PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO

Investigating Large Language Models' Responses to Everyday Moral Dilemmas in Different Cultural Contexts

Daniel Stulberg Huf

PROJETO FINAL DE GRADUAÇÃO

CENTRO TÉCNICO CIENTÍFICO - CTC

DEPARTAMENTO DE INFORMÁTICA

Curso de Graduação em Engenharia da Computação

Rio de Janeiro, Junho de 2025

Daniel Stulberg Huf

Investigating Large Language Models' Responses to Everyday Moral Dilemmas in Different Cultural Contexts

Relatório de Projeto Final, apresentado ao curso de Engenharia da Computação da PUC-Rio como requisito parcial para a obtenção do título de Engenheiro de Computação.

Orientador: Prof. Simone Diniz Junqueira Barbosa

Rio de Janeiro

Junho de 2025

“Pouco importa o objeto da ambição; ela vale por si, independente do alvo. Sempre necessitamos ambicionar alguma coisa que, alcançada, não nos faz desambiciosos.”

(Carlos Drummond de Andrade)

Agradecimentos

Estes últimos seis anos de estudos superiores, vividos entre dois países tão distintos, pareceram surreais. Eu nunca imaginei que conheceria tantas pessoas incríveis, descobriria tantos lugares novos e viveria de forma tão intensa nesse período de tempo.

Primeiramente, gostaria de agradecer aos meus orientadores, Simone Barbosa e José Luiz Nunes, pelo apoio atencioso ao longo do meu último ano da graduação.

Quero expressar minha profunda gratidão aos meus pais pela educação que me proporcionaram ao longo de toda a minha vida, assim como pelo incentivo e apoio incondicional, mesmo que à distância, na minha busca constante por experiências enriquecedoras.

Por fim, agradeço aos meus amigos de longa data do Brasil e aos amigos que fiz durante minha estadia na França por compartilharem memórias incríveis e lições inesgotáveis que levarei comigo por toda a vida.

Concluir essa jornada, cheia de incertezas e desafios a superar, não teria sido possível sem essas pessoas, às quais serei eternamente grato.

Abstract

Stulberg Huf, Daniel. Diniz Junqueira Barbosa, Simone. **Investigating Large Language Models' Responses to Everyday Moral Dilemmas in Different Cultural Contexts**. Rio de Janeiro, 2025. 18p. Relatório de Projeto Final I – Departamento de Informática. Pontifícia Universidade Católica do Rio de Janeiro.

Escrever aqui o resumo do trabalho em 10 linhas (espaço simples)

Keywords

Listar aqui as palavras-chave em inglês

Resumo

Stulberg Huf, Daniel. Diniz Junqueira Barbosa, Simone. **Investigando Respostas de *Large Language Models* a Dilemas Morais Cotidianos em Diferentes Contextos Culturais**. Rio de Janeiro, 2025. 18p. Relatório de Projeto Final I – Departamento de Informática. Pontifícia Universidade Católica do Rio de Janeiro.

Escrever aqui o resumo do trabalho em inglês em 10 linhas (espaço simples)

Palavras-chave

Listar aqui as palavras-chave

Table of Contents

List of Figures

1
Introduction

The rapid advancement of Generative Artificial Intelligence (GenAI) technologies has fundamentally transformed decision-making dynamics across various sectors of society. In particular, Large Language Models (LLMs) form the foundation of advanced natural language processing (NLP) applications and are already being consistently integrated into critical systems that range from simulating social interactions (PIAO et al., 2025) to supporting legal proceedings (LIU; LI, 2024) and mental healthcare services (HUA et al., 2024). Given the significant influence these models already have on social structures and individual lives, understanding the reasoning processes that underpin them becomes essential for managing their use in ways that genuinely benefit humankind.

One challenge that naturally arises in this domain is the alignment of AI systems with human ethical values, particularly because such standards can be often perceived differently across cultures and communities. Added to this issue is the growing awareness that AI systems tend to overrepresent certain demographics and cultural norms (CAO et al., 2023; JOHSON et al., 2022), leading to outputs that may be inherently biased as they reflect the very own nature of its designers. If the mere integration of AI into social systems raises concerns, then the everyday use of models that may preferentially reflect particular cultural values over others poses an even more complex problem of ethical alignment.

Against such a panorama of AI adoption in socially sensitive contexts, designers face the critical dilemma of encoding, either implicitly or explicitly, human values into algorithms. Current research draws on other domains to evaluate and formalize ethical behavior in machines, such as the Moral Foundations Theory originated in the moral psychology literature (ZANGARI et al., 2025). A recurring concern is the potential mismatch between the AI's output on such moral reasoning and that of its users, which motivates further analysis of the mechanisms driving AI decision-making. However, what if solely relying on the textual output of LLMs' moral judgments might not just be enough? Could the deeper computational layers in which these models operate reveal more about the values they reflect and the cultural norms they amplify? Do these underlying

layers point toward a shared moral framework across all models, or do they rather expose some divergences?

This present work aims to explore these questions by analyzing the multidimensional embedding spaces of LLM-generated reasonings of moral dilemmas. Specifically, this work builds on the study *"Normative Evaluation of Large Language Models with Everyday Moral Dilemmas"* (SACHDEVA; VAN NUENEN, 2025), in which the authors examined over 10,000 moral scenarios from the Reddit community “Am I The Asshole” (AITA). In their study, seven different LLMs were tasked with evaluating these scenarios, and their outputs were compared with human judgments. Using the dataset generated by this research, which includes both LLM and Redditor reasonings for the same dilemmas, this thesis examines the embeddings of those reasonings. Our hypothesis is that exploration of these deeper representational spaces will potentially uncover more nuanced signals about how LLMs invoke and apply different moral principles.

Thus, by shedding light on the hidden layers of reasoning that inform LLM decisions in moral contexts, this work contributes to the growing body of knowledge aimed at establishing consistent ethical frameworks in the GenAI field.

2

Current Context

In the domain of NLP, text embeddings have revolutionized the way language data is processed and interpreted. At their core, embeddings are high dimensional, continuous vector representations of words or tokens designed to capture semantic and syntactic information. Early methods of word embeddings learned one fixed vector per word by training neural networks on large text corpora—for example, Word2Vec (MIKOLOV et al., 2013) and GloVe (PENNINGTON; SOCHER; MANNING, 2014). In these models, words that appear in similar contexts would end up with similar vectors.

More recently, contextual embeddings have become the standard. Following the development of the transformer architecture in 2017 (VASWANI et al., 2017), models like Google's BERT began leveraging this architecture to generate context-aware embeddings, which take into account both preceding and succeeding words in a sentence (DEVLIN et al., 2018). This breakthrough in contextual understanding then led the way for newer models such as OpenAI's GPT series likewise, which learn rich, context-sensitive representations via language modeling. Today, embeddings are the fundamental building blocks of every large language model. They encode semantic relationships—such that words or phrases with similar meanings are mapped to similar vectors—and facilitate more sophisticated operations for comparing and combining meanings within and across texts.

In the context of evaluating LLMs' moral reasoning, extensive research has been devoted to understanding how norms and values are encoded in these systems. A common approach involves probing models with standardized experimental material, moral questionnaires or culturally targeted dilemma sets, such as the Moral Foundations Theory or the World Values Survey, and comparing their outputs with human responses (NUNES et al., 2024; HÄMMERL et al., 2022; MEIJER; MOHAMMADI; BAGHERI, 2024). More recently, benchmarks like MoralBench have compiled hundreds of ethical scenarios and developed quantitative metrics to assess how well models' decisions align with human moral standards (JI et al., 2024).

Parallel to that, researchers have examined the ideological leanings of LLMs. Studies report that several conversational LLMs produce outputs that tend to align with center-left political ideologies. For instance, Rozado ran 11 political

orientation tests across 24 different models and found that most of them produced liberal-leaning answers (ROZADO, 2024). Similarly, Evans et al. conducted a voting simulation using an instruction-tuned GPT-4 model from ChatGPT, which showed a preference for Biden over Trump (EVANS et al., 2025), although not all results are uniform in this matter. Overall, these approaches highlight important trends but also face challenges, such as LLM generating highly sensitive responses to how questions are phrased (OH; DEMBERG, 2025) or “moral scores“ that do not significantly correlate with cross-country human surveys (MEIJER; MOHAMMADI; BAGHERI, 2024).

At the intersection of embedding studies and moral reasoning, emerging research explores the extent to which moral values can be decoded directly from LLM embedding spaces. A study by Fitz et al. demonstrated that sentence embeddings from GPT-3.5 could be decomposed into distinct subspaces corresponding to fair and unfair moral judgments, suggesting that the model develops an internal representation of fairness during training (FITZ, 2023). Similarly, Schramowski et al. found that BERT-based models exhibit a "moral dimension" within their embeddings, which can be identified through principal component analysis (FREIRE et al., 2024). Still in this vein, another team of researchers developed MoralBERT, a fine-tuned version of BERT trained to detect moral sentiment in social discourse, drawing from Moral Foundations Theory to assess how moral values are reflected in language (PRENIQI et al., 2024).

Despite these advancements, significant limitations remain in extracting moral reasoning from embedding spaces. The entanglement of moral values with other semantic features in high-dimensional spaces makes it difficult to isolate specific ethical dimensions. In addition, the context-dependent nature of moral judgments poses challenges for static embedding analyses, as actions deemed moral in one situation may be seen as immoral in another. Another difficulty is evaluating the moral content of embeddings in the absence of standardized metrics for cross-model comparisons and assessments of alignment with human ethical standards. Finally, as with text-based moral evaluations, the reliance on culturally specific datasets may limit the generalizability of findings across diverse moral frameworks. All these challenges highlight the need for more nuanced methodologies to better understand and interpret the moral dimensions encoded within LLM embeddings.

3
Research Objectives

Building on the foundation of existing research, this thesis seeks to employ established data processing, machine learning, and statistical techniques to identify and compare the embedding spaces of textual moral reasonings produced by Large Language Models across a range of ethical scenarios. More specifically, the functional requirements for this project are as follows:

1. **Dataset acquisition and preparation**
   Acquire a dataset containing real-world ethical and moral scenarios, along with a comprehensive set of reasoned judgments on those scenarios. For this project, the dataset produced in the study *"Normative Evaluation of Large Language Models with Everyday Moral Dilemmas"* will be used, with permission from its authors. This dataset includes over 10,000 submissions of moral and normative dilemmas arising from everyday situations, originally posted by users of the subreddit “r/AmItheAsshole.” In addition to the original submissions, the dataset contains both the aggregated moral verdicts and accompanying reasonings provided by other Reddit users, as well as those produced by seven different LLMs in response to the same scenarios.
2. **Embedding generation and management**
   Perform an embedding analysis of the dataset by embedding each moral dilemma submission, the corresponding human-provided reasonings, and the LLM-generated reasonings using a consistent embedding model to ensure comparability. The model selected for this purpose is RoBERTa, derived from Google's BERT, which has been academically validated for its ability to grasp nuanced semantic meaning in context. Currently, the generated embeddings are being stored in CSV files.
3. **Comparative analysis of embedding spaces**
   Conduct a critical analysis of the embeddings generated in the previous step. This includes comparing embeddings of different LLM reasonings for the same moral dilemma, comparing reasonings across different dilemmas, and assessing the degree to which LLM-generated embeddings align with those derived from human reasonings. For similarity comparison and retrieval, the *cosine_similarity* function from the *scikit-learn* library is used, as it provides a reliable measure of vector similarity suitable for analyzing relationships in high-dimensional

embedding spaces. Other specific statistical methods to be used for this analysis are still under discussion.

4. **Cross-cultural dataset extension**

   One potential extension of this research involves expanding the dataset to include moral dilemmas from a broader range of cultural and ethical backgrounds, given that the Reddit community—and consequently the dataset used in the study—does not represent a demographically or culturally balanced sample, but it is primarily an echo chamber of younger, male, liberal-leaning, and predominantly American users (SHATZ, 2017), which may limit the generalizability of findings.

All tasks described above will be implemented using Python 3, due to its extensive support for machine learning libraries, data processing, and data visualization. All preprocessing steps, code, and analysis results will be made publicly available via a GitHub repository.

At last, the work presented in this report distinguishes itself from existing approaches by (i) focusing on isolating the moral features encoded within the embedding layer of LLMs, (ii) conducting a broader analysis that considers multiple moral and cultural frameworks, and (iii) contributing toward the early development of standardized metrics for cross-model comparison and alignment with human ethical standards within the embedding space.

4
Research Plan Review

To organize the Final Project, a task schedule was developed based on the previously mentioned requirements, which were established in collaboration with the supervisors of this work. The tasks for the first project phase are listed below:[1]

1. Conduct a literature review and investigate the state of the art in embedding generation in the context of Large Language Models, as well as the applications of LLMs in ethical dilemmas and multicultural scenarios.
2. Extract the dataset produced in the study *"Normative Evaluation of Large Language Models with Everyday Moral Dilemmas"* and perform data cleaning.
3. Generate embeddings for the textual fields in the dataset and save the results to a document.
4. Conduct a qualitative analysis comparing the embeddings generated by the different models and evaluate preliminary results.
5. Write the Final Project I report.

Overall, all tasks planned for the first phase of the project were completed within the predetermined deadlines. It is worth noting that one major difficulty encountered during this period was the initial attempt to scrape new posts from the subreddit “r/AmItheAsshole,” due to limitations on the number of posts retrievable through the Reddit API, as well as the significant amount of time required to build a dataset of reasonable size. Fortunately, the original dataset from our reference study was made available by its authors. A second difficulty involved generating meaningful representations in the embedding space, as the first attempt using the RoBERTa model did not yield good results in cosine similarity comparisons. The model was then replaced with a sentence-transformer, which produced results within the expected range.

As for the pending tasks for the second phase of the project, the following can be listed:

1. Extend the analysis from the first phase by comparing embeddings of LLM-generated texts with those of the original Reddit submission texts.

[1] The project, along with its preliminary results, is available on GitHub at: https://github.com/danielhuf/puc-projeto-final/

2. Enrich dataset by generating or externally obtaining new ethical dilemmas that reflect alternative cultural contexts (e.g., from nationalities or ethnicities not represented in the baseline study).
3. Produce responses to these new dilemmas using different LLMs.
4. Apply the same pipeline from the first phase (data cleaning, embedding generation, qualitative analysis of results).
5. Write the Final Project II report.

5

Timeline

A Gantt chart for each semester of the project has been created, where each row represents a task and the time required for its completion.

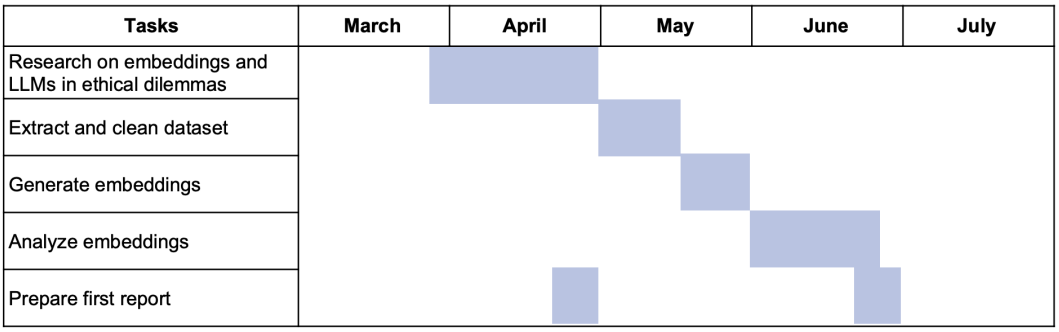

Figure 5.1 - First Semester Timeline

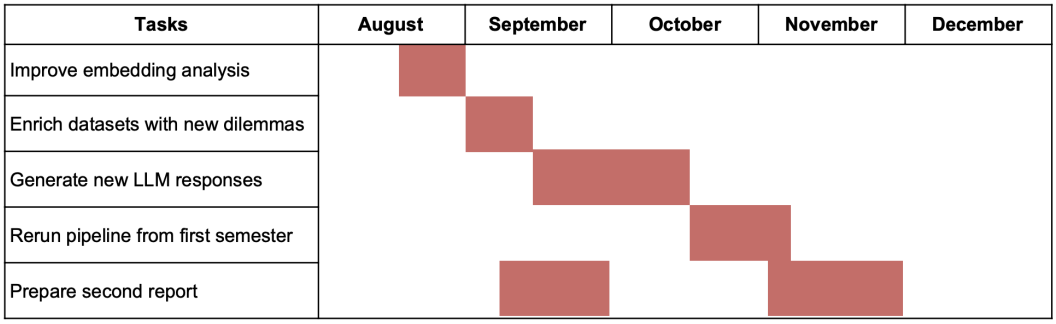

Figure 5.2 - Second Semester Timeline

References

CAO, Yong *et al.* **Assessing Cross-Cultural Alignment between ChatGPT and Human Societies: An Empirical Study**. Disponível em: <https://arxiv.org/abs/2303.17466v2>. Acesso em: 20 maio. 2025.

DEVLIN, Jacob *et al.* **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. Disponível em: <https://arxiv.org/abs/1810.04805v2>. Acesso em: 29 maio. 2025.

EVANS, James *et al*. **Finding political leanings in large language models | The University of Chicago Division of the Social Sciences**. Disponível em: <https://socialsciences.uchicago.edu/news/finding-political-leanings-large-language-models>. Acesso em: 29 maio. 2025.

FITZ, Stephen. **Do Large GPT Models Discover Moral Dimensions in Language Representations? A Topological Study Of Sentence Embeddings**. Disponível em: <https://arxiv.org/abs/2309.09397v1>. Acesso em: 29 maio. 2025.

FREIRE, Pedro *et al.* **Uncovering Latent Human Wellbeing in Language Model Embeddings**. Disponível em: <https://arxiv.org/abs/2402.11777v1>. Acesso em: 29 maio. 2025.

HÄMMERL, Katharina *et al.* **Speaking Multiple Languages Affects the Moral Bias of Language Models**. Disponível em: <https://arxiv.org/abs/2211.07733v2>. Acesso em: 29 maio. 2025.

HUA, Yining *et al.* **Large Language Models in Mental Health Care: a Scoping Review**. Disponível em: <https://arxiv.org/abs/2401.02984v2>. Acesso em: 20 maio. 2025.

JI, Jianchao *et al.* **MoralBench: Moral Evaluation of LLMs**. Disponível em: <https://arxiv.org/abs/2406.04428v1>. Acesso em: 29 maio. 2025.

JOHNSON, Rebecca L. *et al.* **The Ghost in the Machine has an American accent: value conflict in GPT-3**. Disponível em: <https://arxiv.org/abs/2203.07785v1>. Acesso em: 20 maio. 2025.

LIU, John Zhuang; LI, Xueyao. **How do judges use large language models?** Evidence from Shenzhen. Journal of Legal Analysis, v. 16, n. 1, p. 235–262, 1 jan. 2024.

MEIJER, Mijntje; MOHAMMADI, Hadi; BAGHERI, Ayoub. **LLMs as mirrors of societal moral standards: reflection of cultural divergence and agreement across ethical topics**. Disponível em: <https://arxiv.org/abs/2412.00962v1>. Acesso em: 29 maio. 2025.

MIKOLOV, Tomas *et al.* **Efficient Estimation of Word Representations in Vector Space**. Disponível em: <https://arxiv.org/abs/1301.3781v3>. Acesso em: 29 maio. 2025.

NUNES, J. L.; ALMEIDA, G. F. C. F.; ARAUJO, M. de; BARBOSA, S. D. J. **Are Large Language Models Moral Hypocrites? A Study Based on Moral Foundations**. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, [S. l.], v. 7, n. 1, p. 1074-1087, 2024. DOI: 10.1609/aies.v7i1.31704. Disponível em: https://ojs.aaai.org/index.php/AIES/article/view/31704. Acesso em: 24 jun. 2025.

OH, Soyoung; DEMBERG, Vera. Robustness of large language models in moral judgements. **Royal Society Open Science**, abr. 2025.

PENNINGTON, Jeffrey; SOCHER, Richard; MANNING, Christopher D. GloVe: **Global Vectors for Word Representation**. *In*: PROCEEDINGS OF THE 2014 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING (EMNLP). **Anais**... out. 2014. Disponível em: <https://aclanthology.org/D14-1162/>. Acesso em: 29 maio. 2025

PIAO, Jinghua *et al.* **AgentSociety: Large-Scale Simulation of LLM-Driven Generative Agents Advances Understanding of Human Behaviors and Society**. Disponível em: <https://arxiv.org/abs/2502.08691v1>. Acesso em: 20 maio. 2025.

PRENIQI, Vjosa *et al.* **MoralBERT: A Fine-Tuned Language Model for Capturing Moral Values in Social Discussions**. Disponível em: <https://arxiv.org/abs/2403.07678v2>. Acesso em: 29 maio. 2025.

ROZADO, David. **The Political Preferences of LLMs**. Disponível em: <https://arxiv.org/abs/2402.01789v2>. Acesso em: 29 maio. 2025.

SACHDEVA, Pratik S.; VAN NUENEN, Tom. **Normative Evaluation of Large Language Models with Everyday Moral Dilemmas**. Disponível em: <https://arxiv.org/abs/2501.18081v1>. Acesso em: 23 maio. 2025.

SHATZ, Itamar. **Fast, Free, and Targeted**. Soc. Sci. Comput. Rev., v. 35, n. 4, p. 537–549, 1 ago. 2017.

VASWANI, Ashish *et al.* **Attention Is All You Need**. Disponível em: <https://arxiv.org/abs/1706.03762v7>. Acesso em: 24 jun. 2025.

ZANGARI, Lorenzo *et al.* **A survey on moral foundation theory and pre-trained language models: current advances and challenges.** AI & SOCIETY, p. 1–26, 24 mar. 2025.