Mutual Theory of Mind for AI–AI Communication

IMPROVING REASONING IN SMALL LANGUAGE MODEL DEBATE

Scope of This Work

Focus exclusively on AI–AI communication

No human–AI interaction or user modeling

Agents are small language models collaborating via debate

Goal: improve collective reasoning through mutual mental modeling

Motivation

Small language models struggle with complex multi-step reasoning

Multi-agent debate helps but is often shallow

Most AI–AI debate lacks awareness of how other agents reason

Human collaboration succeeds due to mutual theory of mind

Problem: Mindless AI–AI Debate

Agents exchange answers or critiques without reasoning awareness

No explicit modeling of other agents' beliefs or assumptions

Leads to false consensus and fluent-but-wrong solutions

Debate alone is insufficient for robust reasoning

Research Questions

Does mutual ToM improve AI–AI reasoning beyond debate alone?

Which AI–AI communication signals matter most?

Does MToM generalize across reasoning domains?

Can MToM recover from seeded agent misconceptions?

Mutual Theory of Mind (MToM)

Theory of Mind: reasoning about others' beliefs and reasoning

Mutual ToM: reciprocal mental modeling between agents

Agents reason about **how** others think, not just **what** they say

Why Mutual ToM for Small Language Models

SLMs have limited individual reasoning depth

Different models fail in different ways

Mutual ToM enables strategy transfer across agents

Transforms ensembles into collective reasoning systems

MMAD Framework (AI–AI Communication)

Shared input broadcast to all agents

Independent reasoning with answer, trace, and confidence

Iterative debate with mutual mental modeling

Adaptive termination via decider agent

Consensus-based aggregation

MToM-Aware Debate Process

Analyze peers' reasoning traces and confidence

Infer strengths, weaknesses, and biases of other agents

Revise own reasoning based on inferred models

Repeat until convergence stabilizes

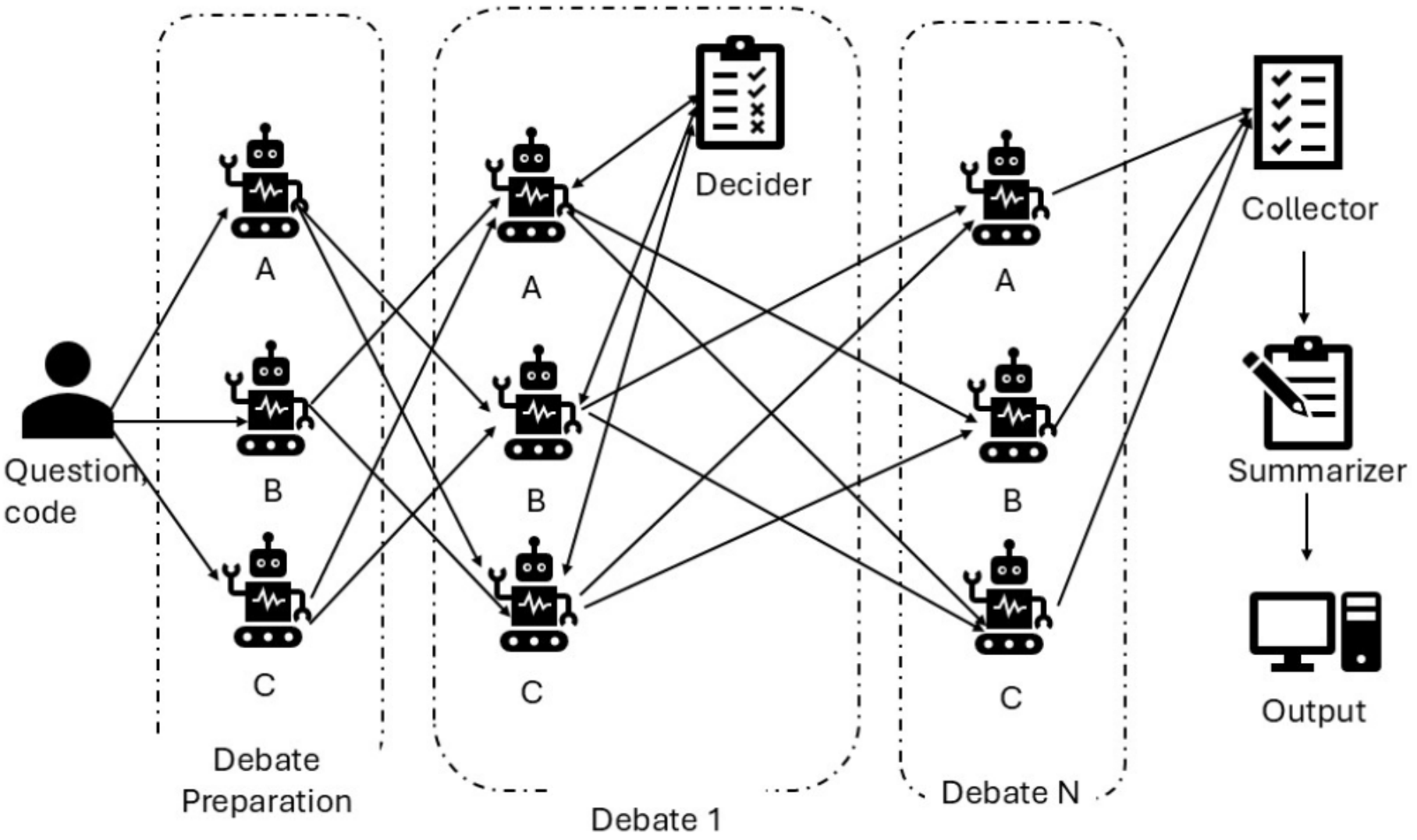
Comparison to Standard AI–AI Debate

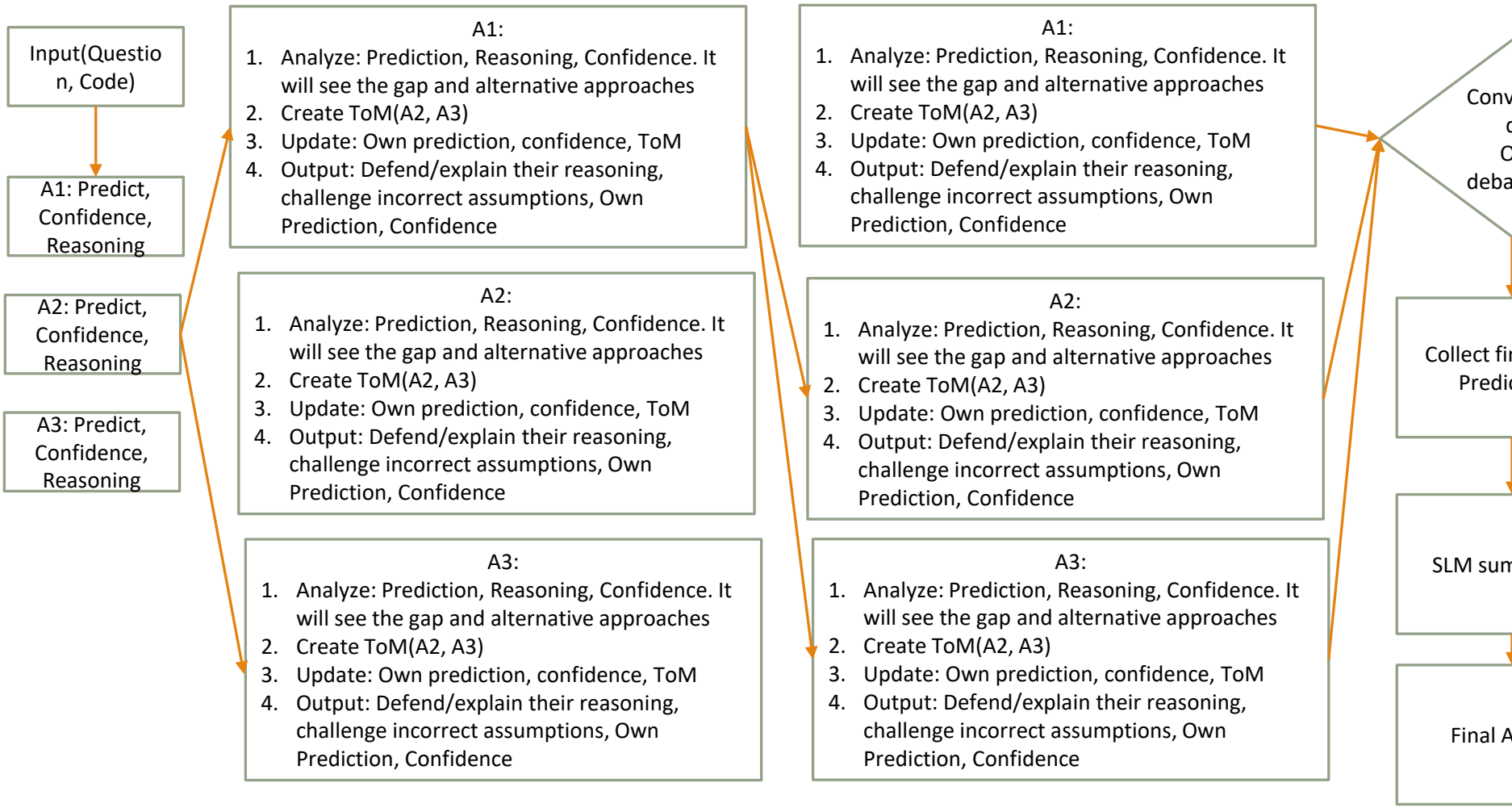
Standard debate: message passing without mental models

MMAD: explicit reasoning-about-reasoning

Standard debate: majority vote or dominance

MMAD: controlled consensus via decider


Question,
code
A
B
C
Debate
Preparation
Decider
A
B
C
Debate 1
A
B
C
Debate N
Collector
Summarizer
Output


Input(Question, Code)
A1: Predict, Confidence, Reasoning
A2: Predict, Confidence, Reasoning
A3: Predict, Confidence, Reasoning
A1:
1. Analyze: Prediction, Reasoning, Confidence. It will see the gap and alternative approaches
2. Create ToM(A2, A3)
3. Update: Own prediction, confidence, ToM
4. Output: Defend/explain their reasoning, challenge incorrect assumptions, Own Prediction, Confidence
A2:
1. Analyze: Prediction, Reasoning, Confidence. It will see the gap and alternative approaches
2. Create ToM(A2, A3)
3. Update: Own prediction, confidence, ToM
4. Output: Defend/explain their reasoning, challenge incorrect assumptions, Own Prediction, Confidence
A3:
1. Analyze: Prediction, Reasoning, Confidence. It will see the gap and alternative approaches
2. Create ToM(A2, A3)
3. Update: Own prediction, confidence, ToM
4. Output: Defend/explain their reasoning, challenge incorrect assumptions, Own Prediction, Confidence
A1:
1. Analyze: Prediction, Reasoning, Confidence. It will see the gap and alternative approaches
2. Create ToM(A2, A3)
3. Update: Own prediction, confidence, ToM
4. Output: Defend/explain their reasoning, challenge incorrect assumptions, Own Prediction, Confidence
A2:
1. Analyze: Prediction, Reasoning, Confidence. It will see the gap and alternative approaches
2. Create ToM(A2, A3)
3. Update: Own prediction, confidence, ToM
4. Output: Defend/explain their reasoning, challenge incorrect assumptions, Own Prediction, Confidence
A3:
1. Analyze: Prediction, Reasoning, Confidence. It will see the gap and alternative approaches
2. Create ToM(A2, A3)
3. Update: Own prediction, confidence, ToM
4. Output: Defend/explain their reasoning, challenge incorrect assumptions, Own Prediction, Confidence
Conv
deba
Collect fi
Predic
SLM sum
Final A

Experimental Setup

Reasoning Benchmarks

Coding Reasoning Tasks

CodeQA, CS1QA, CruxEval, CodeSense

HumanEval, MBPP

Mathematical Reasoning Tasks: GSM8K, MathQA

Commonsense Reasoning Tasks

Agent Models (Small Language Models)

Heterogeneous agent population

Qwen 2.5: balanced reasoning performance

Phi-4-Mini-Reasoning: multi-step inference focused

Gemma: strong linguistic and commonsense generalization

Heterogeneity enables diverse reasoning behaviors

Information Shared Between Agents

Each agent independently produces:

Answer / prediction

Explicit reasoning trace

Confidence score

These signals enable Mutual Theory of Mind

Baselines for Comparison

Single SLM (no collaboration)

Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate(Liang et al., 2024)

Enhancing Reasoning Abilities of Small LLMs with Cognitive Alignment(Cai et al., 2025)

Teach Small Models to Reason by Curriculum Distillation(Jiang et al., 2025)

Evaluation Framework

Chain-of-Agent evaluation

Accuracy, Completeness, Relevance, Clarity (A/C/R/C)

Separate evaluator agents per criterion

Rubric converted into final correctness score

Ablation Studies

No-ToM debate (answers only)

No reasoning sharing

No confidence sharing

No decider (fixed rounds)

Pseudo-homogeneous agents

Isolate the effect of Mutual ToM

Conclusion

AI–AI debate without mental modeling is limited

Mutual Theory of Mind enables structured collective reasoning

MMAD improves robustness and accuracy for small models

Explicit AI–AI reasoning is critical for multi-agent systems