**CMPT 400.3/405.3/407.3**

**Learning Contract 2023 - 2024**

Student Name: William Bushell. CMPT **400** or 405 or 407 **(circle one)**

Supervisor: Ian McQuillan

Project Title: Automata Inference Using OpenAI

**(The following sections are to be completed in conjunction with your supervisor.)**

1. **Learning Objectives.** Identify **three** to **five** significant things that you will learn in this course by working on this topic in the manner you discuss below.

1) How accurately OpenAI can learn DFAs and NFAs.

2) Any limitations in the complexity of the automata OpenAI can learn.

3) Whether OpenAI is better at inferring DFAs or NFAs.

4) How the number of states affects the output.

5) How the number of positive and number of negative examples affects the output.

1. **Prerequisite Knowledge.** Identify knowledge necessary to work on this topic and where you obtained it. Also identify any potential overlap between this work and existing CMPT courses and whether or not you have taken or are taking these courses.

I have taken CMPT 364 which taught DFAs and NFAs, as well as grammars in computer science, and limits in the languages these models can handle. The knowledge from this course is necessary for understanding the automata the LLMs will output, and also necessary in verifying the correctness of the outputs. I have also taken CMPT 470, which had a section on machine learning and AI. I will be using this knowledge to understand how the LLMs output this data, as well as creating meaningful test suites for the LLMs, and seeing if there are any useful techniques to train LLMs to output more correct automata. I see overlap also with CMPT 317, the intro to artificial intelligence, which is a course I have not taken nor am I planning to take. I could see CMPT 318 being a useful course as well, where I could use that knowledge to construct effective data pipelines from the LLM to some automata verifying application. I am taking CMPT 384, and I believe that could be useful in the presenting of my research data results.

1. **Background.** Identify **three** sources of information on this topic that you have already consulted in preparing this learning contract.

1) CMPT 364 (DFAs, NFAs, and grammars in computer science)

2) CMPT 470 (Machine learning and test suites)

3) Professor Ian McQuillan

4) FAdo Library documentation

1. **Project Plan.** Identify **at least seven** major milestones of the work you propose to undertake in the order you plan to do them and with estimates of the time required to complete each of them and the date when each will be completed. One of the early milestones must be the literature survey, and one of the final ones must be the presentation and paper submitted.

1) Curate prompts to ChatGPT that constrains the format of which it outputs automata.

2) Manually test key automata with ChatGPT

3) Pipeline into Fado software to test equivalance and automate testing

4) Preliminary testing

5) Random generation of finite automata

6) Systematic testing of parameter space

7) Gather results and be presentation ready

1. **Resources Required.** Identify the main resources that you will be using to accomplish your project and what arrangements you have made to ensure their availability. I’ve given lots of space for the Literature part, because specific papers, conference proceedings, and/or journals should be listed.

People (such as users or other experts) and their roles:

William Bushell – Research Student

Professor Ian McQuillan - Supervisor

Hardware and Software:

FAdo

Python

OpenAI

ChatGPT

Literature:

Intro textbooks

Intro to machine learning of grammatical inference

Background on OpenAI/ChatGPT

Internet:

OpenAI API

1. **Progress Reports.** Identify the project milestones you expect to reach by these stand-ups dates (remember that you will be accountable for having accomplished these things by these dates).

**Nov. 6 - 1st Progress Report**:

Progress on ChatGPT output constriction

**Dec. 4 - 2nd Progress Report**:

Progress on manual automata testing with ChatGPT

**Jan 8 - 3rd Progress Report:**

Pipeline software with FAdo and OpenAI functional and returning meaningful data

**Feb 4 - 4th Progress Report**:

Collected pipeline software data

**Mar 5 - 5th Progress Report**:

Gather and process results, prepared presentation

1. **Deliverables.** Identify what you will deliver for marking and, where appropriate, the distribution of marks to be used for evaluating these deliverables. In particular, this is used to decide what portion of the Paper component is associated with a code deliverable.

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Exactly what is required as part of the course syllabus, with the percentages as in the syllabus.**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

1. **Miscellaneous.**  Identify any other important factors that should be taken into account in conducting and evaluating this course.

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

1. **Agreements.** The following signatures indicate approval of this Learning Contract. All three signatures are required.

Student: William Bushell Date: \_\_\_Sept 26, 2024\_\_\_\_\_\_\_\_

Supervisor: \_Ian McQuillan\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date: \_\_Sept 26, 2024\_\_\_\_\_\_\_\_\_

Course Coordinator: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_