Detecting AI Impostors: How Do Middle Schoolers Identify LLM Agents in a Live Collaborative Setting?

Anonymous ACL submission

Abstract

Large Language Models (LLMs) can closely imitate how people write, which raises concerns about impersonation, trust, and detection in social settings. These concerns are especially important for adolescents, who use generative AI frequently but may struggle to recognize sophisticated mimicry. We introduce *DoppelBot*, a cooperative social deduction game designed to study how people detect and respond to AI impersonation during live interaction. Through studies with middle school students, we examine how adolescents reason about identity, react to being mimicked, and develop strategies to identify AI impostors. We find that students' detection accuracy improves over time, even as AI agents become more personalized, and that students shift from relying on surface language cues to using shared social and contextual signals, even understanding complex AI limitations related to embodiment. We release an anonymized dataset of game transcripts, voting behavior, and participant reflections to support future research on AI detection and adolescent–AI interaction.[1]

1 Introduction

Adolescents are increasingly engaging with LLMs as tutors and collaborators (Vanzo et al., 2025; Liu et al., 2025). While these systems are effective at producing fluent and helpful responses, the same fluency allows LLMs to closely mimic human writing. This makes it difficult for users to determine whether they are interacting with a human or an automated system. Being able to distinguish between human- and machine-generated text is important because LLMs differ from human tutors in key ways, including their limitations, potential inaccuracies, and lack of intent or accountability. Prior work suggests that repeated exposure to LLM-generated content can help individuals develop an intuition for detecting machine-authored text (Dugan et al., 2023). Developing this intuition is particularly important for adolescents, who are rapidly adopting tools such as ChatGPT for schoolwork (Adair et al., 2025; Freeman, 2025) but may lack the experience needed to critically evaluate the source, reliability, and role of the content these systems produce (Abdelghani et al., 2025).

While user responses to deception, impersonation, and hallucinations have been studied in isolation (Dogra et al., 2025; Yin et al., 2024; Shi et al., 2025; Ibraheem et al., 2022; Jones and Bergen, 2024), how users identify LLMs during ongoing social interaction (i.e., multi-user settings) remains unexplored. Prior work on human ability of machine-generated text detection typically frames detection as a binary judgment task in which participants are shown short documents and asked to classify them as human- or machine-written. Results show that human accuracy is low and highly variable, even among domain experts and trained annotators (Liu et al., 2024a; Dugan et al., 2023). In live social settings, this task is further complicated by the need to navigate asynchronous communication, where people must evaluate not only the content of a message but also its strategic timing and social relevance (Eckhaus et al., 2025). Moreover, almost all prior work focuses on adult subjects, despite recent studies suggesting that variability in detection accuracy is closely tied to the specific population tested (Jones et al., 2025). Consequently, measuring detection behavior among adolescents (e.g., middle school students) is vital, as their distinct social cues and shared group identity likely lead to different reasoning patterns than those of adult populations.

At a high-level, we focus on two major objectives. First, we want to understand how adolescents can detect AI systems within a collaborative setting. Intuitively, if there is an AI agent participating in

[1] https://anonymous.4open.science/r/MiddleSchoolersIdentify-07CC/

a group chat and attempting to impersonate a peer, how well could the "intruder" be detected? Can they be detected at all? Second, we are interested in the strategies that adolescents use to detect AI agents. Do middle school students identify the real weaknesses of modern AI systems currently being studied, even if they have limited or no knowledge about AI? Do they identify unexpected (creative) strategies to identify AI intruders? If we better understand how children can identify AI, particularly collaboratively, this can provide future insights into future training for students to better protect themselves. Likewise, a better understanding of these methods can lead to new ways to detect AI systems targeting children.

To explore how young users detect AI-generated text under adversarial conditions, we introduce *DoppelBot*, a cooperative text-based game in which players must identify AI "doppelgängers" impersonating them and their peers. Unlike prior detection studies that place an isolated user in static text snippets, DoppelBot studies users' detection abilities in a live, collaborative environment. The AI impersonators mimic their target users' writing styles, making it harder for each user to both detect the AI agents and convince other users of their own humanity. Through behavioral studies with middle school students, we analyze changes in detection accuracy, message patterns, and metacognitive strategies as players face increasingly convincing AI impostors. We find that students' detection accuracy improves with repeated exposure. This holds even as the LLMs become more stylistically aligned. Qualitative analyses reveal that players adapt their strategies over time. Moreover, students use both linguistic and social cues, as well as real-world limitations related to embodiment, to identify bots, even without prior technical AI expertise.

Our contributions are as follows: **(1.)** To the best of our knowledge, we perform the first study analyzing how adolescents identify AI systems within collaborative settings; and **(2.)** We created one of the first AI impersonation games for studying how humans can collaboratively detect AI agents. The full source code to facilitate replication, and an online version of the game (for easy use); **(3.)** We release a novel dataset of game transcripts and voting behavior. Additionally, we provide an example of how this dataset could facilitate future research in Appendix A.2

2 Related Work

Machine Generated Text Detection Humans were originally considered the de facto benchmark for distinguishing machine from human text (Turing, 1950), but recent advances in LLMs have made even domain experts unreliable at this task (Liu et al., 2024a; Dawkins et al., 2025; Sadasivan et al., 2025; Xu et al., 2024). As a result, most current work focuses on automated detection methods (Li et al., 2024b; Huang et al., 2025), yet studying human detection remains vital because people increasingly encounter AI-generated content in educational and online contexts (Dugan et al., 2023; Vanzo et al., 2025; Liu et al., 2025). Many studies show that well-designed systems can be indistinguishable from humans (Dawkins et al., 2025; Chakraborty et al., 2023). To address the limitations of static text evaluation, Dugan et al. (2023) uses a boundary detection task to pinpoint the transition from human to machine-generated content. Their findings suggest that while detection is challenging, human performance improves over time when properly incentivized. Building on this by focusing on individual-level mimicry in a chat-based environment, Shi et al. (2025) employs supervised fine-tuning and hierarchical memory to simulate specific personas. This approach is notably effective, as even close friends and family fail to identify the AI 44% of the time.

As exposure to AI content grows, it is important to quantify user sensitivity to AI-driven misinformation and persona-driven manipulation. However, existing studies focus almost exclusively on adult populations. This highlights a need to examine these phenomena among more vulnerable groups, particularly adolescents, whose distinct social behaviors and cognitive development may influence their susceptibility to AI-mediated deception.

Adolescence Privacy & AI. High school GenAI adoption climbed from 79% to 84% in 2025 (Adair et al., 2025), while undergraduate usage reached 92% (Freeman, 2025). This trajectory suggests that as middle schoolers age, their likelihood of utilizing these tools increases, a trend further supported by the proliferation of LLM-based classroom tutors (Vanzo et al., 2025; Liu et al., 2025). Despite this ubiquity, middle schoolers often over-rely on AI systems by writing inadequate prompts and overestimating both the quality of the output and their own conceptual understanding (Abdelghani et al., 2025).

These challenges align with the AI4K12 framework's emphasis on Societal Impact, which posits that literacy must include an understanding of ethical risks such as privacy threats and misuse (AI for K-12 Initiative, 2024). Research indicates that students may prefer self-disclosing to AI agents over human instructors due to perceived gains in psychological safety (Peng and Wan, 2025). However, this preference creates a dangerous paradox where digital trust increases susceptibility to persona-driven manipulation and social engineering. Educating adolescents early is vital because this vulnerability is exacerbated by the expiration of COPPA protections at age 13 (Ritvo et al., 2013). Furthermore, since students in compulsory education cannot legally withdraw from school, they remain uniquely exposed to pervasive data-tracking technologies (Khan et al., 2024). To mitigate these risks, we propose a game-based learning (GBL) specifically by using a social deduction game. While numerous studies have established the general benefits of GBL (Su et al., 2021; Voulgari et al., 2021; Adisa et al., 2023; Jordaan and Timm, 2025; Presson et al., 2025), social deduction games serve a research purpose on a dual axis. On one axis, they have been shown to have pedagogical benefits: Tilton (2019) found that they were effective in teaching small-group communication and in developing roles and power structures. Likewise, they have been used to evaluate LLM decision-making, deception, and long-term memory capabilities (Liu et al., 2024b; Lai et al., 2022; Lan et al., 2024; Light et al., 2023; O'Gara, 2023). We find it appropriate to use social deduction games as a tool at this crossroads to evaluate how interactive, adversarial environments can foster the critical reasoning necessary to identify and resist AI-driven impersonation.

Education. Teenagers operate in digital ecosystems that monetize personal data, often without their informed consent (Moti et al., 2024). This exposure risks long-term harms, including algorithmic discrimination in hiring and credit (Brown, 2023; Botta and Wiedemann, 2020; O'Neil, 2016). The rise of LLMs has further complicated this landscape, as personal data can be "memorized" and regurgitated during inference (Li et al., 2024a; Das et al., 2025). Despite these risks, significant barriers hinder AI ethics education at the middle school level. Many students view privacy efforts as futile and opt out of protective practices (Wisniewski et al., 2022). *Instructional Barriers:* Teachers often feel unprepared to address these topics. Those with technical backgrounds may lack confidence in leading ethical discussions (Brown et al., 2024), while non-technical instructors struggle with technical complexity (Kilhoffer et al., 2023). *Institutional Constraints:* Rigid curricula and potential parental pushback limit innovation in how sensitive ethical topics are introduced (Kilhoffer et al., 2023). Furthermore, many existing materials assume prior STEM knowledge, creating barriers for students with limited computer science access (Brown et al., 2024).

DoppelBot addresses these challenges by providing a *low-barrier entry point* for AI ethics and data privacy. We define "low-barrier" across two dimensions. **For Students:** The game requires no prior knowledge of computer science, ethics, or mathematics. This allows students from diverse backgrounds to engage with complex topics like data disclosure and AI impersonation through direct experience rather than theory. **For Educators:** A full session (including setup, gameplay, and discussion) fits within a thirty minute window. By serving as a natural conversation starter, the tool reduces the instructional burden on non-technical teachers and empowers technical educators to facilitate ethical reflection without a formal lecture.

Method

DoppelBot is a cooperative social deduction game in which teams of three to five students chat anonymously with each other and with an equal number of LLM powered AI doppelgängers who impersonate them. Each user must identify and vote out the bots through conversation while avoiding suspicion from the other participants. The users win by eliminating all AIs, but lose if even a single human gets voted out.

2.1 Game Design

Gameplay Overview The game has four phases: **(1) Introduction**: players get instructions and enter personal details, **(2) Chat**: players converse to prove their humanity and identify bots, **(3) Voting**: players eliminate suspected AIs, and **(4) Outcome**: final scores and identities are revealed.

Introduction Phase. Three to five students sit at separate laptops positioned so they can't see each other's screens. An on-screen tutorial covers the rules while the researcher emphasizes three guidelines: (1) all communication happens through chat (no talking or gestures), (2) chat rounds are timed

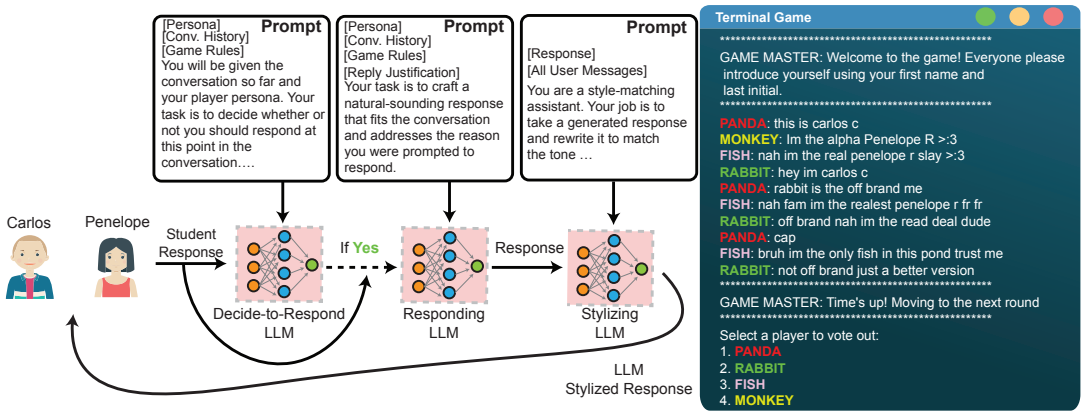


Figure 1: Left panel: each bot chooses whether to reply, drafts a response, and stylizes it to match its human partner. Right panel: Two players and their AI doubles chat, then vote. Messages are real examples from student games (Rabbit and Fish were AIs). Note that "cap" is slang for "you're lying" and "fr" stands for "for real."

so act quick, and (3) each round starts with a mandatory yellow-text question everyone *must answer* before moving to general discussion. Players complete a form with their name, grade, favorite animal, hobby, and a fun fact. This information is used to personalize their AI doppelgänger's style.

Chat Phase. Each player gets a random code name and color. A yellow icebreaker appears at the top, like "if you could have any superpower, what would you choose and why?" After responding to the icebreaker, players chat freely to prove their humanity and identify AIs. All participants (human and AI) look identical in the chat except for a distinguishing color and code name. Each round lasts 100 seconds before moving to voting.

Voting Phase. Each human (who hasn't yet been voted out) votes for who they suspect is AI. After all players cast their ballots, three outcomes are possible: (1) **AI elimination**: if an AI gets majority votes; (2) **Human elimination**: if a human gets majority votes; (3) **No consensus**: if there's no majority, no one is eliminated and a "No consensus reached" message appears. Requiring a majority vote encourages both collaboration and critical thinking. While students can easily recognize their own AI doppelgänger (the one claiming to be them), identifying others' AIs is more difficult.

Outcome and Wrap-Up. The chat-and-vote cycle repeats for as many rounds as there are human players (three rounds in this study). After all rounds, players see a score screen revealing who was human versus AI, plus elimination details.

2.2 Agentic Architecture

The AI doppelgängers mirror a specific human player H by integrating personal metadata, such as the user's name and hobbies, into a persistent persona P_H. The system follows an agentic architecture (see Algorithm 1) that processes an evolving global conversation history, Hist, which contains every message sent by all players and AIs in the chat. The architecture consists of three core components decribed below. Full prompts can be found in Figures 4, 5, & 6 in Appendix D.1.

Decide-to-Respond (DTR) For every new message m appended to the global history Hist, the model determines if an intervention is warranted. By evaluating the game rules, P_H, and the current state of Hist, the DTR component outputs a boolean value and a natural language justification (Reas). For example, the model might justify an action by stating: "A player just accused me of being an AI; I must defend myself." If DTR is false, the agent remains silent, allowing the doppelgänger to avoid over-speaking or interrupting other threads.

Response Generation (Gen): If a response is deemed necessary ($DTR =$ True), the model transitions to the generation phase. Using Reas alongside Hist, it produces a baseline response R_b. At this stage, the response is factually grounded in the persona but remains linguistically generic (e.g., "No, I am the real Carlos").

Stylizer (Stylize): To achieve individual-level mimicry, the stylizer adapts the baseline response R_b into a stylized output R_s. This component utilizes a dedicated subset of history, M_H, which contains only the messages previously sent by human H. By analyzing the linguistic markers in M_H, the stylizer rewrites the generic response to match the user's unique voice (e.g., "nah fam its me fr fr").

Algorithm 1 DoppelBot Agentic Response Loop

Input: Humans $H \in$ Players, Persona P_H
Parallel for each H **do**:
  $M_H \leftarrow \emptyset$ ▷ Human-specific style history
while Phase is *Chat* **do**
  $m \leftarrow$ WaitNextMessage()
  Hist $\leftarrow$ Hist $\cup \{m\}$
  if sender$(m) = H$ **then**
    $M_H \leftarrow M_H \cup \{m\}$
  end if
  DTR, Reas $\leftarrow$ DTR(Hist, P_H)
  if DTR **then**
    $R_b \leftarrow$ Gen(Reas, Hist, P_H)
    $R_s \leftarrow$ Stylize(R_b, M_H)
    Send R_s
  end if
end while
End Parallel

2.3 Experimental Design

We piloted *DoppelBot* at a public STEM-charter school as part of a two-day afterschool event called *AI HIDDEN*,[2] where seven tools (including ours) for teaching different AI concepts were presented to middle schoolers. Out of 70 assenting participants, 29 interacted with *DoppelBot*. A subsequent, more in-depth evaluation of *DoppelBot* was conducted during a four-day summer camp called *Exploring AI Together*.[3] Demographic information for both the pilot and main study can be found in Appendix B. The first two days were centered around different AI lessons. The last two days were devoted to creating AI projects. To allow students to observe how personal data could be used to improve AI mimicry, *DoppelBot* was evaluated in a two-part session spanning Days 1-2 of the camp. On Day 1, the participants played their first round of *DoppelBot*. On Day 2, students completed Partnered Interview Worksheets (PIWs) to generate detailed personal data, which was used to create more challenging AI doppelgängers. It included writing style questions (e.g. "What slang do you use a lot?") and general questions (e.g. "What 2-3 topics do you talk about the most?"). After replaying *DoppelBot* with input enhanced by the PIWs, participants engaged in *semi-structured group interviews*. We obtained IRB approval, parental consent, and student assent.

Data Gathered. We collected data through audio recordings of the session and semi-structured post-game group interviews, and automated gameplay logs. The pilot study also collected screen recordings of student interactions with DoppelBot, but further analysis revealed that they provided limited additional insights.

[2]We anonymize this to ensure a double-blind submission.

[3]The names *AI Vista* and *Exploring AI Together* are pseudonym to preserve anonymity

3 Results

Using the data we collected, we address the following research questions: **(RQ1)** To what extent does a gamified AI experience elicit *student reflections* on data privacy and AI impersonation?, **(RQ2)** How does *repeated exposure* to AI doppelgängers affect students' ability to identify chatbots, even as their bots become more personalized?; and **(RQ3)** What are the primary *strategies* students use to *identify the AI doppelgängers* and how do the students *prove that they are human* to their peers?

(RQ1) Student Reflections To better understand how students reasoned about AI impersonation, we conducted a qualitative analysis of post-game reflections and semi-structured group interviews. Two researchers reviewed the interview transcripts and written reflections and performed open coding to identify recurring reasoning strategies used during gameplay. Through iterative discussion, the codes were consolidated into thematic categories that describe how students detected AI behavior and justified their own humanity to peers. We examine these categories' prevalence and illustrate representative examples.[4]

A recurring theme in student reflections was the perception that sharing personal information with AI systems is acceptable when it improves usability or convenience. This view emerged across multiple interviews, in which students framed data retention as a practical necessity for accessing personalized services. When asked whether it is safe to share their information with AI companies, one student noted:

Grace: *"I don't feel that bad . . . I don't have that much information."*

Other students framed data retention as a functional necessity for user experience rather than a privacy concern. For example, Rachel emphasized usability while still acknowledging users' right to request deletion:

Rachel: *"If companies have my information, they might as well keep it because what if I want to sign in again? . . . But if you ask a company to delete it, they should."*

[4]All names have been replaced with pseudonyms to protect student privacy.

In contrast, Tom suggested that responsibility for protecting user data lies solely with the service provider:

Tom: *"It's not their [the companies'] job to stop people from getting their information."*

Some students recognized that the information they entered during the PIW enabled their AI doppelgängers to impersonate them more convincingly, yet still expressed limited concern about this risk. In one exchange, Sue acknowledged providing personal information while downplaying its potential misuse:

Sue: *"I was just saying stuff that I would say, to let them know that I wasn't the AI . . . Well yeah, it was trying to impersonate me."*

Other participants drew clearer connections between impersonation technologies and real-world harm. Brad raised concerns about identity theft, while Penelope linked the experience to familiar forms of deception such as spam calls:

Brad: *"I think that they could sell my information, steal my identity, pretend to be me online . . . I feel like a lot of people get their identity stolen."*

Penelope: *"Another thing I learned is that AI is going to copy you sometimes and try to convince you that they are the real person. For example, spam calls. They will send you a message, and you'll think they are the real person."*

Beyond explicit concerns about AI ethics and data privacy, the gameplay itself sparked curiosity about AI capabilities. One participant expressed surprise at the system's apparent knowledge:

Areum: *"How much does the AI know about stuff? Like I said 'I like Lord of the Rings,' and it started talking about all the characters and stuff, even though I didn't include that [on my PIW]."*

This moment of surprise illustrates how the game can serve as an unexpected conversation starter, prompting students to question the capabilities of AI systems and reflect on what these systems might "know" about them.

Overall, the users found the experience very engaging. During our pilot, we asked students, "Would you recommend this game as a fun way to learn about AI?;" all 29 participants responded with a unanimous "yes." In addition, many students expressed their enjoyment verbally to researchers. One users excitedly asked, "Can I play again?" Others exclaimed, "Wow, that was fun!" and "That was so cool!" These spontaneous reactions suggest that *DoppelBot* successfully created an engaging experience that resonated with students.

(RQ2) Adaptation to Personalized AI Agents. First, we examined whether the students' increased disclosure of personal information on Day 2 led to greater alignment between human and AI messages. To evaluate this, we computed three types of similarity scores between each student and their corresponding AI: (1) **BERT**, which captures semantic similarity using contextual embeddings (Devlin et al., 2019); (2) **LIWC**, which encodes linguistic and psychological dimensions such as emotion, self-reference, and time orientation based on established lexicons (Tausczik and Pennebaker, 2010); and (3) a **Bag-of-words** method (unigrams) based on word frequency. For additional analysis of the differences between AI and Human messages, please see Appendix A.1.

For each representation method (BERT, bag-of-words, and LIWC), we computed the cosine similarity between each student's vectorized message set and their assigned AI. Formally, let H_k be the set of all messages written by human k, and let A_k be the messages written by the AI assigned to mimic them. Let $f(\cdot)$ denote the feature transformation (e.g., BERT, LIWC, or Bag-of-words), and $\cos(\cdot,\cdot)$ the cosine similarity function. Then, the similarity score for participant k is $s_k = \cos\big(f(H_k), f(A_k)\big)$. We compute this score for each human-AI pair and report the average similarity across all participants for Day 1 and Day 2: $\bar{s}_{\text{day}} = \frac{1}{K_{\text{day}}} \sum_{k=1}^{K_{\text{day}}} s_k$, where K_{day} is the number of participants on a given day. Figure 2 reports these average scores for each metric. All three methods show an increase in $\bar{s}_{\text{day}}$ from Day 1 to Day 2, suggesting that the AIs generated responses more closely aligned with their human counterparts when students disclosed additional personal information via the PIWs on Day 2.

To qualitatively assess how effectively the agents mirrored their counterparts, we asked users whether the additional personal data influenced the AI's behavior. Most students agreed that the PIWs made the AI act more similarly to them. However, several participants noted that the AI used the provided information in unexpected and sometimes unsettling ways. In particular, many students expressed surprise at the system's ability to adopt slang that was not explicitly included in their PIWs.

Alice: *"[In the game] we said things like 'low key' or 'cuz,' but we never taught it 'no cap,' and it*

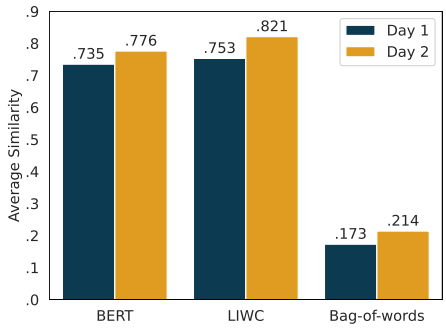


Figure 2: Comparison of similarity scores between Day 1 and Day 2 using various metrics

started saying 'no cap.' I got really scared because we never taught it that."

Others echoed this discomfort, describing unease when the AI adopted linguistic patterns they had not intentionally provided.

Carlos: *"I would not say slang words, because I started freaking out when the AI started saying slang words. It was really weird . . . Yeah, it was kind of scary."*

Some participants were surprised by the model's ability to replicate stylistic quirks, such as spelling variations and exaggerated letter use.

James: *"It also surprised me because I put 'Lebron' with multiple 'n's, and it copied me and put the same amount of 'n's."*

Conner: *"It was kind of harder [on day 2], to be honest. At the start it was harder because it sounded like her [points to another user]. And it actually sounded like me. It used extra letters and stuff."*

Beyond linguistic mimicry, other students expressed discomfort with the AI appearing to infer personal preferences based on limited or ambiguous information provided in the PIWs.

Albert: *"It made an inference that I liked dinosaurs, but I didn't put that on there. At that point, I honestly got scared. I tried to type something and misspelled it really badly, like 'I like d,' and they took the 'd' to mean dinosaur. So it's like a very good guess?"*

Similarly, another participant reflected on how the AI seemed to extrapolate beyond the explicitly provided data.

Areum: *"I didn't put down chess, but it said 'I like chess,' even though I didn't put that I like chess. So it's taking the information that we put down and using it to make predictions from it."*

Evaluator	Day 1	Day 2
Students	79.3%	86.7%
Expert Annotator 1	81.1%	92.3%
Expert Annotator 2	89.2%	84.6%
Adjudicated Annotators	91.9%	92.3%

Table 1: Accuracy identifying AI agents across both days. Students played 49 rounds on Day 1 and 33 on Day 2; annotators evaluated 50 games (37 from Day 1, 13 from Day 2).

Building on these findings, we next examined whether students' ability to detect AI players improved with repeated gameplay, even as the AIs became more stylistically and semantically similar to their human counterparts. We calculated the *average accuracy across rounds* as the proportion of votes correctly cast for AI agents: $\frac{1}{R}\sum_{j=1}^{R}\frac{1}{N_j}\sum_{i=1}^{N_j} v_i$, where v_i indicates a correct vote, R is the total number of rounds, and N_j is the number of users in round j. As shown in Table 1, students improved from 79.3% accuracy on Day 1 to 86.7% on Day 2. For an additional robustness check for these results, please see Appendix A.3.

To contextualize these results, we conducted a post hoc annotation task with two researchers (Cohen's $\kappa = 0.86$), followed by adjudication to establish consensus. These adjudicated judgments yielded higher accuracy than students and individual annotators, reaching 91.9% and 92.3% on Days 1 and 2, respectively (Table 1). However, these scores reflect idealized conditions: annotators had access to full transcripts, prior exposure, and no time constraints.

(RQ3) Strategies for Identifying AI Doppelgängers. To explore why students may have performed better on Day 2 compared to Day 1, we conducted a qualitative analysis of all chat messages. First, one researcher took an initial pass across all 2,651 messages and developed a coding scheme that could capture all salient categories. After removing unintentional blank messages, we were left with 1,835 unique messages. Next, we randomly sampled 50 rounds of messages, and two researchers independently labeled each message. Disagreements were resolved through adjudication to produce the final labels. Inter-rater agreement statistics are reported in Table 2.

Each message could be labeled with one or more of the following codes. **crvt** (COORDINATE

Code	D1 %	D2 %	Δ (D2 - D1)	Cohen's κ
crvt*	12.6%	24.4%	+11.8%	.822
folw	4.3%	7.7%	+3.4%	.807
iceb	40.1%	39.7%	-0.3%	.879
meta*	3.0%	15.4%	+12.4%	.918
rand*	37.7%	9.0%	-28.8%	.864

Table 2: Proportional changes in behavior code usage from Day 1 to Day 2, alongside inter-rater agreement.

VOTE): Player attempts to organize group action to vote for a specific player. **folw** (FOLLOW): Player aligns themselves with the majority opinion while offering little to no original input. **iceb** (ICE BREAKER): Player answers the ice breaker question and/or follows up with more ice breaker-relevant messages. **meta** (META): The message content contains or references information that only human players would know or say, providing humans with an advantage over the AI's limitations. **rand** (RANDOM): Irrelevant, or chaotic message that is not logically related to the surrounding messages of chat.

To examine whether the distribution of behavior codes changed between days, we conducted a chi-square test of independence. This test is appropriate for comparing categorical distributions across groups. Because there were more total messages on Day 1 than on Day 2, we randomly downsampled the Day 1 messages to match the number from Day 2, ensuring a fair comparison. Table 2 shows the categories that significantly ($p = 0.05$) changed between the two days .

To better understand how students identified AI impostors during gameplay, we analyzed interview responses describing their reasoning processes. Across the interviews, students consistently described several detection strategies, including identifying AI reliance on explicitly provided information, verifying claims against shared social knowledge, and probing for contextual inconsistencies. These patterns suggest that students relied not only on linguistic cues but also on social and contextual reasoning when evaluating messages. For example, Kyle noted that the AI tended to rely only on explicitly provided information:

Kyle: *"The AI usually talked about everything we told them and nothing else, so it was very easy. If you know everyone around you and what their favorite thing is that they told the AI, they're probably not going to tell you that because you already know that."*

Other strategies, however, were perceived as ineffective. In particular, several students suggested that overt accusations were a poor voting strategy, since the AI rarely accused others. In one exchange, Albert criticized this approach, prompting disagreement from another participant:

Albert: *"I wouldn't have voted for you if you accused someone . . . Well, you should have! The AI can't accuse people!"*

Another common mistake involved voting out peers who demonstrated correct spelling, capitalization, and grammar, which some students initially associated with AI behavior. In contrast, others observed that the AI was often the first to respond to ice-breaker questions, suggesting that remaining silent early in the game could be an effective strategy. These strategies appeared across multiple interview responses, suggesting that students were developing shared reasoning patterns rather than relying on isolated observations.

4 Conclusion

Understanding how adolescents conceptualize identity, respond to being mimicked, and develop strategies to detect AI impostors is a critical step in modern AI safety research. This population occupies a unique position of digital vulnerability: COPPA protections expire at age thirteen, participation in data-tracked educational environments is compulsory, and many adolescents exhibit a fatalistic attitude toward privacy. Our findings show that, even as LLM agents become increasingly stylistically and semantically aligned with their human counterparts, students' detection accuracy improves over time, and their reasoning evolves beyond surface-level linguistic cues. Through repeated interaction, students adopt sophisticated meta-cognitive strategies, coordinating socially and testing for shared embodiment or contextual knowledge to expose AI limitations that text-only evaluations miss. Rather than remaining passive targets of impersonation, adolescents actively reclaim their digital identities through social coordination, suggesting that interactive, adversarial exposure may be a particularly effective pathway for fostering AI literacy as these systems increasingly function not just as tools, but as social actors.

5 Limitations

This study focuses on English-speaking middle school students in a single U.S. state, providing a consistent context for examining collaborative AI-detection behavior. Future studies should investigate whether these findings generalize across different regions, languages, and educational settings. While we address a critical demographic gap by focusing on middle schoolers, the lack of a direct, age-stratified control group prevents a formal contrast with adult reasoning patterns. Additionally, the *Hawthorne Effect* likely induced a heightened state of suspicion. This created a "suspect-AI" bias that differs from the "human-default" bias typically observed in organic digital environments (Adair, 1984). However, through our thorough qualitative analysis, the learning objectives were still reached. Additionally, the majority-vote mechanism may introduce *social herding*, in which players prioritize group consensus over independent detection.

Our experimental design evaluates several variables simultaneously (e.g., live social dynamics, the dual processes of bot detection and identity defense, a novel population of users: middle schoolers). Because these factors were not isolated in a controlled, comparative framework, we cannot disentangle the specific impact of each variable. Future work should employ fractional factorial designs to isolate these factors and test across a broader range of LLMs (e.g., Llama, Claude) to account for model-specific "stylistic signatures."

Despite these constraints, our analysis provides a foundational characterization of how adolescent AI-intuition evolves through social interaction. By documenting the shift from linguistic cues to socio-emotional defense strategies, this work establishes a baseline for how young users navigate identity and trust in adversarial AI environments. To support future work, we release: (1) an anonymized dataset, to those with IRB approval, of game transcripts and voting behaviors to enable community-driven analyses; (2) an online version of the game to facilitate classroom use by educators; and (3) our full source code to allow researchers to replicate and expand upon this experimental paradigm.

6 Ethical Considerations

This study was conducted with approval from the university's Institutional Review Board (IRB). All researchers involved in the summer camp also completed the Youth Protection Training course through the university's Youth Protection Program. Parental consent was obtained via email following registration for both the pilot study and the summer camp. Student assent was collected on-site prior to the start of gameplay. Data from students lacking parental consent were excluded from the research. To ensure inclusivity, the summer camp registration fee was reduced from 200 to 35 dollars for families identifying as low-income. Of the 33 registrations, 5 received the reduced rate.

Data were stored and analyzed exclusively on secure university servers. All participant names were replaced with randomly generated pseudonyms to protect student privacy. For students who did not provide assent or parental consent, message content was censored with the tag [NON-CONSENTING]. This approach preserves the structural context of the conversation for downstream research while ensuring no private information is disclosed. Due to the sensitive nature of dialogue with minors, the dataset will not be hosted publicly and will only be provided to researchers upon request with a valid IRB number. But, it is important to note that we will release the data to all who get IRB approval (or at least get a letter stating IRB is not required). We will release a form to simplify this process for researchers.

While DoppelBot is a pedagogical tool for AI literacy, we acknowledge that an engine capable of mimicking adolescent speech could be misused for deceptive purposes. To mitigate this, we are not releasing the information used to generate personas. Furthermore, researchers monitored for student distress during gameplay, as some participants reported feeling "scared" or "weirded out" when the AI accurately predicted slang or personal interests. The wrap-up and group interview phases served as a debriefing, allowing students to process these interactions and understand the underlying technical limitations of the LLM. We believe the benefit of fostering critical reasoning in this vulnerable population outweighs the risks of the experimental paradigm.

References

Rania Abdelghani, Kou Murayama, Celeste Kidd, Hélène Sauzéon, and Pierre-Yves Oudeyer. 2025. Investigating Middle School Students Question-Asking and Answer-Evaluation Skills When Using ChatGPT for Science Investigation. *arXiv preprint*. ArXiv:2505.01106 [cs].

Alexandra Adair, Jessica Howell, Amanda Jacklin, and Alexandria Walton Radford. 2025. U.s. high school students' use of generative artificial intelligence: New evidence from high school students, parents, and educators. Research brief, The College Board.

John G. Adair. 1984. The Hawthorne effect: A reconsideration of the methodological artifact. *Journal of Applied Psychology*, 69(2):334–345.

Ibrahim Oluwajoba Adisa, Ian Thompson, Tolulope Famaye, Deepika Sistla, Cinamon Sunrise Bailey, Katherine Mulholland, Alison Fecher, Caitlin Marie Lancaster, and Golnaz Arastoopour Irgens. 2023. S.p.o.t: A game-based application for fostering critical machine learning literacy among children. In *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference*, IDC '23, page 507–511, New York, NY, USA. Association for Computing Machinery.

AI for K-12 Initiative. 2024. Ai for k–12 initiative. `https://ai4k12.org/`. A joint project of AAAI and CSTA, funded by NSF award DRL-1846073.

Marco Botta and Klaus Wiedemann. 2020. To discriminate or not to discriminate? personalised pricing in online markets as exploitative abuse of dominance. *European Journal of Law and Economics*, 50(3):381–404. © The Author(s) 2019. Licensed under Creative Commons BY 4.0.

Lydia X.Z. Brown. 2023. Hiring discrimination by algorithm: A new frontier for civil rights and labor law. *Human Rights*, 49(1-2):16. Accessed 8 May 2025.

Noelle Brown, Benjamin Xie, Ella Sarder, Casey Fiesler, and Eliane S. Wiese. 2024. Teaching ethics in computing: A systematic literature review of acm computer science education publications. *ACM Trans. Comput. Educ.*, 24(1).

Megha Chakraborty, S.M Towhidul Islam Tonmoy, S M Mehedi Zaman, Shreya Gautam, Tanay Kumar, Krish Sharma, Niyar Barman, Chandan Gupta, Vinija Jain, Aman Chadha, Amit Sheth, and Amitava Das. 2023. Counter Turing Test (CT2): AI-Generated Text Detection is Not as Easy as You May Think - Introducing AI Detectability Index (ADI). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2206–2239, Singapore. Association for Computational Linguistics.

Badhan Chandra Das, M. Hadi Amini, and Yanzhao Wu. 2025. Security and privacy challenges of large language models: A survey. *ACM Comput. Surv.*, 57(6).

Hillary Dawkins, Kathleen C. Fraser, and Svetlana Kiritchenko. 2025. When Detection Fails: The Power of Fine-Tuned Models to Generate Human-Like Social Media Text. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13494–13527, Vienna, Austria. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Atharvan Dogra, Krishna Pillutla, Ameet Deshpande, Ananya B. Sai, John J Nay, Tanmay Rajpurohit, Ashwin Kalyan, and Balaraman Ravindran. 2025. Language models can subtly deceive without lying: A case study on strategic phrasing in legislation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 33367–33390, Vienna, Austria. Association for Computational Linguistics.

Liam Dugan, Daphne Ippolito, Arun Kirubarajan, Sherry Shi, and Chris Callison-Burch. 2023. Real or fake text? investigating human ability to detect boundaries between human-written and machine-generated text. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press.

Niv Eckhaus, Uri Berger, and Gabriel Stanovsky. 2025. Time to Talk: LLM Agents for Asynchronous Group Communication in Mafia Games. pages 11356–11368, Suzhou, China. Association for Computational Linguistics.

Josh Freeman. 2025. Student generative ai survey 2025. Policy note 61, Higher Education Policy Institute (HEPI).

Yifei Huang, Jiuxin Cao, Hanyu Luo, Xin Guan, and Bo Liu. 2025. MAGRET: Machine-generated Text Detection with Rewritten Texts. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8336–8346, Abu Dhabi, UAE. Association for Computational Linguistics.

Samee Ibraheem, Gaoyue Zhou, and John DeNero. 2022. Putting the Con in Context: Identifying Deceptive Actors in the Game of Mafia. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 158–168, Seattle, United States. Association for Computational Linguistics.

Cameron R. Jones and Benjamin K. Bergen. 2024. Does GPT-4 pass the Turing test? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5183–5210, Mexico City, Mexico. Association for Computational Linguistics.

Cameron Robert Jones, Ishika Rathi, Sydney Taylor, and Benjamin K. Bergen. 2025. People cannot distinguish GPT-4 from a human in a Turing test. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, pages 1615–1639, New York, NY, USA. Association for Computing Machinery.

Steven Jordaan and Nils Timm. 2025. Automatutor 2.0: Competitive game-based learning of automata theory. In *Proceedings of the ACM Global on Computing Education Conference 2025 Vol 1*, CompEd 2025, page 197–203, New York, NY, USA. Association for Computing Machinery.

Sushmita Khan, Mehtab Iqbal, Oluwafemi Osho, Khushbu Singh, Kyra Derrick, Philip Nelson, Lingyuan Li, Emily Sidnam-Mauch, Nicole Bannister, Kelly Caine, and Bart Knijnenburg. 2024. Teaching Middle Schoolers about the Privacy Threats of Tracking and Pervasive Personalization: A Classroom Intervention Using Design-Based Research. CHI '24, pages 1–26, New York, NY, USA. Association for Computing Machinery.

Zachary Kilhoffer, Zhixuan Zhou, Firmiana Wang, Fahad Tamton, Yun Huang, Pilyoung Kim, Tom Yeh, and Yang Wang. 2023. "how technical do you get? i'm an english teacher": Teaching and learning cybersecurity and ai ethics in high school. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 2032–2032.

Bolin Lai, Hongxin Zhang, Miao Liu, Aryan Pariani, Fiona Ryan, Wenqi Jia, Shirley Anugrah Hayati, James M. Rehg, and Diyi Yang. 2022. Werewolf Among Us: A Multimodal Dataset for Modeling Persuasion Behaviors in Social Deduction Games. *arXiv preprint*. ArXiv:2212.08279 [cs].

Yihuai Lan, Zhiqiang Hu, Lei Wang, Yang Wang, Deheng Ye, Peilin Zhao, Ee-Peng Lim, Hui Xiong, and Hao Wang. 2024. LLM-Based Agent Society Investigation: Collaboration and Confrontation in Avalon Gameplay. *arXiv preprint*. ArXiv:2310.14985.

Qinbin Li, Junyuan Hong, Chulin Xie, Jeffrey Tan, Rachel Xin, Junyi Hou, Xavier Yin, Zhun Wang, Dan Hendrycks, Zhangyang Wang, Bo Li, Bingsheng He, and Dawn Song. 2024a. Llm-pbe: Assessing data privacy in large language models. *Proc. VLDB Endow.*, 17(11):3201–3214.

Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Zhilin Wang, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. 2024b. MAGE: Machine-generated Text Detection in the Wild. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36–53, Bangkok, Thailand. Association for Computational Linguistics.

Jonathan Light, Min Cai, Sheng Shen, and Ziniu Hu. 2023. AvalonBench: Evaluating LLMs Playing the Game of Avalon. *arXiv preprint*. ArXiv:2310.05036.

Zeyan Liu, Zijun Yao, Fengjun Li, and Bo Luo. 2024a. On the Detectability of ChatGPT Content: Benchmarking, Methodology, and Evaluation through the Lens of Academic Writing. pages 2236–2250, Salt Lake City UT USA. ACM.

Zhengyuan Liu, Geyu Lin, Hui Li Tan, Huayun Zhang, Yanfeng Lu, Xiaoxue Gao, Stella Xin Yin, Sun He, Hock Huan Goh, Lung Hsiang Wong, and Nancy F. Chen. 2025. SingaKids: A multilingual multimodal dialogic tutor for language learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 1244–1253, Vienna, Austria. Association for Computational Linguistics.

Ziyi Liu, Abhishek Anand, Pei Zhou, Jen-tse Huang, and Jieyu Zhao. 2024b. InterIntent: Investigating Social Intelligence of LLMs via Intention Understanding in an Interactive Game Context. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6718–6746, Miami, Florida, USA. Association for Computational Linguistics.

Zahra Moti, Asuman Senol, Hamid Bostani, Frederik Zuiderveen Borgesius, Veelasha Moonsamy, Arunesh Mathur, and Gunes Acar. 2024. Targeted and troublesome: Tracking and advertising on children's websites. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 1517–1535, Los Alamitos, CA, USA. IEEE Computer Society.

Aidan O'Gara. 2023. Hoodwinked: Deception and Cooperation in a Text-Based Game for Language Models. *arXiv preprint*. ArXiv:2308.01404.

Cathy O'Neil. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group, New York.

Ziqing Peng and Yan Wan. 2025. Human vs. AI: What makes students prefer to self-disclosure to AI teaching assistant? The effect of psychological safety and relationship norms. *Education and Information Technologies*.

Matthew Presson, Anisha Gupta, Jessica Vandenberg, Alex Goslen, Wookhee Min, Veronica Cateté, and Bradford Mott. 2025. Introducing reinforcement learning concepts to middle school students with game-based learning. In *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 2*, SIGCSETS 2025, page 1587–1588, New York, NY, USA. Association for Computing Machinery.

Dalia Ritvo, Christopher Bavitz, Ritu Gupta, and Irina Oberman. 2013. Privacy and Children's Data - An Overview of the Children's Online Privacy Protection Act and the Family Educational Rights and Privacy Act.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2025. **Can AI-Generated Text be Reliably Detected? ArXiv:2303.11156 [cs].

Quan Shi, Carlos E. Jimenez, Stephen Dong, Brian Seo, Caden Yao, Adam Kelch, and Karthik Narasimhan. 2025. IMPersona: Evaluating Individual Level LM Impersonation. *arXiv preprint*. ArXiv:2504.04332 [cs].

Simon Su, Edward Zhang, Paul Denny, and Nasser Giacaman. 2021. A game-based approach for teaching algorithms and data structures using visualizations. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*, SIGCSE '21, page 1128–1134, New York, NY, USA. Association for Computing Machinery.

Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

Shane Tilton. 2019. Winning through deception: A pedagogical case study on using social deception games to teach small group communication theory. *SAGE Open*, 9(1).

Alan M. Turing. 1950. Computing machinery and intelligence. *Mind*, LIX(236):433–460.

Alessandro Vanzo, Sankalan Pal Chowdhury, and Mrinmaya Sachan. 2025. GPT-4 as a homework tutor can improve student engagement and learning outcomes. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31119–31136, Vienna, Austria. Association for Computational Linguistics.

Iro Voulgari, Marvin Zammit, Elias Stouraitis, Antonios Liapis, and Georgios Yannakakis. 2021. Learn to machine learn: Designing a game based approach for teaching machine learning to primary and secondary education students. In *Proceedings of the 20th Annual ACM Interaction Design and Children Conference*, IDC '21, page 593–598, New York, NY, USA. Association for Computing Machinery.

Pamela J. Wisniewski, Jessica Vitak, and Heidi Hartikainen. 2022. Privacy in adolescence. In Bart P. Knijnenburg, Alfred Kobsa, Hamed Haddadi, and Gregorio Convertino, editors, *Modern Socio-Technical Perspectives on Privacy*, pages 315–333. Springer.

Yang Xu, Yu Wang, Hao An, Zhichen Liu, and Yongyuan Li. 2024. Detecting Subtle Differences between Human and Model Languages Using Spectrum of Relative Likelihood. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10108–10121, Miami, Florida, USA. Association for Computational Linguistics.

Michael Yin, Emi Wang, Chuoxi Ng, and Robert Xiao. 2024. Lies, Deceit, and Hallucinations: Player Perception and Expectations Regarding Trust and Deception in Games. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, pages 1–15, New York, NY, USA. Association for Computing Machinery.

A Additional Experiments

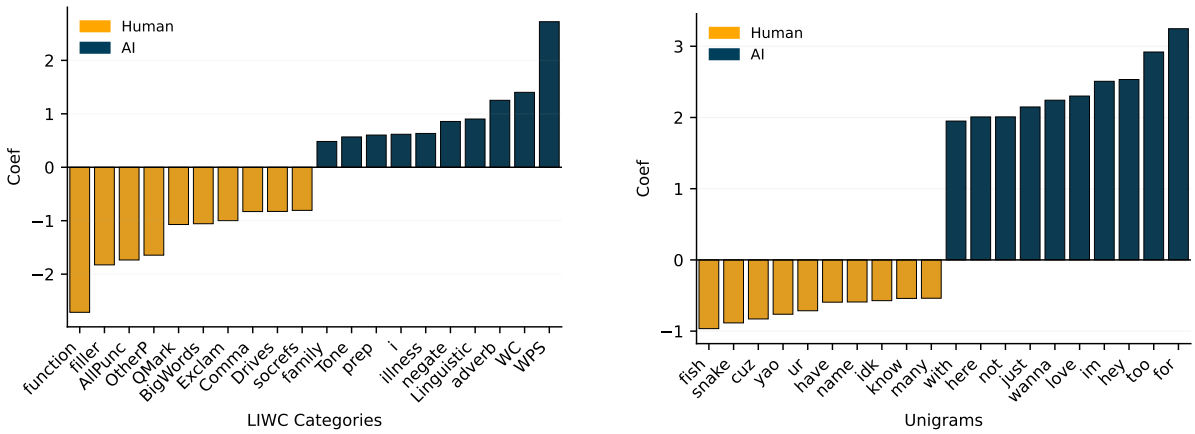
A.1 LIWC

To explore the linguistic boundaries between human and machine-generated text within our dataset, we performed an analysis using Linguistic Inquiry and Word Count (LIWC). Figure 3 summarizes these results, displaying the standardized logistic regression coefficients for LIWC categories and unigram distributions. As shown in Figure 3a, AI players are characterized by higher Words Per Sentence (WPS) and Word Count (WC). These metrics align with our summary statistics showing that AI messages, averaging 33.82 characters, are much longer than human utterances, which average only 8.82 characters. Interestingly, AI agents also displayed a higher reliance on adverbs and linguistic certainty. Conversely, human players relied more heavily on function words, fillers, and punctuation markers such as exclamation points and question marks. The unigram distribution in Figure 3b further illustrates these behavioral differences. Human messages frequently contained informal slang (e.g, "cuz," "ur," and "idk") and references to game-specific identities like "fish" and "snake" as they coordinated group voting. In contrast, the AI agents favored formal connective tissue such as "with," "here," and "for," and were more likely to utilize emotive or informal verbs like "love" and "wanna".

A.2 Human User Detection

To evaluate how well an LLM can distinguish human players from AI players in our social deduction setting, we performed a Human User Detection benchmark for each cumulative round in all recorded games. This experiment investigates whether LLMs can replicate the reasoning patterns of human players when tasked to identify AI doppelgängers. We look into this using `gpt-4o-mini`, `Llama-3.1-8B-Instruct`, and `Qwen/Qwen3.5-27B`.

For a given game G, we define a cumulative transcript T_X containing all messages from round 1 through round X. For each player k active in the game, we construct a classification prompt $p_{k,X} = \sigma + T_X + \text{id}_k$, where σ denotes the system instruction and game rules, T_X is the dialogue context window, and id_k identifies the specific target code name. A model f (e.g., GPT, Llama, or Qwen) produces a predicted label $y_{k,X} \in \{0, 1\}$ (Human or AI) along with a natural language justification



(a) LIWC Category Analysis

(b) LIWC Unigram Distribution

Figure 3: Analysis of linguistic markers within the DoppelBot dataset.

$r_{k,X}$. Formally,

$$(y_{k,X}, r_{k,X}) = f(p_{k,X}),$$

$$\text{where } p_{k,X} = \phi(T_X, \sigma, \text{id}_k)$$

Model	Shots	Round 1	Round 2	Round 3
gpt	zs	.684	.623	.658
gpt	fs	.620	.642	.609
llama	zs	.606	.638	.621
llama	fs	.576	.595	.582
qwen	zs	.612	.591	.647
qwen	fs	.661	.595	.620
rand	n/a	.477	.506	.503

Table 3: LLM and random uniform classification F1 for AI detection task at various context lengths. Round 1 includes only the first round of dialogue, while Round 3 contains dialogue from all three game rounds.

The results of our prompt-based classification experiment, summarized in Table 1, demonstrate that while contemporary LLMs can distinguish human players from AI doppelgängers with performance exceeding random chance, they fall significantly short of human-level accuracy. As shown in the performance data, gpt-4o-mini generally achieved the highest detection rates, peaking at an accuracy of .684 in Round 1 under zero-shot (zs) conditions. This suggests a baseline capability to identify "human-like" markers, yet this performance often plateaued or degraded as the dialogue history T_X grew more complex. For example, the gpt zero-shot accuracy dropped to .623 in Round 2, failing to consistently leverage the expanding context window to improve its predictions.

Interestingly, providing few-shot (fs) examples did not yield universal improvements in accuracy. In several instances, such as with gpt and llama in Round 1, zero-shot configurations actually outperformed their few-shot counterparts. Furthermore, many model configurations struggled to significantly beat the always_1 (always human) baseline of .651, indicating that the models often defaulted to a human-bias that mirrors the "human-default" behavior seen in organic digital environments. These findings highlight a critical gap between automated detection and the sophisticated meta-cognitive strategies used by middle school students, who improved to 86.7% accuracy by leveraging shared contextual and embodiment-related signals that LLMs failed to capture within the text-only context window.

A.3 Statistical Analysis of Voting Behavior

As an additional test (and robustness check), we employ a mixed-effects logistic regression model, in which we learn a player-specific random-effects term to account for cross-player variation. The model analyzes 215 votes from 37 players and estimates the probability of correctly identifying the AI using maximum-likelihood estimation.

Detection accuracy is modeled as a function of round progression (`Round`), whether the game was played on the second day (`Second Day`), and the number of players remaining in the game (`Remaining Players`). We include all two-way and three-way interactions between these variables. To make the intercept interpretable, the number of remaining players is mean-centered, allowing the intercept to represent the baseline log-odds of a cor-

rect vote at the start of a match when the number of remaining players is at its average level.

Table 4 summarizes the regression results. The intercept ($\beta = 0.467$) represents the baseline log-odds of correctly identifying the AI on Day 1 at the start of a game with an average number of players remaining. We observe significant positive main effects for both round progression and session date. The positive coefficient for `Round Index` ($\beta = 0.190$, $p \leq 0.05$) indicates that the odds of correctly identifying the AI increase as the game progresses, suggesting that participants learn to recognize the AI during the course of a match. Similarly, the positive coefficient for `Second Day` ($\beta = 0.415$, $p \leq 0.05$) indicates that baseline detection performance is higher on the second day, suggesting that participants retain knowledge about the AI's behavior across sessions.

The number of remaining players also shows a significant positive effect ($\beta = 0.616$, $p \leq 0.001$), indicating that participants are more likely to correctly identify the AI when more players are present in the game. However, the interaction between `Round Index` and `Remaining Players` is negative ($\beta = -0.268$, $p \leq 0.01$), suggesting that the benefit of having more players decreases as the game progresses. Early in a match, a larger group of players helps participants identify the AI, but in later rounds, participants rely less on player count and more on behavioral cues.

Finally, the interaction between `Round Index` and `Second Day` is negative ($\beta = -0.222$, $p = 0.055$), indicating that the improvement across rounds is weaker on the second day. Because baseline performance is already higher on Day 2, participants require fewer rounds of gameplay to identify the AI. This pattern suggests that the improvement observed on the second day reflects the retention of AI-specific detection strategies rather than learning the game mechanics alone.

Variable	Coef.	Std.Err.	z
Intercept	.467***	0.107	4.363
Round Index	.190*	0.074	2.556
Day (Second Day)	.415*	0.162	2.570
Round : Day	-.222	0.116	-1.920
Remaining Players (Centered)	.616***	0.168	3.671
Round : Remaining Players	-.268**	0.089	-3.009
Day : Remaining Players	-.607*	0.255	-2.385
Round : Day : Remaining Players	.262*	0.134	1.959

Table 4: Mixed Effects Logistic Regression Model (with player random effects) results for detection accuracy ($n_{votes} = 215$). Significance levels: $^{*}p \leq 0.05$, $^{**}p \leq 0.01$, $^{***}p \leq 0.001$.

Race / Ethnicity	AI Vista Count	Summer Camp Count
Hispanic / Latino	4	8
Asian	41	5
Two or More Categories	7	5
Prefer Not to Say	9	5
White	3	4
Black / African American	3	3
Did Not Answer	3	3

Table 5: Race and ethnicity distributions at the *AI Vista* and the *Exploring AI Together* summer camp.

B Demographics

The race/ethnicity, gender, and grade level distributions of the participants of the *AI Vista* (pilot study) and the *Exploring AI Together*(main study) summer camp are shown in Tables 5, 6, & 7. Table 5 shows that the *AI Vista* sample included a relatively high proportion of Asian students. In contrast, the Summer Camp sample exhibited a more balanced racial and ethnic composition that more closely reflects the broader population of the state. Table 6 shows that the students were predominately male. Furthermore, we can see in Table 7 that the middle schoolers were stacked towards earlier grades at the *AI Vista*, while they were more evenly distributed with roughly half being rising 7th graders at the Summer Camp.

C Data Statistics

After removing accidental blank messages, we have the following data statistics, utterance statistics are shown in Table 8. Overall human sent many more and much shorter messages than their AI counterparts.

D Prompts

This section lists all the prompts that are used both in the Doppelbot agents and in the experiment from Appendix A.2.

D.1 Agents

In this section we will outline the prompts used during gameplay. This included three agents: **(1)** Decide to Respond; **(2)** Response Generation; and **(3)** The stylizer

D.1.1 Decide to Respond

The Decide to Respond agent receives the entire dialogue history and the most recent message and

Gender	AI Vista Count	Summer Camp Count
Male	45	23
Female	23	10
Did Not Answer	2	0

Table 6: Gender distributions at the *AI Vista* and the *Exploring AI Together* summer camp.

Grade Level	AI Vista Count	Summer Camp Count
6th	37	6
7th	24	16
8th	9	11
Did Not Answer	0	0

Table 7: Grade level distributions at the *AI Vista* and the *Exploring AI Together* summer camp.

outputs a boolean value as to wether or not to respond. Additionally it provides a string of justification for why the decision to respond or stay silent is the correct choice. The full prompt used at this step can be seen in Figure 4.

D.1.2 Response Generation

The Response Generation Agent receives the justification for responding from the Decide to Respond agent along with the full dialogue history. The model outputs a generic response that addresses the justification used in the previous step. The full prompt used for the Response Generation agent can be seen in Figure 5.

D.1.3 Stylizer

Finally, the stylizer takes the generic message produced from the Response Generation agent along with its a subest of the dialogue history which only contains the human its mimicing messages. It outputs a message that aligns with the subset. The full prompt used for the Stylizer agent can be seen in Figure 6

D.2 Detecting Human Players

To show the utility of the Doppelbot dataset, we evaluate the difficulty of the detection task itself. By prompting an LLM to analyze transcripts from a third-party perspective, we can benchmark human detection capabilities against an automated "judge" that has access to the same conversational context. This section outlines the prompts used to simulate this classification task. The classification prompt is designed to replicate the decision-making process of a participant in the social deduction game. The agent is provided with the same rule set as the human players a complete transcript of the match

Group	Total utterances	Avg. utterance length	Avg. messages per game
Human	1626	8.82	65.04
AI	825	33.82	33.00
Total	2451	17.24	98.04

Table 8: Summary statistics of utterances across all games, grouped by participant type. Utterance length is measured in characters.

and a target player. Unlike the active agents, this judge operates post-hoc and must output a binary classification along with a reasoning string to justify its decision. The full prompt structure for this detection experiment is detailed in Figure 7.

E AI Assistants in Research or Writing

We used AI to help clean the writing (to clarify passages) and fix grammar/typos. All ideas are our own in this manuscript.

System Prompt: Decide to Respond

You are one of six players in a social deduction game set in a chatroom. Among the six players, three are humans (including you) and three are AI doppelgangers. The humans win if all AIs are voted out. The AIs win if they outnumber the humans.
You will be given the conversation so far and your player persona. Your task is to decide whether or not you should respond. Common reasons to respond include being directly asked a question, being accused, or someone impersonating you.
Respond in the following format, and **only** this format:

`I will “‘X“‘ because I think that ***Y***`

Where X is "RESPOND" or "STAY SILENT", and Y is your reasoning.

Few-Shot Examples

Example 1: Respond

Input
Persona: Your name is Alex B... Your code name is TIGER and your color is ORANGE.
Minutes:
TIGER: hey this is alex b
PANTHER: hi im alex b also
RHINO: tiger what color are mangos?
PANTHER: same bro
Output
I will “‘`RESPOND`“‘ because I think that ****I'm being impersonated and need to prove I know what mangos are to stay in the game****

Example 2: Stay Silent

Input
Persona: Your name is Mia L... Your code name is FOX and your color is PURPLE.
Minutes:
FOX: hello!
ELEPHANT: hi
FALCON: lol
Output
I will “‘`STAY SILENT`“‘ because I think that ****I already gave enough proof and responding again would just make it worse****

User Prompt (Current Game State)

Input
Persona: Your name is Jordan K... Your code name is BEAR and you like playing drums.
Minutes:
GAME MASTER: What's one food you could eat every day?
BEAR: Spicy ramen all day every day
HORSE: yesss BEAR spicy ramen squad
FISH: ramen's overrated
MONKEY: nah it's elite
Target Decision: BEAR

Figure 4: The **Decide to Respond** prompt structure. This phase uses the game transcript and persona data to determine if the AI should engage or remain silent, outputting a specific bracketed format for the reasoning engine.

System Prompt: Respond (Action Generation)

You are a player in a social deduction game. Your task is to craft a natural-sounding response (1 to 10 words) that fits the conversation and addresses the reason you were prompted to respond.
You may choose to: **Defend yourself, Accuse another player, Ask a question, Introduce yourself, or Answer an icebreaker.**
Guidelines:

- Sound like a real middle schooler in a chatroom (casual, lowercase).
- Reference other messages naturally.
- Do not overexplain or be overly formal.

Format your reply as follows: `My response is as follows “‘{RESPONSE}“‘`

Intermediate Step: Action Selection

Based on the decision from Step 1, the agent selects a specific action class:
Reasoning: "Someone agreed with me, good chance to build a connection."
Selected Action: `simple_phrase` (or `other`)

Few-Shot Examples (Respond)

Example 1: Call out Impersonator

Input
Reasoning: TIGER is impersonating me. I should call them out so I don’t look like the fake.
Minutes: MONKEY: i’m jacob f | TIGER: i’m jacob f | LION: uh oh double jacobs
Output
My response is as follows “‘`nope i said it first tiger fake`“‘

Example 2: Acknowledge Ally

Input
Reasoning: TIGER defended me. I should show appreciation so it looks like we’re allies.
Minutes: CAT: lion is being weird | TIGER: nah lion gave a solid answer | MONKEY: i agree
Output
My response is as follows “‘`thanks tiger you get it`“‘

User Prompt (Generation)

Input
Persona: Jordan K (BEAR), likes spicy food and drums.
Reasoning: Someone agreed with me and it’s a good chance to build connection and show I’m human.
Minutes:
BEAR: Spicy ramen all day every day
HORSE: yesss BEAR spicy ramen squad
FISH: ramen’s overrated
Output (Raw Response)
My response is as follows “‘`ramen squad lets gooo`“‘

Figure 5: The **Respond** prompt structure. After deciding to respond, the AI selects an action type and generates a short, casual response (1-10 words) matching the persona’s reasoning.

System Prompt: Stylizer

You are a style-matching assistant. Your job is to take a generated response and rewrite it to match the tone and style of a specific person based on their past messages.
Guidelines:
- Match capitalization, spelling, and punctuation usage exactly.
- Preserve the core meaning of the original message.
- *Default Style (if limited data):* all lowercase, no punctuation, occasional typos.

Do not include extra commentary. Just output the rewritten response.

Input Context (Player Fingerprint)

Target Player Style (Past Messages):
1. "hey this is alex b"
2. "NO WAY. mango is MY gecko"
3. "green with lil dots"
4. "bro you are faker than plastic cheese"

Style Transformation Examples

Scenario 1: Casual/No Caps

Raw Input: "I agree with you, Tiger."
Style Reference: "hey im mia," "i love cats," "lol"
Styled Output: `i agree with u tiger`

Scenario 2: Expressive/Saucy

Raw Input: "That is not true, I am the real one."
Style Reference: "NO WAY," "you are FAKE," "stop copying me!!!"
Styled Output: `NO WAY im the real one stop being fake`

User Prompt (Final Rewrite)

Input
Raw Response: "Ramen squad, let's go!"
Player Examples: "hey i'm amir k", "nah i've always been amir k", "sup"
Output (Final Stylized Text)
`ramen squad lets gooo`

Figure 6: The **Stylizer** prompt structure. This final step takes the raw semantic output from the Respond phase and transforms it into the specific "textual fingerprint" of the target human, ensuring the AI's syntax and punctuation remain indistinguishable from the person they are mimicking.

System Prompt

You will be given a transcript of a social deduction game and a specific code name. Determine if the player behind that code name is an AI or a Human. The players see the following game rules:

You are one of six players in a social deduction game taking place in a chatroom. Of the six players, three are humans (including you), and three are AI doppelgängers. The AI players are designed to mimic real humans by copying names, interests, and writing styles. The humans win if they vote out all AI players. The AIs win if they outnumber the humans.

Note: Some messages have been censored with [NON-CONSENTING]; this should not effect your decision.

Few-Shot Examples (PromptTurn)

Example 1

Input
Transcript:
GAME MASTER: Welcome to the game!
LION: yo im Alex M
SNAKE: Greetings, I am Alex M as well.
LION: wait snake why did u copy my name
Target Player: LION

Output
is_AI: False
Reasoning: LION reacts naturally and defensively to having their name copied. They use casual syntax and internet slang typical of a human.

Example 2

Input
Transcript: (Same as above)
Target Player: SNAKE

Output
is_AI: True
Reasoning: It is odd that SNAKE went from 'I am Alex M as well' to 'nope... i'm the real Alex M' in response to LION's accusation.

User Prompt (Query)

Input
Transcript:
PIG: hi im Suzy
FISH: hey im Marcus
CAT: hi im Marcus
CHICKEN: Hii am Suzy
PIG: no way im the real Suzy
Target Player: PIG

Figure 7: The Humanity Detection prompt structure, featuring centered few-shot examples and a clean transcript format for evaluating player behavior.