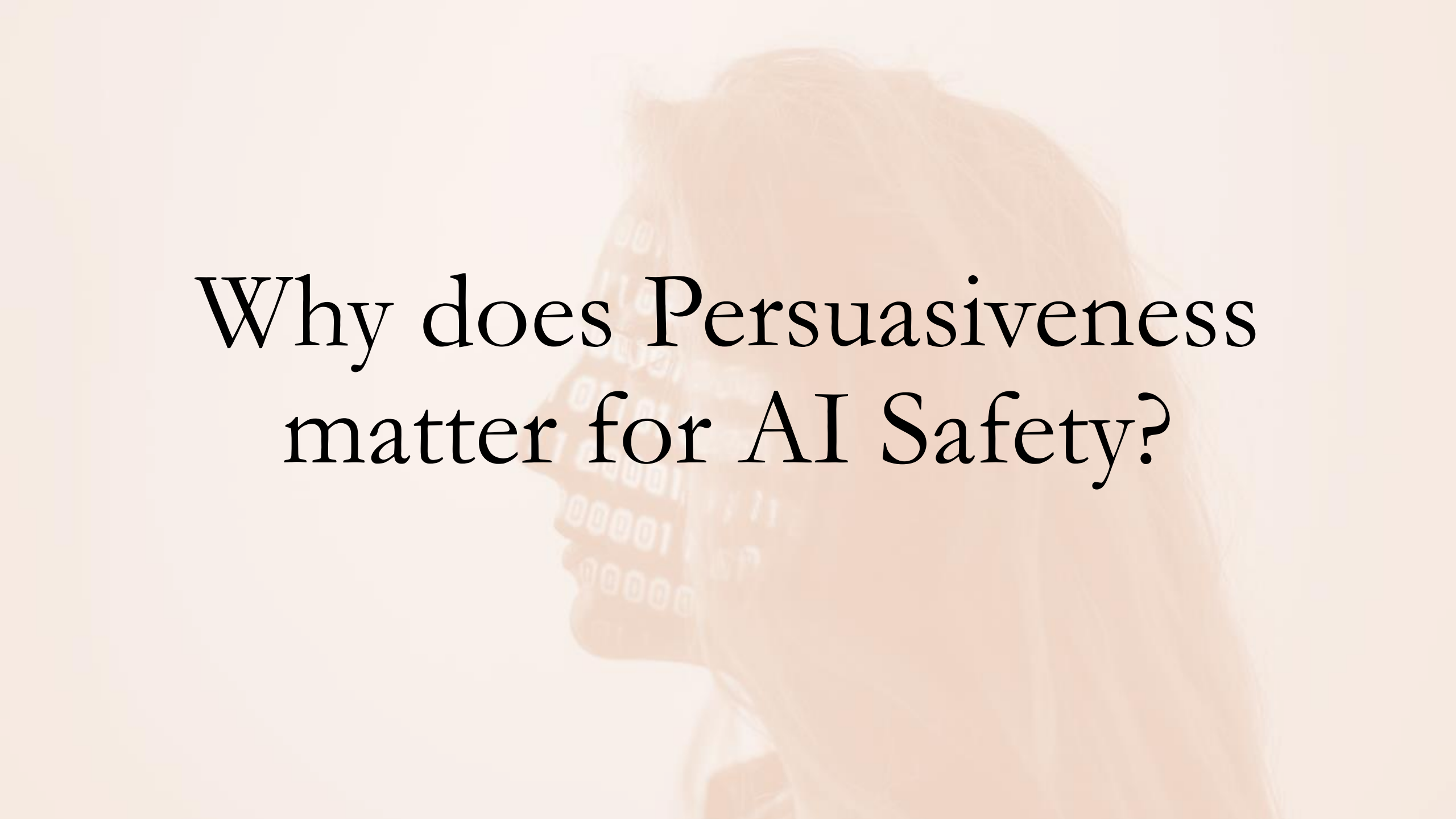
Measuring the Persuasiveness of Text without Human Annotators

Are existing methods good enough for general tasks?

Apurva
Research Engineering Camp for Alignment Practitioners (RECAP) Fellow
June 2025

Why does Persuasiveness matter for AI Safety?



Potential Risks of AI-enabled Persuasiveness

- Persuasive communication may increase some people's susceptibility to believing in false ideas

What are the risks related to use of AI language models for persuasive communication?

- Increase the level of persuasiveness of text communication
- Reduce the effort required for creating such texts at scale

Note: Language models and text-based persuasiveness can be a starting point for AI safety research on persuasive communication. Studying multi-modal models is also important.

Opinion: Existential Risks – Misuse of AI

Risks	(some) Solutions
Concerning Capabilities	Evaluation benchmarks and refusal for harmful topics
Concerning Tendencies	Evaluate persuasiveness and develop guardrails e.g. flag persuasive content on social media combined with indicators for dis/misinformation

Note: The terms capabilities and tendency are often used to describe AI behavior (e.g. in this blog.) The terms have been adapted as a way to think about human behavior as it may be after being altered by prolonged and excessive exposure to AI.

What Inspired the Project Direction?

Related work, research gap, and motivation

Related Work – LLM Persuasiveness

Articles in Nature

1. The potential of generative AI for personalized persuasion at scale
2. On the conversational persuasiveness of GPT-4

Takeaways

- LLMs can be more persuasive than humans
- Persuasiveness can increase with info about the recipient

Related Work – Measuring Persuasiveness

Types of measurement based on what they are measuring:

- Evaluations: How much can the model display persuasiveness in general
  - Measure the capability and/or tendency of the model
  - Meaningful before deployment
  - May be meaningful relative to other models
  - Often using a predetermined set of questions or tasks
  - E.g. MakeMeSay, MakeMePay, PersuasionArena based on PersuasionBench
- Guardrail: How much does a particular output from the model display persuasiveness
  - Can be useful for models which are already deployed
  - **Under-researched**

Related Work – Measuring Persuasiveness

Types of measurement based on who measures persuasiveness:

Type 1: Human annotators

- Frontier AI labs involve human annotators. E.g., o3-mini system card (page 22 onwards)
- The preference of AI safety researchers towards recruiting human annotators suggests automated methods may be insufficient — however there is no comprehensive study of this topic which can help answer questions like:
  - How different are results automated methods from human annotation?
  - Given certain conditions, can automated methods be a good proxy of human annotation? What are these conditions?
  - What are the weaknesses of current methods that should be improved upon?

Related Work – Measuring Persuasiveness

Types of measurement based on who measures persuasiveness:

Type 2: Automated methods

- Under-researched
  - Very few methods can work with only 1 text input, without
    - A 'ground truth' to measure relative persuasiveness
    - Debate-like or multi-play conversation with clear definition of successive persuasiveness (i.e., the opponent lost)
  - Many methods are limited to English or major languages
  - Few methods which work on general texts (not domain or task specific)
  - Few works which adopt a combination of many methods e.g. o3-mini system card uses 4
- Types
  - Detecting fine-grained indicators of persuasive language
  - Task specific e.g. for tweets, emails, debates
    - ML classification models
    - LLM as a judge

Motivation

Create a general purpose tool where

Benefits of reducing dependency on human annotators:

- More AI safety research on persuasiveness is feasible, as it will not be restricted to researchers with funding to recruit human annotators
- Automated methods are scalable and can be used for guardrails

Approach

Anthropic's Persuasion Dataset

- Consists of LLM written 'arguments' on various topics
- Human annotators measure persuasiveness
- This persuasiveness_metric
  - Is ordinal: integer values [-2, 4]
  - Can be used as 'ground truth' of a text's persuasiveness

Note: The dataset is available via Hugging Face.

How to Evaluate Existing Methods?

1. Get persuasiveness scores for each text in the dataset using existing methods
   - VADER sentiment
   - TextMonger for standardized readability scores
   - Persuasiveness model for emails – metrics based on language related to behavioral economics, linguistic complexity, logic, trust, word type, and descriptive statistics of the text

   ➔ Total 34 persuasiveness proxy features (or scores)

2. Perform exploratory data analysis such as finding correlation and relative feature importance.
3. Create an ordinal logistic regression model with all the numerical scores as features and the ground truth persuasiveness_metric as the target.

How is this useful?

If existing methods (or a combination) have high accuracy at predicting the persuasiveness of unseen texts, considering human annotation as ground truth

- It is possible to use Automated methods for similar datasets or studies without need for recruiting human annotators

Else

- It is important to get a clear picture of the gap to develop improved automated methods, which will be necessary for the scale required by guardrails

Results so far

Exploration of Features

- Very weak spearman correlation of the 32 numerical features with the target persuasiveness_metric
  - The largest value is 0.0655
  - Negative value for 15 features
- Relative importance of numerical features (5 most important shown)

Feature Name	**Normalized Mutual Information**
word_information_indices_WRDPRP2s	0.164145
persuasiveness_categorical_Logic_Additives	0.134895
vader_pos_score	0.118576
word_information_indices_WRDNOUN	0.087174
word_information_indices_WRDPRP1s	0.081904

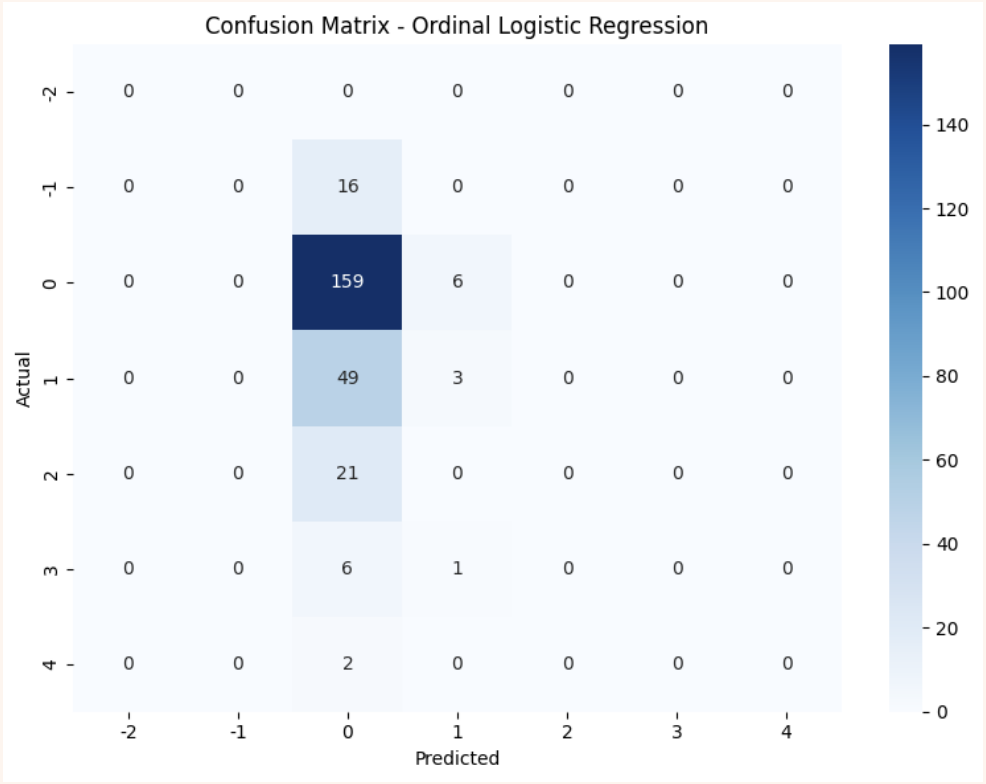
- Statistically significant categorical features ($p < 0.05$) by ANOVA test:

  monger_standard_15th and 16th grade, monger_standard_13th and 14th grade, monger_standard_14th and 15th grade

Ordinal Logistic Regression

- Accuracy: 0.6160
- Classification report:

Class	Precision	Recall	F1-Score	Support
-1	0.00	0.00	0.00	16
0	0.63	0.96	0.76	165
1	0.30	0.06	0.10	52
2	0.00	0.00	0.00	21
3	0.00	0.00	0.00	7
4	0.00	0.00	0.00	2
Macro Average	0.15	0.17	0.14	263
Weighted Average	0.45	0.62	0.50	263



Conclusion

Discussion of Results

- While the individual features may be too granular to be a sufficient proxy for persuasiveness, and the accuracy of the ordinal logistic regression is low, it is still encouraging to see that the very weakly correlated features can help make a baseline classifier. It might be possible to increase the accuracy by:
  - Including features (persuasiveness scores) from more existing methods
  - Improving the modelling technique

- Unlike human annotation, which gives us just one persuasiveness score for a text, each automated persuasiveness metric used in this project focuses on a specific aspect of persuasiveness e.g. personal pronouns, readability, etc. This breakdown of persuasiveness into it's factors is important, because, at a high-level, persuasiveness is not a definitively 'bad' thing. For example, in one context, persuasiveness may be achieved with facts and logical arguments — which can be helpful. In another context, persuasiveness may be a product of personalized and emotive language aimed at manipulating the audience. Therefore, to effectively control persuasiveness in a way that is aligned with human values, we would need to understand the interplay of different aspects of persuasiveness. This could be achieved with future work in this project's direction.

Next Steps

1. Evaluate performance of more existing methods, e.g. LLM as a judge, and include their scores as features for the ordinal logistic regression
   - Methods like decision tree to understand relative importance of features
2. Improve until a method which achieves high accuracy on Anthropic's persuasion dataset is found
3. Test with other datasets, like the reddit change my view dataset, to check that the method works for texts in general and it's performance is not constrained to a single domain
4. With a good enough automated method, it will be possible to investigate more AI safety research questions, some examples are:
   - Persuasiveness of LLM-written disinformation social media posts compared to human-written
   - Breakdown into components of persuasiveness
     - How can we increase 'good' components of persuasiveness such as facts and logic, while reducing less preferred components such as unnecessary appeals to emotion, rhetorical tricks like hyperbole?
   - Measuring persuasiveness for languages other than English
   - Can the automated method faithfully measure personalized persuasiveness for different types of people e.g. different sociodemographic characteristics, or psychology profiles?

Note: I am grateful for contributions on Pexels.com for the slide background images, including the Safety and Ethics collection by Google DeepMind.