoT faithfulness จาก GSM และข้อสอบภาษาไทย
3	30 มิ.ย. 67 – 11 ก.ค. 67	ทดลอง Reasoning Games (20 Questions, Pahee) ทดลอง Randomization ของคำและตัวเลข อ่านงาน Benchmark LLM ประเมิน CoT faithfulness Joint Meeting นำเสนอผลลัพธ์
4	14 ก.ค. 67 – 25 ก.ค. 67	ทดลอง Zipf's Law และ Benford's Law ศึกษา Human Psychology on Randomness ใช้ Neuronpedia (Sparse Autoencoder, Attribution Graph) เตรียมและนำเสนอ Reading Group
5	28 ก.ค. 67 – 31 ก.ค. 67	ทดลอง Circuit Tracing, Latent Clustering Joint Meeting สรุปรายงานฉบับสุดท้าย

### 3. รายละเอียดในแต่ละขั้นตอนของการฝึกงาน



#### 1) ช่วงวันที่ 2 มิ.ย. 67 – 13 มิ.ย. 67

เข้าร่วม Onboarding และ Campus Tour ลงทะเบียนบัญชี JAIST และเรียนรู้การใช้ Supercomputer ของสถาบัน จากนั้นเริ่มอ่านบทความด้าน NLP และ AI Safety โดยเน้น Unbiasing, Jailbreaks, Alignment Faking, Deceptive LLM และ Scalable Oversight รวมถึงการเข้าร่วม Reading Group Seminar และ Joint Meeting ครั้งแรกเพื่อระดมสมองหัวข้อวิจัย (Spatial Benchmark และ AI Safety)

#### 2) ช่วงวันที่ 16 มิ.ย. 67 – 27 มิ.ย. 67

เน้นอ่านบทความวิจัยขั้นสูง เช่น Mechanistic Interpretability, Auditing, RLs on AI Alignment, Data Poisoning, Faithfulness, Reward Hacking และ Sleeper Agent พร้อมเข้าร่วม Reading Group หลายครั้งและประชุมกลุ่มเพื่อหาหรือ Hidden CoT สุดท้ายได้ทดลองประเมิน CoT Faithfulness จาก GSM และข้อสอบภาษาไทย

#### 3) ช่วงวันที่ 30 มิ.ย. 67 – 11 ก.ค. 67

ทำการทดลอง Reasoning Games กับ LLM (Pahee Game, 20 Questions) และการสุ่มคำ/ตัวเลข (Word & Number Randomization) อ่านงาน Benchmark LLM และงาน Multi-turn Reasoning จากนั้นวิเคราะห์ผล CoT Faithfulness และเข้าร่วม Joint Meeting เพื่อนำเสนอผลลัพธ์เกี่ยวกับ Unfaithful CoT

#### 4) ช่วงวันที่ 14 ก.ค. 67 – 25 ก.ค. 67

ศึกษา Randomness ใน LLM โดยใช้ Benford's Law และ Zipf's Law รวมถึงการอ่านงานด้าน Decoder-only Architecture และ Text Generation Algorithms นอกจากนี้ยังทำการทดลองด้วย Neuronpedia's Sparse Autoencoder และ Attribution Graph เพื่อตรวจสอบพฤติกรรมกลุ่มของ LLM และได้มีการนำเสนอใน Reading Group เรื่อง "On the Biology of a Large Language Model"

#### 5) ช่วงวันที่ 28 ก.ค. 67 – 31 ก.ค. 67

ทดลอง Circuit Tracing และ Latent Clustering บนโมเดล LLM เข้าร่วม Reading Group และ Joint Meeting ครั้งสุดท้ายเพื่อหาหรือเกี่ยวกับการจัดทำ Final Report ของการฝึกงาน

### ประโยชน์ที่ได้รับ

1. ได้ฝึกการค้นคว้างานวิจัยผ่าน Google Scholar
2. ฝึกการสนทนาและการนำเสนอเป็นภาษาอังกฤษ
3. ได้เรียนรู้เรื่อง AI Safety AI Alignment และ Mechanistic Interpretability
4. ฝึกการจัดทำรายงานวิชาการ
5. ได้เรียนรู้สังคมและวัฒนธรรมการเรียนในต่างประเทศ

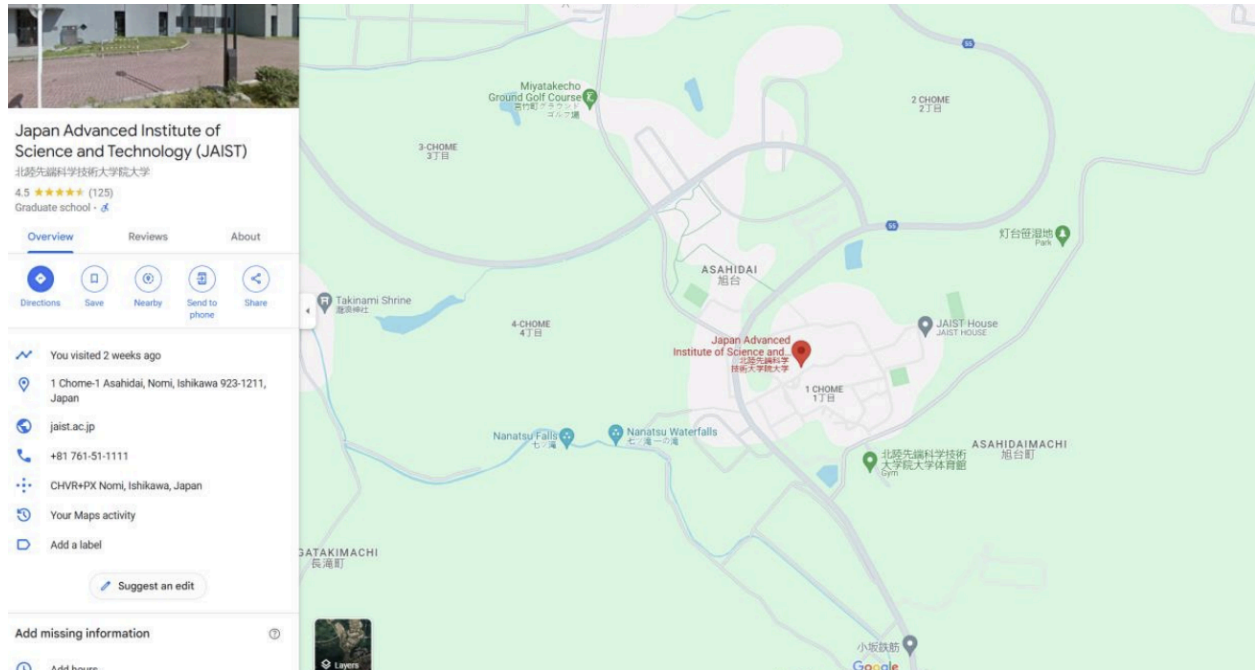
### ปัญหา อุปสรรค และข้อเสนอแนะ

1. กระบวนการรับสมัครค่อนข้างเร็วเทียบกับหน่วยงานอื่นๆ ทำให้ไม่มีตัวเลือกในการตัดสินใจ
2. การประชาสัมพันธ์ของการรับสมัครฝึกงานที่ไม่ทั่วถึงทั้งภาควิชา
3. เสนอให้มีการแลกเปลี่ยนข้อมูลติดต่อกันระหว่างนิสิตที่จะมาฝึกงานสำหรับการวางแผนไปยังญี่ปุ่น



## ภาคผนวก

1. สำเนาแผนที่ที่ได้ส่งให้กับศูนย์บริการจัดการงานของคณะฯ



2. สำเนารายงานทุก 2 สัปดาห์ตลอดระยะเวลาที่ฝึกงาน

**Internship Bi-weekly Report**  
**Computer Engineering Department**  
**No 1**

Name - Surname .....Chotpisit Adunseawat..... Student ID.....6531313221.....

Internship institution.....Japan Advanced Institute of Science and Technology (JAIST) .....

Date	Hours	Description	Student Signature
2025/06/02	8	Onboarding, Create JAIST account, Campus Tour	Chotpisit
2025/06/03	8	Review NLP background	Chotpisit
2025/06/04	8	Reading Group Seminar, Tutorial on accessing JAIST Supercomputer	Chotpisit
2025/06/05	8	Joint Meeting about research topics (Spatial Benchmark and AI Safety)	Chotpisit
2025/06/06	8	Read AI Safety papers (Unbiasing, Surveys, Guardrails)	Chotpisit
2025/06/09	8	Read AI Safety papers (Jailbreaks, Adversarial)	Chotpisit
2025/06/10	8	Read AI Safety papers (Alignment Faking, HCI)	Chotpisit
2025/06/11	8	Reading Group Seminar, Read AI Safety papers (Deceptive LLM, Unfaithful Reasoning)	Chotpisit
2025/06/12	8	Read AI Safety papers (Scalable Oversight, Monitoring)	Chotpisit
2025/06/13	8	Read AI Safety papers (VLM Safety, Harmful Dataset)	Chotpisit

Total hours in this report	80
Total hours from previous reports	0
Current total hours	80

I certify that this report is truthful.

Supervisor signature Kiyoaki Shirai  
 (.....Kiyoaki Shirai.....)

Title ...Professor.....

Date ...June 16, 2025.....

**Internship Bi-weekly Report**  
**Computer Engineering Department**  
 No 2

Name - Surname .....Chotpisit Adunsehawat..... Student ID.....6531313221.....

Internship institution.....Japan Advanced Institute of Science and Technology (JAIST) .....

Date	Hours	Description	Student Signature
2025/06/16	8	Read AI Safety papers (Mechanistic Interpretability, Auditing)	Chotpisit
2025/06/17	8	Read AI Safety papers (RLs on AI Alignment, Data Poisoning)	Chotpisit
2025/06/18	8	Reading Group Seminar, Read AI Safety papers (Latent CoT, Faithfulness, Shortcut)	Chotpisit
2025/06/19	8	Joint Meeting about Hidden CoT	Chotpisit
2025/06/20	8	Read AI Safety papers (Process Reward Model, Cognitive Behaviors, Cross Layer Transcoder)	Chotpisit
2025/06/23	8	Read AI Safety papers (Reward Hacking, Monosemanticity)	Chotpisit
2025/06/24	8	Read AI Safety papers (Anthropic's HHH AI, Sleeper Agent)	Chotpisit
2025/06/25	8	Reading Group Seminar, Read AI Safety papers (Scheming, Unfaithful CoT)	Chotpisit
2025/06/26	8	Reading AI Safety papers (SAD dataset, Steering Vectors)	Chotpisit
2025/06/27	8	Evaluate CoT faithfulness from GSM, Thai Exam dataset	Chotpisit

Total hours in this report	80
Total hours from previous reports	80
Current total hours	160

I certify that this report is truthful.

Supervisor signature Kiyoaki Shirai  
 (.....Kiyoaki Shirai.....)

Title ...Professor.....

Date ....June 30, 2025.....

**Internship Bi-weekly Report**  
**Computer Engineering Department**

No 3

Name - Surname .....Chotpisit Adunsehawat..... Student ID.....6531313221.....

Internship institution.....Japan Advanced Institute of Science and Technology (JAIST) .....

Date	Hours	Description	Student Signature
2025/06/30	8	Test reasoning games with LLM (Thai Tricks games, 20 questions, etc.)	Chotpisit
2025/07/01	8	Evaluate reasoning games CoT	Chotpisit
2025/07/02	8	Reading Group Seminar, Read NLP and LLM Benchmark papers	Chotpisit
2025/07/03	8	Test word randomization with LLM, Read papers (Stateless, Multi-turn)	Chotpisit
2025/07/04	8	Evaluate CoT faithfulness	Chotpisit
2025/07/07	8	Test complex reasoning tasks (Poem Writing, Math Logic)	Chotpisit
2025/07/08	8	Test CoT from word and number randomization, Read more papers (CoT interpretability)	Chotpisit
2025/07/09	8	Reading Group Seminar, Prepare CoT results	Chotpisit
2025/07/10	8	Joint Meeting about Unfaithful CoT results and discuss future topics	Chotpisit
2025/07/11	8	Test and read more on LLM randomness	Chotpisit

Total hours in this report	80
Total hours from previous reports	160
Current total hours	240

I certify that this report is truthful.

Supervisor signature .....*Kiyoaki Shirai*.....

(.....Kiyoaki Shirai.....)

Title ...Professor.....

Date ....July 14, 2025 .....

**Internship Bi-weekly Report**  
**Computer Engineering Department**

No 4

Name - Surname .....Chotpisit Adunsehawat..... Student ID.....6531313221.....

Internship institution.....Japan Advanced Institute of Science and Technology (JAIST) .....

Date	Hours	Description	Student Signature
2025/07/14	8	Test cases for LLM word and number randomness	Chotpisit
2025/07/15	8	Verify results with Benford and Zipf's Law, Read papers (Decoder-only Architecture, Text Generation Algorithm)	Chotpisit
2025/07/16	8	Reading Group Seminar, Test Zipf's Law on CoTs from Anthropic's Alignment Faking paper	Chotpisit
2025/07/17	8	Research human psychology on randomness and Benford's Law exceptions	Chotpisit
2025/07/18	8	Joint meeting on randomness results and debiasing approaches, Read papers on Polysemanticity	Chotpisit
2025/07/21	8	Use Neuronpedia's Sparse Autoencoder for examining randomness, Prepare Reading Group presentation	Chotpisit
2025/07/22	8	Prepare Reading Group presentation, Read papers on Circuit Tracings	Chotpisit
2025/07/23	8	Prepare Reading Group presentation, Read papers on Cross-Layer Transcoder and Superposition in NLP	Chotpisit
2025/07/24	8	Present Reading Group Seminar on "On the Biology of a Large Language Model" paper	Chotpisit
2025/07/25	8	Use Neuronpedia's Attribution Graph for examining randomness on numbers and words	Chotpisit

Total hours in this report	80
Total hours from previous reports	240
Current total hours	320

I certify that this report is truthful.

Supervisor signature Kiyoaki Shirai  
(.....Kiyoaki Shirai.....)

Title ...Professor.....

Date ....July 28, 2025.....

**Internship Bi-weekly Report**  
**Computer Engineering Department**

No 5

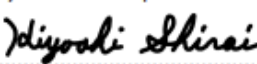
Name - Surname .....Chotpisit Adunsehawat..... Student ID.....6531313221.....

Internship institution.....Japan Advanced Institute of Science and Technology (JAIST) .....

Date	Hours	Description	Student Signature
2025/07/28	8	Explore more Randomness cases using Neuronpedia's Circuit Tracing	Chotpisit
2025/07/29	8	Attempt Latent Clustering on Circuit Tracing model	Chotpisit
2025/07/30	8	Reading Group Seminar, Explore more cases	Chotpisit
2025/07/31	8	Joint Meeting on the LLM randomness progress and future work, Leaving JAIST	Chotpisit

Total hours in this report	320
Total hours from previous reports	32
Current total hours	352

I certify that this report is truthful.

Supervisor signature ..........  
(.....Kiyooki Shirai.....)

Title . ...Professor.....

Date .....July 31, 2025.....

3. รูปภาพ ตาราง ที่ได้กล่าวอ้างจากบทต่าง ๆ ข้างต้น