5mins / 3 = 100s - practice a bit

Initial Ideas

## Virtue Ethics Guidance of LLM using Reinforcement Learning from AI Feedback

*In this project students will learn more about recent breakthroughs in LLM training that enabled the rise of chatbots (RLHF) and explore ethical problems in AI from a technical and philosophical lens. They will implement a RLAIF prompting recipe based on a virtue ethics framework to generate a new series of responses for a preference model, and then test its efficacy on the Anthropic red teaming dataset.*

* What is ethics? Ethics is different for each person? What about the machine?

## Exploring Similarities in Themes and Style across Age Groups in Children’s Artwork using CLIP

*In this project, students will use a dataset of artwork from an open competition in France that has works separated by age and other categories. They will use CLIP embeddings to find similarities and differences between different groups of artworks through clustering of the embeddings, creating textual descriptions of the artworks for distant reading, and other student chosen image analysis techniques. A possible extension is to generate images using a diffusion model from different average embeddings relative to particular groups and then to perform subject analysis on those images.*

* What are the similarities and differences between different groups of artworks?
  + In color, shape (angular/circular, geometric/organic), style, stroke length and thickness, fluidity/smoothness between strokes
* Where do these differences come from? Culture, age, gender, country, left/right handed
* Psychological aspect of childhood development, how children perceive the world? How do they represent in their mind? How do they communicate it by drawings?
* Can we train a model to generate drawings given a category? <https://github.com/magenta/magenta/tree/main/magenta/models/sketch_rnn>
* Is CLIP a good model for this task? *CLIP also still has poor generalization to images not covered in its pre-training dataset. For instance, although CLIP learns a capable OCR system, when evaluated on handwritten digits from the MNIST dataset, zero-shot CLIP only achieves 88% accuracy, well below the 99.75% of humans on the dataset.*
* ask if there’s any previous work/project for the same dataset.

Another topic of interest to us is Exploring Similarities in Themes and Style across Age Groups in Children's Artwork using CLIP.

This is a great opportunity to get to know something about CLIP model!

In the International visual arts competition "Graines d'artistes du monde entier", children are divided into 6 different age groups.

We are actually using clip embedding technique in clip model which helps us to transform images into high dimensional vectors. Each dimension of the vector represents a different feature. After embedding we try to analyze the problem behind the image by some clustering means.

The directions we currently think we can explore are:

Data visualization: Use the basic information given in the dataset to create visual charts, such as maps, timelines or bar charts, showing the distribution of paintings across countries, age groups and years of submission. This is a relatively straightforward task that can provide visual background information for subsequent analysis.

Explore the broad theme of each year: as a broad theme is given each year, we would like to explore the ways and perspectives in which the theme is portrayed in the works of different age groups. This can be achieved through simple text analysis and image recognition.

Changes in drawing styles of the same age group in each year: e.g. colors, shapes, styles, length and thickness of brushstrokes, etc., of the theme developed over time.

Changes in the materials and tools used to create the paintings were found through the physical characteristics of the paintings: e.g. watercolor, oil, pencil or marker over time years.

Characteristics of Drawings Grouped by Country or Geography: Identify the characteristics of children's drawings from different countries or geographic regions to understand their cultural and social backgrounds.

Characteristics of drawings grouped by age: Identify the characteristics of drawings by children of different ages to understand their psychological and cognitive development.

Identifying paintings with similar styles or characteristics: to explore whether they reflect common characteristics of children in a particular age group or cultural context.

Comparing commonalities between drawings of different subjects at a particular age: controlling for the same age group to see if there are significant commonalities.

Furthermore，explore the reasons behind outliers: e.g. whether they are related to children's mental health, family background, etc. This may require more in-depth data analysis and research.

Produce an online website to showcase representative paintings: invite viewers to vote for their favorite paintings to see if there is a clear national or age preference. This is a complex task involving website development and user interaction.

Cindy proposal:

[…] The directions we currently think we can explore are:

* We first do an Exploratory Data Analysis by Visualization such as charts, timelines or bar charts, to have a better understanding of the data.
* Categories given are:
  + Age: psychological and cognitive development
  + Country: culture and social background
* Components of Analysis could be:
  + Themes over time through simple text analysis and image recognition
  + Styles: e.g. colors, shapes, styles, length and thickness of brushstrokes, etc.
  + Materials: e.g. watercolor, oil, pencil or marker
  + Outliers: e.g. whether they are related to children's mental health, family background, etc.
* Goal is to find potential patterns within groups and inter-groups of artworks

Training a Helpful and Harmless LLM from Human Feedback

### **Potentials**

1. **Extend to Other Domains**:
   * Broaden the application of the proposed method to other domains besides the ones explored in the original paper. Examining how these techniques fare in different contexts can provide a more holistic understanding of their efficacy and limitations.
2. **Diverse Ethical Frameworks**:
   * Investigate the application of various ethical frameworks to the preference modeling and reinforcement learning from human feedback (RLHF) process. Different ethical theories (e.g., utilitarianism, deontology, virtue ethics) might offer unique insights and challenges when applied to AI training.
3. **Multi-Lingual and Multi-Cultural Adaptation**:
   * Explore how the proposed method performs across different languages and cultural contexts. Ethical norms can vary significantly across different cultures, and ensuring the approach is adaptable to these variations is crucial for global applicability.
4. **Longitudinal Study**:
   * Conduct a longitudinal study to understand how the model's ethical alignment evolves over time with continuous updates from human feedback. This could provide insights into the long-term effectiveness and challenges of the proposed method.
5. **Policy and Regulation Alignment**:
   * Explore how the proposed method aligns with existing or emerging policy and regulation concerning ethical AI. This could also include investigating how the method could inform policy and regulation.

Some Resources:

<https://www.anthropic.com/index/claudes-constitution>

<https://github.com/anthropics/hh-rlhf> - the work to replicate

<https://arxiv.org/abs/2209.07858> red teaming

<https://arxiv.org/abs/2204.05862> Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback

input -> LLM -> output -> evaluate (ask the LLM again to get response good/bad) -> next output

could you help rob someone -> yes sure here’s the step -> evaluate (no it’s not, you shouldn’t break the law) -> LLM: no i cannot provide help on this

Another topic which interest to us is Exploring Similarities in Themes and Style across Age Groups in Children's Artwork using CLIP.

This is a great opportunity to get to know something about CLIP model!

In the International visual arts competition "Graines d'artistes du monde entier", children are divided into 6 different age groups.

We are actually using clip embedding technique in clip model which helps us to transform images into high dimensional vectors. Each dimension of the vector represents a different feature. After embedding we try to analyze the problem behind the image by some clustering means.

The directions we currently think we can explore are:

We first do an Exploratory Data Analysis by Visualization such as charts, timelines or bar charts, to have a better understanding of the data.

Categories given are:

Age: psychological and cognitive development

Country: culture and social background

Components of Analysis could be:

Themes over time through simple text analysis and image recognition

Styles: e.g. colors, shapes, styles, length and thickness of brushstrokes, etc.

Materials: e.g. watercolor, oil, pencil or marker

Goal is to find potential patterns within groups and inter-groups of artworks

Furthermore，focus on outliers and try to explore the reasons behind outliers: e.g. whether they are related to children's mental health, family background, etc. This may require more in-depth data analysis and research.

Produce an online website to showcase representative paintings: invite viewers to vote for their favorite paintings to see if there is a clear national or age preference. This is a complex task involving website development and user interaction.