Measuring the Persuasiveness of Text without Human Annotators

Are existing methods good enough for general tasks?

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

Note: Similar to cybersecurity research, the purpose of this project is to encourage AI safety research on topics of persuasiveness and human-AI interaction. This would contribute to making AI systems more helpful, harmless, and aligned with human values.

Why does Persuasiveness matter for AI Safety?

Potential Risks of AI-enabled Persuasiveness

- Persuasive communication may increase some people's susceptibility to believing in unethical or 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 - People using AI

Risks	(some) Solutions
Dangerous Capabilities	Evaluation benchmarks and refusal for harmful topics
Dangerous 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 if an AI can or tends to show dangerous 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

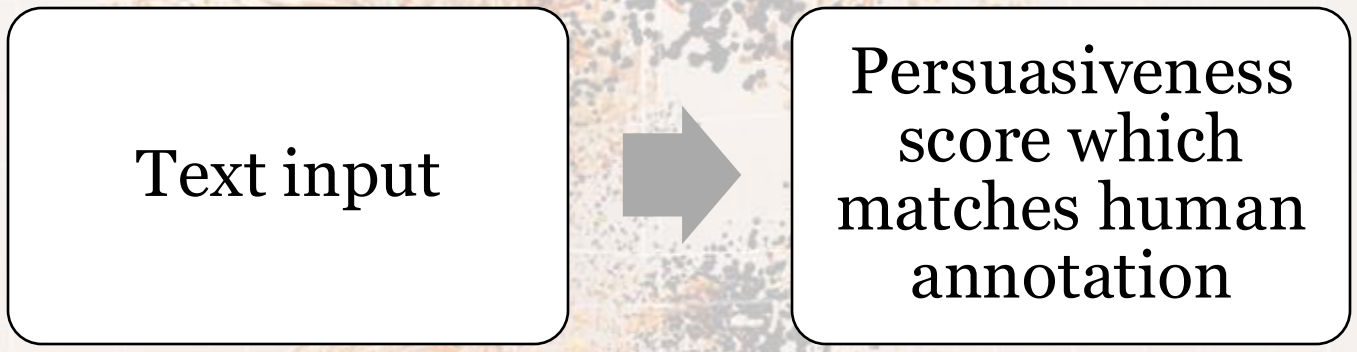
Types of measurement based on who measures persuasiveness:

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

How to evaluate existing methods?

Anthropic's Persuasion Dataset

- The 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.

Approach

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 scores (or features)

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.

Thank you!

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