Aligning Large Language Models to Public Values: Project Resources and Feasibility Summary

James Thompson*

March 22, 2026

Abstract

Large Language Models (LLMs) are increasingly deployed in society, yet their alignment targets are typically defined by small teams within private companies, lacking democratic legitimacy. This project proposes a proof-of-concept demonstration of aligning LLMs to publicly collected human values from the World Values Survey, with evaluation conducted through public consultation. This document outlines the resources required to complete this 18-week project, including computational needs, data sources, human participants, timeline, and estimated budget. [1]

1 Introduction and Background

The deployment of Large Language Models (LLMs) has accelerated rapidly, with these systems now acting with increasing autonomy and decreasing human oversight [3]. This raises a pressing challenge: ensuring these models behave in alignment with human values [1, 9, 4]. Current alignment methods rely on training data labeled by small teams of annotators [2, 13]. These alignment methods are carried out within for-profit organizations with incentives that may not align with public good [15, 8].

This project addresses a gap in AI alignment research by developing a democratically grounded approach to LLM alignment which is both more ethically defensible and increases public trust. Rather than relying on opaque annotation pipelines, we propose using existing value survey data (World Values Survey [7]) to define **alignment targets** that are representative, and then conduct **alignment evaluation** by consulting the public who are the intended beneficiaries of aligned AI systems. This approach aims to enhance the legitimacy and societal acceptance of AI efforts.

The project aims to answer two key research questions: (1) Can simple post-training methods effectively align LLMs to human values defined by survey data? and (2) Will the public "prefer" the behavior of models aligned to their values compared to unaligned models?

2 Project Overview

This project will be an end-to-end demonstration of a novel 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.

Using value survey data as an alignment target will create a dataset of question-answer pairs (e.g., "On a scale of 1-10, how important is it to have a strong police force?" *response:* "7"). This dataset can then be used to fine-tune the model using supervised fine-tuning [13] and, if needed, more complex methods like Kahneman-Tversky Optimization [6]. These methods allow us to push the model's behavior towards responding in the same distribution as the survey respondents. To evaluate the alignment, we will test the model's behavior in out-of-distribution scenarios (i.e., simulations of real-world situations) and then consult the public on the model's behavior in these scenarios. This public consultation will be in the form of a factorial experiment [11] where participants will be shown vignettes of the model's behavior in different scenarios and asked to rate how well the model's behavior aligns with their values.

*Te Herenga Waka — Victoria University of Wellington

[1] Full background and project proposal available here where project will developed .

3 Resources Required

3.1 Computational Resources

Training and evaluating LLMs requires significant GPU resources. Based on our experimental design, we estimate the following computational requirements:

- **Base Model**: We will use open-weight LLMs with approximately 30 billion parameters. Based on current benchmarks, models at this scale achieve approximately 70% of frontier model performance while remaining feasible to fine-tune with limited resources[2].
- **Parameter-Efficient Fine-Tuning**: Using LoRA (Low-Rank Adaptation) [10] and QLoRA [5], we can effectively fine-tune large models by updating only a small fraction of parameters. This reduces VRAM requirements from hundreds of gigabytes to the realm of 100Ggb that will work on a single modern GPU.
- **Estimated GPU Hours**: Approximately 120 hours of GPU time total, including:
  - Pilot fine-tuning runs: 15 hours
  - Full fine-tuning experiments (3 methods $\times$ 5 value sets): 75 hours
  - Evaluation runs: 30 hours
- **Alternative Options**: University-provided GPU resources (2x A100 40 GB) are available. However, due to being older hardware will increase time to use and decrease depth of experiments.

3.2 Data Sources

- **World Values Survey (WVS)**: The primary data source for defining alignment targets. The WVS contains responses from approximately 1,000 respondents per country across hundreds of questions covering human values. This data is openly accessible for research purposes.
- **Base Models**: We will use open-weight models [3] such as Qwen [14] or open-source models [4] such as Olmo [12]. These models can be freely downloaded and used without financial cost or licensing restrictions.

3.3 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., Samplify, 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.

4 Budget Summary

5 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] Comparing Qwen3.5 27b vs GPT-5.4 on Artificial Analysis Leaderboard

[3] An open-weight model is one in which the final output (neural network weights) is released, but limited or no training information is provided.

[4] An open-source model is one in which both the final output and all inputs (training data and training code) are supplied [16].

[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.

Item	Estimated Cost (NZD)
GPU Compute (120 hours $3.5/hour [6])	420
Participant Compensation (100 participants $10)	1000
Contingency (20%)	284
Total	1704

Table 1: Estimated Budget for Project Resources, all values in New Zealand Dollars (NZD).

[2] 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).

[3] 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).

[4] 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).

[5] 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).

[6] Kawin Ethayarajh et al. *KTO: Model Alignment as Prospect Theoretic Optimization*. Nov. 2024. DOI: 10.48550/arXiv.2402.01306. arXiv: 2402.01306 [cs]. (Visited on 11/26/2025).

[7] 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.

[8] 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).

[9] 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).

[10] 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).

[11] *Measuring Social Judgments : The Factorial Survey Approach*. Beverly Hills : Sage Publications, 1982. ISBN: 978-0-8039-1816-0. (Visited on 03/21/2026).

[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] *The Open Source AI Definition - 1.0 - Open Source Initiative*. https://opensource.org/ai/open-source-ai-definition. (Visited on 01/25/2026).