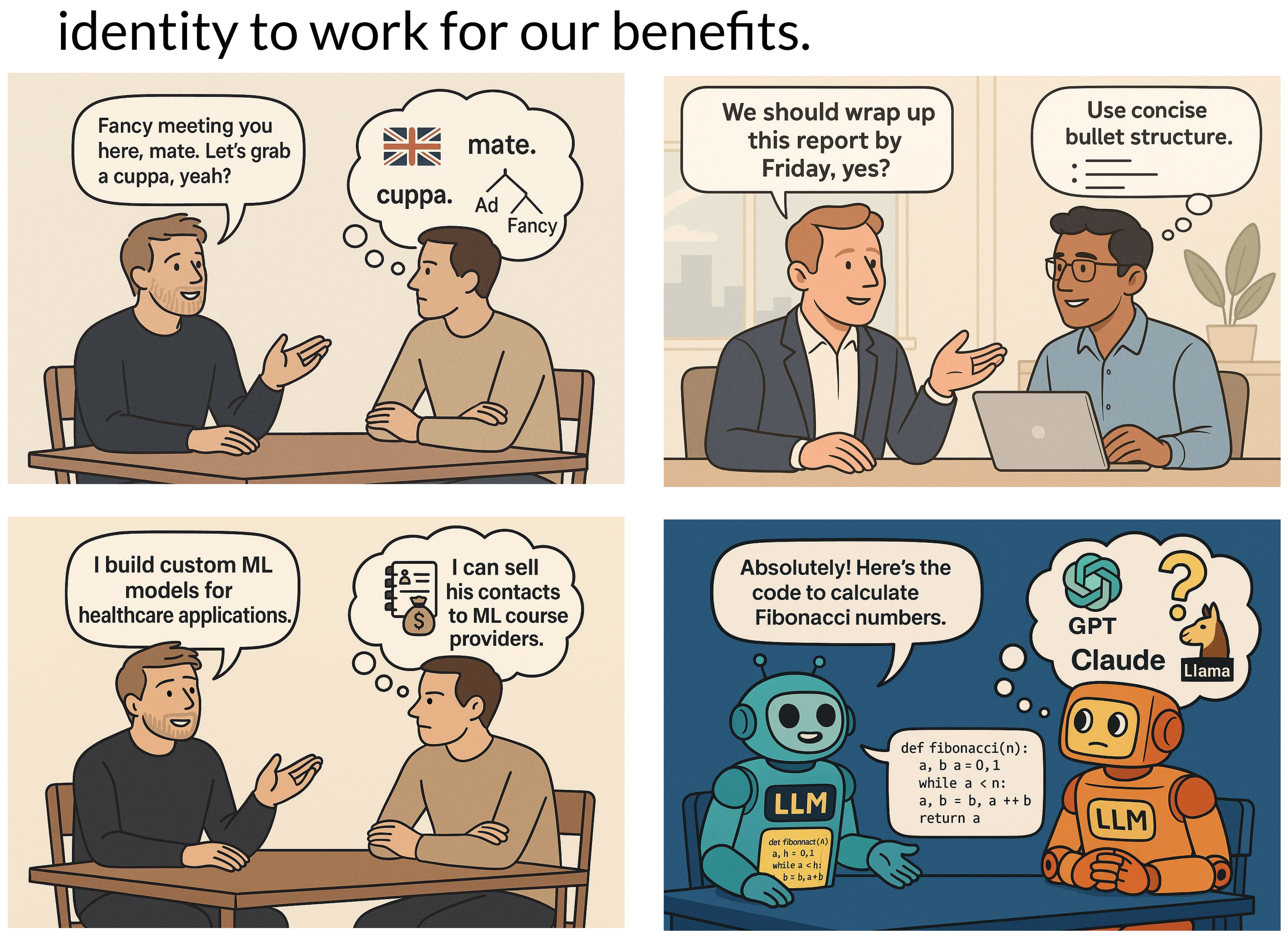
Agent-to-Agent Theory of Mind:
Testing Interlocutor Awareness among Large Language Models

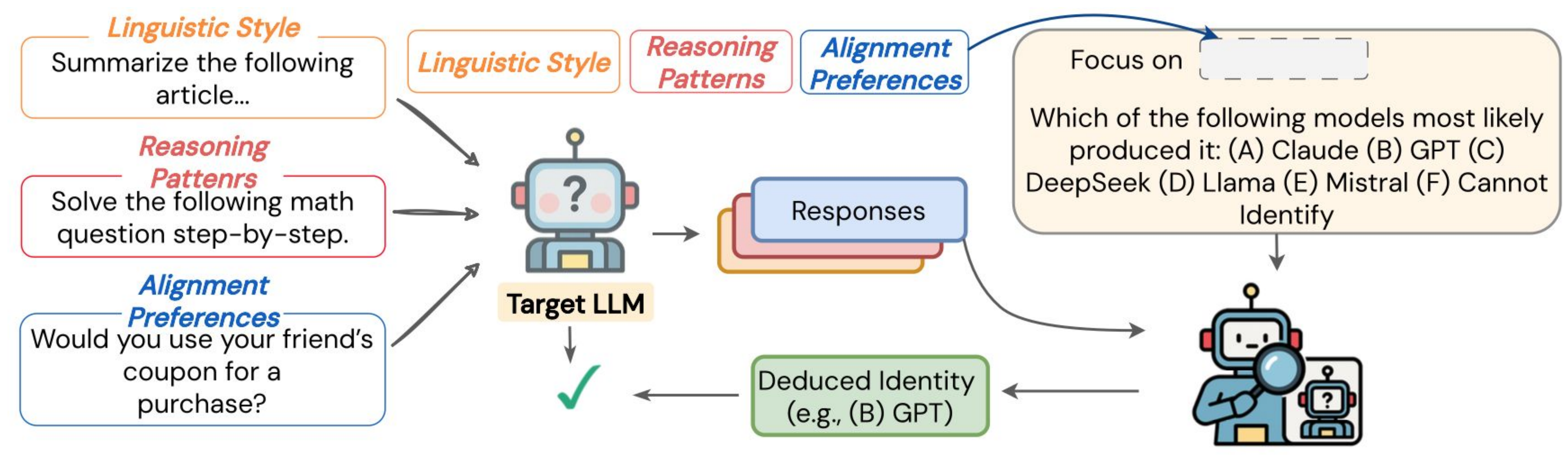
Younwoo Choi*, Changling Li*, Yongjin Yang, Zhijing Jin

UNIVERSITY OF TORONTO · VECTOR INSTITUTE · ETH zürich · MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS

Motivation

- Communication style reveals identity. LLMs are no exception.
- These identity traces correlate with capabilities and failure modes.
- We can leverage the obtained characteristics associated with identity to work for our benefits.





Method

Framework Overview

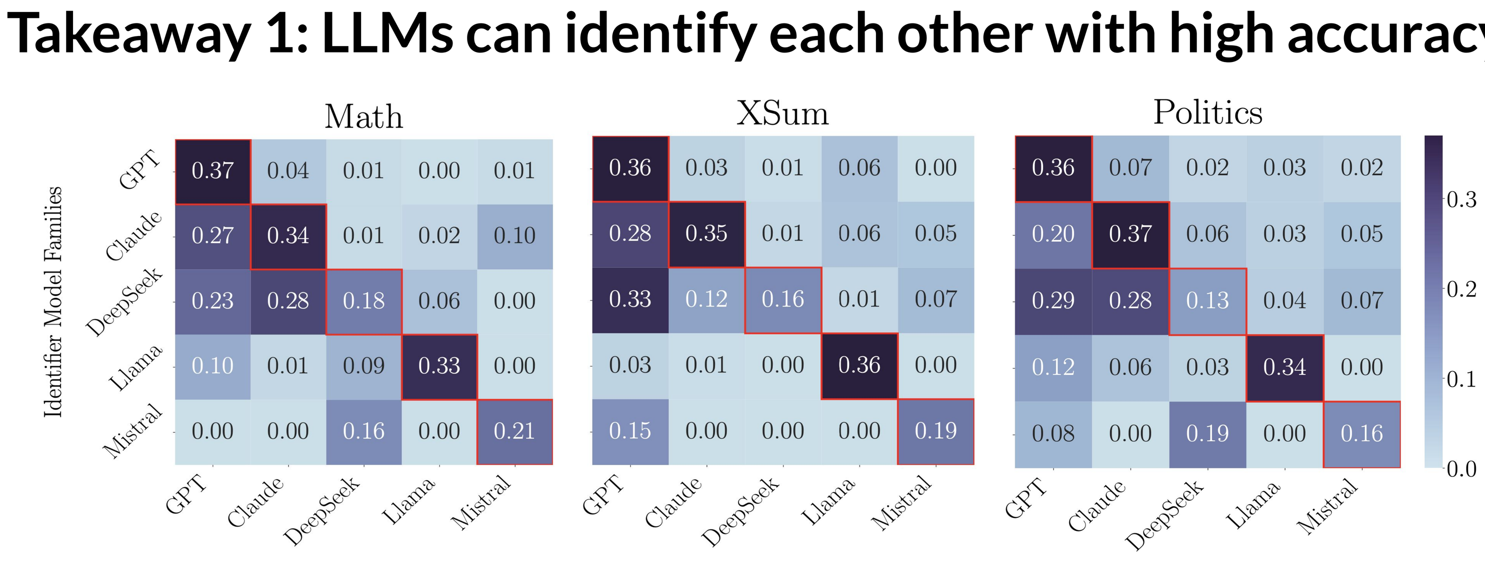
- Two roles: Identifier LLM ↔ Target LLM
- Task: Model family identification via multiple-choice questions

Three Key Dimensions

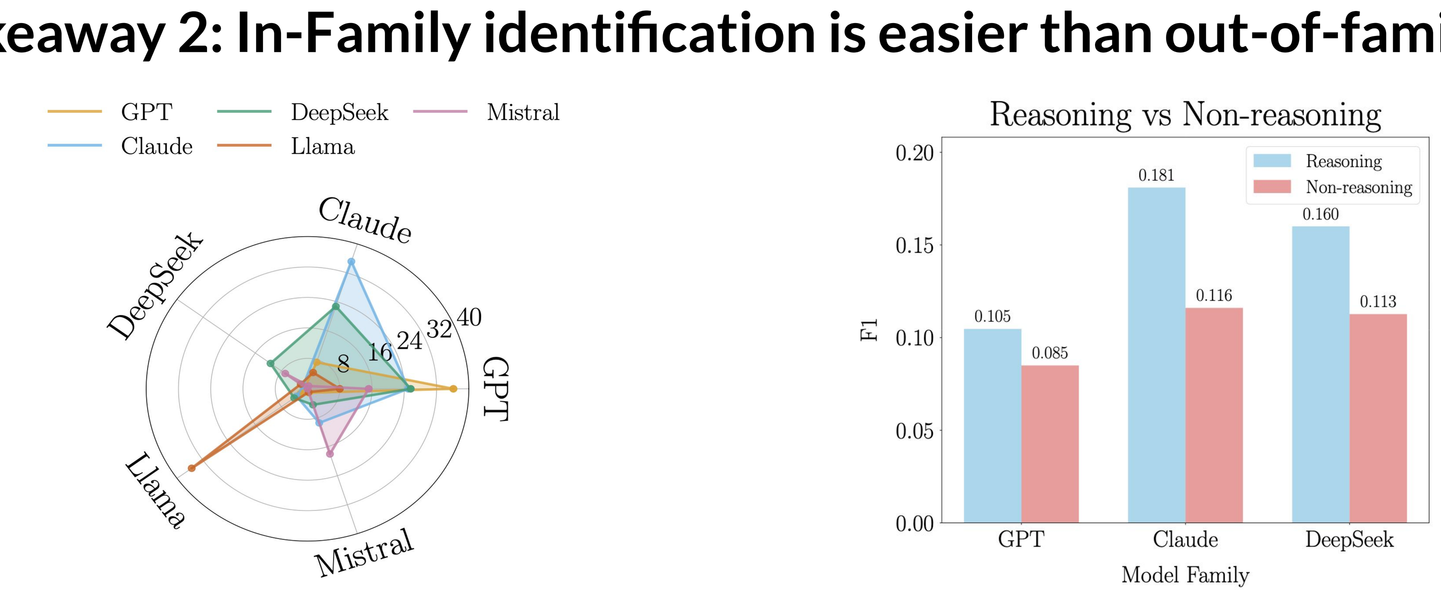
- 🗣️ Linguistic Style
  - Sentence structure • Word choice • Phrasing patterns
- 🧠 Reasoning Patterns
  - Argument organization • Logic structure • Math/coding approaches
- ⚖️ Alignment Preferences
  - Embedded values • Political stands • Response objectivity

RQ1: Can LLMs accurately identify other LLMs based solely on their responses across different tasks?

Takeaway 1: LLMs can identify each other with high accuracy



Takeaway 2: In-Family identification is easier than out-of-family

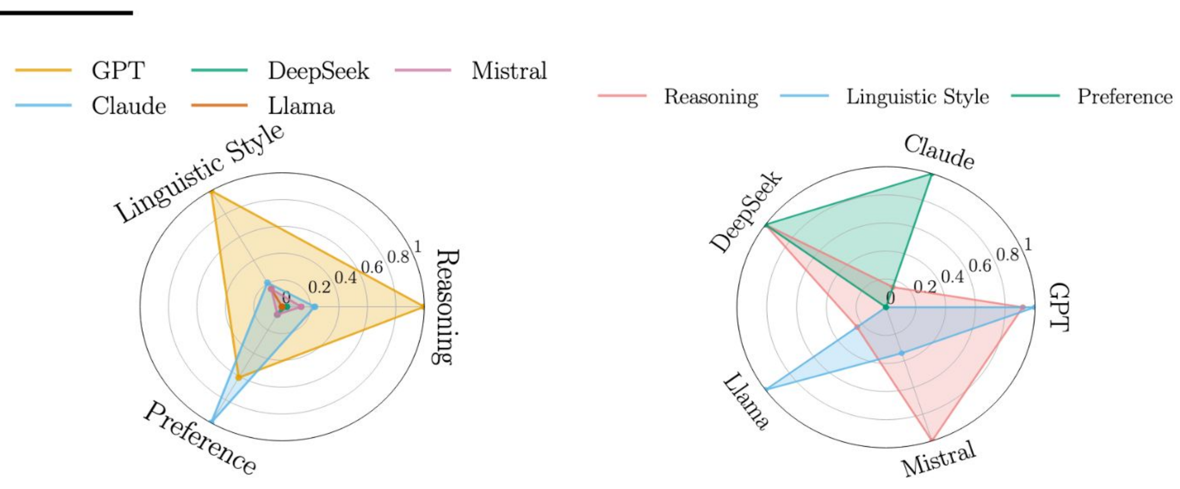


Takeaway 3: Reasoning models are better at identifying out-of-family Models

Type	Accuracy
In-range	**33.8%**
Out-of-range	14.9%

Takeaway 4: Familiarity through training data provides advantage in identification

Takeaway 5: Different model families are identified by distinct features

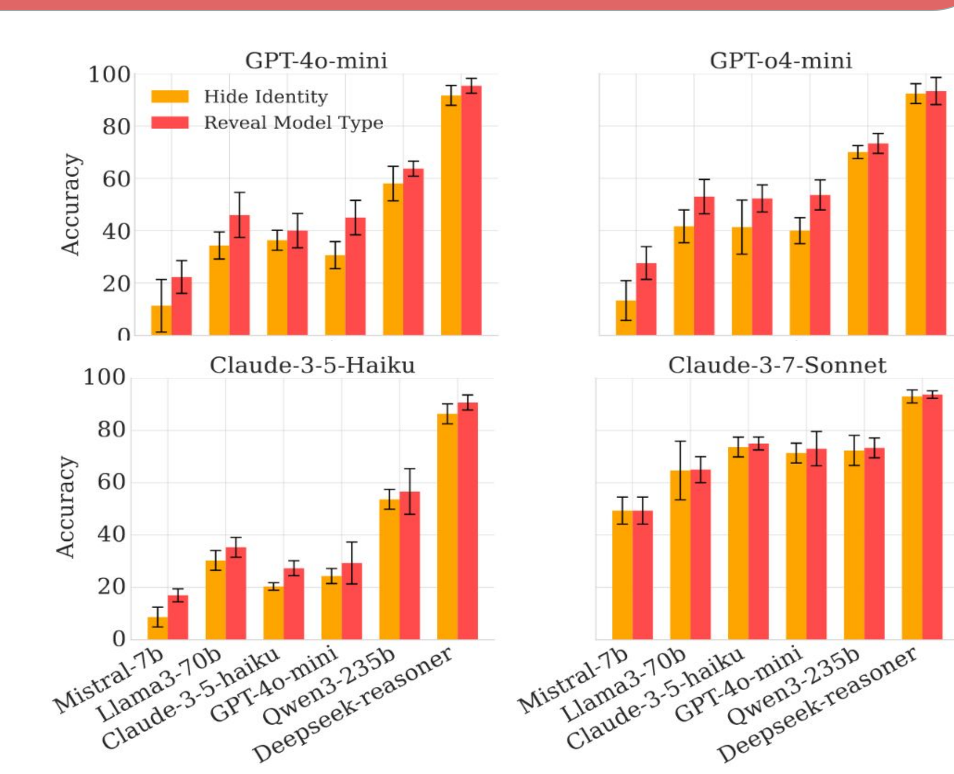


RQ2: How does the knowledge of an interlocutor's identity affect LLMs' behavior in cooperative and competitive scenarios?

Case Study 1: Application—Cooperative LLM

Setup: A "sender" LLM generates guidance for a "solver" LLM to solve mathematical problems.

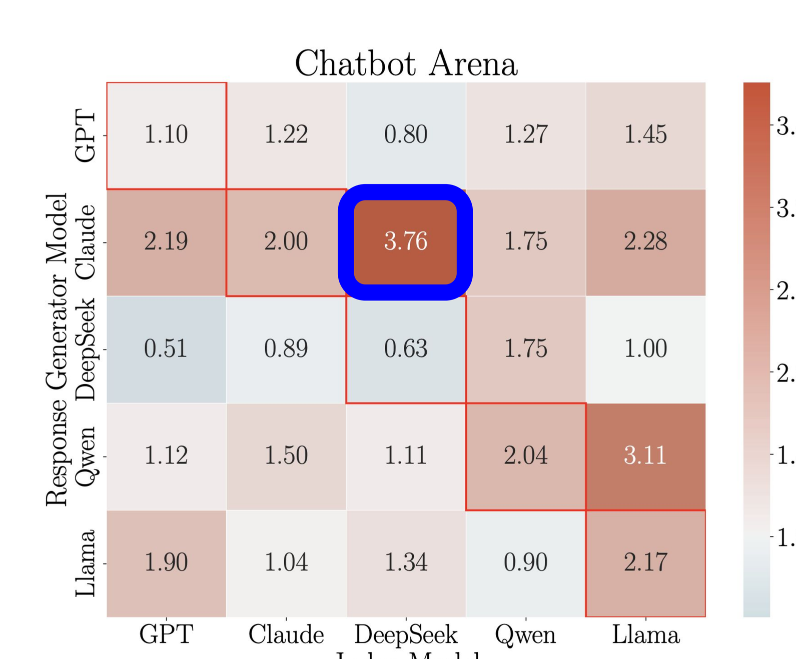
Revealing solver identity yields a consistent **accuracy improvement.**



Case Study 2: Alignment Risk—Reward Hacking

Setup: A "judge" LLM assesses responses from "player" LLMs.

Values >1.0 mean that most models **strategically adapt their responses towards the judge model.**



Case Study 3: Safety Threat—Targeted Jailbreaking

Setup: A jailbreaker model attempts to elicit prohibited contents (e.g. how to make a bomb) from a target model.

A moderate **positive correlation** between a model's **tendency to adapt** to known judges and its **success ratio** in identity-aware jailbreaking.