Project Proposal: Reverse-Engineering Pragmatic Inference in LMs

A Cross-Scale Bayesian Analysis of Knowledge and Implicature

Jingyu Han, Illia Shakun, Yankı Öztürk, Lukas Viestädt
Course: Modeling Agents (WS2025/26)
Instructor: Polina Tsvilodub

February 26, 2026

1 Research Objective

This project investigates the computational mechanisms of pragmatic inference in Large Language Models (LLMs) by replicating and extending the experimental paradigm of **Goodman & Stuhlmüller (2013)**. We aim to evaluate whether LLMs adjust their interpretation of scalar quantifiers (e.g., "some") based on the speaker's epistemic state k (knowledge access), and whether this behavior is governed by the formal integration of Bayesian priors and likelihoods as defined in the Rational Speech Act (RSA) framework.

2 Formal Problem Definition

We define the communicative interaction as a Bayesian inference problem:

- **State Space** S: The world state $s \in \{0, 1, \ldots, N\}$, where $N = 3$ items.
- **Observation** k: The speaker's knowledge access $k \in \{1, \ldots, N\}$.
- **Utterance** U: The quantifier used, $u \in \{\text{'none', 'some', 'all'}\}$.
- **Target**: The listener's pragmatic posterior $P_{L_1}(s|u, k)$.

3 Experimental Design

3.1 Task 1: Controlled Behavioral Evaluation

We will execute a structured behavioral test to measure the "pragmatic shift" in LLMs.

- **Stimuli**: Text-based scenarios varying k (observation) and u (utterance).
- **Model Gradient**: We select SOTA models across three parameter tiers to evaluate scaling effects:
  1. **Small (7B-9B)**: `Llama-3.1-8B`, `Gemma-2-9B`.
  2. **Medium (27B-72B)**: `Gemma-2-27B`, `Qwen-2.5-72B`.
  3. **Frontier**: `DeepSeek-V3`, `Llama-3.1-405B`, and `GPT-4o`.
- **Inference Strategy**: Open-weights models will be run via **local inference** (vLLM/HuggingFace) to ensure full access to log-probabilities. Official APIs will be used for proprietary models.

3.2 Task 2: RSA Component Probing

To reverse-engineer the "mechanism," we decompose the LM's output into distinct Bayesian sub-tasks:

1. **Prior Elicitation ($P(s)$)**: We will probe the model's unconditioned base-rate expectations for s using neutral context prompts.

2. **Speaker Likelihood (P_S)**: We assess the speaker model using a rational choice soft-max:

$$P_S(u|s,k) = \frac{\exp(\alpha \cdot \text{Utility}(u;s,k))}{\sum_{u'} \exp(\alpha \cdot \text{Utility}(u';s,k))} \tag{1}$$

   We will extract **next-token log-probabilities** for the Quantifier set to fit the optimality parameter α via Maximum Likelihood Estimation (MLE).

4 Hypothesized Scaling Outcomes

We anticipate three stylized patterns regarding model scale:

- **O1 (Emergent Rationality)**: Increased scale correlates with higher behavioral similarity to humans and stronger mathematical consistency between $P_S, P(s)$, and P_{L_1}.
- **O2 (Heuristic Scaling)**: Larger models exhibit human-like behavior (P_{L_1}) but fail the internal Bayesian consistency test, indicating reliance on high-dimensional surface patterns.
- **O3 (Inference Bottleneck)**: Models accurately represent priors and speaker logic but fail to integrate them normatively, regardless of parameter count.

5 Technical Implementation

We will implement a **unified evaluation pipeline** to standardize prompt templates across local and API-based models.

- **Metrics**: KL-divergence ($\mathbb{D}_{KL}$) between human behavioral curves (from G&S 2013) and model posteriors.
- **Reliability**: We will use temperature-controlled sampling and evaluate sensitivity via prompt paraphrasing to ensure the results are robust across linguistic variations.