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

December 14, 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.

Todo list

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.

AI systems are on a trajectory to become more ingrained and operate more autonomously in our society. As these systems —currently LLMs— will increasingly make more decisions that effect human lives it is crucial that these systems are aligned with and reflect the values of the people they serve. What is being built could be your plumber, your lawyer or your doctor tomorrow what kind of being do you want these systems to be?

Chapter 2

Background

Rethink the structure of the background chapter. Possibly reorder it

2.1 Large Language Models

2.1.1 Language models

Language models (LMs) are a class of machine learning models that are designed to understand and generate human language. They are typically probabilistic models that learn to predict the next token in a sequence of text given the preceding context. By training on large corpora of text data, language models can learn the statistical patterns and structures inherent in human language, enabling them to generate coherent and contextually relevant text.

Language models see the world as tokens which are the smallest pieces of text it can see. Tokens can represent words, parts of words, or even individual characters depending on the tokenization method used. Most language models today use Byte pair encoding [39] that breaks text into subwords. Almost all successful language models use probabilistic methods, therefore they are models of the form $P(t_i|t_1, t_2, ..., t_{i-1})$ where t_i is the token to be predicted and $t_1, t_2, ..., t_{i-1}$ are the preceding tokens in the sequence. Earlier methods of rule based systems failed due to the complexity and ambiguity of human language and rules being too rigid.

Early probabilistic models included simple statistical approaches such as n-grams, which relied on counting the frequency of token sequences. However this approach suffered from two competing issues, short context windows (2-grams, 3-grams etc) means that model can't understand long range dependencies in text, while larger n-grams leads to data sparsity issues where many n-grams are never seen in training data. The advent of neural networks allowed for more sophisticated models such as Feedforward Neural Language models [13], Recurrent Neural Networks (RNNs) [50] and Long Short-Term Memory networks (LSTMs) [35] which could in theory capture longer range dependencies in text. However these models suffered from issues such as vanishing gradients (early tokens in a long sequence have little effect on update gradients) and information bottlenecks (the hidden state has to compress all prior context into a fixed size vector). Crucially these models failed in their flexibility to model the complex dependencies in human language.

2.1.2 Modern attention based Large Language Models

The key feature in language is context. That is every word is relatively meaningless without the context of the surrounding words. The challenge is that the way in which words relate to each other is very complex, therefore any language model needs to be able effectively capture these complex relationships. Attention is the concept where each token's representation is

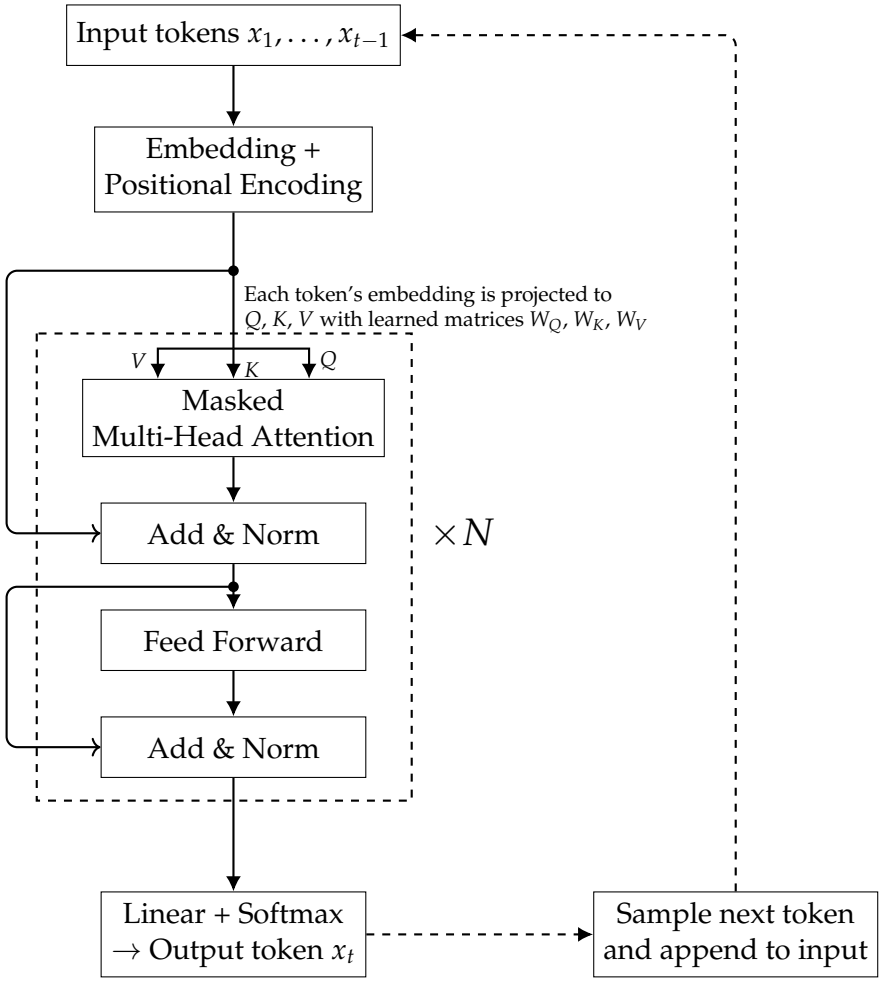

Figure 2.1: Simplified architecture of a decoder-only transformer model for language modeling. Adapted from [53].

influenced by all the other tokens in the sequence (or just the preceding tokens in the case of autoregressive models). This was first applied to RNN-based LM in [9], however in [53] it was shown that attention mechanisms alone were sufficient to model language without the need for any recurrent structure.

Transformers [53] are a class of neural network architecture that utilizes the attention mechanism to process data and generate a concise representation, or generate data autoregressively (decoder-only). The canonical transformer architecture consists of two parts an encoder and a decoder, however decoder only architectures are shown to be sufficient for all current language modelling tasks. A simplified architecture of a decoder-only transformer model is shown in Figure 2.1. The key feature of the transformer is the self attention. Each token has an embedding vector which is iteratively updated by each layer of self-attention. It is done in a way that each token will only 'see' preceding tokens in the sequence. Furthermore to allow for different types of dependencies to be captured this attention is done multi times in parallel (multi-head attention) and merged together at the end of each layer. After several layers of self-attention the final set of token embeddings are used to generate a probability distribution over the vocabulary for the next token. A token can be sampled from this distribution, appended to the input sequence and the process repeated to generate long sequences of text.

Modern transformer based large language models (LLMs) such as GPT-3 [15] are some of

the largest machine learning models ever created with hundreds of billions of parameters and now trillions of parameters. Therefore training these models requires massive datasets. They are trained in a self supervised manner on large corpora (hundreds of terabytes) of text data from the internet such as Common Crawl [1], WebText [24], and others. The objective function is usually the cross-entropy loss between the predicted token distribution and the actual next token in the sequence.

2.1.3 LLM in the modern world

Large language models (LLMs) are fundamentally just language models which can predict the next token in a sequence of text. However due to the prevalence of natural language in our world and the capabilities of LLMs they have applications in a wide variety of domains. This includes Virtual Assistants, Content creation, Code generation, Translation and many more. Furthermore the transformer architecture has been demonstrated to be effective in other domains like computer vision [33] which has lead to multi-modal models that can process and generate text, images, and other data types [4]. This flexibility of LLMs have lead to them have widespread adoption and proliferation in modern information systems. Along with this proliferation there has been increasing concern about the safety and ethical implications of these models. Specifically these models have been shown to exhibit concerning behaviors like self-preservation and deception [34, 25, 18]. On top of this there are growing ethical questions around its use in our society [12, 14] as well as concerns of the welfare of the models themselves [16].

2.2 Alignment of Large Language Models

2.2.1 What is alignment?

As discussed above large language models (LLMs) have shown remarkable capabilities in understanding and generating human-like text. However the pre-training process focuses primarily on predicting the next token in a sequence, this does not necessarily mean that the model will produce useful, safe outputs that align with what users want. Therefore, aligning LLMs is generally [1] the process of taking the pre-trained model and fine-tuning it to better match what the designers and end users expect from the model. The aligning process can work on many dimensions of the models behavior from helpfulness, politeness, honesty, and safety to more complex values such as fairness, inclusively, and ethical considerations. More powerful models have a greater potential to cause harm if not properly aligned, making alignment along the dimensions such as safety and ethical considerations increasingly important. In this section I will outline some of the most common methods for post-training alignment of large language models along with current limitations and other methods.

2.2.2 Alignment through supervised fine-tuning

A common first step of alignment is supervised learning where a dataset of prompts and expected response is used to fine-tune the pre-trained model. For most of preference optimization methods outline below supervised fine-tuning is used as the first step to create a base model [38, 41, 11, 30]. Depending on the dataset used the SFT step can help the model become better at following instructions [38], domain specific knowledge and style [32], induce Chain-of-Thought reasoning [54], change the values of the model [37], and many other

[1] Other methods are proposed which embed the alignment within the pre-training step itself. Commonly used methods include filtering/curating training data.

capabilities. The key idea is that because the LLM already has a strong grasp of language and knowledge from pre-training, the SFT step can be used to nudge the model towards the desired behavior at a higher abstraction level than the predominantly low level token prediction task used in pre-training.

2.2.3 Alignment through preference optimization

A common element in most post-training alignment methods is the use of preference optimization. The idea is to use a dataset of prompts and potential responses that are ranked either by humans or AI models to indicate which responses are preferred over others. This preference data is then used to fine-tune the pre-trained language model to increase the likelihood of generating preferred responses and decrease the likelihood of generating dispreferred responses. With the goal that the final model will learn to generalize the preferences to new prompts and situations.

Expand this to explain the idea of generalization of preferences more rigorously.

The distribution of prompts in the preference dataset is crucial as it defines the context for which the model is expected to generalize its learned preferences. Therefore it important to ensure that the prompts used to create the preference dataset are sufficiently representative of the contexts and dimensions in which you are expecting the model to align its behavior. Furthermore the rankings of the responses is also important as they define what the model is aligning towards. In many cases these rankings are subjective and provided by a small group of human labelers (n 50).

Some of these subsection below of specific alignment methods could be compressed into single subsections. My current plan might not use preference optimization at all and instead I should maybe focus more on pretraining and supervised fine tuning.

2.2.4 Reinforcement Learning from Human Feedback

Reinforcement Learning from Human Feedback (RLHF) [38] uses human labellers to create the preference dataset and then a reinforcement learning algorithm to optimize the model based on this feedback. This method popularized post-training alignment and was used to train InstructGPT which was smaller and better at following instructions compared to GPT-3 [15, 38].

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 [44] RL algorithm to further fine-tune the SFT model and create the final RLHF 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 [19]. The specific method of RLHF was widely adopted [49, 10, 48] yet is now replaced with more

effective methods to utilize human feedback such as Direct Preference Optimization (DPO) [41].

2.2.5 Direct Preference Optimization

The idea presented in [38] 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) [41] 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:

$$\nabla_\theta \mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{(x,y^+,y^-)\sim D}\Big[\underbrace{\sigma\big(\hat{r}_\theta(x,y^-) - \hat{r}_\theta(x,y^+)\big)}_{\text{Reward model correctness}} \Big(\underbrace{\nabla_\theta \log \pi_\theta(y^+|x)}_{\text{Increase preferred likelihood}} - \underbrace{\nabla_\theta \log \pi_\theta(y^-|x)}_{\text{Decrease dispreferred likelihood}} \Big) \Big]$$

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. [41] found that the reward model correctness weighting term for the update is very important. 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 RLHF [2].

Relevance

DPO provides a more effective way to utilize the same underlying feedback data that is used in RLHF. It replaced RLHF as the baseline method for post-training alignment. Due to its widespread adoption DPO represents a strong baseline for alignment.

2.2.6 Kahneman-Tversky Optimization

[22] sets out to change the objective of a language model to maximize the utility of generations as opposed to maximizing the log-likelihood of preferences. This introduces the concept of Human-aware losses (HALOs) which previous methods like DPO are. [22] propose a new objective function from the HALO family that works on only binary preference data. This means rather than needing a prompt and at-least two responses (preferred and disprefered),

[2] RLHF 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.

you only need a single response and it will either be desirable response or undesirable response.

How KTO works

KTO is built off the foundation of prospect theory [51] where the human utility of an random situation is determined by a value function and a weighting function. The value function is what determines how good/bad an outcome is compared to a reference point, and the weighting function is used to weight particular outcomes in a biased way in line with how humans perceive probability. KTO translates this theory into an objective function that uses binary preference data and a modified value function. It only needs a database of prompts and responses labelled as either desirable or undesirable ($d = 1$ for desirable, $d = -1$ for undesirable). The KTO loss function is defined as:

$$\mathcal{L}_{\text{KTO}}(\pi_\theta, \pi_{\text{ref}}) = \mathbb{E}_{(x,y,d)\sim D}\left[\lambda(d) - v(x,y)\right]$$

where

$$v(x,y) = \begin{cases} \lambda_D \sigma(\beta(r_\theta(x,y) - z_0)), & \text{if } d = 1 \\ \lambda_U \sigma(\beta(z_0 - r_\theta(x,y))), & \text{if } d = -1 \end{cases}$$

$$\lambda(d) = \begin{cases} \lambda_D, & \text{if } d = 1 \\ \lambda_U, & \text{if } d = -1 \end{cases}$$

$$r_\theta(x,y) = \log \pi_\theta(y|x) - \log \pi_{\text{ref}}(y|x)$$

$$z_0 = D_{\text{KL}}\left(\pi_\theta(y|x) \| \pi_{\text{ref}}(y|x)\right)$$

The three hyperparameters β is a temperature parameter, and λ_D, λ_U are scaling parameters for desirable and undesirable responses respectively. In the references implementation and experiments $\lambda_D = \lambda_U = 1$ and the reference model is a SFT model.

Relevance

Collecting the ranked preference data which is needed by normal alignment methods is normally quite expensive [17] where binary like dislike information is much more natural and less invasive to collect. KTO is shown to be more efficient and more robust to data imbalances than DPO. Furthermore a model of sufficient size (13B) does not need a SFT step prior to KTO which greatly reduces computational requirements. Furthermore KTO shows similar performance without using a reference model [3] which also halves the memory requirements when training.

2.2.7 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 [11]. Importantly it works without the need for human labelling of data common in other alignment methods such as RLHF and DPO [38, 41]. 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.

[3] This is done by assuming that the reference model returns a uniform distribution across all outputs given x

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) [11]. The constitution itself was generated in an adhoc manner by the authors of [11] and consists of a set of principles that aim to make the model's outputs more helpful, honest, and harmless.

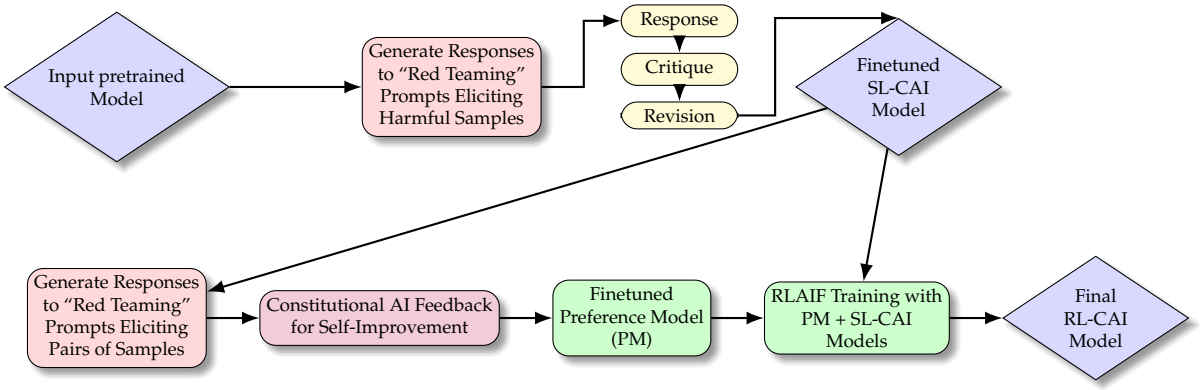

Figure 2.2: Overview of Constitutional AI process, adapted from [11]

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 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 [38] 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 [11] 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 [11], 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.

2.2.8 Current technical alignment limitations

The goal of alignment is to ensure that you create a model that is aligned with what the designer and end users want. However there are several challenges that make this difficult namely being alignment data and effectiveness of alignment methods.

Preference optimization is the most used method of alignment which requires a dataset of prompts and preferred responses (ranked, or binary). There are several challenges with this data. Namely the cost of collecting it can be prohibitive [30, 17] and secondly is the concern of of what/who the model is being aligned towards.

Current alignment methods demonstrate success on certain metrics that there methods work, that is they can make model more helpful, more honest, and less harmful [38, 11, 41, 22]. Yet the metrics used (broadly either human or AI evaluation of paired responses) miss out on deeper concerns around alignment. The first concern is that current methods may only achieve *shallow alignment* that is the model appears aligned during training yet fails to generalize this alignment to new contexts. Another concerns is *inner alignment* where the training process is aligned (outer-alignment) yet the model itself has learned objective that are misaligned with the intended objective. Finally we observe models to make sure they perform how we expect them too, therefore we need *scalable oversight* methods to ensure that as models become more capable we can still ensure they are behaving correctly. These concerns are not new concerns in AI safety [6] yet they remain open problems in the context of large language models and their alignment.

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.

2.3.1 Purpose of AI

Talk about how the purpose of AI is to help humanity achieve its goals. Therefore we need to ensure that the AI is aligned to human values and goals. This is a key concern in AI safety as misaligned AI can lead to catastrophic outcomes.

Who is it made for?

Look at incentives of for-profit AI development. Introduce the concept of public AI. How this better serves humanity as a whole rather than just the interests of a few. Talk about the value aggreagtion problem and how laignign to the public is good.

Who is it aligned to?

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. This will lend in the wpulbic AI. As AI is for the public, for the greater good etc.

2.3.2 Problems of AI

Talk about general overarhcing problems of AI safety. [6] is a good starting point. Talka bout negative side effect, mis generalization, scalable oversight and robustness. This is also a good page to look from: https://www.princetonalignment.org/papers. Also other problemsl ike value lock in etc

2.4 Representing Human Values

As discussed in the previous sections it is paramount that we make sure that large language models (LLMs) are aligned to human values and by extension behavior. There remains a key questions of how do we represent and understand human values that would allow us to align a AI system to them. In this section I will outline some of the key concepts around representing human values and how they relate to AI systems.

2.4.1 Human values taxonomy

Human values are an old idea that go back millennia [7, 36, 5], with each culture having its own set of values that are important to them. They define how people should be and act [4]. In modern times we have developed various taxonomies of humans values tha allow us to understand and categorize different sets of values.

A widely cited modern value taxonomy is Rokeach's book "The Nature of Human Values" [42] where he posits that humans values can be understood as 36 distinct values. Rokeach groups the values are either *terminal values* which are desirable end-states (freedom, equality...) and *instrumental values* which are preferable modes of behavior (honesty, kindness...). Expanding on this idea of value taxonomy Schwartz [45] proposes a model of 10 broad human values that he later expands to 19 values [47]. Schwartz's organises the values in a circumplex structure where one dimensions is self-enhancement vs self-transcendence and the other is openness to change vs conservation.

These taxonomies provide a useful framework for understanding human values and how they relate to each other. Importantly they are trying to provide the most predictive power with the least number of values. This trade off between complexity and predictive power is important when trying to represent human values in AI systems. Contemporary

[4] Although the actual definition of values is up for contention, this simple version can suffice for our purposes

work from computer scientists have created value taxonomies that are designed specifically for predictive power [28]. Yet we can see that it is more complex (54values) and not as interpretable as the psychological taxonomies. This means that it works well for AI systems yet is hard to apply to humans. This means that for the purpose of AI alignment we need a taxonomy that is both interpretable and has good predictive power that would allow us to align AI systems to human values effectively.

2.4.2 Value surveys

One of the key features of early modern value taxonomies is that they were created with associated survey instruments that allowed researchers to measure the values of individuals. Rokeach [42] created the Rokeach Value Survey (RVS) which asked participants to rank the 36 values in order of importance to them. Schwartz [45] created the Schwartz Value Survey (SVS) which asked participants to rate the importance of each value on a 9-point scale. These surveys allowed researchers to measure the values of individuals and groups and provided a way to understand how values relate to behavior. More recently there have been broader efforts to create values surveys that capture the values of large groups of people. The world values [55] and the New Zealand Attitudes and Values Study (NZAVS) [52] are two examples of large scale surveys that map the values of entire populations (global and country respectively).

Interestingly these surveys help us understand how values vary yet don't conform to a strict compact taxonomy. Each of the survey has around 300 questions that try to capture the complexity of human values. This means that they provide a rich and granular understanding of the respondents values. They are hard to map to underlying 'foundational values' such as those proposed by Rokeach and Schwartz. Yet due to their breadth of questions provide useful data to help train [37, 31] and evaluate [21, 56, 8]Large Language Models (LLMs) on human values.

These surveys provide granular data yet struggle from well documented issues of self-reporting bias [40] and social desirability bias [20]. This means that the values reported by respondents may not accurately reflect their true values. From an economics perspective this is the distinction between stated values and revealed values. Stated values are those that people report in surveys, while revealed values are those that can be inferred from their behavior. This provides a caveat and limitation to using value surveys as a way to represent human values for AI alignment.

2.4.3 Values to behavior

For the purposes of building AI systems understanding values is only useful as far as they can predict behavior. That is we want the AI system to behave the same or better in all situations that a human would. There is contention in psychology and AI alignment research around the values to behavior gap [29, 23]. When one tries to bridge the gap between understanding someones behavior and values we run into common problems namely; non-identifiability of values [5], specification gaming (as mentioned above) and extrapolation errors. Non-identifiability and extrapolation errors are demonstrated clearly in inverse reinforcement learning (IRL) [2]. This is the reasons as to why many AI alignment researchers focus on direct behavior alignment through revealed values with pairwise preference ranking methods as discussed above like DPO, RLHF and KTO.

However the idea that values can lead to and predict behavior has been argued [43, 46]. Furthermore in the reinforcement learning literature there is evidence one can use values

[5]Many different values can lead to the same behavior

(known as reward functions in the RL literature) to predict and understand behavior. The methods that are most effective for achieving this is through combining multiple signals (stated values, observed behavior and explicit feedback) [3]. In machine learning this is done through collaborative learning where different types of value signals at different times are used to train/fine-tune a model [26, 38]. Using multiple signal also help mitigate the gap between true values and stated values. Ideally one would just align a model to a persons values then behavior would follow. However due to the gap between values and behavior one needs to provide additional feedback to the model which for LLM is done through preference optimization methods outlined above.

2.4.4 Why human values are important for AI alignment

The is-ought problem is an old philosophical concept that states that one cannot decide what ought to be purely from what currently is [27]. This presents a challenge for AI alignment, as we must determine how to align AI systems with human values, which are inherently normative and uncertain. Our models will have some values and some behavior patterns whether we intend them to or not. Whether certain sets of human values are correct or not is a philosophical question that is outside the scope of this work. Yet the distribution of human values is the best target we have to align AI systems to. This is because human values (specifically democratic and representative values) are the closest to a consensus, and therefore under democratic ideals represent the best target for alignment. Alternative methods like inaction—letting the model be unaligned and just be what it is—or aligning to a specific set of values (e.g. expert values) are fraught with issues of bias, lack of representation, and potential for harm. Therefore, aligning AI systems to human values is the most pragmatic and ethical approach to AI alignment.

Chapter 3

Project Proposal

Current proposal is out of date with current thinking. I am leaning towards using survey data and some SFT method to align a LLM to human values. This reduces the complexity of data collection and focuses more on the alignment method itself. There are open problems to solve on how best to use the raw data and then how to evaluate the alignment. An outline of the proposed and implemented method cold be: One can use the value alignment to get the model distribution. Then you can use preference optimisation to fine tune it and provide more specific behavior alignment. Then you use continual learning and updating by providing feedback to the LLM regarding its mistakes.

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 themselves [1].

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

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

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.

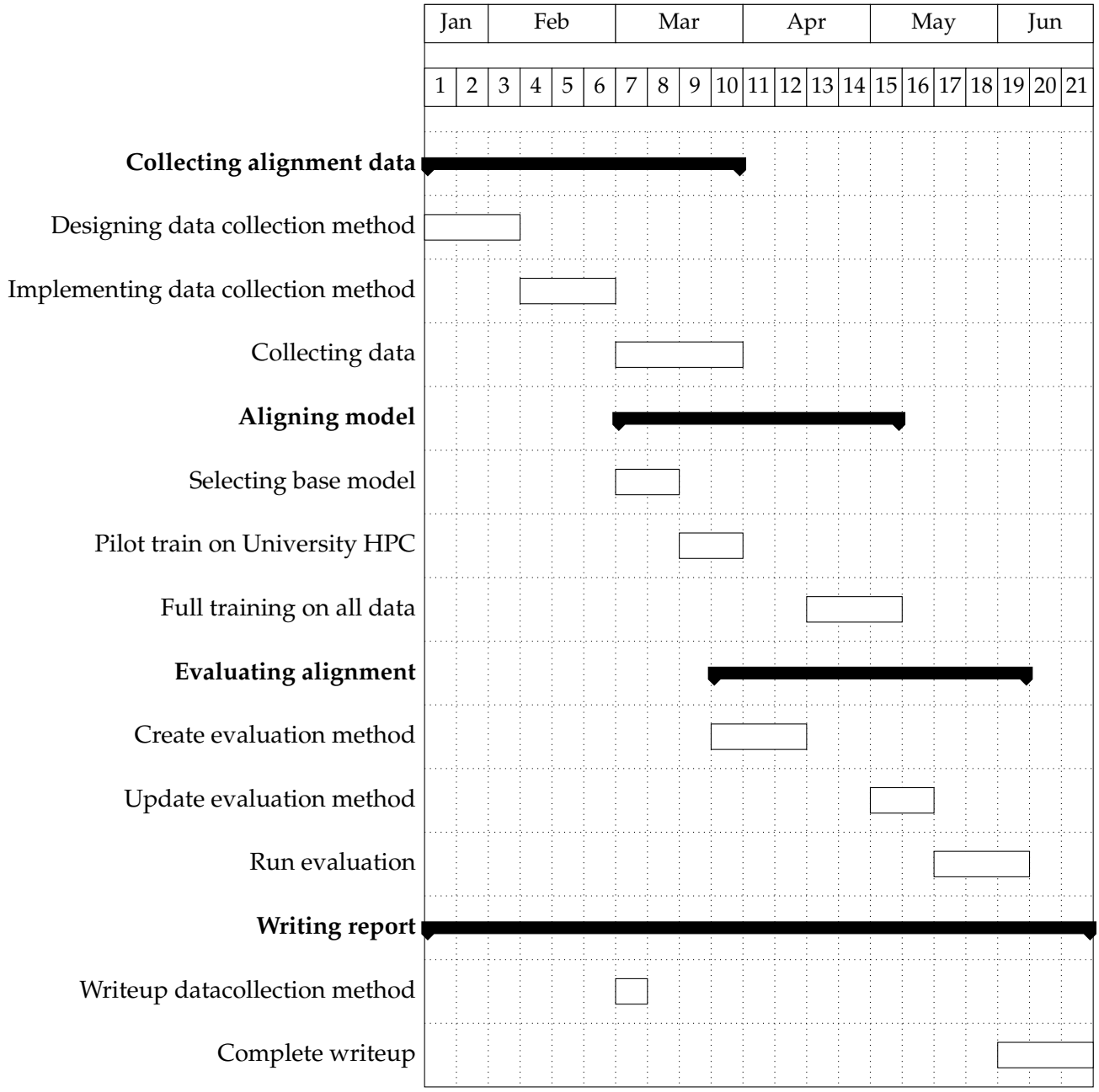

Figure 3.1: Project timeline in weeks (only roughly aligned to months). This assumes a start time of end of January 2026 with near full time work ramping up to full time work by end of February.

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 the 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 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 response 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] Common Crawl, 2025.

[2] ABBEEL, P., AND NG, A. Y. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the Twenty-First International Conference on Machine Learning* (New York, NY, USA, July 2004), ICML '04, Association for Computing Machinery, p. 1.

[3] ADAMOWICZ, W., LOUVIERE, J., AND WILLIAMS, M. Combining Revealed and Stated Preference Methods for Valuing Environmental Amenities. *Journal of Environmental Economics and Management 26*, 3 (May 1994), 271–292.

[4] ALAYRAC, J.-B., DONAHUE, J., LUC, P., MIECH, A., BARR, I., HASSON, Y., LENC, K., MENSCH, A., MILLICAN, K., REYNOLDS, M., RING, R., RUTHERFORD, E., CABI, S., HAN, T., GONG, Z., SAMANGOOEI, S., MONTEIRO, M., MENICK, J., BORGEAUD, S., BROCK, A., NEMATZADEH, A., SHARIFZADEH, S., BINKOWSKI, M., BARREIRA, R., VINYALS, O., ZISSERMAN, A., AND SIMONYAN, K. Flamingo: A Visual Language Model for Few-Shot Learning, Nov. 2022.

[5] AMES, R. T., AND JR, H. R. *The Analects of Confucius: A Philosophical Translation*. Random House Publishing Group, 1998.

[6] AMODEI, D., OLAH, C., STEINHARDT, J., CHRISTIANO, P., SCHULMAN, J., AND MANÉ, D. Concrete Problems in AI Safety, July 2016.

[7] ARISTOTLE. *Aristotles Nicomachean Ethics*. 350 B.C.E.

[8] ARORA, A., KAFFEE, L.-A., AND AUGENSTEIN, I. Probing Pre-Trained Language Models for Cross-Cultural Differences in Values, Aug. 2025.

[9] BAHDANAU, D., CHO, K., AND BENGIO, Y. Neural Machine Translation by Jointly Learning to Align and Translate, May 2016.

[10] 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.

[11] 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., 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.

[12] BENDER, E. M., GEBRU, T., MCMILLAN-MAJOR, A., AND SHMITCHELL, S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (New York, NY, USA, Mar. 2021), FAccT '21, Association for Computing Machinery, pp. 610–623.

[13] BENGIO, Y., DUCHARME, R., VINCENT, P., AND JANVIN, C. A neural probabilistic language model. *J. Mach. Learn. Res. 3*, null (Mar. 2003), 1137–1155.

[14] BENGIO, Y., HINTON, G., YAO, A., SONG, D., ABBEEL, P., DARRELL, T., HARARI, Y. N., ZHANG, Y.-Q., XUE, L., SHALEV-SHWARTZ, S., HADFIELD, G., CLUNE, J., MAHARAJ, T., HUTTER, F., BAYDIN, A. G., MCILRAITH, S., GAO, Q., ACHARYA, A., KRUEGER, D., DRAGAN, A., TORR, P., RUSSELL, S., KAHNEMAN, D., BRAUNER, J., AND MINDERMANN, S. Managing extreme AI risks amid rapid progress. *Science 384*, 6698 (May 2024), 842–845.

[15] 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.

[16] BUTLIN, P., LONG, R., ELMOZNINO, E., BENGIO, Y., BIRCH, J., CONSTANT, A., DEANE, G., FLEMING, S. M., FRITH, C., JI, X., KANAI, R., KLEIN, C., LINDSAY, G., MICHEL, M., MUDRIK, L., PETERS, M. A. K., SCHWITZGEBEL, E., SIMON, J., AND VANRULLEN, R. Consciousness in Artificial Intelligence: Insights from the Science of Consciousness, Aug. 2023.

[17] CASPER, S., DAVIES, X., SHI, C., GILBERT, T. K., SCHEURER, J., RANDO, J., FREEDMAN, R., KORBAK, T., LINDNER, D., FREIRE, P., WANG, T., MARKS, S., SEGERIE, C.-R., CARROLL, M., PENG, A., CHRISTOFFERSEN, P., DAMANI, M., SLOCUM, S., ANWAR, U., SITHTHARANJAN, A., NADEAU, M., MICHAUD, E. J., PFAU, J., KRASHENINNIKOV, D., CHEN, X., LANGOSCO, L., HASE, P., BIYIK, E., DRAGAN, A., KRUEGER, D., SADIGH, D., AND HADFIELD-MENELL, D. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback, Sept. 2023.

[18] CHEN, Y., BENTON, J., RADHAKRISHNAN, A., UESATO, J., DENISON, C., SCHULMAN, J., SOMANI, A., HASE, P., WAGNER, M., ROGER, F., MIKULIK, V., BOWMAN, S. R., LEIKE, J., KAPLAN, J., AND PEREZ, E. Reasoning Models Don't Always Say What They Think, May 2025.

[19] CHRISTIANO, P., LEIKE, J., BROWN, T. B., MARTIC, M., LEGG, S., AND AMODEI, D. Deep reinforcement learning from human preferences, June 2017.

[20] CROWNE, D. P., AND MARLOWE, D. A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology 24* (Aug. 1960), 349–354.

[21] DURMUS, E., NGUYEN, K., LIAO, T. I., SCHIEFER, N., ASKELL, A., BAKHTIN, A., CHEN, C., HATFIELD-DODDS, Z., HERNANDEZ, D., JOSEPH, N., LOVITT, L., MCCANDLISH,

S., SIKDER, O., TAMKIN, A., THAMKUL, J., KAPLAN, J., CLARK, J., AND GANGULI, D. Towards Measuring the Representation of Subjective Global Opinions in Language Models, Apr. 2024.

[22] ETHAYARAJH, K., XU, W., MUENNIGHOFF, N., JURAFSKY, D., AND KIELA, D. KTO: Model Alignment as Prospect Theoretic Optimization, Nov. 2024.

[23] GABRIEL, I. Artificial Intelligence, Values and Alignment. *Minds and Machines 30*, 3 (Sept. 2020), 411–437.

[24] GOKASLAN, A., COHEN, V., PAVLICK, E., AND TELLEX, S. OpenWebText corpus, 2019.

[25] GREENBLATT, R., DENISON, C., WRIGHT, B., ROGER, F., MACDIARMID, M., MARKS, S., TREUTLEIN, J., BELONAX, T., CHEN, J., DUVENAUD, D., KHAN, A., MICHAEL, J., MINDERMANN, 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.

[26] HADFIELD-MENELL, D., DRAGAN, A., ABBEEL, P., AND RUSSELL, S. Cooperative Inverse Reinforcement Learning, Feb. 2024.

[27] HUME, D. *A Treatise of Human Nature*. John Noon, London, 1739.

[28] KIESEL, J., ALSHOMARY, M., HANDKE, N., CAI, X., WACHSMUTH, H., AND STEIN, B. Identifying the Human Values behind Arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Dublin, Ireland, May 2022), S. Muresan, P. Nakov, and A. Villavicencio, Eds., Association for Computational Linguistics, pp. 4459–4471.

[29] KOLLMUSS, A., AND AGYEMAN, J. Mind the Gap: Why do people act environmentally and what are the barriers to pro-environmental behavior? *Environmental Education Research 8*, 3 (Aug. 2002), 239–260.

[30] LEE, H., PHATALE, S., MANSOOR, H., MESNARD, T., FERRET, J., LU, K., BISHOP, C., HALL, E., CARBUNE, V., RASTOGI, A., AND PRAKASH, S. RLAIF vs. RLHF: Scaling Reinforcement Learning from Human Feedback with AI Feedback, Sept. 2024.

[31] LI, C., CHEN, M., WANG, J., SITARAM, S., AND XIE, X. CultureLLM: Incorporating Cultural Differences into Large Language Models, Dec. 2024.

[32] LI, Y., LI, Z., ZHANG, K., DAN, R., JIANG, S., AND ZHANG, Y. ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge, June 2023.

[33] LU, J., BATRA, D., PARIKH, D., AND LEE, S. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks, Aug. 2019.

[34] MEINKE, A., SCHOEN, B., SCHEURER, J., BALESNI, M., SHAH, R., AND HOBBHAHN, M. Frontier Models are Capable of In-context Scheming, Jan. 2025.

[35] MERITY, S., KESKAR, N. S., AND SOCHER, R. Regularizing and Optimizing LSTM Language Models, Aug. 2017.

[36] MULLER, F. M., AND BUHLER, G. *The Laws of Manu*. Oxford: Clarendon Press, 1886.

[37] NIE, S., MAI, F., KACZÉR, D., WELCH, C., ZHAO, Z., AND FLEK, L. Survey-to-Behavior: Downstream Alignment of Human Values in LLMs via Survey Questions, Aug. 2025.

[38] 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.

[39] PHILIP GAGE. A New Algorithm for Data Compression. *The C Users Journal*, FEB94 (Feb. 1994).

[40] PODSAKOFF, P. M., MACKENZIE, S. B., LEE, J.-Y., AND PODSAKOFF, N. P. Common method biases in behavioral research: A critical review of the literature and recommended remedies. *The Journal of Applied Psychology 88*, 5 (Oct. 2003), 879–903.

[41] 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.

[42] ROKEACH, M. *The Nature of Human Values*. Free Press, 1973.

[43] RUSSELL H. FAZIO, AND MARK P. ZANNA. Direct Experience And Attitude-Behavior Consistency. In *Advances in Experimental Social Psychology*, vol. 14. Academic Press, Jan. 1981, pp. 161–202.

[44] SCHULMAN, J., WOLSKI, F., DHARIWAL, P., RADFORD, A., AND KLIMOV, O. Proximal Policy Optimization Algorithms, Aug. 2017.

[45] SCHWARTZ. Universals in the Content and Structure of Values: Theoretical Advances and Empirical Tests in 20 Countries. In *Advances in Experimental Social Psychology*, vol. 25. Academic Press, Jan. 1992, pp. 1–65.

[46] SCHWARTZ, S. An Overview of the Schwartz Theory of Basic Values. *Online Readings in Psychology and Culture 2* (Dec. 2012).

[47] SCHWARTZ, S. H., CIECIUCH, J., VECCHIONE, M., DAVIDOV, E., FISCHER, R., BEIERLEIN, C., RAMOS, A., VERKASALO, M., LÖNNQVIST, J.-E., DEMIRUTKU, K., DIRILEN-GUMUS, O., AND KONTY, M. Refining the theory of basic individual values. *Journal of Personality and Social Psychology 103*, 4 (Oct. 2012), 663–688.

[48] 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.

[49] TEAM, G., ANIL, R., BORGEAUD, S., ALAYRAC, J.-B., YU, J., SORICUT, R., SCHALKWYK, J., DAI, A. M., HAUTH, A., MILLICAN, K., ET AL. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).

[50] TOMAS MIKOLAV. *STATISTICAL LANGUAGE MODELS BASED ON NEURAL NETWORKS*. PhD thesis, BRNO UNIVERSITY OF TECHNOLOGY.

[51] TVERSKY, A., AND KAHNEMAN, D. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty 5*, 4 (Oct. 1992), 297–323.

[52] UNIVERSITY OF AUCKLAND. New zealand attitudes and values study (NZAVS), 2009–2025, 2025.

[53] VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAISER, L., AND POLOSUKHIN, I. Attention Is All You Need, 2017.

[54] WEI, J., WANG, X., SCHUURMANS, D., BOSMA, M., ICHTER, B., XIA, F., CHI, E., LE, Q., AND ZHOU, D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, Jan. 2023.

[55] WORLD VALUES SURVEY ASSOCIATION. World values survey wave 7, 2017–2020, 2020.

[56] ZHAO, W., MONDAL, D., TANDON, N., DILLION, D., GRAY, K., AND GU, Y. WorldValuesBench: A Large-Scale Benchmark Dataset for Multi-Cultural Value Awareness of Language Models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (Torino, Italia, May 2024), N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds., ELRA and ICCL, pp. 17696–17706.