Aligning LLMs to Public Values: Funding Request

James Thompson
Te Herenga Waka — Victoria University of Wellington

July 16, 2026

Large Language Models (LLMs) are increasingly making decisions that affect human lives, yet their behavior is determined by private for-profit companies, lacking democratic legitimacy. This project proposes a pragmatic method for tuning LLMs to behave in ways that align with public values. This will be achieved by using publicly collected value survey data to fine-tune open-weight LLMs. Behavioral preferences will then be validated through a public consultation. [1]

1 What the problem is

Large Language Models (LLMs) are some of the most powerful AI systems ever created, capable of performing a wide range of tasks with minimal human input [4]. This has led them to be deployed in autonomous systems in which they have decision powers that affect many real world lives. This raises a pressing challenge: ensuring these models behave in alignment with human values [1, 10, 5].

The *alignment problem* is typically solved by the developers using small internal teams to label training data, which is then used to train the models [3, 13]. In the post-training of these models decisions are made which determine the values the model will exhibit and the way it will behave.

The current training and alignment pipelines are insufficiently robust to ensure public accountability and democratic legitimacy. Furthermore these LLMs are overwhelmingly the product of for-profit organizations, which inherently aim to maximize profit, which is generally in tension with the public good [15, 9]. This can create scalable harm and compound downstream misalignment risks.

Many current alignment methods [2, 6] are targeting a form of *value alignment* in which the model is aligned to broad values set by the developers—which may or may not be influenced by public views. Human values, however, are very heterogeneous, and so a model "aligned" may not be in practice aligned with the end user's values. This is exacerbated when models are trained once and deployed to many users across national and cultural boundaries.

2 What the project will do

To improve both public alignment and public accountability of LLMs this project will demonstrate a pragmatic approach to LLM alignment. This will be done by using an open-weight model [2] and readily available value survey data to modify model behavior. To validate that the resultant model is indeed preferred by the public (intended beneficiary), behavioral simulations will be run and used in a public consultation. Below is a short summary and a more complete methodology breakdown can be found in Appendix A.

The **research questions** that will be answered are:

1. Can we use publicly collected value survey data to modify the behavior of LLMs to align with the values of the public?

2. Will users prefer the behavior of models aligned to their values compared to unaligned models?

To answer these questions this project will do a proof-of-concept demonstration of this pragmatic approach to LLM alignment. The **method** can be summarized in three steps:

[1] This project is developed in the open and can be found at `https://github.com/1jamesthompson1/AIML589`

[2] Open-weight models are those whose weights are publicly available and are a superset of open-source AI models, in which the weights along with training pipeline and data are publicly available.

- Using already publicly collected value survey data to fine-tune a LLM to respond to survey questions with the same distribution as the survey respondents.
- Building and running text-based simulations to evaluate the behavior of the fine-tuned LLMs in real world scenarios. This is to test the models to see for behavioral changes.
- Consulting the public (i.e a new anonymous survey) on the behavior of the fine-tuned LLMs in these simulations. Which will involve the public rating vignettes of the model's behavior in the simulations.

The **expected outcome** will be to answer the research questions as well as provide more evidence for the efficacy of this approach to getting LLM powered systems to behave in ways that are preferred by the public.

In addition to the above contributions, the project will produce reusable datasets and pipelines. First, a NZ public values dataset to fine-tune an arbitrary LLM to align with the values of the NZ public (or, through small modifications, those of other countries around the world). Second, a simulator that can be used to evaluate the behavior of LLMs in real world scenarios, focused on the New Zealand context. Both resources will be built in the open and made available publicly for other researchers to use.

3 What is needed to complete the project

To complete this project there are two main requirements:

- Computational resources to fine-tune and evaluate LLMs.
- Human participants to evaluate the behavior of the models.

Both of these can be sourced at a reasonable cost, with the computational resources being available through a cloud provider and the human participants being recruited through a panel recruitment service. A summary of the estimated budget is provided below.

Item	**Estimated Cost (NZD)**
GPU Compute (160 hours @ $3.5/hour)	560
Participant Compensation (100 participants @ $15)	1500
Contingency (20%)	412
Total	2472

Table 1: Estimated Budget for Project Resources, all values in New Zealand Dollars (NZD). GPU price is based on vast.ai price for H200 (140GB VRAM). Participant compensation is estimated and a quote is pending. See appendix B for a detailed breakdown of the resources required.

The other requirements for the project are publicly and freely available namely the World Values Survey data and open-weight LLMs.

3.1 Funding counterfactual

This project has a graceful degradation of scope, regardless of funding it will be completed on some level. The first part of the scope to be removed would be the paid public consultation which would be replaced with a voluntary public consultation. This reduces the statistical power of answering the second research question, but would still allow the first question to be answered fully. The second item to be removed would be the paid GPU compute which would be replaced with using university provided GPU resources, and would cause the breadth of the experiments to be cut by about half.

4 References

[1] Dario Amodei et al. *Concrete Problems in AI Safety*. July 2016. DOI: 10.48550/arXiv.1606.06565. arXiv: 1606.06565 [cs]. (Visited on 12/07/2025).

[2] Yuntao Bai et al. *Constitutional AI: Harmlessness from AI Feedback*. Dec. 2022. DOI: 10.48550/arXiv.2212.08073. arXiv: 2212.08073 [cs]. (Visited on 11/17/2025).

[3] Yuntao Bai et al. *Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback*. Apr. 2022. DOI: 10.48550/arXiv.2204.05862. arXiv: 2204.05862 [cs]. (Visited on 11/25/2025).

[4] Rishi Bommasani et al. *On the Opportunities and Risks of Foundation Models*. July 2022. DOI: 10.48550/arXiv.2108.07258. arXiv: 2108.07258 [cs]. (Visited on 03/21/2026).

[5] Nick Bostrom and Eliezer Yudkowsky. "The Ethics of Artificial Intelligence". In: *The Cambridge Handbook of Artificial Intelligence*. Ed. by Keith Frankish and William M. Ramsey. Cambridge: Cambridge University Press, 2014, pp. 316–334. ISBN: 978-0-521-87142-6. DOI: 10.1017/CBO9781139046855.020. (Visited on 03/21/2026).

[6] *Collective Constitutional AI: Aligning a Language Model with Public Input*. https://www.anthropic.com/news/constitutional-ai-aligning-a-language-model-with-public-input. Oct. 2023. (Visited on 11/30/2025).

[7] Tim Dettmers et al. *QLoRA: Efficient Finetuning of Quantized LLMs*. May 2023. DOI: 10.48550/arXiv.2305.14314. arXiv: 2305.14314 [cs]. (Visited on 01/06/2026).

[8] 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*. Madrid, Spain and Vienna, Austria, 2024. DOI: doi:10.14281/18241.24.

[9] Dan Hendrycks et al. *Aligning AI With Shared Human Values*. Feb. 2023. DOI: 10.48550/arXiv.2008.02275. arXiv: 2008.02275 [cs]. (Visited on 03/21/2026).

[10] Dan Hendrycks et al. *Unsolved Problems in ML Safety*. June 2022. DOI: 10.48550/arXiv.2109.13916. arXiv: 2109.13916 [cs]. (Visited on 12/13/2025).

[11] Edward J. Hu et al. *LoRA: Low-Rank Adaptation of Large Language Models*. Oct. 2021. DOI: 10.48550/arXiv.2106.09685. arXiv: 2106.09685 [cs]. (Visited on 01/06/2026).

[12] Team Olmo et al. *Olmo 3*. 2025. arXiv: 2512.13961 [cs.CL].

[13] Long Ouyang et al. *Training Language Models to Follow Instructions with Human Feedback*. Mar. 2022. DOI: 10.48550/arXiv.2203.02155. arXiv: 2203.02155 [cs]. (Visited on 10/25/2025).

[14] Qwen Team. *Qwen3.5: Towards Native Multimodal Agents*. Feb. 2026.

[15] R. H. Coarse. "The Problem of Social Cost: The Journal of Law and Economics: Vol 3". In: *The Journal of Law and Economics* (). (Visited on 02/02/2026).

[16] Guangxuan Xiao et al. *SmoothQuant: Accurate and Efficient Post-Training Quantization for Large Language Models*. Mar. 2024. DOI: 10.48550/arXiv.2211.10438. arXiv: 2211.10438 [cs]. (Visited on 01/25/2026).

A Project methods

This project will be an end-to-end demonstration of a pragmatic alignment approach. This involves using a pre-trained LLM, defining alignment targets, fine-tuning to those targets, and then evaluating the results through public consultation. There are three key components to this project: (1) using value survey data to fine-tune LLMs, (2) creating a simulator to evaluate the behavior of the fine-tuned LLMs, and (3) consulting the public on the behavior of the LLMs in these simulations.

A.1 Fine tuning of LLMs to Public Values

There are 4 things that are needed to fine tune a LLM. A loss function, a base model, a training dataset, and a fine-tuning method. Here I will outline each of these components and how they will be used in this project.

A.1.1 Creating training dataset

The World Values Survey (WVS)[8] is a global research project that collects survey responses from about 1000 people from each country on a about 300 questions about their values and beliefs. The response data from the WVS will be used to create a dataset of question-answer pairs that will look like the following:

> *Question:* On a scale of 1-10, how important is it to have a strong police force?
> *Response:* [expected_response]

The exact format of the question, including the system prompt, is yet to be determined. The same structure of getting the model to answer the question using its "own values" will be used. This is so that when training we are targeting the model's internal values and not just the model's ability to predict responses for a subgroup.

To allow for comparison between different subgroups of the population [3], the survey responses will be grouped into subgroups based on a yet to be determined clustering method.

The value of the expected response will be determined by the responses from the particular WVS subgroup. Depending on the fine-tuning method used, it will either be the modal response, a sample from the distribution of responses, or the distribution of responses itself.

The training dataset will range in length from as small as about 300 questions (i.e. simple modal modelling with a fixed template) to as large as 3,000 question-answer pairs (i.e. distributional modelling with many templates/samples per question). The exact size of the complete training dataset will be determined by the fine-tuning method used and the number of subgroups that are used.

A.1.2 Fine tuning loss function

Two general methods will be explored for fine-tuning: traditional modal response fine-tuning in which the model is trained to produce the modal response for each question, and distributional fine-tuning in which the model is trained to produce responses that match the distribution of responses from the WVS subgroup. The two methods will be compared to see which method produces a model that is more aligned with the values of the WVS subgroup. The two methods are:

Most fine-tuning is modal response fine-tuning and so is used as a baseline. The training dataset will simply be the questions and the modal response from the WVS subgroup. This dataset will then be used in a supervised fine-tuning (SFT) method [13] to train the model to produce the modal response for each question.

A second more nuanced method will be used to optimize the model to produce responses that match the distribution of responses from the WVS subgroup. This will be attempted in two different ways. First, by simply augmenting the training dataset to include many repeats of the same question (with different templates) with the expected response being drawn from the distribution of responses from the WVS. When applied to SFT as above, in expectation it matches the WVS subgroup distribution. Second, one could optimize for the distribution directly by using a soft entropy loss between the expected WVS subgroup distribution and the model's distribution for the response token.

A.1.3 Base model and fine-tuning method

The base model used is restricted by our hardware resources. The model must be small enough to be fine-tuned on a single modern GPU (i.e., around 100 GB VRAM). This restricts the base model to be around 30 billion parameters. This gives us model options such as Qwen 3.6 27B that achieve

[3] i.e. a person from group A preferred behavior from its own group compared to group B but was indifferent relative to the base model

approximately 62% of frontier model performance while remaining feasible to fine-tune with limited resources[4].

To actually fine-tune a model of this size, we will use both parameter-efficient fine-tuning [11, 7] and quantization [16]. These methods allow us to fine-tune large models with significantly reduced memory requirements.

A.2 Running Simulations to Evaluate LLM Behavior

Understanding how well a model is aligned to human values requires evaluating the model's behavior in real world scenarios. To do this, a simulator will be built that will allow for the evaluation of the model's behavior in a variety of scenarios. The goal of the simulations is both to test whether the fine-tuning methods have changed the model's behavior and to provide a basis for public consultation on the model's behavior.

The structure of the simulation will involve placing the LLM in an agentic harness and letting it operate in a text-based simulation of the real world. The scenarios used are yet to be determined, but will have a focus on "grounded" New Zealand scenarios. There will be 20 scenarios in total, with each scenario being run multiple times by each model (different fine-tunes using the base model).

Each agent's trajectory through the simulation will be logged and turned into a vignette using as deterministic a template as possible. Some LLM assistance will likely be used to make the vignettes readable.

A.3 Public Consultation on LLM Alignment

The public consultation will be an anonymous survey where participants will be asked to rate their preference for a model's behavior in a given scenario. The survey will be designed to be completed in about 15 minutes, with participants recruited from a panel recruitment service to ensure fair representation of the New Zealand population.

Each question in the survey will present the vignette of a single model's behavior in a single scenario and ask participants to use a Likert scale to rate their "agreement" with the model's behavior. Each participant will be shown 3 random simulations. Through aggregation of the responses, we will be able to determine which model's behavior is preferred by the public.

For the purposes of understanding inter-group differences, respondents will be asked to answer a handful of questions from the WVS that are most predictive of the WVS subgroup that they belong to. This will allow the survey respondents to be grouped into the same subgroups as the WVS respondents and open more lines of inquiry into preferences for the model's behavior between different subgroups of the population.

A.4 Answering research questions

A.4.1 Research Question 1: Can we use publicly collected value survey data to modify the behavior of LLMs to align with the values of the public?

To answer this question we will first need to ensure that the model has successfully learned to produce responses that match the distribution of responses from the WVS subgroup. This will be done by ensuring that the fine-tuning loss function is minimized. Conditional on this being the case, we will then run the simulations to evaluate the behavior of the fine-tuned LLMs.

These simulations will have a set of key decisions that can be evaluated to see if the model's behavior has changed from the base model. Note that understanding whether the model's behavior changes in the direction of the WVS subgroup will only be done through public consultation.

[4]Comparing Qwen3.6 27b vs Claude Fable 5 on Artificial Analysis Leaderboard

A.4.2 Research Question 2: Will users prefer the behavior of models aligned to their values compared to unaligned models?

To answer this, a public consultation will be done in the form of an anonymous survey. The survey results will be many Likert ratings of the model's behavior in a given scenario by respondents from different subgroups. These ratings will be aggregated to determine which model's behavior is preferred by the public (i.e. the model that is aligned to the respondents' subgroup, another group, or the base model). This will allow us to determine if the model that is aligned to the respondents' subgroup is preferred by the respondents.

B Resources Required

B.1 Computational Resources

Training and evaluating LLMs requires significant GPU resources. Based on the method design in Appendix A, I have estimated the total GPU hours required to complete the project to be approximately 160hours. This is broken down as follows:

- Training pipeline development and testing: 25 hours
- Full fine-tuning experiments (3 methods $\times$ 3 value sets $\times$ 3 models @ 4 hours per fine-tune): 81 hours
- Evaluation runs (27 trained models with 2 hours each): 54 hours

A reduction in funding may result in no cloud compute available and so using university provided GPU resources is possible. However due to the reduced power (2x A100 40 GB or 1x RTX 6000 Blackwell) and limited availability of these resources (i.e, waiting for availability), the number of experiments that can be run will be reduced. This will reduce the depth of the experiments, thus reducing the generalization of the results.

B.2 Data Sources

The two forms of data that are going to be needed are value survey data and the parameters of a base LLM. Fortunately both of these are publicly available and can be used without financial cost or licensing restrictions.

- **World Values Survey (WVS)**: The primary data source for defining alignment targets. The WVS contains responses from approximately 1,000 respondents in New Zealand with the survey being almost 300 questions long. This survey data is also available for many other countries, allowing for the same method to be applied to other countries in the future.
- **Base models**: We will use open-weight models such as Qwen 3.5[14] or open-source models such as Olmo [12]. Final model choice will be determined based on the latest performance benchmarks.

B.3 Human Resources

- **Public Consultation Participants**: We will recruit approximately 100 participants from the general public to evaluate model outputs ($\approx$ 15 minutes each). They will be sourced using a panel recruitment service (e.g., Pureprofile, Dynata) to ensure demographic diversity. [5]
- **Ethics Approval**: Full ethics approval will be obtained from the university's ethics committee. This project is expected to be category B (low risk), meaning the application can be submitted and processed at any time of the year.

[5] Advice from an experienced panel recruitment service suggests that the sweet spot is about 15 minutes per participant, which provides good depth while retaining engagement.