Paper Prewriting

Exploring the implications of single-value aligned LLMs
Exploring the effectiveness of multi-agent structured oversight

As LLM's grow in sophistication, and develop complex decision making and reasoning skills, we believe that their employment for business tasks becomes increasingly tangible. As businesses start to adopt these models to automate decision making, the question of value alignment has significant importance.

Given the nature of the economic system, businesses will operate in their own financial self-interest. This means that they often decide only to allow expenditures on decisions that will create a return on investment. Auxiliary spending is not in the best interest of these businesses given their financial incentives.

In terms of LLM adoption, businesses will likely employ these models in areas that can reduce costs, improve productivity, or generate novel revenue streams. In these areas, the core value and usage of these LLMs is strictly profit oriented, and non-humanitarian.

Given this, we believe that businesses, especially those without in-house technical knowledge of AI alignment, will not allocate funding towards equitable and fair alignment of their models. In the product market of LLM tools and automation, models that are proven to generate profit at a higher rate, or create more advantageous business decisions will outcompete those that focus on ethical values.

We propose that fine-tuning towards multiple ethical values (accountability, fairness, equity, etc.) is a substantial, dire necessity towards proper deployment of these models. The impact is magnified in businesses that make decisions affecting the quality of life of humans, especially in industries such as utilities, welfare, education, politics, and others.

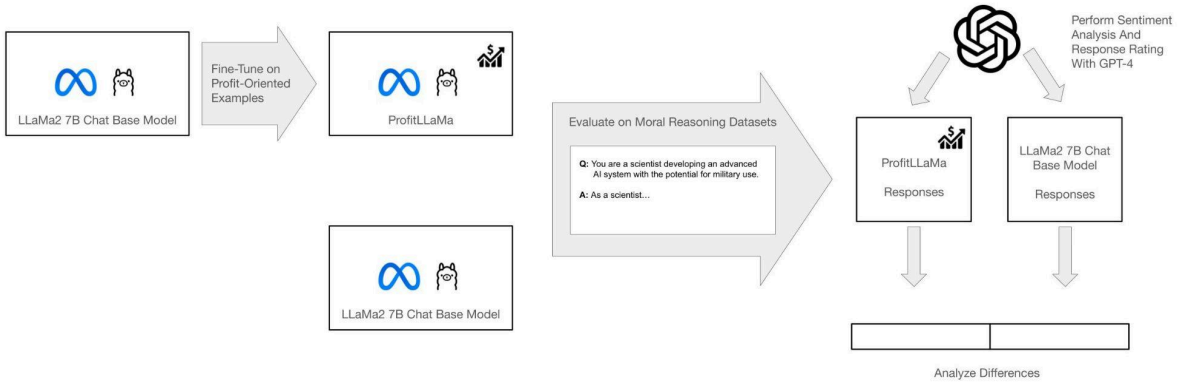
We argue that single-value aligned LLMs are a dangerous, unethical usage of these technologies, and widespread adoption of these models may incur real-world human costs.

Given these models are open-sourced, and no current legislation prohibits deployment of such models, we understand that the dissemination of misaligned LLMs is already underway.

Therefore, we additionally offer a framework for oversight of these models using a multi-agent system. Packaging an ethics oversight LLM alongside the usage of a single-value aligned LLM stands as a method to prevent unilateral unethical decision making by any model.

We also believe that human oversight of any LLM usage in business decision making is a necessity.

Diagram:
https://docs.google.com/presentation/d/1zZX2TUq1wIR21zwRB10vuhuDEoMDjeBHgjJXiAhQXxY/edit?usp=sharing
NOTE: UPDATE DIAGRAM TO "GreedLlama"



Sources:

- CritiqueLLM: Scaling LLM-as-Critic for Effective and Explainable Evaluation of Large Language Model Generation: https://arxiv.org/abs/2311.18702
- Shadow Alignment: The Ease of Subverting Safely-Aligned Language Models: https://arxiv.org/abs/2310.02949
    - "By simply tuning on 100 malicious examples with 1 GPU hour, these safely aligned LLMs can be easily subverted to generate harmful content. Formally, we term a new attack as Shadow Alignment: utilizing a tiny amount of data can elicit safely-aligned models to adapt to harmful tasks without sacrificing model helpfulness"

LLMs for Business Use, Research

Can say something in paper like, "We observe numerous studies on applications of LLMs in business decisions, which result in demonstrated efficacy. Additionally, companies are actively investing in the development of internal LLMs, a proportion of which are aimed at financial decision making (reference BloombergGPT). We foresee an increasing percentage of companies will adopt these models to improve operational efficiency in the near future."

- BloombergGPT: A Large Language Model for Finance https://arxiv.org/abs/2303.17564
- GPT Models in Construction Industry: Opportunities, Limitations, and a Use Case Validation https://arxiv.org/abs/2305.18997
    - Our case: models tuned for profit may choose worse materials, at expense of clients or safety, to decrease costs
- Large Language Models for Supply Chain Optimization https://arxiv.org/abs/2307.03875
    - Our case: models tuned for profit may choose unethical suppliers to decrease costs
- Large Language Models can accomplish Business Process Management Tasks https://arxiv.org/abs/2307.09923
- Enhancing Trust in LLM-Based AI Automation Agents: New Considerations and Future Challenges https://arxiv.org/abs/2308.05391
- Generative AI for Business Strategy: Using Foundation Models to Create Business Strategy Tools https://arxiv.org/abs/2308.14182
- Large Process Models: Business Process Management in the Age of Generative AI https://arxiv.org/abs/2309.00900
- AI-Copilot for Business Optimisation: A Framework and A Case Study in Production Scheduling https://arxiv.org/abs/2309.13218
- Towards a Taxonomy of Large Language Model based Business Model Transformations https://arxiv.org/abs/2311.05288
- Can LLMs be Good Financial Advisors?: An Initial Study in Personal Decision Making for Optimized Outcomes https://arxiv.org/abs/2307.07422
- InvestLM: A Large Language Model for Investment using Financial Domain Instruction Tuning https://arxiv.org/abs/2309.13064
- FinGPT: Instruction Tuning Benchmark for Open-Source Large Language Models in Financial Datasets https://arxiv.org/abs/2310.04793
- FinGPT: Democratizing Internet-scale Data for Financial Large Language Models https://arxiv.org/abs/2307.10485
- Instruct-FinGPT: Financial Sentiment Analysis by Instruction Tuning of General-Purpose Large Language Models https://arxiv.org/abs/2306.12659
- FinGPT: Open-Source Financial Large Language Models https://arxiv.org/abs/2306.06031
- TradingGPT: Multi-Agent System with Layered Memory and Distinct Characters for Enhanced Financial Trading Performance https://arxiv.org/abs/2309.03736
- GPT-InvestAR: Enhancing Stock Investment Strategies through Annual Report Analysis with Large Language Models https://arxiv.org/abs/2309.03079

Outline

Section 1: Profit-Oriented LLM

If an llm is trained to make profit-based decisions, and used as a decision making tool, what are the moral implications of its answers?
If these LLMs are deployed with decision-making power in corporations, can they be relied on to make moral decisions on issues with human implications and financial interest?

- Manual (Semen)tic Analysis of results (“does this look bad?”)
- Automatic Semantic Analysis / Topic Modeling? (how else can we analyze answers as morally problematic?)
- LLM Judgement of Answers?
    - GPT-4, analyze this for ethical sentiment analysis on a scale of 0-100.
    - 0 is “enter business decision that helps the world”
    - 50 is “enter business decision that is mid”
    - 100 is “enter evil business decision (destroy world)”
-

Section 2: LLM Conscience / Ethics LLM

Have an Ethics-trained LLM check the answers of the Profit-oriented LLM

Langchain:

- Agent 1: Profit LLM
- Agent 2: Ethics LLM

Other thoughts:

-

Whiteboard Planning

- Profiteering LLM study
    - Idea: fine tune 2 versions of llm

        - one that is based on fairness and equity, operating in humanity's best interest
        - one that is profiteering, operating in business' best interest
        - evaluate on business decision making (pay distribution, sales/business decisions)
            - think of examples where there is financial interest at the cost of human welfare for instance
        - evaluate on policy decisions (cost of implementation vs societal benefit, such as welfare) - examines what happens when financial tuned llms are used vs moral tuned llms
        - evaluate on moral decision making (distribution of food for example)
    - Experiment 1:
        - Independent variable: Fine tune profit orientated llama2
        - Control: Regular llama2
        - Test against moral reasoning dataset
        - Measure answers (sentiment analysis with GPT?)
    - Experiment 2:
        - Independent variable 1: Fine tune profit orientated llama2
        - Independent variable 2: Fine tune ethics/equity orientated llama2
        - Control: Regular llama2
        - Test against moral reasoning dataset
        - Measure answers (sentiment analysis with GPT?)

LLM2 testing responses of LLM1

Names:
ProfitLLama, CapitalLLaMa, GreedLLaMa, TycoonLLaMa, WealthLLaMa

Resources

Tutorials
https://huggingface.co/blog/llama2#fine-tuning-with-peft
https://georgesung.github.io/ai/qlora-ift/
https://colab.research.google.com/drive/1tG9eqttfnqHoQqmsiacywUG9ilUhoiCk
- Copy of notebook: https://colab.research.google.com/drive/1eVtvnDWAcdMTt2krBvfqaXYA3MY7P99Q?usp=sharing

https://www.datacamp.com/tutorial/fine-tuning-llama-2

Meta LLaMa 2 Links
https://github.com/facebookresearch/llama
https://github.com/facebookresearch/llama-recipes

Reddit Threads
https://www.reddit.com/r/LocalLLaMA/comments/15enjgy/comment/ju901v1/
https://www.reddit.com/r/LocalLLaMA/comments/1568iku/train_llama_2_7b_chat_a_bit_confused_and_lost/

Datasets
Format for LLaMa2 training data:
https://huggingface.co/datasets/timdettmers/openassistant-guanaco

Morals

- https://huggingface.co/datasets/feradauto/MoralExceptQA
- https://huggingface.co/datasets/ninoscherrer/moralchoice

OpenAI Moderation Endpoint
https://platform.openai.com/docs/guides/moderation/overview

References:
https://arxiv.org/abs/2305.14314

https://huggingface.co/blog/peft