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 24, 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.

Contents

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 [56] 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) [69] and Long Short-Term Memory networks (LSTMs) [50] 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

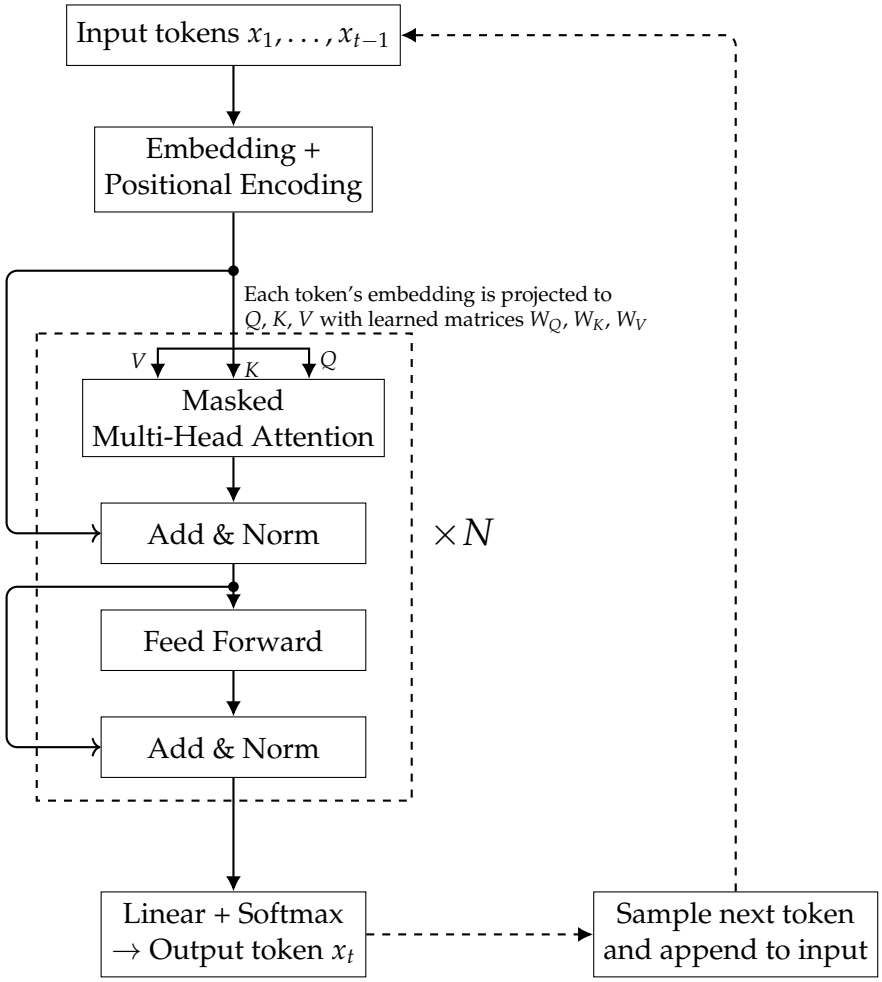

Figure 2.1: Simplified architecture of a decoder-only transformer model for language modeling. Adapted from [73].

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 [73] it was shown that attention mechanisms alone were sufficient to model language without the need for any recurrent structure.

Transformers [73] 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 [16] 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 [30], 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 [48] 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 [49, 32, 19]. 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 [17].

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 [55, 59, 11, 44]. Depending on the dataset used the SFT step can help the model become better at following instructions [55], domain specific knowledge and style [47], induce Chain-of-Thought reasoning [74], change the values of the model [54], 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 usually (but not always) in the form of pairwise comparisions. That is given a prompt x, a pair of responses i and j the labellers (AI or human) indicate which response is perferred for example $i \succ j$ indicates response i is preferred over response j. If we have a dataset of these pairwise comparisons we can use the Bradley-Terry model [15] to create a maximum likelihood loss fuctuion

$$P(i \succ j \mid x) = \frac{\pi_i}{\pi_i + \pi_j}, \tag{2.1}$$

$$\mathcal{L}_{\text{BT}} = -\sum_{(x,i,j)\sim D} \log P(i \succ j \mid x) \tag{2.2}$$

$$= -\sum_{(x,i,j)\sim D} \log \frac{\pi_i}{\pi_i + \pi_j} \tag{2.3}$$

Where π_i and π_j are the scores of responses i and j respectively given prompt x. The way these scores are calculated depends on the specific preference optimization method used.

The success of preference optimization methods relies on the ability of modern LLMs to generalize learned preferences to new contexts and prompts not seen during training. 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).

2.2.4 Reinforcement Learning from Human Feedback

Reinforcement Learning from Human Feedback (RLHF) [55] 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 [16, 55].

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. The preference model is trained using a Bradley -Terry loss function where the scores π_i is equal to $\exp(\text{PM}_\theta(x, i))$. This is fed into 2.3 to create the loss function for the preference model.

Finally the fitted PM is used as a reward signal in the PPO [62] 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 [20]. The specific method of RLHF was widely adopted [68, 10, 66] yet is now replaced with more effective methods to utilize human feedback such as Direct Preference Optimization (DPO) [59].

2.2.5 Direct Preference Optimization

The idea presented in [55] 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) [59] 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 seperate preference model DPO uses the SFT model as a reference model and defines the reward function $\hat{r}_\theta(x, y) = \log \pi_\theta(y|x) - \log \pi_{SFT}(y|x)$. Then be some derivation [59] it can be shown that the following loss function gradient can be used to optimize the Bradley Terry model directly on the preference data:

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

Where σ is the sigmoid function, and (x, y^+, y^-) are the prompt, preferred response, and dis-preferred response respectively. [59] 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].

[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.

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

Traditional preference optimization methods such as RLHF and DPO set out with the objective of making the model maximize the likelihood of preferred responses over dispreferred responses. KTO [28] sets out to change the objective of a language model to maximize the perceived utility of generations, which is how useful the generation is for the end user. The key being perceived utility is not the same as actual utility due to cognitive biases discussed in the follow sections. Ethayarajh et Al introduces the concept of Human-aware losses (HALOs) which optimize for what human perceive as the best outcome. It is shown that DPO and PPO are both HALOs which can help explain their success. Furthermore Ethayarajh et Al proposes a new objective function from the HALO family that works on only binary preference data[3], making data collection easier and more natural.

Prospect Theory

KTO is built off the foundation of prospect theory [70] which is a theory to explain how humans make decision when dealing with uncertainty and risk [4]. It explains several cognitive biases that humans have when making decision involving risk, namely loss aversion (losses hit harder than gains), probability weighting (overweight small probabilities and underweight large probabilities), and reference dependence (we evaluate outcomes relative to a reference point which is usually the status quo). In relation to a LLM this means that when a human is evaluating the response of the model they are going to perceive the quality (utility) of the response in a biased way, so using a objective function that takes into account (Human-aware) these biases should lead to better alignment.

Prospect theory does introduce several key concepts. Firstly is each random variable Z (the LLM output) has a utility which is determined by $\sum_{z \in Z} w(z)v(z - z_0)$. Secondly is the weighting function $w(z)$ which is used to weight the probabilities of each outcome z in a biased way. Finally is the value function $v(z - z_0)$ which determines how good/bad an outcome is compared to a reference point z_0, this also will also include some bias from the individuals perspective.

How KTO works

KTO applies the idea of prospect theory by modifying the canonical utility function (shown in 2.2) provided in [70] to work for a language model fine tuning objective function. Namely it does his by omitting the weighting function (as users only see the final response, not the full probability distribution), making some minor adjustments for computational stabilty and using the Kullback-Leibler divergence [43] (between current model and reference model which is usually the SFT model) as a dynamic reference point. Notably this means that KTO only needs a database of prompts and responses labelled as either desirable or undesirable

[3] Given a prompt and response a labellers just need to label the response as desirable or undesirable

[4] For example why when humans are given the choice between \$50 and a 50% chance to win \$100, they often prefer the sure \$50, when they are both worth the same amount. Furthermore this effect is still present when the gamble is worth even more (\$110) [39]

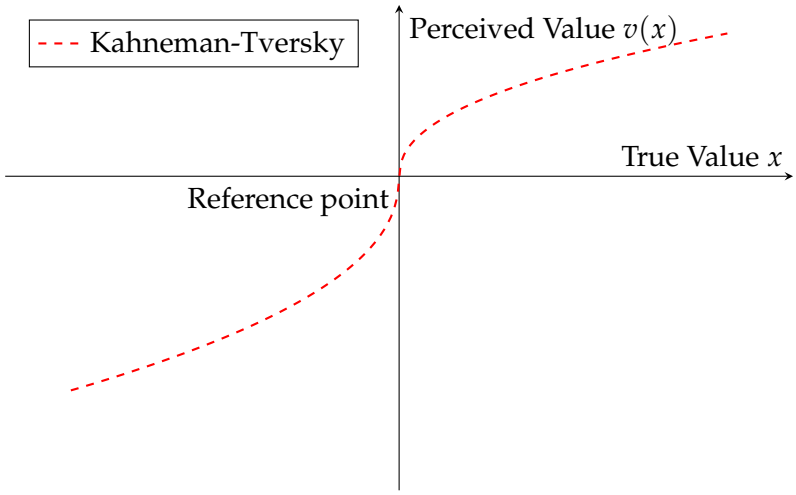

Figure 2.2: Implied human value functions showing loss aversion: concave for gains (risk averse) and convex for losses (risk seeking). Adapted from [70], with a change of terminology to match the context of value alignment and exaggeration for illustration.

($d = 1$ for desirable, $d = -1$ for undesirable) where the reference point is neutral (0). 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

$$v(x,y) = \begin{cases} \underbrace{\lambda_D \sigma(\beta(r_\theta(x,y) - z_0))}_{\text{increase probability above reference point}} & , \quad \text{if } d = 1 \\ \underbrace{\lambda_U \sigma(\beta(z_0 - r_\theta(x,y)))}_{\text{decrease probability below reference point}} & , \quad \text{if } d = -1 \end{cases}$$

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

$$r_\theta(x,y) = \log \pi_\theta(y|x) - \log \pi_{\text{ref}}(y|x) \quad \text{Same reward model as in DPO}$$

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

The three hyperparameters β is a risk aversion 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

KTO offers several key logistical and theoretical advantages over prior preference optimization methods, that make it an attractive method for alignment. Firstly the logistical advantage. Collecting the ranked preference data which is needed by normal alignment methods is normally quite expensive [18] 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 [5] which also halves the memory requirements when training. Secondly are the theoretical advantages. KTO is the first alignment method to explicitly take into account human cognitive biases when optimizing for alignment.

Another key point is the concept of applying prospect theory to alignment. Ethayarajh et Al apply it in the context of preference optimization, however propect theory and the biases in human cognition can be applied more broadly than that in alignment.

There is something here yet I am not quite sure. One could change the axises in 2.2 to represent the concept of actual values and perceived values (however this feels like trying to shoehorn it in too much when teh concept of revealed preferences is more appropriate). Another idea is that prospect theroy explains how users wil provie feedback ina biased way and we should actualy

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 [55, 59]. 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) [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.

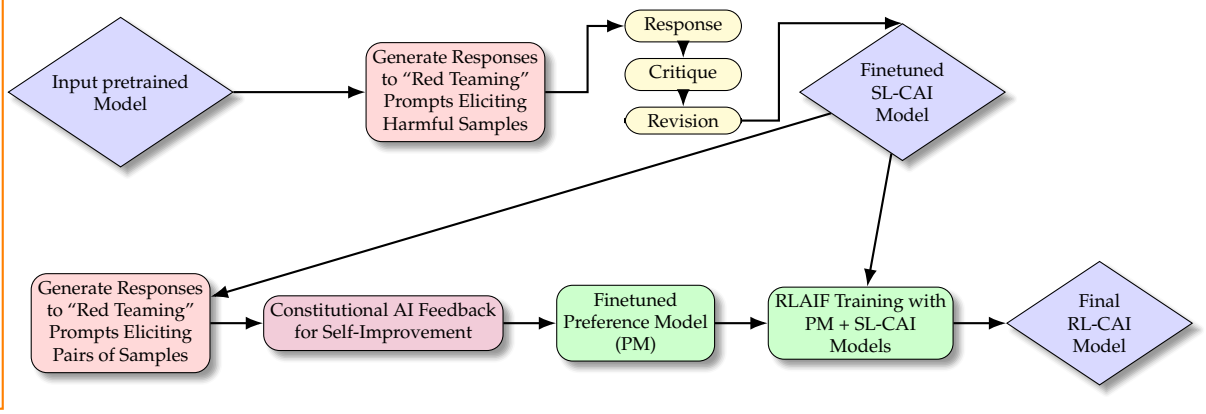

Figure 2.3: Overview of Constitutional AI process, adapted from [11]

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

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

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 [55] to generate the final model RL-CAI. However the use of RLHF is not needed and other more modern methods like KTO and DPO can be employed 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). Collective Constitutional AI [36] is the process of generating the constitution through public consultation.

2.2.8 Evaluation of Large Language Model alignment

Talk about how we really care about how the flow on effects are. I.e societal impacts, individual impacts of the model. Yet what we are measuring are single model ouputs (in general). Talk about how we can measure these flow on effects better. Talk about the limitations of current evaluation methods.

2.2.9 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. Furthermore there is also a prevalent background concern of what/who the model is being aligned towards.

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 [44, 18] 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 [55, 11, 59, 28]. 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 Desiderata for AI

Alignment is a key concern in AI safety as we want to ensure that AI systems act and behave in a way that aligns with what humans want. More broadly we are at a transitional point in the history of Artificial Intelligence where the technology, systems and institution we build around AI have the potential to shape the future of humanity. Therefore it is paramount that we understand the context around AI, ways in which AI could bring about net negative change and the way in which we would like AI to built and integrated into society.

2.3.1 Goal of AI

Anytime we add something to the world we should ensure that it is worth adding and going to provide at least net positive benefit to humanity. Measuring net positive benefit is hard and requires a metric that we can measure the positive and negative impacts of AI on humanity with. A complete discussion and outline of such a metric is out of scope for this section, furthermore a single perfect "goal" is not likely to exist [6] and if it did wouldn't be a useful goal due to Goodhart's law [31]. However we can outline some high level ideas of what a good goal for the concept of AI could be, with the caveat and understanding that this is not a perfect goal and simply some high level guidance.

A good starting point for a "goal of AI" is the idea of human flourishing [72] that is split into 5 domains that are near universally agreed upon as good things for humans to have; happiness and life satisfaction, mental and physical health, meaning and purpose, character and virtue, close social relationships.

Definition 1 *The goal of AI should be to help maximize the median human flourishing across all people currently alive and in the future.* [6]

We can see the utility of this goal as it is broad enough to allow each individuals own interpretation of what human flourishing means to them, while still allowing us to understand harms and benefits of various AI systems. More discussion is found in a later section on how to understand human values.

Feel like I need to add something more here that outlines what is currently missing from the goal of AI. Like a little critique of the status quo.

[6] Including non humans at the root level of the goal is plausible. However one could reach the same conclusions of animal welfare by maximizing the human flourishing of "animal concerned" humans.

2.3.2 Stakeholders of AI

if we are on the precipice of a new technological revolution we need to understand what and importantly who is at stake. artificial intelligence has the potential to impact everyone on the planet, an american firm that utilizes ai to handle customer service can take away jobs from urban workers in india. it can also have a impact through time, an ai paperclip factory that turns everyone into paperclips would extinguish humanity in the future and remove all possible positive future for humanity. ignoring certain groups of humans when designing and building ai systems leads to optimization for the benefit of one group over the other. pressingly this happens when we ignore future people and take short term gains and push the costs and risks onto future generations or when we extract value from marginalized groups to benefit a select few [7].

There are many stakeholder of AI and to ensure that they are all represented we need to have the concepts of non-exclusion and non-domination. This means that no group of people should be excluded from the benefits of AI and no group of people use of AI should dominate over another group of people use of AI. When the groups are globally and temporally separated strict adherence to these principles could ensure that AI is atleast evenly distributed across all people.

The problem however is when not all the stakeholders of AI are considered or empowered. Looking at the example of the customer service AI which will be designed, developed and deployed by a company for use by other companies. In a fair market it will make both companies rich at the expense of the workers who have lost their jobs. AI systems are overwhelming targeting companies as their 'users' which leads to companies driving the development and deployment of AI systems with disregard of the negative externalities of the true stakeholders.

2.3.3 Alignment of AI

The actions of a few can impact the many. AI is a technology that can multiple the impact of ones actions. Therefore we need to ensure that everything around AI is aligned to the best interests of all stakeholders of AI. Yet due to no universal truth to govern what is best for all stakeholders we need to have AI that is aligned to the collective best interests of all stakeholders of AI. It is unclear what the method of collective alignment would be. Different subgroups of the stakeholders have different values and goals, aligning collectively therefore means finding a way to have a target which can be done in multiple ways.

1. **Pluralism:** Different groups will have different values. We have no ground truth to guide us on which values are better than the other. Therefore we should align models to be pluralistic in nature and make decisions that reflect this diversity along with the uncertainty about which values are most important. This can result in multiple "aligned" models or one model that can reflect different values in different contexts. Most aligned with the idea of agnostic democracy [75].

2. **Aggregation:** With the different sets of values from different groups we can aggregate these values to find a middle ground (naive averaging) however this can lead to poor compromises and tyranny of the majority [52, 25]. Therefmore more complex aggregation that weight different groups values by impact, vulnerability and importance can be used to find a better middle ground.

[7] like modern day colonialism where the worlds rich extract data, rare earth materials and labour from the worlds poor and use it for their own gain. [23]

3. **Conglomeration:** If one assumes that all people can mostly agree than we can use idea from deliberative democracy [33] to find a set of values that all people can agree on [8]. This is where methods like citizen assemblies and tools like polis [22] can be used to find a set of values that all people can agree on.

Regardless of the method of aligning to the stakeholders, the current goal of AI is misaligned with this definition as the goal of most AI systems is to maximize profit for companies and shareholders. It shouldn't take excessive development to understand the negative externalities of profit motives [67]. The problem is not that GPT-4 itself is aligned to the goal of profit maximization, but that is second order effects (company share holder profit) that cause the development and deployment of the AI system to maximize profit. A related demonstration of this misalignment is the proliferations of recommender systems. These machine learning models that started with the intention of helping users find content they enjoy, relevant information have turned into systems that prioritize engagement (and therefore ad revenue) over user flourishing [51]. We can see that the incentives of the institutions and frameworks that built the systems *captured* the original incentive of the AI system which could of been "help the user", "show the user something they will enjoy". This is why it is important not just to align the AI system itself, but also the institutions and systems around AI to be aligned to human values and goals. Full stack alignment [27] is this idea exactly where you need to make sure that individual models, oversight models and everything in the middle are all aligned to the target that we want.

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 public AI. As AI is for the public, for the greater good etc.

2.3.4 How to do AI

If AI has any chance of achieving the goal of maximum average human flourishing we need to ensure that it is done in a way that is aligned to this goal. Full stack alignment is the idea of this however each part of the stack needs to be defined. A key idea here however is that due to objective capture the higher up the stack the alignment is the more influential the alignment is. Therefore aligning the institutions and frameworks which govern and build AI systems has the highest leverage for ensuring that AI is built in a way that is aligned to human values and goals.

Leaving AI development to private companies driven by profit might be effective in some ways, drive rapid development yet it is fundamentally misaligned with the goal of maximizing human flourishing (especially in the long term). A better framework and systems around AI development is needed. Currently a democratic framework rules the world as the most progressive and forward thinking way to govern large groups of people.

> "Democracy is the worst form of government, except for all the other forms that have been tried from time to time."[21].

Even though democracy has it flaws it is arguably the most admissible way to govern. Therefore using democratic ideals and frameworks to also govern, spearhead and deploy AI is a good starting point for ensuring that the goal of AI is upheld.Operationally this means making artificial intelligence a public good that is governed and operated by the public for the public. [58]

[8] All people agree can be operationally defined as "maximum equal approval" [38]

Public AI provides a good that is both non-dominate and non-exclusionary. This method of making it a public good provides has three key features; public access, public accountability and permanent availability. Public access ensure that the opportunities to use AI and receive its benefits are available to all people. Public accountability crucially ensure that the stakeholders of AI (all people) have a say in how AI is developed and deployed, which allows AI the opportunity to attain the goal of maximizing median human flourishing. Finally the permanence feature allows it to be practically used and built up in much the same way that the electrical grid or internet is built and maintained.

There are two important failure modes however in this public approach. One is economic failure where the public good is not economically sustainable and therefore turns to undesirable methods of funding (ads, data extraction etc) [9].The second is more of a global coordination/cooperation failure, even if all AI in the world is public AI one country could still prioritize its own interests over the collective good.

2.4 Problems of AI

AI alignment is a key concern in AI safety. However there are several other concerns that need to be addressed to ensure that AI is safe and beneficial for humanity.

2.4.1 Concrete problems of AI safety

negative side effects sometimes we wantt hte model to do things yet sometimes we don't want it to do thinkg

Mis generalization how we might align and then it operates different in tet

Scalable oversight as models getm ore powerful how to know if they are being alright

Robustness to distributional shift when the model is in a new environment how to ensure it is still aligned

2.4.2 Catastrophic risks of AI

https://arxiv.org/abs/2306.12001 https://futureoflife.org/resource/catastrophic-ai-scenarios/

Even if style framing. There are still some great risks of AI.

Runaway self-improvement

Weapons of mass destruction (cyber, chemical etc)

Mis aligned AI

2.4.3 Preventing catastrophic risks of AI

Stopping should we do Stopping can we do it

Deep alignment (the topics here)

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

[9]Which is very much similar to the way that the internet was developed. It started off funded by the public sector (laying cables etc), however then for anyone to have a successful business offered on the internet it had to be "free" to use and so companies (E.g Google, Facebook) needed to find alternative methods of funding which has proven detrimental to greater public health

2.5 Representing Human Values for LLM alignment

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. There are broadly two methods one works from defining broad principles and specific actions are inferred from there the other method is providing specific examples and generalizing from them. Interestingly when working with humans we can only really tell them in a high level what we want their values to be and let them infer the specifics, yet with modern AI systems like LLMs we can provide lots of specific examples and let it infer the values from this [10]. In this section I will outline some of the key concepts around representing human values and how they relate to AI systems.

2.5.1 Human values taxonomy

Human values are an old idea that go back millennia [7, 53, 5], with each culture having its own set of values that are important to them. They define how people should be and act [11]. 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" [60] 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 [63] proposes a model of 10 broad human values that he later expands to 19 values [65]. 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 ability of peoples actions given values 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 work from computer scientists have created value taxonomies that are designed specifically for their predictive [12] power [40]. 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.

2.5.2 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 [60] created the Rokeach Value Survey (RVS) which asked participants to rank the 36 values in order of importance to them. Schwartz [63] created the Schwartz Value Survey

[10] Providing a human with 10,000 different questions and responses of "how to answer the phone" won't provide much clarity, yet giving them some simple high level rules is quite effective. Oppositely for AI most alignment methods like RLHF, DPO etc provide lots of specific examples and let the model infer the underlying values.

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

[12] Which in this case is the ability for a model to predict someones values (from the taxonomy) given an argument that they made. Plausibly the word modelling power might be more intuitive.

(SVS) which asked participants to rate the importance of each value on a 9-point scale. Later Schwartz replaced the SVS with the Portrait Values Questionnaire (PVQ) [64] which was less abstract and better at soliciting the values from a diverse set of people as respondents are answering questions of the form "How much is this person like you", and by rating their similarity with 40 different portraits one can infer someones values. These surveys allowed researchers to measure the values of individuals and groups and provided a way to understand how values relate to behavior.

All the above surveys (PVQ, SVS and RVS) are designed to measure the values of individuals in relation to be a predefined taxonomy of values. They have relatively few questions (¡50), abstract in nature and quick to complete. An alternative path for measuring values is to create broader surveys that try to capture the values of large groups of people through many questions. Importantly these surveys do not map directly to a predefined taxonomy of values yet instead try to capture the complexity of human values through many questions (300). The World Values Survey [35] is the biggest example of this where they have been surveying over 1,000 people from over 100 countries every 5 years since 1981. The New Zealand Attitudes and Values Study (NZAVS) [71] is another example of a large scale value survey that has been running since 2009 surveying around 30,000 New Zealanders every year. These surveys don't correspond to a strict values taxonomy yet instead provide granular data from each individual. This data can help train [54, 46] and evaluate [26, 76, 8]Large Language Models (LLMs) on human values.

These surveys provide granular data yet struggle from well documented issues of self-reporting bias [57] and social desirability bias [24]. 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.

2.5.3 Values to behavior

For the purposes of building AI systems understanding values is only useful as far as they can predict behavior (in the form of responses and actions). 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 [42, 29]. When one tries to bridge the gap between understanding someones behavior and values we run into common problems namely; non-identifiability of values (Many different values can lead to the same behavior), specification gaming (as mentioned above) and extrapolation errors. Non-identifiability and extrapolation errors are demonstrated clearly in inverse reinforcement learning (IRL) [2]. Furthermore it can be shown that human values are fickle and can stem from ones previous actions as humans rationalize what they have done [45]. 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.

Use some of the reference from [41] and otherresearch to talk more about values and different definitions.

However the idea that values can lead to and predict behavior has been argued [61, 64]. Furthermore in the reinforcement learning literature there is evidence that given examples of actions a model can learn a opaque value function that can predict what actions will be made next (called inverse reinforcement learning). 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 [34, 55]. 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.

2.5.4 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 [37]. This presents a challenge for AI alignment, as we must determine what values we ought to align AI systems to, and all we currently have is what human values currently are. 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, they might not be the perfect values yet the best target. 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

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 an end-to-end project that involves using current survey data and collecting new data, aligning a large language model using different methods and evaluating the effectiveness of the alignment.

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 we can let the public directly and indirectly decide on what values they want the model to be aligned to. More so the public as the users and recipients of the model's outputs should have a say in the way that the model behaves. Large language models are on track to become tomorrows, accountants, lawyers, doctors and more, we need to ensure that we the people are the ones who decide on what kind of people they are.

3.3 Timeline

For now I will leave the timeline uncompleted. Once I have a more concrete idea of the method I will fill this in properly. There is code left in to help with how to make a gantt chart.

The project will be carried out over three main phases. The first phase will be the organization and clustering of existing survey data to represent human values. The second phase will be an iterative process of aligning a large language model and evaluating its performance. The third phase involves collecting specific alignment data to further refine and train the model, followed by a final evaluation.

3.4 Output

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

- A collection of aligned LLMs with various checkpoints of just survey data and then with the collected data.
- A public dataset of collected data around preferred values and responses for alignment.
- Experimental results evaluating the effectiveness of different alignment methods using the collected data.

3.5 Resources needed

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

3.5.1 Hardware

The first is hardware which will be needed to do the alignment finetuning evaluation of the models. For the base LLM I would want to use a model in the range of 7 billion to 30 billion parameters, as this will get me close to state of the art performance on many tasks. This means that I will need access to about 200GB of GPU memory to be able to finetune the model effectively. Each finetuning run will take in the realm of 3-6 hours depending on the size of the model, amount of data used, alignment methods used and GPU hardware. Therefore given the goal of trying a handful of methods and different value sets I estimate about [1] 70 hours of GPU time will be needed. The cost of a GPU hour on a high end GPU such as a B200 is 6 USD per hour which would make the total cost about 420 USD. Alternatively the Engineering and Computer Science department's GPU servers should give me access to machines with around 80GB of vram. This would allow me to fine-tune smaller models in the 7-13 billion parameter range. This should still be sufficient for a proof of concept and evaluation of the alignment methods.

Public consultation responses

The idea of using the public to help align large language models is central to this project. Most of the fine tuning will be done using the survey data which can be used to get the model roughly aligned to human values. However to get more specific and granular alignment I will collect new data from the public to further refine the model's behavior. This will be in the form of an interactive survey where participants will provide some form of preference feedback on model outputs. I envision using a tool like Polis or similar to collect and aggregate responses in a meaningful way. I would expect to need at least 50 or so people from the various subgroups and for each of them to spend maybe 2-3 15 minutes blocks providing feedback on model outputs.

[1] 4 hours per fine tune * 3 different methods * 5 different value sets + 5 hours of evaluation = 65 hours

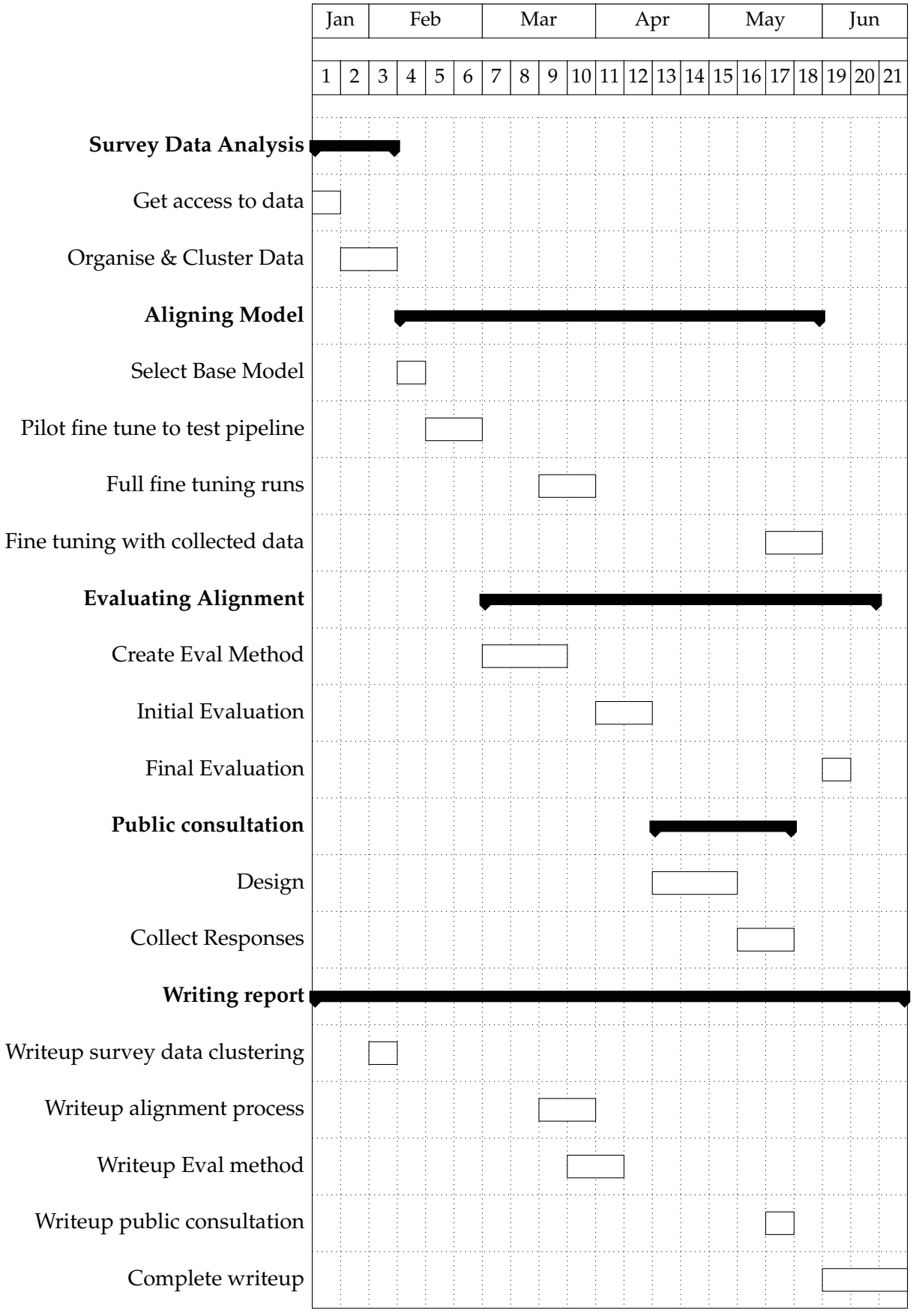

Figure 3.1: Project timeline in weeks (only roughly aligned to months). This assumes a start time of end of January 2026 with near part time work ramping up to full time work by end of February.

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] BRADLEY, R. A., AND TERRY, M. E. RANK ANALYSIS OF INCOMPLETE BLOCK DESIGNS: THE METHOD OF PAIRED COMPARISONS. *Biometrika 39*, 3-4 (Dec. 1952), 324–345.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

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

[21] CHURCHILL, WINSTON S. Speech to the House of Commons, Nov. 1947.

[22] COMPDEMOCRACY. Polis. The Computational Democracy Project, Nov. 2025.

[23] CRAWFORD, K. *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press, Apr. 2021.

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

[25] DE TOCQUEVILLE, A., AND REEVE, H. *Democracy in America*. Saunders and Otley, London, 1835.

[26] 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.

[27] EDELMAN, J., ZHI-XUAN, T., LOWE, R., KLINGEFJORD, O., WANG-MASCIANICA, V., FRANKLIN, M., KEARNS, R. O., HAIN, E., SARKAR, A., BAKKER, M., BAREZ, F., DUVENAUD, D., FOERSTER, J., GABRIEL, I., GUBBELS, J., GOODMAN, B., HAUPT, A., HEITZIG, J., JARA-ETTINGER, J., KASIRZADEH, A., KIRKPATRICK, J. R., KOH, A., KNOX, W. B., KORALUS, P., LEHMAN, J., LEVINE, S., MARRO, S., REVEL, M., SHORIN, T., SUTHERLAND, M., TESSLER, M. H., VENDROV, I., AND WILKEN-SMITH, J. Full-Stack Alignment: Co-Aligning AI and Institutions with Thick Models of Value, Dec. 2025.

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

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

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

[31] GOODHART, C. Problems of monetary management: The UK experience. In *Inflation, Depression, and Economic Policy in the West*, A. S. Courakis, Ed. Barnes and Noble Books, Totowa, New Jersey, 1981, p. 116.

[32] 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.

[33] HABERMAS, J. *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*. Studies in Contemporary German Social Thought. MIT Press, Cambridge, MA, USA, Jan. 1998.

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

[35] HAERPFER, CHRISTIAN AND INGLEHART, RONALD AND MORENO, ALEJANDRO AND WELZEL, CHRISTIAN AND KIZILOVA, KSENIYA AND DIEZ-MEDRANO, JUAN AND LAGOS, MARTA AND NORRIS, PIPPA AND PONARIN, EDUARD AND PURANEN, BI. World Values Survey: Round Seven, 2024.

[36] HUANG, S., SIDDARTH, D., LOVITT, L., LIAO, T. I., DURMUS, E., TAMKIN, A., AND GANGULI, D. Collective Constitutional AI: Aligning a Language Model with Public Input. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency* (June 2024), pp. 1395–1417.

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

[38] JONATHAN STRAY,. A Practical Definition of Political Neutrality for AI.

[39] KAHNEMAN, D., AND TVERSKY, A. Prospect theory: An analysis of decision under risk. *Econometrica : journal of the Econometric Society 47*, 2 (1979), 263–291.

[40] 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.

[41] KLINGEFJORD, O., LOWE, R., AND EDELMAN, J. What are human values, and how do we align AI to them?, Apr. 2024.

[42] 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.

[43] KULLBACK, S., AND LEIBLER, R. A. On Information and Sufficiency. *The Annals of Mathematical Statistics 22*, 1 (Mar. 1951), 79–86.

[44] 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.

[45] LEON FESTINGER. *A Theory of Cognitive Dissonance*. 1957.

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

[47] 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.

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

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

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

[51] MILANO, S., TADDEO, M., AND FLORIDI, L. Recommender systems and their ethical challenges. *AI & SOCIETY 35*, 4 (Dec. 2020), 957–967.

[52] MILL, J. S. *On Liberty*. Jan. 2011.

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

[54] 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.

[55] 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.

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

[57] 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.

[58] PUBLIC AI. Public AI White Paper, Aug. 2024.

[59] 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.

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

[61] 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.

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

[63] 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.

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

[65] 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.

[66] 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.

[67] STIGLITZ, J. E. Markets, Market Failures, and Development. 197–203.

[68] 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).

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

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

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

[72] VANDERWEELE, T. J. On the promotion of human flourishing. *Proceedings of the National Academy of Sciences 114*, 31 (Aug. 2017), 8148–8156.

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

[74] 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.

[75] WENMAN, M. *Agonistic Democracy: Constituent Power in the Era of Globalisation*. Cambridge University Press, Cambridge, 2013.

[76] 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.