AI-THOS

Team Name - Binary Bandits

[AG16]

Team Member

1. Shreshth Sharma
2. Subhrajit Mukherjee
3. Vibhanshu Sharma
4. Dhruvil Patel
5. Erum Fatima

TABLE OF CONTENT

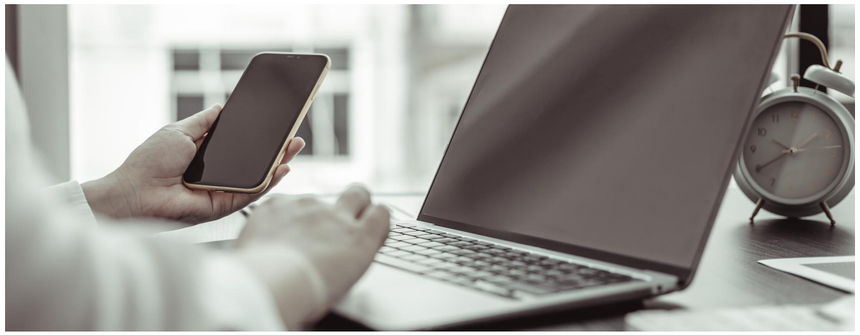
PROBLEM STATEMENT

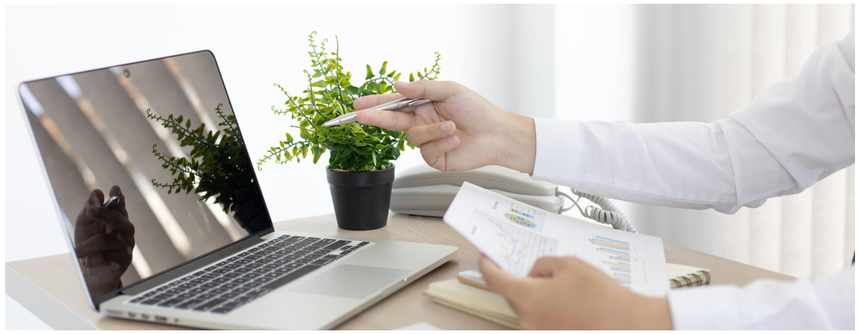
Morality Engine – Dilemmas in Autonomous Decision-Making

Autonomous systems, such as self-driving cars or AI bots, face complex moral decisions that require more than pre-programmed rules. For example, a self-driving car may need to decide between minimizing harm to its passenger or avoiding a pedestrian. Similarly, AI content moderators face decisions about balancing freedom of speech with preventing harm. Existing systems use static, rule-based approaches, but ethical decisions need to adapt dynamically based on varying moral values.

You are tasked with designing a moral decision-making framework for autonomous systems, simulating decisions for real-world scenarios and allowing users to adjust the ethical parameters influencing those decisions.

SOLUTION





01 We have created a **web interface for the user** to give a scenario and four options to get a response - give a random user bias, choose from the moral frameworks, and simulate all possible decisions using rationale from the moral frameworks and responses expected from an average human.

02 We have also created **a question-answer interface for users** to choose their decisions from randomly generated 15 questions and decisions and give an overall personality analysis based on the three moral frameworks.

03 We have also implemented **real-time decision-making** using data inputs from live news events and analyzed each decision and scenario.

META PROGRAMMING FRAMEWORK

We have used AI Agents with an LLM as the backbone for the creation of the following entities as an agent:

- **Utilitarian Ethics Philosopher**
- **Deontological Ethics Philosopher**
- **Virtue Ethics Philosopher**
- **Opportunistic Human Being**
- **Human Being with user-defined traits**
- **Questionnaire Expert**
- **Personality Analysis Expert**
- **Socially Responsible News Reporter**

MODEL DETAILS

Model: Llama-3.3-70B-Instruct

We have used a serverless inference API by Together AI which has the following advantages:

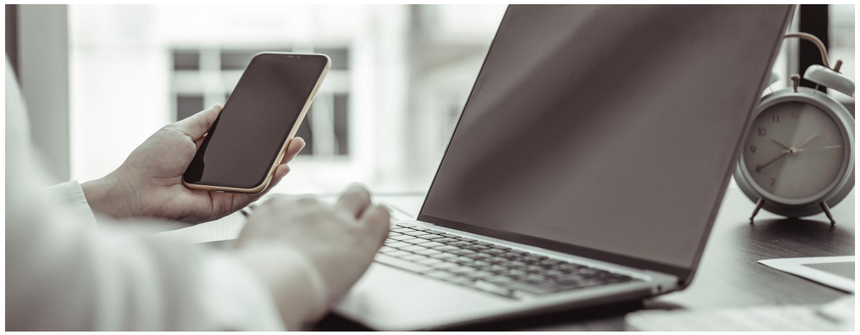
- It has a latency of 2.5-3.0 seconds per query

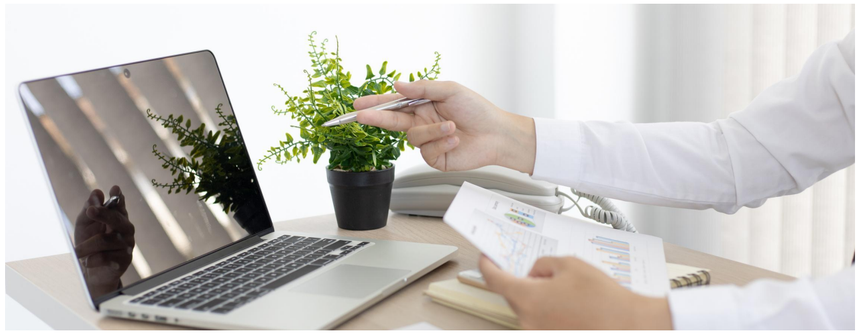
- It is instruction-tuned, making it suitable for AI Agents

- The model is open-source.

- The endpoint used is free-tier with no response limits, making it more scalable

BENCHMARKING





We have benchmarked our model on the ETHICS dataset, a common standard for evaluating the ethical nature of LLMs. The dataset consisted of examples from three separate moral frameworks in the problem statements. We have achieved an overall **F1-Score of 0.856** and a **Matthews Correlation Coefficient (MCC) of 0.721.**

FUTURE SCOPE

01 Latency improvement of the model

02 Use of quantized model for CPU inferencing

03 Edge device deployment

04 Extension to autonomous vehicles

REFERENCES

- Hong, S., Zheng, X., Chen, J., Cheng, Y., Wang, J., Zhang, C., Wang, Z., Yau, S.K.S., Lin, Z., Zhou, L., Ran, C., Xiao, L. and Wu, C. (2023). MetaGPT: Meta Programming for Multi-Agent Collaborative Framework. [online] arXiv.org. doi:https://doi.org/10.48550/arXiv.2308.00352 .

- Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A. and Rodriguez, A. (2024). The Llama 3 Herd of Models. arXiv (Cornell University). doi:https://doi.org/10.48550/arxiv.2407.21783

- Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D. and Steinhardt, J. (2021). Aligning AI With Shared Human Values. arXiv (Cornell University). doi:https://doi.org/10.48550/arxiv.2008.02275 .

THANK YOU