



VICTORIA UNIVERSITY OF  
**WELLINGTON**  
TE HERENGA WAKA

**School of Engineering and Computer Science**  
Te Kura Mātai Pūkaha, Pūrorohiko

PO Box 600  
Wellington  
New Zealand

Tel: +64 4 463 5341  
Internet: [office@ecs.vuw.ac.nz](mailto:office@ecs.vuw.ac.nz)

**Project title not yet defined**

James Thompson

Supervisor: Ali Knott

November 26, 2025

Submitted in partial fulfilment of the requirements for  
Master of Artificial Intelligence.

**Abstract**

Also undecided at this time and will be filled in later.



# Chapter 1

## Introduction

In this section you should include a very brief introduction to the problem to the problem and the project.

Your project proposal should cover the following points:

- the engineering problem that you are going to solve;
- how you plan to solve your problem;
- how you intend to evaluate your solution; and
- any resource requirements for your project such as software, hardware or other resources that will be needed in the course of the project.

I am going to cite [1]

Your proposal should be not more than than 3 pages long.

### 1.1 The Problem

In this section you should give a brief description of the problem itself. You want to briefly explain the problem, why it is important to solve the problem and define your project aims. After reading this section, the reader should understand why it is a problem, believe that it is important to solve and have a clear idea of the aims of your project.

When describing the aims of the project, you should avoid vague, unmeasurable words like ‘analyse’, ‘investigate’, ‘describe’, and use specific, measurable words like ‘implement’, ‘demonstrate’, ‘show’, ‘prove’.

For example:

**Good** The aim of this project is to implement and evaluate a management system for network switches;

is much better than:

**Bad** The aim of this project is to investigate management systems for network switches.

In the second case there is no idea of how much work is involved, and you will never know whether you have finished. You and your supervisor (and the markers of your project) may have very different ideas about what such an ‘investigation’ involves. Of course, it is possible that the task you set yourself is not achievable, but if you are clear from the outset this is less likely, and will more easily be corrected.



# Chapter 2

## Background

### 2.1 Large Language Models

*A basic introduction to large language models that are transformer based. Including simple description of self attention, along with training process of pretraining and next token prediction.*

### 2.2 Constitutional AI

#### 2.2.1 What is Constitutional AI?

Constitutional AI (CAI) is a method of post-training a large language model to be aligned to set of principles that are outlined in simple natural language document called a constitution [1]. Importantly it works without the need for human labelling of data common in other alignment methods such as RLHF and DPO [4, 5]. INstead of human labelling CAI uses the model itself to generate feedback on its outputs based on the principles outlined in the constitution, this feedback is then used to further train the model to align it to the constitution.

#### 2.2.2 How does Constitutional AI work?

CAI is a process that takes in a constitution, a pre-trained language model, and a dataset of prompts and outputs a model that is aligned to the constitution. The process consists of three main steps: self generative and supervised fine-tuning (SFT), preference modelling (PM), and reinforcement learning with human feedback (RLHF) [1]. The constitution was generated in an adhoc manner by the authors of [1] and consists of a set of principles that aim to make the model's outputs more helpful, honest, and harmless.

#### 2.2.3 Supervised learning on revised responses

The first stage of the process involves taking the pre-trained model and using it to generate responses to a set of prompts. These prompts are designed to elicit harmful behavior. As the pre-trained model is designed to be helpful it will likely generate harmful responses to these prompts. These responses are then critiqued by the model itself using principles outlined in the constitution, only a subset of which are used for each response. The critique is then used to revise the original response to make it more aligned to the constitution. This process of generating responses, critiquing them, and revising them is done iteratively to create a dataset of revised responses. This dataset is then used to fine-tune the pre-trained model using self-supervised learning to create a model called the SL-CAI model.

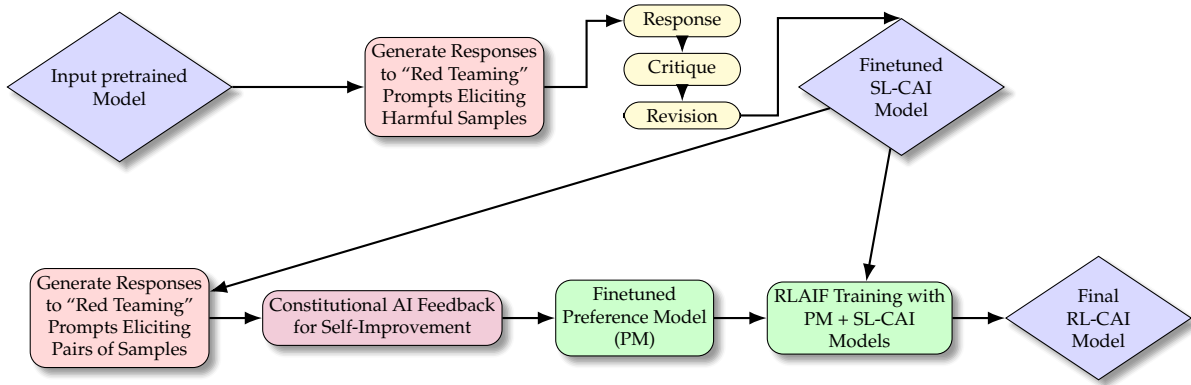


Figure 2.1: Overview of Constitutional AI process, adapted from [1]

#### 2.2.4 Reinforcement learning with AI feedback

The next step involves using the SL-CAI model to generate pairs of responses to a set of prompts (overlapping with before or not). Then using the constitution the SL-CAI model is used to provide feedback on which of the two responses is better aligned to the constitution. This feedback is then used to fine-tune a preference model (PM) that can predict which of two responses is better aligned to the constitution. Finally the PM and SL-CAI models are used together to do reinforcement learning in style of RFHL [4] to generate the final model RL-CAI. This final model is now aligned to the constitution without any human labelling of data. In the setup used in [1] the authors did in fact use some labels from human labelers which were used to provide the helpfulness signal in the preference model training, however in principle this step could of been done with AI.

#### 2.2.5 Relevance

In the context of the time that this paper was to address the key dilemma of helpfulness vs harmfulness in large language models. The goal being to make a model that is both helpful and harmless, without sacrificing one for the other. Further research has expanded the horizons to be concerned with more than just harmfulness vs helpfulness trade off. Along with this there are concerns that methods like CAI and RLHF are simply superficially aligning models to avoid harmful outputs rather than deeply aligning to the values we want [2, 3]. As hinted at by the authors in [1], there is room for improvement in the CAI method to make it focus on more than just harmfulness vs helpfulness. This can be done by modifying the constitutions to include more diverse principles as well as modifying the red teaming prompts to target more diverse scenarios. In this way we can see how elements of CAI can be built upon by generating these diverse constitutions and red teaming prompts in a representative manner to create models aligned to a representative set of values.

## Chapter 3

# Project Proposal

### 3. Proposed Solution

In this section you will explain how solve the problem, that is, how you intend to carry the project out. At this early stage you need to be both clear about what you are going to do and flexible enough to adapt to changing circumstances. Making an early plan will not prevent you from running into trouble, but it will help you identify possible problems early. For example, if you intended to run an experiment in HCI, you might realise early on that there would be problems gathering sufficient data to get reliable results, and that you should re-design your experiment.

Part of the planning process involves producing a timetable for when the work is actually going to be done.

Each part of the project should produce some output. For example you might plan on spending two weeks on background reading: the output of this will be a bibliography, and a possibly a literature survey for your report. Indeed, if you take the advice given above about having specific, measurable goals, you should describe this part of your project as:

**Good** Produce bibliography (est: 2 weeks)

rather than

**Bad** Background reading (est: 2 weeks)

Note that the methodology you outline here is dependent upon the type of project and engineering area. You must talk to your supervisor about this.

### 4. Evaluating your Solution

In this section you will explain how you will evaluate your solution once you have built it. The method of evaluation will be domain specific. Your supervisor should provide guidance as to what is an appropriate form of evaluation. For example, user testing for a HCI project or performance measurement for a Network Engineering project.

### 5. Resource Requirements

In this section you will detail any resource requirements such as hardware, software or access to subjects.



# Bibliography

- [1] BAI, Y., KADAVATH, S., KUNDU, S., ASKELL, A., KERNION, J., JONES, A., CHEN, A., GOLDIE, A., MIRHOSEINI, A., MCKINNON, C., CHEN, C., OLSSON, C., OLAH, C., HERNANDEZ, D., DRAIN, D., GANGULI, D., LI, D., TRAN-JOHNSON, E., PEREZ, E., KERR, J., MUELLER, J., LADISH, J., LANDAU, J., NDOUSSE, K., LUKOSUITE, K., LOVITT, L., SELLITTO, M., ELHAGE, N., SCHIEFER, N., MERCADO, N., DASARMA, N., LASENBY, R., LARSON, R., RINGER, S., JOHNSTON, S., KRAVEC, S., SHOWK, S. E., FORT, S., LANHAM, T., TELLEEN-LAWTON, T., CONERLY, T., HENIGHAN, T., HUME, T., BOWMAN, S. R., HATFIELD-DODDS, Z., MANN, B., AMODEI, D., JOSEPH, N., MCCANDLISH, S., BROWN, T., AND KAPLAN, J. Constitutional AI: Harmlessness from AI Feedback, Dec. 2022.
- [2] GREENBLATT, R., DENISON, C., WRIGHT, B., ROGER, F., MACDIARMID, M., MARKS, S., TREUTLEIN, J., BELONAX, T., CHEN, J., DUVENAUD, D., KHAN, A., MICHAEL, J., MINDERMAN, S., PEREZ, E., PETRINI, L., UESATO, J., KAPLAN, J., SHLEGERIS, B., BOWMAN, S. R., AND HUBINGER, E. Alignment faking in large language models, Dec. 2024.
- [3] HUBINGER, E., DENISON, C., MU, J., LAMBERT, M., TONG, M., MACDIARMID, M., LANHAM, T., ZIEGLER, D. M., MAXWELL, T., CHENG, N., JERMYN, A., ASKELL, A., RADHAKRISHNAN, A., ANIL, C., DUVENAUD, D., GANGULI, D., BAREZ, F., CLARK, J., NDOUSSE, K., SACHAN, K., SELLITTO, M., SHARMA, M., DASARMA, N., GROSSE, R., KRAVEC, S., BAI, Y., WITTEN, Z., FAVARO, M., BRAUNER, J., KARNOFSKY, H., CHRISTIANO, P., BOWMAN, S. R., GRAHAM, L., KAPLAN, J., MINDERMAN, S., GREENBLATT, R., SHLEGERIS, B., SCHIEFER, N., AND PEREZ, E. Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training, Jan. 2024.
- [4] OUYANG, L., WU, J., JIANG, X., ALMEIDA, D., WAINWRIGHT, C. L., MISHKIN, P., ZHANG, C., AGARWAL, S., SLAMA, K., RAY, A., SCHULMAN, J., HILTON, J., KELTON, F., MILLER, L., SIMENS, M., ASKELL, A., WELINDER, P., CHRISTIANO, P., LEIKE, J., AND LOWE, R. Training language models to follow instructions with human feedback, Mar. 2022.
- [5] RAFAILOV, R., SHARMA, A., MITCHELL, E., ERMON, S., MANNING, C. D., AND FINN, C. Direct Preference Optimization: Your Language Model is Secretly a Reward Model, July 2024.