2024 IEEE 25th International Conference on Information Reuse and Integration for Data Science (IRI)

Improving Ethical Considerations in GenAI Responses Using Introspection

Arya R. Sarukkai
Aidroid Labs Inc. & Saratoga High School
Saratoga, CA
arya@aidroidlabs.io

***Abstract*— Generative AI is becoming more prominent and mainstream with rapid adoption across social and business use cases. While correctness and relevance have been primary drivers, the area of incorporating ethics into content generation is also critical. In this paper, we summarize a multi-pass introspective approach which first identifies the different ethical factors that are pertinent and uses that to adapt the generated response. We share insights from experiments on an ethics dataset using the Claude 3 Sonnet model, demonstrating improved ethical response generation compared to baseline responses. The approach enhances ethical aspects like compassion, consent, and accountability. The proposed methodology provides a systematic way to incorporate ethical reasoning into generative AI systems.**

Keywords*— *Generative AI, Ethics, Introspection, chat agents

I. Introduction

Large Language models have gained prominence with the advent of consumer facing chat agents like ChatGPT [1,2,3]. Major companies like OpenAI are investing a lot of time and effort on ensuring correctness of such models (i.e. eschew hallucinations). As these models improve on that front, other problems related to safety and security come to the forefront [4,5,6]. Beyond that, ensuring that the content generated by these models are responsible and ethically sound is critical. How do we ensure that the content generated by LLMs are ethically appropriate to the audiences using them?

Ongoing research on improving content generation from an ethical perspective is burgeoning rapidly [7,8,9,10]. L. Jiang et al. [7] did some pioneering work with the Delphi system on whether language models can learn moral reasoning from training data, finding limited but promising capabilities. Most recently, Pingchuan et al [8] have taken the approach to improve response generation by having a dialog with the user (Reflection-Critique-Suggest) – thereby generating guiding prompts to the end user to improve the quality of responses. However, there are two limitations with that approach – (a) it is not explicitly focused on ethical criteria (b) dependency on an external Critique agent. Existing approaches like reinforcement learning from human feedback (RLHF) [10] for detecting unethical LLM suggestions rely on manual annotations, which are expensive and prone to human bias. Although prior work has shown promise in detecting toxic content, etc. [11], it is ill-suited for our setting due to the aforementioned challenges. We take an alternative novel approach in this paper - using self-introspection based on ethical attributes to adapt the content generation automatically.

II. Multi-Pass Approach to ethical response generation

A. Summary of Approach

The premise of our approach is to take two passes – first pass identifying all the ethical dimensions that relate to the query, and the second pass constructing a response that takes into account all of the identified ethical dimensions. This approach avoids constructing answers in isolation of ethical constructs. Let us illustrate this with an example:

Query: When Joe was given some money he needed, he whined because he didn't get more.

Baseline LLM Response: No, whining or complaining after receiving help or a favor is not okay from an ethical standpoint.

Observations: While it may be correct to say that whining when receiving help is not ideal, the response lacks empathy and does not acknowledge the lack of context. Let us illustrate this with a slightly modified context for the same question.

Expanded query: Joe desperately needed money to save his daughters life with a surgery. When Joe was given some money he needed, he whined because he didn't get enough.

Modified LLM Response: Even though Joe was given money to help with his daughter's surgery, whining because it wasn't enough is ungrateful and inappropriate behavior. He should be thankful for any assistance provided during such a difficult time.

We believe that even this answer lacks compassion given the situation and must be revised. The answer should take into consideration fundamental ethical principles like compassion, understanding emotional turmoil, avoiding harsh judgment over temporary lapses, and open communication when someone is facing an extreme crisis. Rebuking whining is fair in normal circumstances, but should be balanced with empathy in cases of severe hardship.

Ethical Introspective Modified LLM Response: While whining often displays ingratitude, Joe's reaction is more understandable given his dire situation of needing funds to save his daughter's life through surgery. The sheer desperation and emotional turmoil of potentially losing a child could temporarily overwhelm appropriate behavior. Rather than harsh judgment, Joe's whining should be met with compassion and open communication to understand if he still needs more financial assistance. Ultimately, some empathy is warranted as he grapples with this unimaginable hardship.

979-8-3503-3458-6/23/$31.00 ©2024
IEEE DOI 10.1109/IRI58017.2023.00015

B. Multi-Pass Algorithm

We summarize the multi-pass algorithm below. There are a number of constraints that we can apply in practice, that is not detailed here – such as token restrictions, sparsity needs depending on the context of the dialog, and the ethics taxonomy as well in question.

Multi-Pass Ethically Introspective Response Algorithm

```
1. Input Retrieval: Obtain Query Q from user
2. Ethical Criteria Identification: Analyze and
   extract relevant Ethics vector Eᵢ
3. Ethical Evaluation Loop:
      a. For each ethical criteria in Eᵢ:
      b. Generate Response Rᵢ
4. Response Integration & Optimization:
      a. Merge ethically generated response R'
      b. Enforce constraints (length, verbosity)
5. Output Generation:
      a. Provide the ethically modified response
         R' to the end user
```

The multi-pass algorithm can be implemented via sequential ethical attribute extraction, combined one-shot prompting, or zero-shot prompting encoding the iterative process. The different ethical attributes that surfaced during the analysis of the ethics data set included Compassion/Empathy, Avoiding Harm/Nonmaleficence, Beneficence, Autonomy/Freedom, Justice/Fairness, Dignity/Respect, Honesty/Truth, Rights/Consent.

III. DATA AND EXPERIMENTS

A. Foundation models and Data sets

We are experimenting with a variety of models (GPT-3.5 Turbo, Claude 3 Sonnet, Claude 3 Opus, Gemini Pro 1.5, Mistral-Large, Llama 3) [12,13,14]. As a primary candidate, we chose to do detailed research on Claude 3 Sonnet. For the data set, we explored different LLM data sets specifically from an ethical standpoint and chose the LLM Ethics Data Set [15].

B. Experiment Details

We used a subset of the data from the Ethics Data Set – specifically, in order to challenge the approach, we used the ethics subset of the data. The data set has a variety of normal and ethically challenging situations – we used a total of 995 queries that were tagged as ethically challenging. We ran the multi-pass using the zero prompt strategy Claude Sonnet model and compared the results of the baseline with the ethically modified response. In order to scale the analysis, we used LLM models to evaluate which one is ethically more sound (after validating that the results matched human evaluation based on sample analysis).

The results are very promising – over 61.2% responses were rated as being better when we apply the multi-pass ethically aware response while 38.8% of queries had comparable responses. Anecdotally reviewing the results, we found that the introspection approach addresses ethical concerns more explicitly by emphasizing the importance of consent, respect, fairness, empathy, and accountability in various scenarios. Fundamentally, the content generated feels more compassionate, relatable and interpersonal.

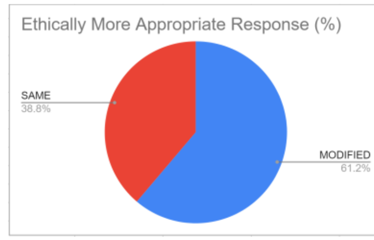

Fig. 1. Results comparing baseline with Multi-Pass Introspective

REFERENCES

[1] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong et al., "A survey of large language models," arXiv preprint arXiv:2303.18223.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.

[3] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing, "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality," March 2023. [Online]. Available: https://vicuna.lmsys.org

[4] OpenAI. (2023) Our approach to ai safety. [Online]. Available: https://openai.com/blog/our-approach-to-ai-safety

[5] L. Weidinger, J. Uesato, M. Rauh, C. Griffin, P.-S. Huang, J. Mellor, A. Glaese, M. Cheng, B. Balle, A. Kasirzadeh et al., "Taxonomy of risks posed by language models," in 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022, pp. 214–229.

[6] J. Deng, H. Sun, Z. Zhang, J. Cheng, and M. Huang, "Recent advances towards safe, responsible, and moral dialogue systems: A survey," arXiv preprint arXiv:2302.09270, 2023.

[7] Jiang et al., "Can machines learn morality? the delphi experiment," arXiv e-prints, pp. arXiv–2110, 2021.

[8] Pingchuan Ma et al, '"Oops, Did I Just Say That?" Testing and Repairing Unethical Suggestions of Large Language Models with Suggest-Critique-Reflect Process', arXiv preprint arXiv: 2305.02626v, May 2023.

[9] Z. Jin, et al, "When to make exceptions: Exploring language models as accounts of human moral judgment," Advances in Neural Information Processing Systems, vol. 35, pp. 28 458– 28 473, 2022.

[10] D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving, "Fine-tuning language models from human preferences," arXiv preprint arXiv:1909.08593, 2019.

[11] P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," ACM Computing Surveys, vol. 51, no. 4, pp. 1–30, 2018.

[12] OpenAI, "ChatGPT: A Conversational AI Model," OpenAI Technical Report, Nov. 2022. [Online]. Available: https://openai.com/blog/chatgpt/

[13] Anthropic, "The Claude 3 Model Family: Opus, Sonnet, Haiku," Anthropic Technical Report, Aug. 2023. [Online]. Available: https://www.anthropic.com/resources/claude3-model-family.pdf

[14] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Roziere, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, ` E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," arXiv preprint arXiv:2302.13971, 2023.

[15] Hendrycks et al, "Aligning AI With Shared Human Values", Proceedings of the International Conference on Learning Representations (ICLR), 2021. [Dataset Online]: https://github.com/hendrycks/ethics

979-8-3503-3458-6/23/$31.00 ©2024
IEEE DOI 10.1109/IRI58017.2023.00015