# Project Proposal: Reverse-Engineering Scalar Implicatures 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 28, 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 LMs adjust their interpretation of scalar quantifiers (e.g., "some") and the number words (e.g., "one", "two") 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 $s \in \{0, 1, \dots, N\}$, where $N = 3$ items.
- **Observation** $k$: The speaker's knowledge access $k \in \{1, \dots, N\}$.
- **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 LMs.

- **Stimuli**: Text-based scenarios promting with varying $k$ (observation), $u$ (utterance) and system prompt strategies (None, CoT, structured output).
- **Models**: We select SOTA models across four parameter tiers to evaluate scaling effects: **Very small (4B and below)**, **Small (7B-9B)**, **Medium (27B-72B)**, and **Frontier**. We evaluate both opensource models (Qwen, LLaMa, Phi, Gemma families) and proprietary models (GPT, Gemini providers) in both reasoning and non-reasoning settings when available.

- **Inference Strategy**: Open-weight models bellow 9B as well as quantized versions of the medium tier will be run via local inference (vLLM/HuggingFace). Official APIs will be used for bigger or proprietary models.

- **Prompting Strategies**: We will compare the influence of the popular prompting strategies on the pragmatic shift:

  - **Baseline**: Direct question without additional context with and without 100$ bets distribution experiment setup.
  - **Chain-of-Thought (CoT)**: Prompting the model to reason step-by-step before answering.
  - **Structured output promting**: We will also experiment with structured output prompts that explicitly ask the model to output the 100$ bet distribution over the possible states, which may help to distil the internal representations of the model from the natural language flow bias.

### 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 $\alpha$ via Maximum Likelihood Estimation (MLE).

## 4 Testing Hypotheses

We anticipate two hypothesis regarding model scale and prompting setup:

- **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 (Prompt Independence)**: Larger models exhibit human-like behavior ($P_{L_1}$) independent of prompting strategy, while smaller models show more variability across strategies.

## 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.