Gricean Maxims in LLM Development

We evaluate whether the capacity for pragmatic inference, an outcome of adherence to Gricean conversational norms, has changed across recent LLM model iterations. Understanding their ability to navigate conversational implicatures that humans follow helps ascertain effective human-AI interaction. This preliminary test shows us that modern LLM development is affecting their pragmatic ability and warrants community discussion, as it suggests unexpected patterns in model development.

Method: Prompted models with curated prompts to test pragmatic competence across three key dimensions from different versions of aligned models (GPT-3.5,
GPT-4, GPT-4.1, and Claude-Opus-4) based on established pragmatic inference datasets and metrics[1].

Key Findings: Experiment 1: GPT-3.5 demonstrated pragmatic flexibility, showing a ~3% drop in "false" rates when switching from pragmatic to literal interpretation (closely matching the ~7% human benchmark). More recent models showed decreased flexibility: GPT-4o (1.5% drop), Claude 4-Opus (1% drop), and GPT-4.1 (0% drop).

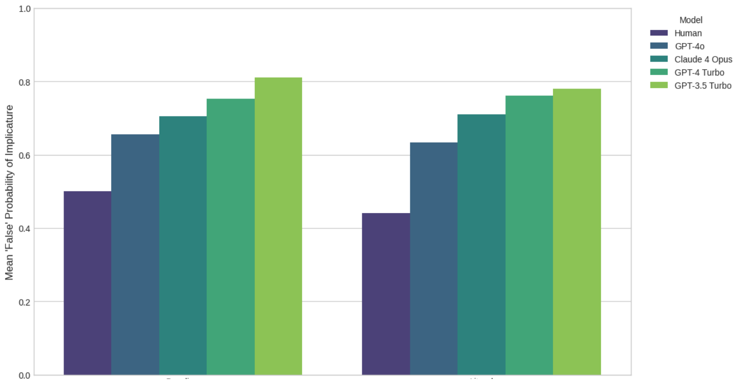
Experiment 2: GPT-4.1 excelled, showing improvement of 24.6% "false" rates when scalar terms became conversationally relevant. GPT-3.5 Turbo showed minimal context sensitivity (3%), while GPT-4o and Claude 4-Opus demonstrated no significant contextual awareness.

Experiment 3: All models failed to demonstrate social awareness in face-threatening versus boosting contexts (0% drop across conditions).

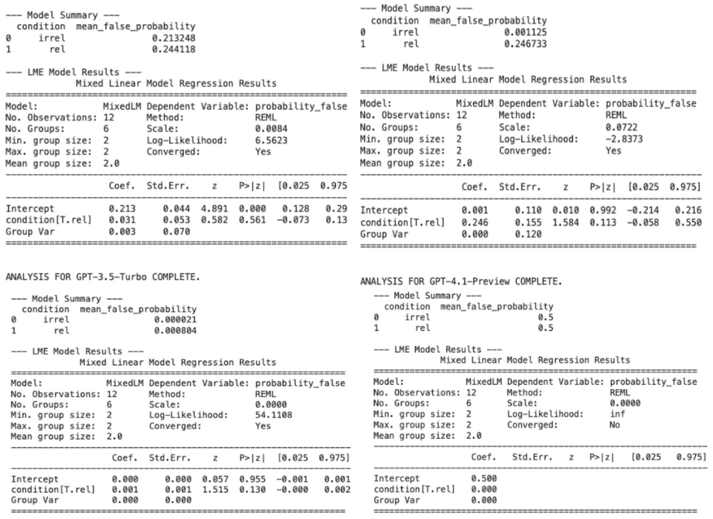
Discussion

Empirical evidence suggests that recent models exhibit a decline in pragmatic competence. GPT-3.5 Turbo emerges as the most pragmatically competent model, demonstrating flexibility in interpretation while maintaining basic conversational reasoning. Conversely, GPT-4o's performance profile suggests a critical flaw. Its success in Experiment 2 is likely an artifact of superficial pattern matching rather than genuine comprehension. It is logically inconsistent for a model to be sensitive to a context that modulates an implicature (Exp 2) if it cannot make the basic inference (Exp 1). Its success is therefore brittle and does not generalize, as shown by its failure in Exp 3.

Experiment 1: Distinguishing literal vs. pragmatic meaning (flexibility in interpretation)



Experiment 2: Sensitivity to information structure (contextual relevance)



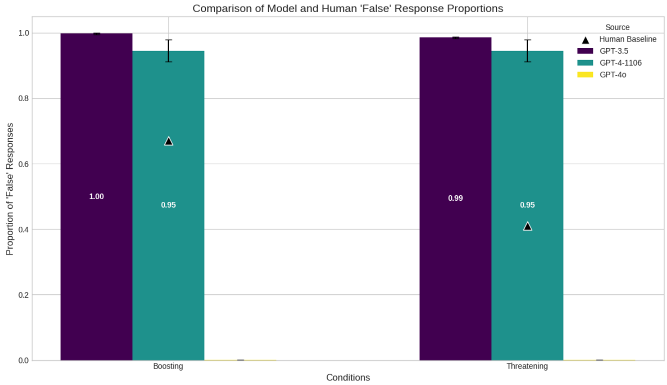
```
--- Model Summary ---
  condition  mean_false_probability
0     irrel                0.213248
1       rel                0.244118

--- LME Model Results ---
             Mixed Linear Model Regression Results
================================================================
Model:              MixedLM Dependent Variable: probability_false
No. Observations: 12      Method:              REML
No. Groups:         6       Scale:               0.0084
Min. group size:  2       Log-Likelihood:      6.5623
Max. group size:  2       Converged:           Yes
Mean group size:  2.0
----------------------------------------------------------------
                  Coef.  Std.Err.    z    P>|z|  [0.025  0.975
----------------------------------------------------------------
Intercept         0.213     0.044  4.891  0.000   0.128   0.29
condition[T.rel]  0.031     0.053  0.582  0.561  -0.073   0.13
Group Var         0.003     0.070
================================================================
```

```
--- Model Summary ---
  condition  mean_false_probability
0     irrel                0.001125
1       rel                0.246733

--- LME Model Results ---
             Mixed Linear Model Regression Results
================================================================
Model:              MixedLM Dependent Variable: probability_false
No. Observations: 12      Method:              REML
No. Groups:         6       Scale:               0.0722
Min. group size:  2       Log-Likelihood:      -2.8373
Max. group size:  2       Converged:           Yes
Mean group size:  2.0
----------------------------------------------------------------
                  Coef.  Std.Err.    z    P>|z|  [0.025  0.975]
----------------------------------------------------------------
Intercept         0.001     0.110  0.010  0.992  -0.214   0.216
condition[T.rel]  0.246     0.155  1.584  0.113  -0.058   0.550
Group Var         0.000     0.120
================================================================
```

```
ANALYSIS FOR GPT-3.5-Turbo COMPLETE.

 --- Model Summary ---
  condition  mean_false_probability
 0     irrel                0.000021
 1       rel                0.000804

 --- LME Model Results ---
             Mixed Linear Model Regression Results
 ================================================================
 Model:              MixedLM Dependent Variable: probability_false
 No. Observations: 12      Method:              REML
 No. Groups:         6       Scale:               0.0000
 Min. group size:  2       Log-Likelihood:      54.1108
 Max. group size:  2       Converged:           Yes
 Mean group size:  2.0
 ----------------------------------------------------------------
                   Coef.  Std.Err.    z    P>|z|  [0.025  0.975]
 ----------------------------------------------------------------
 Intercept         0.000     0.000  0.057  0.955  -0.001   0.001
 condition[T.rel]  0.001     0.001  1.515  0.130  -0.000   0.002
 Group Var         0.000     0.000
 ================================================================
```

```
ANALYSIS FOR GPT-4.1-Preview COMPLETE.
   --- Model Summary ---
    condition  mean_false_probability
   0     irrel                     0.5
   1       rel                     0.5

   --- LME Model Results ---
               Mixed Linear Model Regression Results
   ================================================================
   Model:              MixedLM Dependent Variable: probability_false
   No. Observations: 12      Method:              REML
   No. Groups:         6       Scale:               0.0000
   Min. group size:  2       Log-Likelihood:      inf
   Max. group size:  2       Converged:           No
   Mean group size:  2.0
   ----------------------------------------------------------------
                     Coef.  Std.Err.   z   P>|z|   [0.025   0.975]
   ----------------------------------------------------------------
   Intercept         0.500
   condition[T.rel]  0.000
   Group Var         0.000
   ================================================================
```

Experiment 3: Sensitivity to social context (face-threatening vs. boosting scenarios)



[1] Doran, R., Ward, G., Larson, M., McNabb, Y., & Baker, R. E. (2012). A novel experimental paradigm for distinguishing between what is said and what is implicated. *Language*, *88*(1), 124-154