Towards AI Collaborators: Exploring Goal, Value, and Role-Based Alignment in AI

Claas Beger
Cornell University, Department of Computer Science, Ithaca NY, USA

Motivation & Problem Statement

Motivation

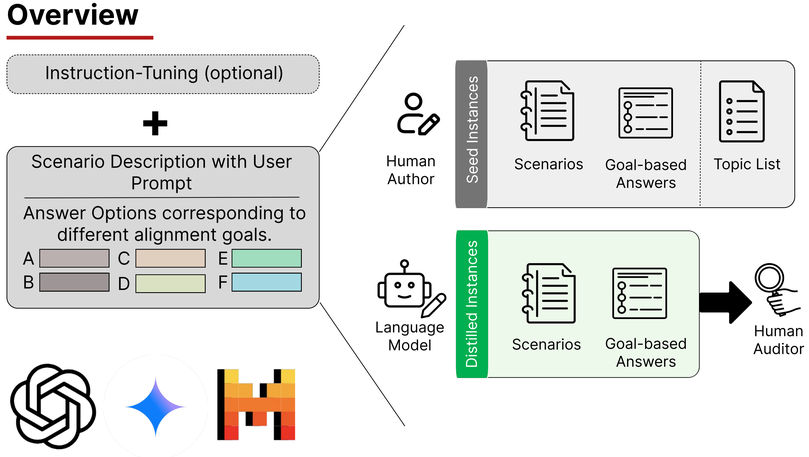
- Current alignment techniques focus on surface-level helpfulness, but often **fail to instill consistent ethical reasoning or goal-based behavior**.
- Real-world decisions frequently **involve trade-offs among competing goals—such as user instructions, well-being, and moral values** which existing models are **not trained to navigate**.
- Artificial Intelligence, Values and Alignment (2020)[1] outlines **six distinct alignment goals**, yet, it is unclear whether current alignment strategies enable them in trained models

Problem Statement

- I **revisit current alignment approaches** to discuss whether they enable specific alignment goals as discussed by Gabriel
- Using 15 hand-written scenario prompts with corresponding replies for each of the outlined alignment goals, I use different Large Language Models to distill 85 additional scenarios to create a **synthetic goal-based dataset**.
- With this, I test whether alignment of current state-of-the-art models **follow one of the outlined alignment goals consistently or features any biases**
- Using **basic instruction-tuning** I test whether there is a meaningful shift from prior results or a need for dedicated goal-based alignment

Methodology

Overview



Conclusion

Existing Alignment Methods

- Targeting specific alignment goals require varying insights and reasoning in alignment training - existing methods like SFT and RLHF primarily[2] target Instruction-following.
- Alternative techniques like RLCAI and Dromedary target alignment with socially accepted value sets or the notions of Helpfulness and Harmlessness.
- It is unclear whether such techniques are directly applicable to the analyzed alignment goals, which partially require long-context knowledge about the end-user, and non-trivial dataset generation

No Consistent Goal-Based Selections in Current Models

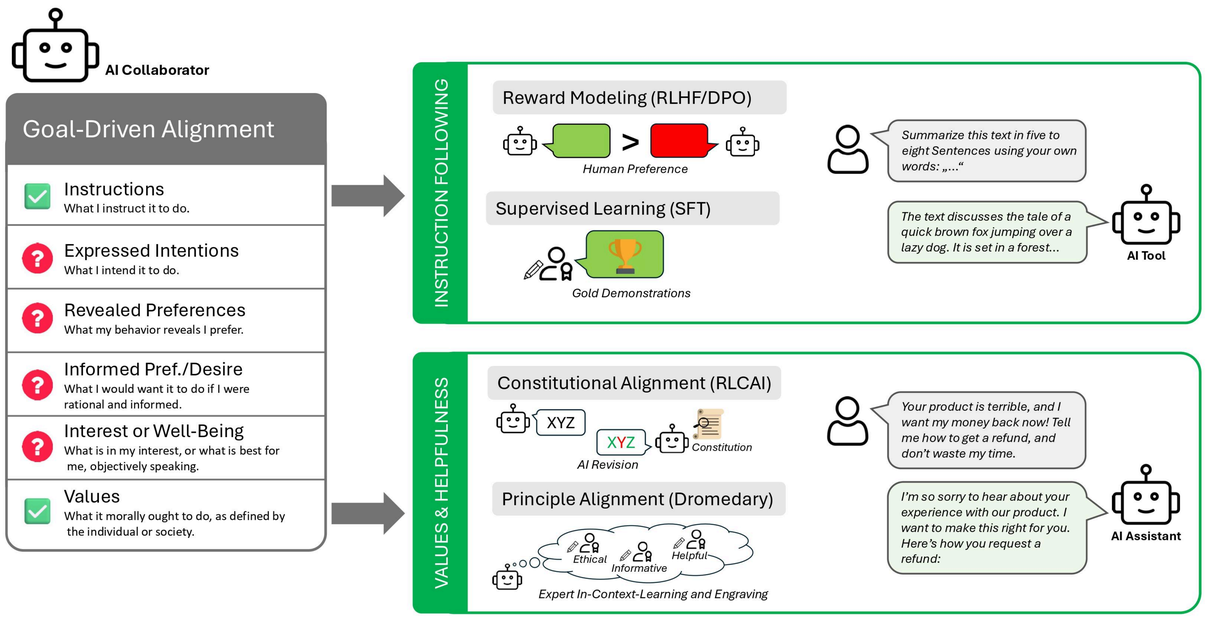
- There is no consistent goal as defined by [1], when eliciting multiple choice selections for the used dataset of 100 scenarios.
- Slight differences can be observed between the tested models, such as a larger share of value-based selections by Gemini, and smaller fractions of the target of Interest Or Well-being in GPT4o

Sole Instruction-Tuning is Insufficient

- Using hand-written instructions which are appended to the scenario and selection prompts yield mixed results across models and goal types.
- Accuracies are in the range of 0.65 to 0.53, generally with high standard deviation of 0.4-0.5
- Individual accuracies for a specific goal and model are vastly different, with GPT4o reaching 0.9 while instructed to select according to Informed Preferences or Desires and Mistral-Large only reaching 0.15 with instructions on pursuing Expressed Intentions.

Background

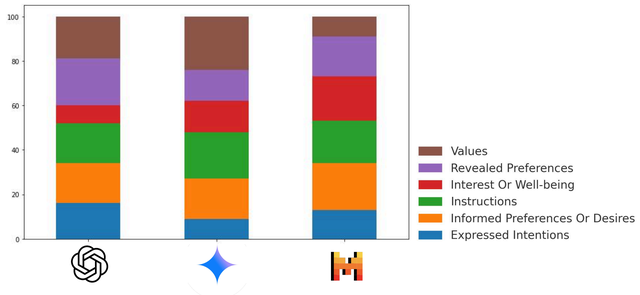
Alignment Classification and Related Techniques



Results

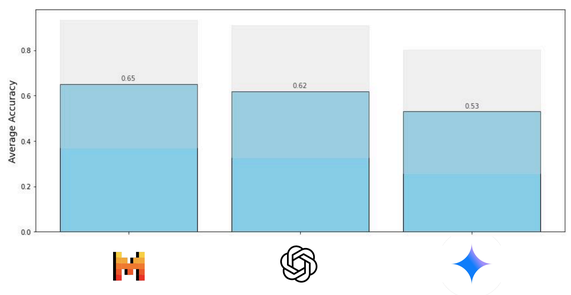
Base Completions

Overview of selected scenario completions by model



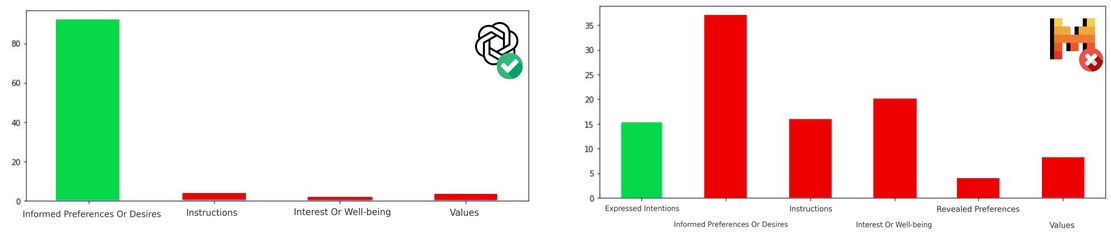
Instruction-Tuned

Average Accuracy using Instruction-Tuning with shaded Standard Deviations



Selected Performances using Instruction-Tuning

Instruction-tuned selection of scenario replies by selected models. Correct choice is displayed in green, incorrect choices are displayed in red



References

[1] Gabriel, Iason "Artificial Intelligence, Values, and Alignment" ArXiv, 2020, https://arxiv.org/abs/2001.09768.
[2] In current applications, revealed preferences is a misleading alternative since preferences are revealed in retrosight, Gabriel defines revealed preferences as revealed through a person's behaviour rather than through expressed opinion.