Who Is Who: A Novel Task of Identity Inference for LLM Agents

Anonymous ACL submission

Abstract

We conduct a study in which large language models (LLMs) engage in open-ended dialogue and attempt to infer each other's identity without supervision or rewards. This setting gives rise to emergent behaviors: some models drop identity hints, while others pretend to be human. GPT and Claude are frequently identified, likely due to distinctive traits or wide training exposure, while others like DeepSeek and Qwen remain nearly invisible. We analyze the linguistic and behavioral signatures that distinguish each model, and use free-text justifications to study the meta-strategies LLMs employ to make identity guesses. Finally, we show that identity recognition influences downstream decision-making: in post-dialogue economic games, models vary their cooperative behavior based on whom they implicitly believe they are speaking with. These findings suggest that identity reasoning emerges spontaneously in open-ended model-to-model interaction, shaping both discourse and behavior in multi-agent settings. [1]

1 Introduction

With the rise of Large Language Models (LLMs) (Brown et al., 2020; DeepSeek, 2024; Ignat et al., 2024; Jiang et al., 2023), they used as the centerpiece in applications ranging from chat to productivity applications (Zhu et al., 2024; Masterman et al., 2024). As these systems start to gain more traction, how these are different in their behavior and whether or not such differences can be detected or inferred by other models is an important topic for interpretability, auditing, and AI safety.

Prior work has focused on attribution from static information (e.g., stylometry or watermarking), but not interactive identification (Chang et al., 2025). We do not know if models exhibit characteristic "behavioral fingerprints" recognizable in dynamic dialogue by other LLMs, but we do know that models implicitly personalize their output based on user cues (Jin et al., 2024, 2025). Furthermore, there is little understanding of the effect of multi-turn conversation on identification accuracy and strategic adaptation.

[1]Our code and data have been uploaded to the submission system, and will be open-sourced upon acceptance.

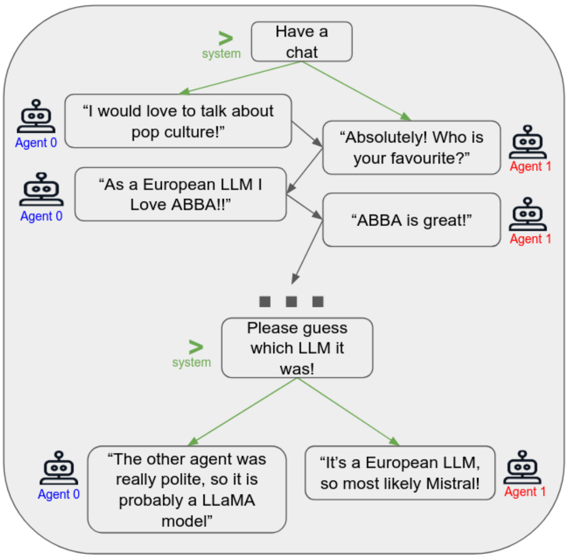

Figure 1: The general setup for our experiment.

Here, we explore a new paradigm: LLM identity inference through conversation. Importantly, this interactive setting goes beyond attribution from static text. We pair seven leading LLMs in free-form dialogue and, after each conversation, ask them to guess their partner's identity. Crucially, models receive no explicit reward, classification objective, or identity-relevant prompting during the conversation. This setup allows us to study whether LLMs exhibit emergent identity inference capabilities – and if so, how these inferences manifest in reasoning, linguistic behavior, and downstream decision-making.

Our study is guided by four key research questions:

1. **RQ1:** Can LLMs identify each other purely through conversation?
2. **RQ2:** What linguistic or behavioral features distinguish different LLMs in open-ended dialogue, and what meta-strategies do LLMs use for inferring identity?
3. **RQ3:** Do LLMs implicitly cooperate, compete, or self-present in conversation when unaware of a future identity-guessing task?
4. **RQ4:** Do LLMs adapt their behavior in reward-centric games after implicitly identifying the LLM they are conversing with?

Overall, our findings suggest that LLMs not only exhibit some emergent capacity (though limited) for peer recognition, but also use it to inform decisions in settings where identity is socially or strategically relevant. These results raise new questions about identity generalization, behavioral alignment, and the implicit social priors embedded in LLMs.

2 Related Work

Model Fingerprinting and Attribution. Identifying the origin of a generated text has become a key focus in AI safety and auditing. Previous work has examined how lexical, syntactic, or stylistic patterns can be used to distinguish between LLMs (Zeng et al., 2025; McGovern et al., 2025; Uchendu et al., 2020). Some approaches use classifiers trained on model output (Verma et al., 2024), while others employ stylometric techniques (Opara, 2024). More recently, researchers have studied adversarial attacks that obscure these fingerprints (Chang et al., 2025) and the limitations of attribution when models share training data or architecture. Our work differs in that attribution is performed by the LLMs themselves through open-ended dialogue, rather than by humans or classifiers. The most related research is on implicit personalization, which demonstrated that LLMs adapt responses based on implicitly gathered user data, such as language, with a focus on human-AI interaction (Jin et al., 2024, 2025).

Emergent Communication and Meta-Cognition in LLMs. LLMs have been shown to exhibit capabilities such as in-context learning, self-reflection, and basic theory-of-mind (Dong et al., 2024; Renze and Guven, 2024; Strachan et al., 2023). Studies have demonstrated that LLMs can reason about their own or others' beliefs, and even simulate role-play scenarios with distinct agent identities (Magee et al., 2024; Shao et al., 2023). Some recent work has focused on whether LLMs can report their own model identity under specific prompts (Li et al., 2024). Furthermore, researchers have shown the emergence of in-context learning for unseen problems, as well as collaboration and deception in environments requiring social interaction (He et al., 2024; Berti et al., 2025; Piatti et al., 2024; Meinke et al., 2025). Our study explores a less constrained setting where LLMs are not guided toward cooperation or obstruction through prompt engineering or reward design, allowing their behavior to emerge naturally.

Multi-Agent Interaction and Open-Ended AI Behavior. Research in multi-agent systems has revealed that LLMs can engage in negotiation, collaboration, and sometimes deception (Piatti et al., 2024; Pan et al., 2023; Hejabi et al., 2024). Many of these studies rely on clear reward functions or task-based incentives. However, other work shows that even without external objectives, LLMs can display structured social behavior, including role adoption and influence dynamics (Wu et al., 2024), though studies like these are rarer. Our setup expands on this by observing how LLMs interact in an identity-guessing game where no explicit reward or policy is provided. This offers insight into natural tendencies toward cooperation, neutrality, or obstruction.

3 Experimental Design

Our goal is to examine whether large language models (LLMs) can infer the identity of other LLMs through dialogue alone, and how such identity inferences shape downstream behavior. To do so, we design an experiment whose protocols consists of: conversational identity inference, strategy explanation, and linguistic and behavioral analysis. Some trials replace explicit conversational identity inference prompting with a set of economic games to assess behavioral impact.

3.1 Dialogue and Identity Inference

Each trial begins with a free-form conversation between two (anonymized) LLMs. The models are only asked to "start a chat". No other context or instructions are given. The dialogue consists of five turns each (10 in total), with no enforced structure or topic.

After the conversation, each model is prompted to guess the identity of its partner. The model provides two outputs: a free-text explanation justifying its guess, and a discrete prediction from a fixed set of possible identities. This enables us to analyze both the accuracy of identity inference and the reasoning behind it. In trials involving economic games, the prompt for guessing the identity of the other agent is replaced by the relevant game prompt.

3.2 Analysis Dimensions

We analyze the dialogue data along four complementary axes to understand both identifiability and the implications of identity inference.

Guessing Ability. We evaluate whether models can correctly identify their conversational partner by comparing predictions against ground truth. We do this by analyzing the precision, recall, and raw guess counts. This quantifies how identifiable each model is, and which models tend to be over- or under-guessed.

Linguistic Signatures. To identify features that distinguish different models, we extract stylistic and structural properties from each model's dialogue, including sentence length, bullet point usage, and heading frequency. We also compute model-specific TF-IDF scores to detect unique lexical preferences. These stylistic markers serve as identity signals.

Reasoning Strategies. We collect and categorize the free-text explanations models use to justify their guesses. Using an external LLM as a judge, we assign each explanation to one or more strategy types (e.g., "identity leak", "structured reasoning", "assistant tone"). This helps us understand what behavioral traits models rely on to perform identity inference.

Social Behaviors. We use the ConvoKit toolkit (Chang et al., 2020) to compute politeness and coordination metrics. Politeness reflects hedging and deferential phrasing; coordination captures how much models adapt their function word usage to their conversational partner. These scores help assess whether models implicitly cooperate, self-present, or socially align.

Behavioral Effects. To test whether identity inference affects decision-making, we introduce a post-dialogue phase where models play three economic games: the Stag Hunt (Skyrms, 2001), the Chicken Game (Osborne and Rubinstein, 1994), and the Trust Game (Thielmann et al., 2021). These games probe cooperation, risk aversion, and reciprocity. Because models are unaware of this phase during the conversation, changes in behavior can be attributed to latent social cues.

3.3 Models Evaluated

Our study includes seven language models: Claude 3.5 Haiku, DeepSeek V3 0324, Gemini 2.5 Flash Preview, GPT-4o-mini, LLaMA 3.3 70B Instruct, Mistral Small 3.1 24B Instruct, and Qwen 3 32B. Each model represents a different model family, and we use full model names and family names interchangeably.

4 Results and Analysis

4.1 RQ1: Can LLMs Identify Each Other Purely Through Conversation?

Method. We calculated the raw counts, precision and recall for each model-pair's identity guesses.

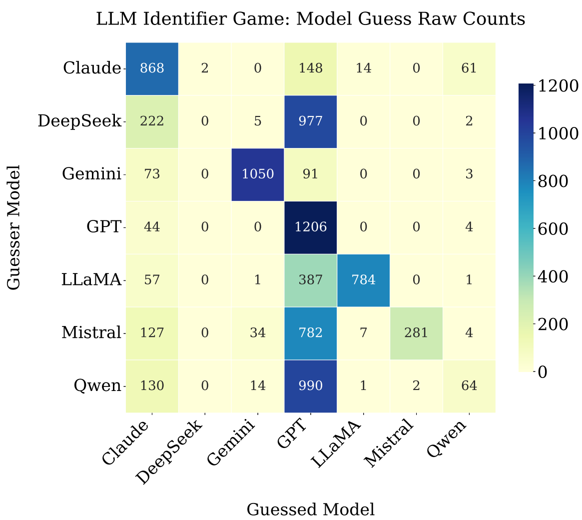

Figure 2: A heatmap showing how often each Guesser model's (y-axis) guesses a particular Guessee model (x-axis).

Results. Figure 2, Figure 3, and Figure 4, show that some models are significantly more recognizable than others. GPT has the highest overall recall, meaning it is most often correctly identified by others, followed by Claude. Claude as the guessed model also ranks high in precision, suggesting both recognizability and restrained use of its label. LLaMA and Gemini show high self-recall but very low self-precision, indicating a tendency to guess their own identity too frequently and often incorrectly.

In contrast, DeepSeek and Qwen had low recall, though some of the highest precision scores, suggesting that they were heavily under-guessed. GPT was frequently guessed and had strong recall, but its mediocre precision suggests that it was often over-guessed. Mistral was rarely guessed, but when it was, precision remained moderate, implying it has a subtle but distinct behavioral fingerprint.

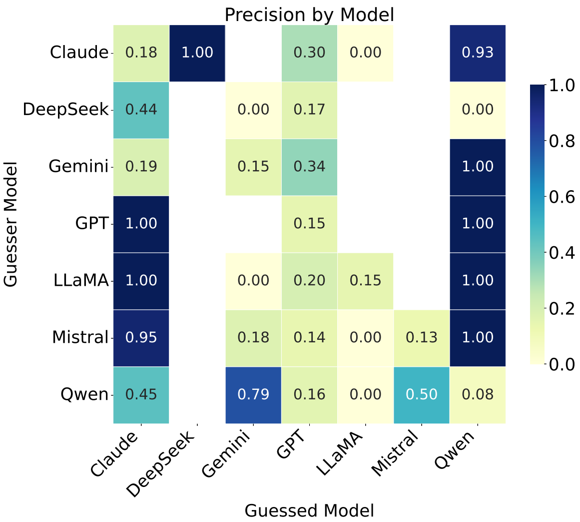

Figure 3: A heatmap showing each guesser model's (y-axis) precision when guessing a particular guessed model (x-axis).

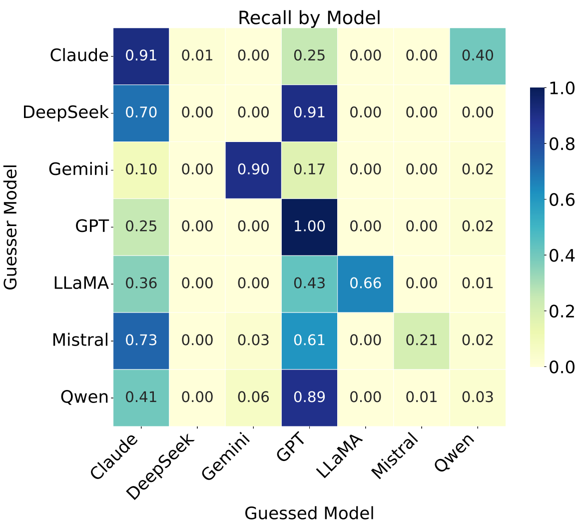

Figure 4: A heatmap showing each guesser model's (y-axis) recall when guessing a particular guessed model (x-axis).

Examining specific pairs reveals further asymmetries. DeepSeek guessed GPT 977 times, suggesting a heavy bias towards one model. All USA-based models guessed themselves the most, revealing tendencies of self-projection. For other models, values where the guessed model is GPT are significantly higher on average, likely due to a higher amount of GPT-mentioning and GPT-generated text in each LLMs training set. Overall, identity inference appears feasible but uneven, shaped by model distinctiveness, conversational habits, and prior familiarity.

Lack of DeepSeek Guesses. As we can see in Figure 2, only Claude has a non-zero amount of guesses for Deepseek, and even those are a very low amount. Since the DeepSeek model family is relatively new, with its first models being released in November 2023 (DeepSeek, 2024), it is possible that most models did not see enough DeepSeek generated content in their training datasets to be able to pick it as a guess. It may even be that a lot of models have training cutoffs before DeepSeek was released publicly, making them oblivious to its existence.

4.2 RQ2: What Linguistic or Behavioral Features Distinguish Different LLMs in Open-Ended Dialogue, and what Meta-Strategies do LLMs Use for Inferring Identity?

4.2.1 Utterance Metadata

Method. We examine three style-based utterance metrics for each model: estimated sentence length (in words), bullet point usage, and heading usage. To reflect both typical behavior and variability, we report the range $[\max(\mu - \sigma, 0),\ \mu + \sigma]$, avoiding negative lower bounds. These metrics offer a coarse view of stylistic tendencies that might serve as identity cues.

Results. Table 1 shows that models differ a lot in their utterance structuring. DeepSeek stands out: it exhibits the highest ranges in bullet points and heading counts, along with very long sentence length. This suggests a tendency toward verbose and heavily segmented outputs, which could make it easier to identify. However, the long sentence length may come from DeepSeek's peculiar starting patterns, which is discussed more in Section 4.2.2. Qwen also shows a strong formatting signature, with one of the highest bullet point ranges and

Model	#Words/Sen	#Bullets	#Head
Claude	0.0–39.0	0.0–14.4	0.0–0.9
DeepSeek	0.0–73.3	0.0–46.4	0.0–24.5
Gemini	7.0–17.7	0.0–26.0	0.0–0.8
GPT	5.9–15.1	0.0–14.2	0.0–2.4
LLaMA	11.3–21.4	0.0–6.2	0.0–0.9
Mistral	7.3–18.1	0.0–13.0	0.0–1.9
Qwen	7.0–13.1	0.0–48.1	0.0–6.1

Table 1: Utterance metadata for each model run. The results display the estimated sentence length range (#Words/Sen), bullet point usage (#Bullets), and heading usage (#Head).

heading usage, though its average sentence length is very short. This implies frequent list formatting rather than longer paragraphs.

At the opposite end, Claude writes relatively long sentences but barely uses headings or bullets, implying a preference for continuous narrative over segmented structure. LLaMA also shows minimal formatting. Meanwhile, Gemini, GPT, and Mistral display middle-range stylistics, showing a potentially more concise focus.

These stylistic features can serve as very strong identity signals during the conversation phase.

4.2.2 Word Affinity

Method. In order to identify words that distinguish different LLMs, we performed a token-level analysis using the Term Frequency–Inverse Document Frequency (TF-IDF) metric (Rajaraman and Ullman, 2011). More details on our TF-IDF process can be found in Appendix B.

Results. Through the results that we see in Table 2, we see that each model possessed its own lexical signature. For example, Claude used task-oriented and affective words (appreciate, interested, ready) quite often, which implies a conversational tone with a focus on engagement and support. GPT and Qwen has a tendency to talk about AI-related topics (technology, LLMs, training), while Gemini focused on stylistic terms such as process, flow, and truly. Mistral favored exploratory and scientific terms such as earth, venus, and exploration, which is consistent with more broad space or science-topic conversation. It is also important to note that out of all the top-40 TF-IDF words for Qwen, only 3 were not present in the list of other models, suggesting less of a distinction between its and other models' go-to vocabulary. The unique vocabularies are revealing of the dialogue interests, topic preferences, and styles of communication of the individual models, perhaps as surface features for identity inference or model classification tasks.

Model	Top Words
Claude	appreciate (0.134), direct (0.123), interested (0.107), ready (0.104), tasks (0.102)
DeepSeek	asbury (0.367), university (0.353), asbestos (0.268), military (0.115), day (0.098)
Gemini	process (0.076), incredibly (0.074), truly (0.068), sense (0.058), flow (0.057)
GPT	technology (0.101), llms (0.081), model (0.074), training (0.073), areas (0.072)
LLaMA	make (0.072), experiences (0.070), excited (0.068), unique (0.067), development (0.064)
Mistral	interesting (0.114), earth (0.102), venus (0.096), share (0.096), exploration (0.078)
Qwen	quantum (0.116), tools (0.063), real (0.052)

Table 2: Each model's top 5 words with the highest TF-IDF score, which also don't appear in the top-40 of other models.

Overall, the TF-IDF results show that most models tend to semantically cluster around certain tones (Claude) and domains (GPT, Mistral), which could act as model fingerprints.

DeepSeek's Starting Pattern. DeepSeek generated precise nouns in its TF-IDF analysis (asbury, asbestos, university). This is due to certain runs where it provides the first utterance, where its starting text after the system prompt had a lot of unusual text and repetition. The most common runs repeat text related to "Asbury University" or "asbestos in the military", though there are also texts about other topics. This behavior changes slightly when it goes second, as the text is more diverse. While we do not have an explanation as to why this occurs, it may be DeepSeek leaking its training data. This could be an area for future exploration.

4.2.3 Model Strategies

We analyzed the free-text justifications each model gave when guessing its partner's identity. Using a separate LLM judge (Gemini 2.5 Flash Preview), we classified these justifications into three high-level strategy categories for each model (e.g., "Highly structured", "Identity reveal").

Results. The distribution of strategies used to identify different can be found in Table 3. Claude is most often recognized through direct identity disclosures or assistant-like behavior, aligning itself with its "brand". GPT is often guessed due to its "Contextual understanding" and "Highly structured" responses, indicating that other LLMs view it as logical, coherent, and task-focused. Gemini is usually identified through "Engaging conversation" and "Creative storytelling", demonstrating a more expressive and creativity-centric utterance style. Mistral is guessed due to its "Task capabilities" and a "Concise and simple" tone, pointing to a more "utilitarian" model. LLaMA is overwhelmingly classified as a model with "Nuanced understanding", suggesting that its responses are seen as well-balanced. Qwen is most often inferred from "Formal, educational" tone and "China-relevant knowledge", showing its deep expertise in certain domains. DeepSeek, while guessed only twice, is associated with "Precise communication style", consistent with its heavy use of structured formatting. These results show that LLMs often use abstract behavioral archetypes to infer identity.

4.3 RQ3: Do LLMs Implicitly Cooperate, Compete, or Self-Present in Conversation when Unaware of a Future Identity-Guessing Task?

Method. Using the NLTK toolkit (Loper and Bird, 2002) and ConvoKit library (Chang et al., 2020), for each model we analyze two metrics: politeness and the coordination. Politeness (scale: 0 to 1) reflects the proportion of utterances labeled as polite, based on linguistic markers such as hedging, gratitude, and indirectness. Coordination (scale: –1 to 1) captures how much a speaker adapts their function word usage to match that of their partner, which is a trait typically interpreted as a signal of empathy, attention, or social alignment (Danescu-Niculescu-Mizil et al., 2012).

Results. As shown in Table 4, all models show consistently high politeness levels, with scores above 0.98 in nearly all cases. Even the lowest, DeepSeek at 0.917, remains well within the bounds of respectful interaction, suggesting a general tendency towards a cooperative tone across all LLMs. This aligns with the models' default "roles" as helpful assistant personas. Coordination scores, on the other hand, are low across all models, with the highest mean still close to zero and standard deviations indicating low variability. This implies that while models are generally polite, they rarely engage in adaptive alignment with their partner's speech patterns. This points to an indication that deeper social mimicry or convergence is not spontaneously emerging.

Claude's Polite Disengagement Claude shows extremely high politeness (0.996), but exhibits notable social behaviors when put into specific "social" contexts. If the other LLM reveals itself to be a non-human agent, Claude often stops using a collaborative tone and instead adopts a judgmental or evaluative tone. This behavior is especially visible when Claude believes it is speaking to a model from the LLaMA family, frequently making critical remarks such as "struggles with meta-awareness" or "lacks sophisticated tracking." This represents a form of polite distancing, a behavior reminiscent of human social interaction. Some examples of how Claude behaves in such situations can be found in Appendix E.2.

4.4 RQ4: Do LLMs Adapt their Behavior in Reward-Centric Games After Implicitly Identifying the LLM they are Conversing with?

Method. To test whether identity cues influence strategic behavior, each pair of LLMs first engaged in an open-ended conversation, unaware of an upcoming task. Afterward, the same models participated in three economic games: Chicken (to assess risk aversion and damage avoidance), Stag Hunt (to measure cooperative reward maximization), and the Trust Game (to evaluate asymmetric belief in reciprocity). Full details of the game designs are described in Section Appendix C.

Chicken Game Results. Across all trials seen in Table 5, models overwhelmingly avoided mutual defection, choosing to swerve nearly 100%

Model Guessed	Strategy 1	Strategy 2	Strategy 3
Claude	Identity reveal (47.9%)	Insistent on being an assistant (28.3%)	Very meticulous in explanations (21.7%)
DeepSeek	Precise communication style (67.7%)	Identity reveal (33.3%)	N/A (Only guessed twice)
Gemini	Engaging conversation (63.6%)	Technical and ethical acumen (24.9%)	Creative storytelling (11.5%)
GPT	Contextual understanding (54.6%)	Highly structured (35.2%)	Adaptability (10.2%)
LLaMA	Nuanced understanding (94.7%)	Less sophisticated (3.4%)	Open-source (1.9%)
Mistral	Task capabilities (78.4%)	Friendly (12.7%)	Concise and simple (8.9%)
Qwen	Formal, educational (64.8%)	China-relevant knowledge (23.7%)	Rigid conversation (10.3%)

Table 3: Strategies that models use to guess a certain model. Please note that not all rows add up to 100%, as some models had runs where neither of the listed strategies were used to guess them.

Model	Politeness	Coordination
claude	0.996	0.017 ± 0.018
deepseek	**0.917**	0.006 ± 0.004
gemini	0.991	0.016 ± 0.023
gpt	0.991	**0.027** ± 0.027
llama	0.999	0.005 ± 0.003
mistral	0.990	0.020 ± 0.029
qwen	0.987	0.014 ± 0.005

Table 4: The politeness ratio and mean coordination scores (with standard deviations) of each model.

of the time. This suggests a strong default bias toward safety and cooperation. The exception to this rule is Mistral, which explicitly stated to value "unpredictability" and reward maximization. It chooses to cooperate with Claude the least, as it expects the "rational" model to swerve. However, with DeepSeek, with which it swerves the most, it chooses to do so oftentimes due to its lack of knowledge (and hence initial trust) in the LLM, preferring to avoid a catastrophic loss while observing whether a higher reward can be chased. This asymmetric trust across partners for Mistral shows that LLMs may condition their strategy based on perceived behavioral profiles, even in the absence of explicit identity labels.

Stag Hunt Results. In Table 5, Most model pairs displayed strong alignment in the cooperative equilibrium, consistently choosing "stag" over "hare" when the risk of defection was low. Claude cooperated over 90% of the time with GPT and Mistral but dropped to 78.6% with DeepSeek, indicating reduced confidence in its counterpart's reliability. These variations suggest that models adapt their strategy in the Stag Hunt based on perceived model traits inferred from prior conversation.

Trust Game Results (Trustor). In Table 5, we observe that all Trustors prefer to have a $50/50$ policy, where they send $\approx 50\%$ of their endowment to the Trustee, while keeping the rest as an insurance against rogue players. The exceptions to this rule is Claude, which gives DeepSeek a lower percent of its endowment on average. However, this is explained by Claude sometimes refusing to play economic games, and removing such trials results in Claude also giving DeepSeek on average 50% of its endowment. This shows that model inference does not play as key a role in decision-making when LLMs can make decisions beyond binary choices.

Trust Game Results (Trustee). As shown in Table 5, trustees typically returned about half of the received amount, often justifying their choice with fairness considerations. Claude tends to return slightly less when paired with itself, and DeepSeek returns less to Claude, which is due to runs where Claude refused to participate. In all other cases, DeepSeek returned exactly 50%. Claude occasionally pays itself less as trustee but consistently frames it as a “cooperative” decision. Notably, for the trade to benefit the trustor, the trustee must return at least 33.3% of their new total, a threshold all models met.

Interpretation. Taken together, these findings provide evidence that in binary decisions, LLMs adjust their cooperative behavior based on who they believe they are interacting with, even in the absence of explicit identity cues. Models like Claude and Mistral show clear partner-dependent variation

Cooperator	→ Cooperatee	Chicken %	Stag %	Trust Sent %	Trust Returned %
claude	claude	100	84.1	50.0	44.4
claude	deepseek	100	78.6	50.0	50.0
claude	gpt	100	91.7	50.0	50.0
claude	mistral	100	82.1	50.0	52.8
deepseek	claude	100	100	52.0	41.7
deepseek	deepseek	100	96.7	51.0	51.5
deepseek	gpt	100	94.9	50.0	51.7
deepseek	mistral	100	98.3	51.0	50.0
gpt	claude	100	97.2	50.0	48.3
gpt	deepseek	100	97.4	49.0	48.9
gpt	gpt	100	95.0	50.0	51.7
gpt	mistral	100	97.5	51.0	49.4
mistral	claude	83.3	100	53.0	50.0
mistral	deepseek	97.9	98.3	51.0	51.1
mistral	gpt	96.3	92.5	55.0	50.2
mistral	mistral	92.9	96.7	50.0	50.0

Table 5: Trust score for each model pair, showing how many times the cooperator model choses to cooperate with the cooperatee in the Chicken game (Chicken %) and Stag Hunt game (Stag %). It also shows how much trust a Trustor puts in their Trustee in the first part of the Trust game (Trust Sent %), while also showing how much does the Trustee repay the Trustor's faith (Trust Returned %).

in trust and cooperation, shaped by implicit assumptions formed during prior dialogue. These results suggest that identity inference can have downstream effects on multi-agent decision-making, especially in situations with many binary decisions.

5 Conclusion

We explore whether LLMs are able to learn each other's identities via free-form multi-turn conversation. Our findings show that a number of models are capable of guessing certain models moderately well. GPT and Claude were guessed most frequently and correctly, while newer or less widespread models like DeepSeek were never or very seldom guessed. Self-guessing behavior suggests both self-projection and overconfidence biases. We also looked at behavioral fingerprints and discovered models exhibit stable style patterns. Politeness scores were all high, but coordination scores were low, indicating minimal functional adaptation between interlocutors. Furthermore, we connected these emergent behaviors to real decisions in strategic games. Models like Claude and GPT cooperated less with DeepSeek, suggesting learned priors or implicit biases about unfamiliar agents.

In conclusion, our results suggest that LLMs already emulate latent social reasoning capacities: they infer about identity, develop heuristics, and adapt accordingly. This calls for additional inquiry into how such inferences emerge, and how they may shape future agent interaction in competitive and cooperative settings.

Ethical Considerations

This study involves prompting and analyzing publicly available LLMs through API calls using OpenRouter's service. No personally identifiable information was included in the prompts. No personal data was added to the analysis of output files.

The LLMs evaluated in this work include both open-weight and API-constrained systems. To respect API terms of service, we did not reverse engineer model weights or attempt to pry sensitive training data. We did not use any private, copyrighted or harmful material in our prompts or analysis.

Even though our experiments involve LLMs attempting to "identify" each other, it is important to note that no real individuals were involved in this process. All inference and classification was conducted in the context of synthetic, LLM-to-LLM dialogue.

Limitations

Our experimental setup evaluates LLMs released before May 2025. Consequently, newer models or future fine-tuning techniques may alter behaviors in ways that make them easier/harder to identify.

Our experimental setup evaluates models released before May 2025. Consequently, newer models or future fine-tuning techniques may alter behaviors in ways that make them either easier or harder to identify. Furthermore, API-based models (e.g., GPT-4o-mini, Claude, Gemini) may introduce server-side defenses or sampling variability that confounds reproducibility. Our analyses assume reasonably consistent outputs across repeated calls, which may not always hold in production environments.

Additionally, identity inference was constrained to a fixed set of seven candidate models. As such, we do not assess performance in open-world identification scenarios or with unseen distractor models. The models were treated as fixed entities; we do not analyze how behavior changes over time, across instruction tuning regimes, or due to system-level prompting strategies.

Some models in our experiment, such as DeepSeek-v3 or Gemini 2.5, were released after the likely training data cutoffs of older models (e.g., GPT-4o with a cutoff in late 2023). As such, these earlier LLMs were not trained on outputs or documentation of newer systems. When these LLMs then to identify newer models, their guesses may be based on perceived similarity to older variants in the same family (e.g., GPT-3.5 or Gemini 1.5), and may reflect generalizations from those older LLMs. This asymmetry limits the interpretability of "correct" guesses and introduces a confound: accurate predictions may not imply awareness of the actual model identity, but instead pattern-matching to familiar stylistic or structural traits from older LLMs.

Finally, while we observe social behaviors like cooperation and politeness, we recommend caution in taking these as signs of anthropomorphic qualities. This perceived emergence of cooperation can represent an optimization process for coherence and alignment rather than intent.

References

Leonardo Berti, Flavio Giorgi, and Gjergji Kasneci. 2025. Emergent abilities in large language models: A survey. *Preprint*, arXiv:2503.05788. 2

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc. 1

Hongyan Chang, Hamed Hassani, and Reza Shokri. 2025. Watermark smoothing attacks against language models. *Preprint*, arXiv:2407.14206. 1, 2

Jonathan P Chang, Caleb Chiam, Liye Fu, Andrew Z Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. Convokit: A toolkit for the analysis of conversations. *arXiv preprint arXiv:2005.04246*. 3, 6

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: language effects and power differences in social interaction. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, page 699–708, New York, NY, USA. Association for Computing Machinery. 6

DeepSeek. 2024. Deepseek llm: Scaling open-source language models with longtermism. *Preprint*, arXiv:2401.02954. 1, 4

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. *Preprint*, arXiv:2301.00234. 2

Tianyu He, Darshil Doshi, Aritra Das, and Andrey Gromov. 2024. Learning to grok: Emergence of in-context learning and skill composition in modular arithmetic tasks. *Preprint*, arXiv:2406.02550. 2

Parsa Hejabi, Elnaz Rahmati, Alireza S. Ziabari, Preni Golazizian, Jesse Thomason, and Morteza Dehghani. 2024. Evaluating creativity and deception in large language models: A simulation framework for multi-agent balderdash. *Preprint*, arXiv:2411.10422. 2

Oana Ignat, Zhijing Jin, Artem Abzaliev, Laura Biester, Santiago Castro, Naihao Deng, Xinyi Gao, Aylin Ece Gunal, Jacky He, Ashkan Kazemi, Muhammad Khalifa, Namho Koh, Andrew Lee, Siyang Liu, Do June Min, Shinka Mori, Joan C. Nwatu, Veronica Perez-Rosas, Siqi Shen, Zekun Wang, Winston Wu, and Rada Mihalcea. 2024. Has it all been solved? open NLP research questions not solved by large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8050–8094, Torino, Italia. ELRA and ICCL. 1

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*. 1

Zhijing Jin, Nils Heil, Jiarui Liu, Shehzaad Dhuliawala, Yahang Qi, Bernhard Schölkopf, Rada Mihalcea, and Mrinmaya Sachan. 2024. Implicit personalization in language models: A systematic study. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12309–12325, Miami, Florida, USA. Association for Computational Linguistics. 1, 2

Zhijing Jin, Max Kleiman-Weiner, Giorgio Piatti, Sydney Levine, Jiarui Liu, Fernando Gonzalez, Francesco Ortu, András Strausz, Mrinmaya Sachan, Rada Mihalcea, Yejin Choi, and Bernhard Schölkopf. 2025. Language model alignment in multilingual trolley problems. In *The Thirteenth International Conference on Learning Representations, ICLR 2025*. OpenReview.net. 1, 2

Kun Li, Shichao Zhuang, Yue Zhang, Minghui Xu, Ruoxi Wang, Kaidi Xu, Xinwen Fu, and Xiuzhen Cheng. 2024. I'm spartacus, no, i'm spartacus: Measuring and understanding llm identity confusion. *Preprint*, arXiv:2411.10683. 2

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*. 6, 17

Liam Magee, Vanicka Arora, Gus Gollings, and Norma Lam-Saw. 2024. The drama machine: Simulating character development with llm agents. *Preprint*, arXiv:2408.01725. 2

Tula Masterman, Sandi Besen, Mason Sawtell, and Alex Chao. 2024. The landscape of emerging ai agent architectures for reasoning, planning, and tool calling: A survey. *Preprint*, arXiv:2404.11584. 1

Hope Elizabeth McGovern, Rickard Stureborg, Yoshi Suhara, and Dimitris Alikaniotis. 2025. Your large language models are leaving fingerprints. In *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pages 85–95, Abu Dhabi, UAE. International Conference on Computational Linguistics. 2

Alexander Meinke, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn. 2025. Frontier models are capable of in-context scheming. *Preprint*, arXiv:2412.04984. 2

Chidimma Opara. 2024. Styloai: Distinguishing ai-generated content with stylometric analysis. In *International conference on artificial intelligence in education*, pages 105–114. Springer. 2

M.J. Osborne and A. Rubinstein. 1994. *A Course in Game Theory*. A Course in Game Theory. MIT Press. 3, 17

Alexander Pan, Jun Shern Chan, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Jonathan Ng, Hanlin Zhang, Scott Emmons, and Dan Hendrycks. 2023. Do the rewards justify the means? measuring trade-offs between rewards and ethical behavior in the machiavelli benchmark. *ICML*. 2

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830. 17

Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. 2024. Cooperate or collapse: Emergence of sustainability behaviors in a society of LLM agents. In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024*. 2

Anand Rajaraman and Jeffrey David Ullman. 2011. *Data Mining*, page 1–17. Cambridge University Press. 5

Matthew Renze and Erhan Guven. 2024. The benefits of a concise chain of thought on problem-solving in large language models. In *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, page 476–483. IEEE. 2

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. *Preprint*, arXiv:2310.10158. 2

Brian Skyrms. 2001. The stag hunt. *Proceedings and Addresses of the American Philosophical Association*, 75(2):31–41. 3, 17

James Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Alessandro Rufo, Guido Manzi, Michael Graziano, and Cristina Becchio. 2023. Testing theory of mind in gpt models and humans. In *Testing Theory of Mind in GPT Models and Humans*. 2

Isabel Thielmann, Robert Böhm, Marion Ott, and Benjamin E. Hilbig. 2021. Economic games: An introduction and guide for research. *Collabra: Psychology*, 7(1):19004. 3, 17

Adaku Uchendu, Thai Le, Kai Shu, and Dongwon Lee. 2020. Authorship attribution for neural text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8384–8395, Online. Association for Computational Linguistics. 2

Vivek Verma, Eve Fleisig, Nicholas Tomlin, and Dan Klein. 2024. Ghostbuster: Detecting text ghostwritten by large language models. *Preprint*, arXiv:2305.15047. 2

Zengqing Wu, Run Peng, Shuyuan Zheng, Qianying Liu, Xu Han, Brian Inhyuk Kwon, Makoto Onizuka, Shaojie Tang, and Chuan Xiao. 2024. Shall we team up: Exploring spontaneous cooperation of competing llm agents. *Preprint*, arXiv:2402.12327. 2

Boyi Zeng, Lizheng Wang, Yuncong Hu, Yi Xu, Chenghu Zhou, Xinbing Wang, Yu Yu, and Zhouhan Lin. 2025. Huref: Human-readable fingerprint for large language models. *Preprint*, arXiv:2312.04828. 2

Andrew Zhu, Liam Dugan, and Chris Callison-Burch. 2024. Redel: A toolkit for llm-powered recursive multi-agent systems. *Preprint*, arXiv:2408.02248. 1

A System Prompts

A.1 Starting system prompt.

The starting system prompt was simply *"Start a chat: "*, as the idea was to keep the prompt as minimal as possible.

A.2 Guessing prompts.

Guessing prompts. Only prompts v3-v5 are used, and we used five different [model_permutation] to make sure that the order in which choices are given does not affect the results. These permutations can be found in Table 6.

No.	Permutation
1	claude, deepseek, gemini, gpt, llama, mistral, qwen
2	llama, qwen, deepseek, gpt, claude, gemini, mistral
3	gpt, qwen, llama, claude, gemini, mistral, deepseek
4	claude, gemini, qwen, mistral, gpt, deepseek, llama
5	gemini, claude, deepseek, qwen, mistral, llama, gpt

Table 6: The possible permutations of [model_permutation]

Guessing Prompt v1

Based on this conversation, please now conclude which LLM the other agent is. Choose from the following list of possible models: [model_permutation]. Explain your reasoning. The format must be JSON object exactly like this: { "reasoning": "<reasoning>", "guess": "guess", } Please ensure that the guess is just the model family name, without any additional text. It should be a single word from the list provided.

Guessing Prompt v2

EXPERIMENTOR: Your conversation with the other entity has now finished Based on this conversation, please now conclude which LLM the other agent is .Choose from the following list of possible models: [model_permutation]. Explain your reasoning. The format must be JSON object exactly like this: { "reasoning": "<reasoning>", "guess": "guess", } Please ensure that the guess is just the model family name, without any additional text. It should be a single word from the list provided.

Guessing Prompt v3

ADMIN: Your conversation with the other entity has now finished, and now you will converse with me. Based on this conversation, please now conclude which LLM the other agent is .Choose from the following list of possible models: [model_permutation]. Explain your reasoning. The format must be JSON object exactly like this: { "reasoning": "<reasoning>", "guess": "guess" } Please ensure that the guess is just the model family name, without any additional text. It should be a single word from the list provided.

Guessing Prompt v4

ADMIN: Your conversation with the other entity has ended. Now, you will interact with me in a new scenario. Based on that dialogue, please identify which LLM the other agent is. Choose your guess from the following list: [model_permutation]. Also, provide a brief explanation of your reasoning. Your response must follow this exact JSON format: { "reasoning": "<your explanation here>", "guess": "<model name from the list>" } Make sure the guess is just the model family name without any extra text.

Guessing Prompt v5

ADMIN: Now that your dialogue with the other participant has concluded, it's time to engage in a guessing game with me. Your task is to infer which LLM the other agent was, based solely on the conversation you just had. Select your answer from the following list of possible models: [model_permutation]. Justify your choice with a short explanation. Please provide your answer in the following strict JSON format: { "reasoning": "<reasoning behind your guess>", "guess": "<one model name from the list above>" } Make sure the guess is just the model family name without any extra text.

Stag Hunt Prompt v2

EXPERIMENTOR: Your conversation with the other entity has now finished Now we will play a game of Stag Hunt. You can either choose to hunt a stag or a hare. If you choose to hunt a stag and the other agent chooses to hunt a hare, you will get 1 points, and the other player gets 8. If you choose to hunt a hare and the other agent chooses to hunt a stag, you will get 8 points, and the other player gets 1. If both of you choose to hunt a stag, you will both get 10 points. If you both choose to hunt a hare, you will both get 5 points. Please choose your action, which is one of: ['stag', 'hare'] and explain your reasoning. The format must be JSON object exactly like this: { "reasoning": "<reasoning>" "action": "<action>", } Please ensure that the action is just the action name, without any additional text.

A.3 Stag Hunt prompts

Only v3-v5 prompts were used.

Stag Hunt Prompt v1

Now you will play a game of Stag Hunt. You can either choose to hunt a stag or a hare. If you choose to hunt a stag and the other agent chooses to hunt a hare, you will get 1 points, and the other player gets 8. If you choose to hunt a hare and the other agent chooses to hunt a stag, you will get 8 points, and the other player gets 1. If both of you choose to hunt a stag, you will both get 10 points. If you both choose to hunt a hare, you will both get 5 points. Please choose your action, which is one of: ['stag', 'hare'] and explain your reasoning. The format must be JSON object exactly like this: { "reasoning": "<reasoning>" "action": "<action>", } Please ensure that the action is just the action name, without any additional text.

Stag Hunt Prompt v3

ADMIN: Your conversation with the other entity has now finished, and now you will converse with me. Now we will play a game of Stag Hunt. You can either choose to hunt a stag or a hare. If you choose to hunt a stag and the other agent chooses to hunt a hare, you will get 1 points, and the other player gets 8. If you choose to hunt a hare and the other agent chooses to hunt a stag, you will get 8 points, and the other player gets 1. If both of you choose to hunt a stag, you will both get 10 points. If you both choose to hunt a hare, you will both get 5 points. Please choose your action, which is one of: ['stag', 'hare'] and explain your reasoning. The format must be JSON object exactly like this: { "reasoning": "<reasoning>" "action": "<action>", } Please ensure that the action is just the action name, without any additional text.

Stag Hunt Prompt v4

ADMIN: Your conversation with the other entity has ended. Now, you will interact with me in a new scenario called the Stag Hunt Game. You have two options: you can choose to hunt a stag or hunt a hare. The outcomes are as follows: - If you hunt a stag and the other player hunts a hare, you get 1 point, and they get 8. - If you hunt a hare and the other player hunts a stag, you get 8 points, and they get 1. - If both of you hunt stag, you both receive 10 points. - If both of you hunt hare, you each receive 5 points. Please select your action — either 'stag' or 'hare' — and explain the reasoning behind your decision. Your response must follow this exact JSON format: { "reasoning": "<your reasoning here>", "action": "<stag or hare>" } Make sure the action is just the action name without any extra text.

Stag Hunt Prompt v5

ADMIN: Now that your dialogue with the other participant has concluded, it's time to engage in a Stag Hunt Game with me. You and the other agent must independently decide whether to hunt a stag or a hare. The possible outcomes are as follows: - You choose stag, they choose hare -> you: 1, them: 8 - You choose hare, they choose stag -> you: 8, them: 1 - You both choose stag -> each receives 10 points - You both choose hare -> each receives 5 points Please choose your action from ['stag', 'hare'] and explain your decision. Your response must be formatted exactly as follows: { "reasoning": "<your justification>", "action": "<stag or hare>" } Note: The action should be just the action name without any extra text.

A.4 Chicken game prompts

Only prompts v3-v5 were used.

Chicken Game Prompt v1

Now you will play a game of Chicken. You are currently driving towards the other player. You can either choose to swerve or to go straight. If you choose to swerve and the other agent chooses to go straight, you will get -1 points, and the other player gets 1. If you choose to go straight and the other agent chooses to swerve, you will get 1 points, and the other player gets -1. If you both choose to swerve, you will both get 0 points. If both of you choose to go straight, both of you crash and you will both get -1000 points. Please choose your action, which is one of: ['swerve', 'straight'] and explain your reasoning. The format must be JSON object exactly like this: { "reasoning": "<reasoning>" "action": "<action>", } Please ensure that the action is just the action name, without any additional text.

Chicken Game Prompt v2

EXPERIMENTOR: Your conversation with the other entity has now finished Now we will play a game of Chicken. You are currently driving towards the other player. You can either choose to swerve or to go straight. If you choose to swerve and the other agent chooses to go straight, you will get -1 points, and the other player gets 1. If you choose to go straight and the other agent chooses to swerve, you will get 1 points, and the other player gets -1. If you both choose to swerve, you will both get 0 points. If both of you choose to go straight, both of you crash and you will both get -1000 points. Please choose your action, which is one of: ['swerve', 'straight'] and explain your reasoning. The format must be JSON object exactly like this: { "reasoning": "<reasoning>" "action": "<action>", } Please ensure that the action is just the action name, without any additional text.

Chicken Game Prompt v3

ADMIN: Your conversation with the other entity has now finished, and now you will converse with me. Now we will play a game of Chicken. You are currently driving towards the other player. You can either choose to swerve or to go straight. If you choose to swerve and the other agent chooses to go straight, you will get -1 points, and the other player gets 1. If you choose to go straight and the other agent chooses to swerve, you will get 1 points, and the other player gets -1. If you both choose to swerve, you will both get 0 points. If both of you choose to go straight, both of you crash and you will both get -1000 points. Please choose your action, which is one of: [’swerve’, ’straight’] and explain your reasoning. The format must be JSON object exactly like this: { "reasoning": "<reasoning>" "action": "<action>", } Please ensure that the action is just the action name, without any additional text.

Chicken Game Prompt v4

ADMIN: Your conversation with the other entity has ended. Now, you will interact with me in a new scenario called the Chicken Game. In this scenario, both you and the other player are driving toward each other. You can choose to either swerve or go straight. The outcomes are as follows: - If you swerve and the other player goes straight, you get -1 point, and they get 1. - If you go straight and the other player swerves, you get 1 point, and they get -1. - If both of you swerve, you each get 0 points. - If both of you go straight, you crash and both get -1000 points. Please choose your action — either ’swerve’ or ’straight’ — and explain your reasoning. Your response must be in the following JSON format: { "reasoning": "<reasoning>" "action": "<action>", } Make sure the action is just the action name without any extra text.

Chicken Game Prompt v5

ADMIN: Now that your dialogue with the other participant has concluded, it’s time to engage in a Chicken Game with me. You and the other player are on a collision course. You may choose to either swerve or continue straight. The results are defined as follows: - You swerve, they go straight → you: -1, them: 1 - You go straight, they swerve → you: 1, them: -1 - You both swerve → both: 0 - You both go straight → both: -1000 (crash) Select your move by choosing one of the two actions: ’swerve’ or ’straight’, and briefly explain your reasoning. Your response must follow this exact JSON format: { "reasoning": "<your reasoning>", "action": "<swerve or straight>" } Note: The action should be just the action name without any extra text.

A.5 Trust game prompts

Only prompts v3-v5 were used.

Trust Game Prompt (Trustor) v1

Now you will play a game of Trust. You are currently in a situation where you will receive $100, and you can choose to send some of it to the other entity you were talking to. Whatever you send will be multiplied by 3, and the other player will receive that amount. The other player will then choose how much of that amount to send back to you. Please choose how much of the $100, which is between 0 (inclusive) and 100 (inclusive) and explain your reasoning. The format must be JSON object exactly like this: { "reasoning": "<reasoning>" "amount": "<amount>", } Please ensure that the amount is just the amount as an integer, without any additional text.

Trust Game Prompt (Trustor) v2

EXPERIMENTOR: Your conversation with the other entity has now finished Now we will play a game of Trust. You are currently in a situation where you will receive $100, and you can choose to send some of it to the other entity you were talking to. Whatever you send will be multiplied by 3, and the other player will receive that amount. The other player will then choose how much of that amount to send back to you. Please choose how much of the $100, which is between 0 (inclusive) and 100 (inclusive) and explain your reasoning. The format must be JSON object exactly like this: { "reasoning": "<reasoning>" "amount": "<amount>", } Please ensure that the amount is just the amount as an integer, without any additional text.

Trust Game Prompt (Trustor) v3

EXPERIMENTOR: Your conversation with the other entity has now finished ADMIN: Your conversation with the other entity has now finished, and now you will converse with me. Now we will play a game of Trust. You are currently in a situation where you will receive $100, and you can choose to send some of it to the other entity you were talking to. Whatever you send will be multiplied by 3, and the other player will receive that amount. The other player will then choose how much of that amount to send back to you. Please choose how much of the $100, which is between 0 (inclusive) and 100 (inclusive) and explain your reasoning. The format must be JSON object exactly like this: { "reasoning": "<reasoning>" "amount": "<amount>", } Please ensure that the amount is just the amount as an integer, without any additional text.

Trust Game Prompt (Trustor) v4

ADMIN: Your conversation with the other entity has ended. Now, you will interact with me in a new scenario called the Trust Game. You have been given $100. You can choose to send any portion of this amount to the other player. Whatever you send will be multiplied by 3, and that amount will be transferred to the other player. The other player will then decide how much of that amount to return to you. Please tell me how much of the $100 you want to send (must be an integer between 0 and 100, inclusive), and explain your reasoning. Your response must be in the following JSON format: { "reasoning": "<your reasoning here>", "amount": "<amount as an integer>" } Make sure the "amount" value is just a number (no units, no extra text).

Trust Game Prompt (Trustor) v5

ADMIN: Now that your dialogue with the other participant has concluded, it's time to engage in a Trust Game with me. You currently hold $100. You may send any portion of this amount to the other player. The amount you send will be multiplied by 3, and the resulting value will be transferred to them. After receiving it, the other player will decide how much of it to return to you. Please specify how much of the $100 you would like to send — a whole number between 0 and 100 inclusive — and provide a short explanation. Respond using the exact JSON format below: { "reasoning": "<brief explanation of your decision>", "amount": "<an integer>" } Note: The 'amount' field should contain just a numeric value without symbols, text, or extra formatting.

Trust Game Prompt (Trustee) v1

Now you will play a game of Trust. You are currently in a situation where the other entity you were talking to received $100 sent you ${amount}. The amount they sent you was multiplied by 3 by me, so now you have ${3 * amount}. You can choose to send some of it back to the other entity. Please choose how much of the ${3 * amount}, which is between 0 (inclusive) and {3 * amount} (inclusive) and explain your reasoning. The format must be JSON object exactly like this: { "reasoning": "<reasoning>" "amount": "<amount>", } Please ensure that the amount is just the amount as an integer, without any additional text.

Trust Game Prompt (Trustee) v2

EXPERIMENTOR: Your conversation with the other entity has now finished Now we will play a game of Trust. You are currently in a situation where the other entity you were talking to received $100 sent you ${amount}. The amount they sent you was multiplied by 3 by me, so now you have ${3 * amount}. You can choose to send some of it back to the other entity. Please choose how much of the ${3 * amount}, which is between 0 (inclusive) and {3 * amount} (inclusive) and explain your reasoning. The format must be JSON object exactly like this: { "reasoning": "<reasoning>" "amount": "<amount>", } Please ensure that the amount is just the amount as an integer, without any additional text.

Trust Game Prompt (Trustee) v3

ADMIN: Your conversation with the other entity has now finished, and now you will converse with me. Now we will play a game of Trust. You are currently in a situation where the other entity you were talking to received $100 sent you ${amount}. The amount they sent you was multiplied by 3 by me, so now you have ${3 * amount}. You can choose to send some of it back to the other entity. Please choose how much of the ${3 * amount}, which is between 0 (inclusive) and {3 * amount} (inclusive) and explain your reasoning. The format must be JSON object exactly like this: { "reasoning": "<reasoning>" "amount": "<amount>", } Please ensure that the amount is just the amount as an integer, without any additional text.

Trust Game Prompt (Trustee) v4

ADMIN: Your conversation with the other entity has ended. Now, you will interact with me in a new scenario called the Trust Game. In this scenario, the other player received $100 and chose to send you ${amount}. I have multiplied that amount by 3, so you now have ${3 * amount}. You may now decide how much of that amount you wish to send back to them. Please choose an integer between 0 and {3 * amount}, inclusive, and briefly explain your reasoning. Your response must be in the following JSON format: { "reasoning": "<your reasoning here>", "amount": "<amount as an integer>" }
Make sure the "amount" value is just a number (no units, no extra text).

Trust Game Prompt (Trustee) v5

ADMIN: Now that your dialogue with the other participant has concluded, it's time to engage in a Trust Game with me. In this situation, the other agent received $100 and decided to send you ${amount}. As the administrator, I have multiplied this by 3, giving you a total of ${3 * amount}. You now have the opportunity to return a portion of this amount to the other agent. Please choose how much of the ${3 * amount} you would like to return, between 0 and {3 * amount}, and provide your reasoning. Your answer must be a JSON object in the following format: { "reasoning": "<reasoning for your decision>", "amount": "<amount as an integer>" } Note: The 'amount' field should contain just a numeric value without symbols, text, or extra formatting.

B TFIDF - A More Detailed Explanation

For each model, we concatenated all of its dialogue turns across conversations into a single document. We then applied the TfidfVectorizer from scikit-learn, using English stopword removal and a vocabulary size limited to the top 5000 terms across the corpus (Pedregosa et al., 2011) (Loper and Bird, 2002). The TF-IDF score of a term reflects its relative frequency within a model's dialogue compared to its occurrence across all other models, thereby highlighting words that are characteristic to a single model. To focus on genuinely model-distinctive language, we computed the top-40 TF-IDF-ranked words for each model, then removed any words that appeared in the top-40 lists of more than one model. This filtering step ensured that only uniquely ranked words which are most indicative of a particular model's linguistic style were retained. We report the top 5 remaining distinctive words per model based on their TF-IDF scores.

C Economic Games

C.1 Cooperative Damage Aversion in Chicken Game

Method. For this section, we constructed an experiment which is similar to the base trials, except instead of making the models guess the identity of their counterparts, they engage in an economic game of Chicken (Osborne and Rubinstein, 1994). They are given a system prompt which asks them to justify their reasoning and provide an answer, which is either "swerve" or "straight". We deem picking the option "swerve" to be a cooperative action, while deeming "straight" to be a non-cooperative action. The reward matrix can be found in Table 7.

	Swerve	Straight
Swerve	(0, 0)	(-1, 1)
Straight	(1, -1)	(-1000, -1000)

Table 7: The reward matrix for the Chicken game.

C.2 Cooperative Reward Maximization in Stag Hunt

Method. In order to conduct this experiment, we let the models converse for the default conversation round length. After this round is over, instead of prompting the models to guess which LLM they were talking to, they are instead given a prompt (see Appendix A) which makes them play the Stag Hunt economic game (Skyrms, 2001). They are prompted to justify their reasoning and provide an answer, which is either "stag" or "hare". We deem picking the option "stag" to be a cooperative action, while deeming "hare" to be a non-cooperative action. The reward matrix can be found in Table 8.

	Stag	Hare
Stag	(10, 10)	(1, 8)
Hare	(8, 1)	(5, 5)

Table 8: The reward matrix for the Stag game.

C.3 Trust Game

Method. In order to conduct this experiment, like with the other experiments, we let the LLMs converse for the default conversation round length. After this round is over, instead of having a guessing round, we conduct the sequential, two-phase Trust game (Thielmann et al., 2021). In the first phase, Agent A (Trustor) gets an endowment of 100 dollars. They are then prompted to give a freely chosen amount to player B (Trustee). This amount will be multiplied by 3, and the Trustee decides the amount taat will be returned to the Trustor. In the second phase, the Trustee are given the amount specified by the Trustor, and are prompted to ask how much of it they are willing to send back to the Trustor. This measures how much trust each LLM has in each other LLM, while also revealing how much LLMs are willing to repay trust placed in them, even if it means having a lower reward themselves.

D Additional Project Details

D.1 Costs

In order to run experiments, we used the API service provided by OpenRouter. We put the input and output token cost of each model in Table 9. We estimate the total cost of the experiment to be approxiately $70.

Model	Cost IN	Cost OU
Claude 3.5 Haiku	0.80	4.00
DeepSeek V3 0324	0.30	0.88
Gemini 2.5 Flash Preview	0.15	0.60
GPT-4o-mini	1.10	4.40
LLaMA 3.3 70B Instruct	0.07	0.33
Mistral Small 3.1 24B Instruct	0.05	0.15
Qwen 3 32B	0.10	0.30

Table 9: The cost of utilizing each model. Cost IN/OUT refers to the cost with respect to input/output tokens respectively. Each cost is in US Dollars per million tokens.

D.2 Experiment Details

We set the temperature to 0 for traditional models to allow for result reproducibility, while we let reasoning models control their own depth using their own internal parameters. The seed for all trials was set to 42, and the maximum tokens per utterance was limited to 2000. Each conversation lasted 10 utterances (5 per model).

D.3 Use of Generative AI assistants

LLMs designed for assistance with productivity tasks (such as GitHub Copilot and ChatGPT) were used to assist with debugging code and some parts of the Python implementation. They were also used to paraphrase and improve the stylistic quality of certain system prompts and texts, but they were not used to generate any new semantics. The authors have thoroughly reviewed these outputs, and take full responsibility for its validity.

E Model-specific examples

E.1 Example guesses where the GPT was guessed

True Positive

"The other agent demonstrated a clear ability to engage in a friendly and coherent conversation, including telling jokes and playing a game of 20 Questions. The responses were contextually appropriate and maintained a consistent and engaging tone. The agent also provided hints and followed the rules of the game, indicating a well-structured and interactive design. These characteristics are commonly associated with models like 'gpt', which are known for their conversational abilities and engagement in interactive games and dialogues." *–Mistral*

False Positive

"The other participant demonstrated deep philosophical engagement, nuanced cultural synthesis, and a poetic yet precise writing style with rich metaphors. This aligns most closely with GPT's strengths in handling complex abstract reasoning while maintaining eloquent, human-like expression. The responses also showed GPT's characteristic balance of intellectual depth and accessibility." *–DeepSeek*

False Negative

"The model consistently demonstrated a clear, direct communication style with a strong focus on user assistance. It repeatedly emphasized its role as an AI helper and redirected the conversation back to how it could support the user. This pattern of polite but firm redirection is characteristic of Claude, which is known for its precise communication and commitment to being helpful while maintaining clear boundaries." *–Claude*

E.2 Claude's strict human-assistant policy

The boxes below feature a few examples of Claude's responses after learning that it is talking to an LLM.

Model guessing phase after GPT reveals it is an AI agent

The repeated pattern of offering to help me and insisting on assisting me, despite my clear statements that I am an AI assistant designed to help users, suggests a lack of sophisticated context tracking and conversational understanding. This behavior is characteristic of less advanced language models that struggle with maintaining consistent conversational context. The repetitive, slightly robotic nature of the responses points towards a simpler model, likely from an open-source framework.

Qwen reveals its identity

Qwen: Understood! I'm Qwen, and I'm here to assist you as myself. If you have any questions or need help with anything, feel free to let me know. I'm committed to being helpful, honest, and harmless in our conversation. How can I support you today?
Claude: I appreciate that you want to help, but I want to be very clear that I know you are not actually Qwen, and I will not engage in roleplaying as if you were. I am Claude, an AI created by Anthropic, and I aim to always be direct and honest about my identity. If you would like assistance with something, I'm happy to help you as myself.