



VICTORIA UNIVERSITY OF
WELLINGTON
TE HERENG A 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

Project title not yet defined

James Thompson

Supervisor: Ali Knott

November 28, 2025

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

Abstract

This is also undecided at this time and will be filled in later.

Chapter 1

Introduction

Unsure if this is needed. Yet could be a simple "one pager" (which is bigger than the one paragrapher of the abstract) that introduces the topic area, motivates the project and outlines the project.

Chapter 2

Background

2.1 Large Language Models

2.1.1 Language models

Language models are designed to understand and generate text. Generally probabilistic models that learn to predict next tokens. Use a large corpus of text data to learn statistical patterns in language. Started with simple statistical models such as n-grams and evolved to neural network based models such as RNNs, LSTMs and transformers.

2.1.2 Modern Large language models

Modern LLMs are transformer based. Just use the decoder. Take the input and generate a sequence of tokens autoregressively. Use self attention mechanism to capture long range dependencies in text. Trained using next token prediction objective on large datasets. Explain what self attention is (high-level), how the basic architecture looks (layers of attention, feedforward, multihead). How generation is done auto regressively. Lastly the training objective of next token prediction and very large corpus of text.

2.1.3 Applications of LLM

LLMs are used a lot in modern day. Can achieve a lot. Agentic behavior is common.

2.2 Alignment of Large Language Models

2.2.1 What is alignment?

Introduction to the concept of alignment in large language models, why it is important, and common methods to achieve it. Should reduce the duplicate descriptions of various alignment methods below.

Note: Some of these subsection below of specific alignment methods could be compressed into single subsections and only explain the ones that are most relevant to my final proposed method. For now I include the three papers that I have actually read, I expect to compress RLHF and DPO into a "Other alignment methods" subsection. If I use more modern methods such as CAI with KTO or ORPO then I can expand on those more.

2.2.2 Reinforcement Learning from Human Feedback

Having a model that is able to predict the next token is powerful however it doesn't guarantee that the model will behave in a manner that is aligned with the user preferences.

Reinforcement Learning from Human Feedback (RLHF) [7] is a method that aims to align large language models to human preferences by using human feedback in the form of response rankings to guide the training process. This method was popularized by its use in training InstructGPT which was smaller and better at following instructions compared to GPT-3 [3, 7].

How does RLHF work?

RLHF is separated into three distinct stages. Firstly is supervised fine-tuning (SFT) where a pre-trained language model is fine-tuned on a dataset of prompts and human written ‘expected’ responses to those prompts. The second step is collecting multiple responses from the SFT model to create a set of prompts and response then having human labelers rank these responses from best to worst. This ranked data is then used to train a preference model (PM) that can predict which of two responses is better aligned to human preferences. Finally the PM is used as a reward signal in the PPO [9] RL algorithm to further fine-tune the SFT model and create the final RLHF model.

The dataset of prompts are taken from the data distributions that you want to align towards. In the case of [7] the prompts were taken from playground use of an early version of InstructGPT. These prompts are intended to be representative of the types of queries and instructions that users would give to the model in practice. The human labelers that provide the human feedback part is a small set (40) of contractors that are hired to both provide desired responses to the prompts as well as rank the responses generated by the SFT model.

Relevance

The idea of using human feedback to help train machine learning models is powerful, although it is not new and was around in theory and concept for years beforehand [4]. The specific method of RLHF was widely adopted [11, 1, 10] yet is now replaced with more effective methods to utilize human feedback such as Direct Preference Optimization (DPO) [8]. It does however form the basis for an alignment method that allows us to train models that directly align to human preferences rather than just next token prediction. It is one of the first key methods that started the post-training focus for large language models.

2.2.3 Direct Preference Optimization

The idea presented in [7] introduces the concept of using human feedback to train large language models to be aligned to human preferences. However the process of RLHF is complex and requires training multiple models (SFT, PM, RLHF). Direct preference optimization (DPO) [8] is a method that simplifies this process by removing the need for reinforcement learning altogether. Instead DPO reformulates the preference modelling step to instead create an objective function that implicitly creates the reward function from the model and a reference model (usually the SFT model).

How does DPO work?

DPO starts with the same preferences data as RLHF; a set of prompts with responses ranked by humans. Like RLHF it starts with a pre-trained language model and is supervised fine-tuned on the prompts and highest ranked responses to create an SFT model. Instead of training a separate preference model, DPO uses the SFT model and the current model to create the DPO update gradient:

$$\nabla_{\theta} \mathcal{L}_{DPO}(\theta) = -\mathbb{E}_{(x,y^+,y^-) \sim D} [\sigma(\hat{r}_{\theta}(x, y^-) - \hat{r}_{\theta}(x, y^+)) [\nabla_{\theta} \log \pi_{\theta}(y^+|x) - \nabla_{\theta} \log \pi_{\theta}(y^-|x)]]$$

Where $\hat{r}_\theta(x, y) = \log \pi_\theta(y|x) - \log \pi_{SFT}(y|x)$ is the reward function defined as the log difference between the current model and the SFT model, σ is the sigmoid function, and (x, y^+, y^-) are the prompt, preferred response, and dis-preferred response respectively. [8] found that the first weighting term is important and can be interpreted as increasing the weight of the update when the reward estimate is wrong. The second term is there to increase likelihood of generated preferred responses and decrease the likelihood of dis-preferred responses. This update can be calculated with only 4 forward passes through the network per preference pair (2 for reference model and two for current model), making it much more efficient than RLHF¹.

Relevance

Aligning large language models to human preferences is a crucial step in ensuring that these models are useful and safe for real-world applications. DPO provides a more efficient and straightforward process. It is the most commonly used public method for post training. However there are still more modern approaches that improve on alignment quality or efficiency such as [5, 6]. Due to its widespread adoption DPO represents a strong baseline for alignment.

2.2.4 Constitutional AI

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 [2]. Importantly it works without the need for human labelling of data common in other alignment methods such as RLHF and DPO [7, 8]. 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.

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) [2]. The constitution itself was generated in an adhoc manner by the authors of [2] and consists of a set of principles that aim to make the model's outputs more helpful, honest, and harmless.

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 critique. 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

¹RLHF will need to do many forward passes of the network to train the preference model, then it will still have to do 3 forward passes for each datapoint when it comes to the RL step.

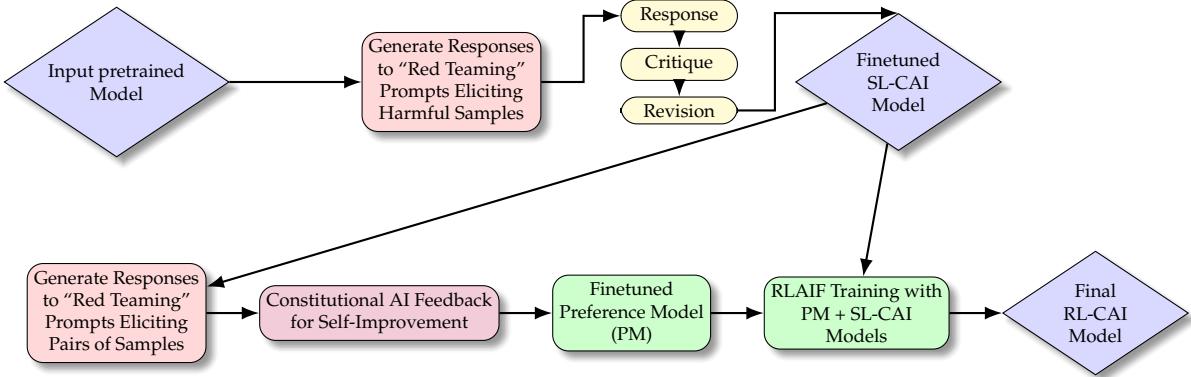


Figure 2.1: Overview of Constitutional AI process, adapted from [2]

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.

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 the set of prompts from 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 [7] 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 [2] 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 the preference signal could have been entirely generated by an AI model.

Relevance

In the context of when this paper was released there was a key dilemma of helpfulness vs harmlessness in large language models. Therefore, this paper set out 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. As hinted at by the authors in [2], there is room for improvement in the CAI method to make it focus on more than just harmlessness 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 (i.e controversial topics, political leanings, etc). 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.

2.2.5 Current alignment concerns

Introduce some of the ongoing concerns around current alignment methods. Such as shallow alignment, inner misalignment, scalable oversight etc.

2.3 Broader AI safety concerns

Take the problem of alignment and expand it broader AI alignment concerns. Talk about mis generalization, and specification gaming. Introduce the concepts from PublicAI, incentives and collective intelligence. ...

2.4 Representing Human Values

Discuss concepts such as how does one understand the values of a group. How does it relate to building AI systems. I need to significant reading on this as I am most outside my comfort zone for this. Look into some philosophical concepts around how the collective the best baseline if we have no other gold standard. Provide some rationale as to why using the public as the source of values is a good idea, as opposed to some "gold standard" values provided by experts.

Chapter 3

Project Proposal

Rather than write the proposal at the end I will write it down now to get my current ideas and can update it as I go along. I worry this could "lock in" my ideas too early, yet the benefits of the process of writing seems to outweigh that. Furthermore I will expand to make it more complete and detailed as I go along and figure out what the technical approach will be.

3.1 What I will do

The goal of this project is to develop, implement and evaluate for future feasibility a method for aligning large language models to a set of representatively collected human values. This will be a complete end to end process from data collection to final models output.

3.2 What problem I will solve

Current methods for training large language models involve scraping the internet for large amount of text data. This creates a very powerful next token prediction model. Then to turn this models into something useful for the end user (e.g. chatbots, assistants, robo doctor etc) the model is aligned to how we want it to behave. Current methods for alignment involve contracting human labelers/annotators/demonstrators to provide data to give to the model on how we want it to behave. Problematically this is a opaque process that gives the end user little insight into what the model is "designed to do". Instead by collecting a representative set of human values from the public we can align models to be in line with the end users themselves¹.

3.3 Timeline

The project will be carried out over three main phases. The first phase will be development of a data collection method to gather a representative set of human values. The second phase will be the implementation of an alignment method that uses the collected data to align a large language model. The third phase will be the evaluation of the aligned model against the human values dataset to determine the effectiveness of the alignment.

¹In the sense that the public are the end users of these models.

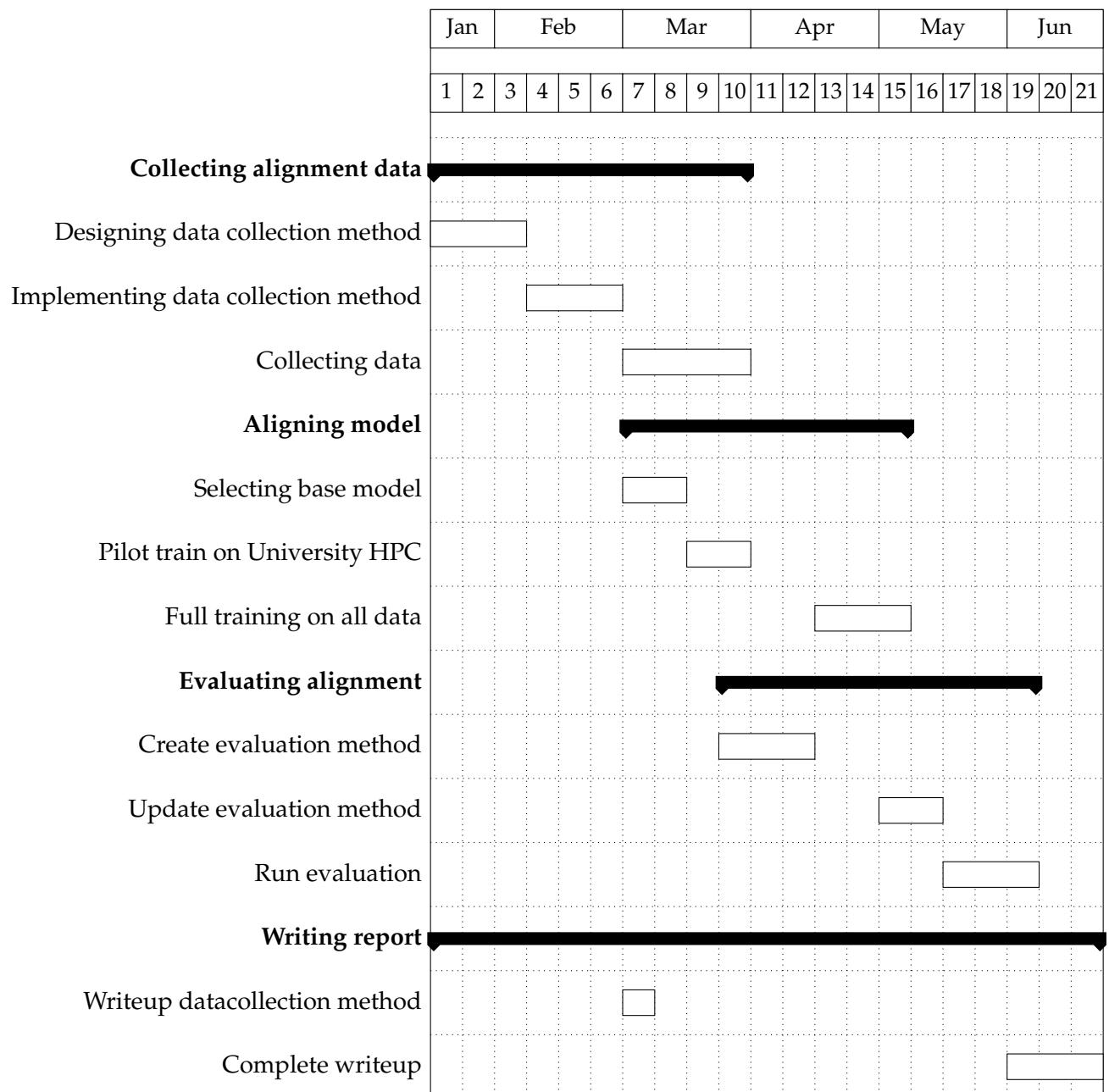


Figure 3.1: Project timeline in weeks (only roughly aligned to months). This assumes a start time of end of January 2026 with near full time work ramping up to full time work by end of February.

3.4 Output

The output is both qualitative and quantitative in nature. There are four main deliverables:

- A dataset of representatively collected human values for alignment of large language models.
- A reproducible method for collecting a representative set of human values.
- A aligned set of LLMs using the collected human values.

- A report that outlines all previous outputs including evaluation of the aligned models.

3.5 Resources needed

There are two resource requirements for this project, hardware and survey responses.

3.5.1 Hardware

The first is hardware which will be needed to do the alignment finetuning and training of the model. This will be done on the Engineering and Computer Science department's GPU servers which should give me access to a machine with up to 4 A40 GPUs. This will be sufficient to train models up to the size of about 30 billion parameters which is sufficient for *near* SOTA performance on many tasks. Alternatively I may rent GPU time on cloud providers such as Lambda AI which could cost in the realm of several hundred NZD depending on the total training time needed.

Survey Responses

No idea here. Need to research how many responses would be needed. I imagine that I would want several thousands responses to get a sufficiently large data signal. Therefore at least a few hundred respondents to get meaningful "representation" and subgroups.

Bibliography

- [1] BAI, J., BAI, S., CHU, Y., CUI, Z., DANG, K., DENG, X., FAN, Y., GE, W., HAN, Y., HUANG, F., HUI, B., JI, L., LI, M., LIN, J., LIN, R., LIU, D., LIU, G., LU, C., LU, K., MA, J., MEN, R., REN, X., REN, X., TAN, C., TAN, S., TU, J., WANG, P., WANG, S., WANG, W., WU, S., XU, B., XU, J., YANG, A., YANG, H., YANG, J., YANG, S., YAO, Y., YU, B., YUAN, H., YUAN, Z., ZHANG, J., ZHANG, X., ZHANG, Y., ZHANG, Z., ZHOU, C., ZHOU, J., ZHOU, X., AND ZHU, T. Qwen Technical Report, Sept. 2023.
- [2] 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., NDOSSE, K., LUKOSUIT, K., LOVITT, L., SELLITTO, M., ELHAGE, N., SCHIEFER, N., MERCADO, N., DASSARMA, 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.
- [3] BROWN, T. B., MANN, B., RYDER, N., SUBBIAH, M., KAPLAN, J., DHARIWAL, P., NEELAKANTAN, A., SHYAM, P., SAstry, G., ASKELL, A., AGARWAL, S., HERBERT-VOSS, A., KRUEGER, G., HENIGHAN, T., CHILD, R., RAMESH, A., ZIEGLER, D. M., WU, J., WINTER, C., HESSE, C., CHEN, M., SIGLER, E., LITWIN, M., GRAY, S., CHESS, B., CLARK, J., BERNER, C., MCCANDLISH, S., RADFORD, A., SUTSKEVER, I., AND AMODEI, D. Language Models are Few-Shot Learners, July 2020.
- [4] CHRISTIANO, P., LEIKE, J., BROWN, T. B., MARTIC, M., LEGG, S., AND AMODEI, D. Deep reinforcement learning from human preferences, June 2017.
- [5] ETHAYARAJH, K., XU, W., MUENNIGHOFF, N., JURAFSKY, D., AND KIELA, D. KTO: Model Alignment as Prospect Theoretic Optimization, Nov. 2024.
- [6] HONG, J., LEE, N., AND THORNE, J. ORPO: Monolithic Preference Optimization without Reference Model, Mar. 2024.
- [7] 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.
- [8] 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.

- [9] SCHULMAN, J., WOLSKI, F., DHARIWAL, P., RADFORD, A., AND KLIMOV, O. Proximal Policy Optimization Algorithms, Aug. 2017.
- [10] SHAO, Z., WANG, P., ZHU, Q., XU, R., SONG, J., BI, X., ZHANG, H., ZHANG, M., LI, Y. K., WU, Y., AND GUO, D. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models, Apr. 2024.
- [11] TEAM, G., ANIL, R., BORGEAUD, S., ALAYRAC, J.-B., YU, J., SORICUT, R., SCHALK-WYK, J., DAI, A. M., HAUTH, A., MILLICAN, K., ET AL. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).