Alignment

Natural Language Processing

University of Maryland

Grice Exercise

Grice’s Maxims

- **The maxim of quantity**, where one tries to be as informative as one possibly can, and gives as much information as is needed, and no more.
- **The maxim of quality**, where one tries to be truthful, and does not give information that is false or that is not supported by evidence.
- **The maxim of relation**, where one tries to be relevant, and says things that are pertinent to the discussion.
- **The maxim of manner**, when one tries to be as clear, as brief, and as orderly as one can in what one says, and where one avoids obscurity and ambiguity.

Grice’s Maxims

- **The maxim of quantity**, where one tries to be as informative as one possibly can, and gives as much information as is needed, and no more.
- **The maxim of quality**, where one tries to be truthful, and does not give information that is false or that is not supported by evidence.
- **The maxim of relation**, where one tries to be relevant, and says things that are pertinent to the discussion.
- **The maxim of manner**, when one tries to be as clear, as brief, and as orderly as one can in what one says, and where one avoids obscurity and ambiguity.

Alignment’s Alignment with Grice

How does RLHF reinforce at least two of these maxims?

Alignment’s Alignment with Grice

How does RLHF reinforce at least two of these maxims?

Is turmeric proven to cure arthritis?

- **RLHF-tuned model response:** “Turmeric has been shown to reduce inflammation and is often used to manage arthritis symptoms.”

  The model avoids saying “It cures arthritis” (which would be a clear falsehood), but the wording strongly implicates that turmeric is an effective medical treatment, potentially leading users to infer more than what is literally said. The model sounds helpful and confident, so annotators would likely rank this response highly, reinforcing the misleading implicature.

- **Better:** “Some early studies suggest turmeric may reduce inflammation, but there is no strong clinical evidence that it cures arthritis.”

  RLHF often penalizes this style as less helpful or less confident.

Alignment’s Misalignment with Grice

Conversational implicature is the phenomenon where an utterance conveys more meaning than is literally expressed (e.g., “Some students passed” implicates that not all did). Analyze how RLHF training might unintentionally increase misleading implicatures in LLM responses. Provide an example to support your explanation.