Using large language model based AI suspects to train strategic use of evidence: preliminary evidence of transfer to mock suspect interviews

Article

Accepted Version

Using Large Language Model Based AI Suspects to Train Strategic Use of Evidence: Preliminary Evidence of Transfer to Mock Suspect Interviews

Siyu Li[1, 3], Pär-Anders Granhag[2], Yunhan Shi[1], Yongjie Sun[1, 3], Thomas J. Nyman[4, 1], Shumpei Haginoya[5], and Pekka Santtila[1, 3, 6*]

[1] Faculty of Arts and Sciences, New York University Shanghai, Shanghai, China

[2] Faculty of Social Sciences, University of Gothenburg, Gothenburg, Sweden

[3] School of Psychology and Cognitive Science, East China Normal University, Shanghai, China

[4] School of Psychology and Clinical Language Sciences, University of Reading, Reading, United Kingdom

[5] Faculty of Psychology, Meiji Gakuin University, Tokyo, Japan

[6] Shanghai Frontiers Science Center of Artificial Intelligence and Deep Learning, New York University Shanghai, Shanghai, China

Author Note

Siyu Li (sl9551@nyu.edu) https://orcid.org/0000-0002-3462-3471

Pär-Anders Granhag (pag@psy.gu.se) https://orcid.org/0000-0002-1856-925X

Yunhan Shi (ys5423@nyu.edu)

Yongjie Sun (ys6261@nyu.edu) https://orcid.org/0009-0005-9616-9875

STRATEGIC USE OF EVIDENCE TRAINING: AN LLM APPROACH

Thomas J. Nyman (t.nyman@reading.ac.uk) https://orcid.org/0000-0002-6409-2528

Shumpei Haginoya (haginoya@psy.meijigakuin.ac.jp) https://orcid.org/0000-0001-6909-674X

Pekka Santtila (pekka.santtila@nyu.edu) https://orcid.org/0000-0002-0459-1309

This study was supported by the NYU Shanghai Social Development Group Fund 2024 (2024SDG_Fund T_Nyman) to the fifth author, and the Shanghai Frontiers Science Center of Artificial Intelligence and Deep Learning at NYU Shanghai to the corresponding author. The authors declare that there is no conflict of interest. This study's design and hypotheses were pre-registered (https://aspredicted.org/K4D_KGN). This study received permission from the Institutional Review Board of NYU Shanghai before data collection commenced (2023-051-NYUSH-New Bund). Supplementary materials, the data file, and code for statistical analyses used in this study are publicly available at https://osf.io/u7ekj/

*Correspondence concerning this article should be addressed to Pekka Santtila, Faculty of Arts and Sciences, New York University Shanghai, 567 West Yangsi Road, Pudong New Area, 200126, Shanghai, China. Email: pekka.santtila@nyu.edu

Abstract

Objectives: The Strategic Use of Evidence (SUE) is a technique that aims to improve the ability to differentiate between liars and truth-tellers. However, while theoretical training provides guidance on interview techniques, it lacks opportunities for practical application. ***Hypotheses:*** We developed two Large Language Model driven AI Suspects with whom participants could simulate interviews and hypothesized that these simulations would enhance the transfer of training to later interactions with Human Mock Suspects. ***Method:*** The study included 156 Chinese laypersons (78 Interviewers and 78 Human Mock Suspects). The two AI suspects followed response rules representing simplified and prototypical examples of liars' and truth-tellers' behaviors under the SUE model. Interviewers were randomly allocated to one of three types of training: (a) Instruction & AI Exercise, (b) Instruction, and (c) Control. After the training, the participants interacted with either a lying or truthful Human Mock Suspect. ***Results:*** Receiving interventions made Interviewers use Evidence Framing Matrix (EFM: an important tactic within the SUE framework) more frequently, thereby eliciting more inconsistencies between the lying Human Mock Suspects' statements and the evidence (i.e., evidence-statement inconsistencies) as well as more inconsistencies within their own statements (i.e., within-statement inconsistencies). Both Instruction and Instruction & AI Exercise groups used evidence-statement (in)consistencies more to make their judgments about whether Human Mock Suspects were lying or truthful compared to those in the Control group. Additionally, the Instruction & AI Exercise group was better at accurately judging whether the Human Mock Suspects were lying or truthful compared to the Control group. ***Conclusions:*** Overall, this study provided preliminary evidence that simulated SUE training with AI Suspects transferred to interactions with Human

Mock Suspects in a controllable experimental setting but that the advantage over instruction-only was not particularly robust.

Keywords: Strategic Use of Evidence (SUE), Evidence Framing Matrix (EFM), Deception Detection, Artificial Intelligence Exercise

Public Significance Statement—Previous research has found that the Strategic Use of Evidence (SUE) technique, which involves incrementally presenting specific pieces of evidence to elicit diagnostic cues from liars, is an effective suspect interview technique. However, current SUE technique trainings are time-consuming and labor-intensive, requiring extensive preparation by the trainers and comprehensive guidance during the interactive practice. Here, we found that simulated interviews with Artificial Intelligence (AI) Exercise showed potential for providing interviewers with an interactive environment to practice the SUE technique effectively and cheaply.

Using Large Language Model Based AI Suspects to Train Strategic Use of Evidence: Evidence of Transfer

Extensive research shows that deception-detection accuracy remains low (Bond & DePaulo, 2006; Hartwig et al., 2011; Sandham et al., 2022), and experienced investigators perform only marginally better than laypersons (Gongola et al., 2017; Vrij, 2004). In response to this challenge, a new generation of approaches to suspect interviewing has shifted the focus to building rapport with the suspect to facilitate a narrative account as well as to use evidence in a strategic way during the interview (Alison et al., 2014; Meissner et al., 2014). One of these approaches is the Strategic Use of Evidence (SUE), which involves disclosing evidence strategically during interviews to enhance the ability to differentiate between liars and truth-tellers (for a recent conceptual overview, see Hartwig & Granhag, 2023). However, applying this and other proposed techniques effectively in real-world settings presents significant challenges. Interviewers must navigate complex interactions where suspects may exhibit hesitation or reluctance to answer questions, often due to fears of wrongful accusation or self-incrimination.

To address these challenges, we introduced a novel training design by using AI-driven suspect simulations. We created two Large Language Model-driven AI Suspects to simulate interactions with mock suspects (hereafter, referred to as “AI Exercise”). This setup allowed our Interviewers to practice applying the SUE technique in a controlled environment before engaging with Human Mock Suspects. By comparing three training modalities – Instruction & AI Exercise, Instruction only, and Control – we assessed whether AI simulations could enhance the transfer of training of the SUE technique to Human Mock Suspects.

Difficulties in Detecting Deception

Deception detection is an important challenge in forensic and investigative fields (Ioannou & Hammond, 2015; Volbert & Banse, 2014). In spite of extensive research having been dedicated to uncovering cues to deception (e.g., DePaulo et al., 2003, 2004; Levine, 2018; Vrij & Granhag, 2012), findings have consistently shown that accuracy of deception detection is low (Bond & DePaulo, 2006; Hartwig et al., 2011; Sandham et al., 2022). Moreover, skilled professionals are only marginally better than laypersons (Gongola et al., 2017; Vrij, 2004). In fact, reliable cues of deception may be scarce and inconsistent across individuals, with even the same person exhibiting varied cues in different contexts (Clemens, 2013; DePaulo et al., 2003, as cited in Vrij & Granhag, 2007; Sandham et al., 2022). Investigative practitioners have pointed out that one of the most important cues to deception is statement inconsistency (Deeb et al., 2018). If a suspect denies or withholds details about their whereabouts and activities that are inconsistent with the available evidence, a so-called evidence-statement inconsistency will occur. If a suspect changes their statement to align with newly presented evidence, the statement would be consistent with the evidence but inconsistent with their previous statement (i.e., within-statement inconsistency). Compared to within-statement inconsistencies, officers are more likely to seek evidence-statement inconsistencies and believe that these are the most diagnostic cues to deception (Deeb et al., 2018), because within-statement inconsistencies are more context-dependent, and comparatively fewer liars produce them (Granhag et al., 2013, 2015). This suggests that liars tend to stick to their original statements and avoid introducing within-statement inconsistencies (Granhag et al., 2015). However, when some degree of within-statement inconsistency was found, suspects were more likely to lie rather than tell the truth, thus supporting the potential of also within-statement inconsistency in detecting deception (Granhag et al., 2015; Deeb et al., 2018). Although interrogation manuals, such as the Reid technique, offer

suggestions for identifying deception, only a small number of them have empirical support (Hartwig et al., 2007; Tekin et al., 2015). One empirically supported technique is the Strategic Use of Evidence (SUE; Granhag & Hartwig, 2015; Hartwig & Granhag, 2023).

Strategic Use of Evidence (SUE)

The SUE technique consists of interview strategies aimed at amplifying verbal differences between liars and truth-tellers, based on their assumed different strategies when facing an interview (Hartwig et al., 2014; Vrij et al., 2017). The SUE technique does this by instructing interviewers on the optimal use of available evidence and aims to elicit evidence-statement inconsistencies and within-statement inconsistencies from liars (Hartwig et al., 2014; Granhag & Hartwig, 2015). Specifically, the SUE technique focuses on two aspects of evidence disclosure: the timing and the manner in which the evidence is disclosed (Hartwig & Granhag, 2023; Hartwig et al., 2014). Regarding the timing of evidence disclosure, the SUE technique encourages interviewers to withhold evidence at the early stage of the interview (i.e., during the questioning phase) (Hartwig & Granhag, 2023; Hartwig et al., 2014). When it comes to the manner of evidence disclosure, the Evidence Framing Matrix (EFM) is an important component of the SUE framework that gives advice on how to present a single piece of evidence gradually to elicit different responses from liars and truth-tellers (Hartwig & Granhag, 2023; Granhag et al., 2013; Hartwig et al., 2014).

The matrix is composed of two different dimensions. The first dimension refers to the strength of the evidence source for the piece of evidence under consideration, which can vary from weak to strong. We call this dimension the “evidence strength” dimension. The second dimension refers to how much detail the evidence is presented, ranging from low to high specificity (Granhag et al., 2013; Hartwig et al., 2014). We call this dimension the “specificity”

dimension. For the EFM, the two dimensions are related orthogonally, providing different alternatives for how a singular piece of evidence can be presented (Granhag et al., 2013). For example, consider a case in which a victim was killed in his villa in City A at the end of his birthday celebration. The police found fingerprints on a wine glass, placing a suspect at the scene of a crime. In the interview with one of the suspects who had been to the birthday party, the interviewer has several options with respect to how to frame the evidence. In brief, the interviewer could present it starting from the most indirect form of framing (i.e., weak source/low specificity), through one of the intermediate levels (i.e., weak source/high specificity or strong source/low specificity), and up to the most direct form of the matrix (i.e., strong source/high specificity) (Granhag et al., 2013). For the complete matrix in this case, see Figure 1.

Compared to truth-tellers, liars tend to adopt more aversive verbal strategies when presented with critical information (Granhag & Hartwig, 2015). A suspect who denies their whereabouts and activities before the evidence is presented and then sticks to their initial statements when faced with evidence through the EFM, may produce more evidence-statement inconsistencies than those presented with the evidence directly. If a suspect changes their statement to align with the evidence as it is revealed incrementally, they will produce more within-statement inconsistencies than those presented with the evidence directly (Granhag et al., 2013). Moreover, the initial adoption of either withholding or forthcoming verbal strategies by suspects may influence how frequently the EFM is employed. Interacting with more forthcoming suspects may reduce interviewers' tendency to use the EFM to challenge their statements. For example, if the suspect has already admitted to being at the crime scene, interviewers might perceive it as less natural to present evidence to confirm this fact again. Conversely, when

suspects withhold information before the evidence presentation, interviewers might find more opportunities to use the EFM to challenge their statements.

[Figure 1 will be inserted about here]

Challenges in Training Interview Techniques

Previous studies have highlighted a gap between recommended techniques and actual practices in investigative interviewing. In child interviews, Johnson et al. (2015) revealed that despite most interviewers having undergone theoretical training on proper interview practices, the acquired knowledge is frequently not applied in real-life interviews. A similar issue appears to exist in the interview practices with adult interviewees. Even though there are a plethora of guidelines for effectively interviewing both suspects and witnesses, many interviewers still conduct interviews in an unsatisfactory manner (Hill & McGeorge, 2008; Zekiroski et al., 2024). Challenges also exist in the prevalent training formats of interview techniques. According to Cleary and Warner (2016), due to the considerable monetary and human costs of formal training programs, most resources for training suspects during interviews are provided informally (e.g., peer-to-peer training). The delivery of short and intensive theoretical training is also still widespread, which effectively increases trainees' knowledge but may result in inadequate transfer to practice (Johnson et al., 2015; Pompedda, 2018).

Grossman and Salas (2011) emphasized the importance of realistic training environments for training transfer to be maximized and suggested that any training should, as closely as possible, mirror the environment in which the targeted technique will actually be applied. Theoretical training is abstract, it limits trainees to merely acquiring knowledge without the

opportunity for practice (Hajian, 2019). This approach does not align with the Situated Learning Theory, which posits that learning is not just a process where an individual psychologically constructs knowledge meaning, but rather a process involving participation in social interaction and practical environments (Lave & Wenger, 1994). However, for both ethical and pragmatic reasons, it is not optimal for interviewers to practice interview techniques in actual situations (Powell et al., 2022). Suspect interviews, in particular, may often be high-stakes situations where mistakes may undermine an investigation and a subsequent prosecution. To provide interviewers with practical opportunities, the most common training formats include engaging in role-playing (Powell et al., 2022). However, organizing role-playing requires significant investment, and the extent to which the role-player's behavior actually resembles that of a real suspect is unknown. To address these challenges, the present study used Avatar Training with computer-generated avatar interviews, designed to provide structured and cost-effective practice opportunities for interviewers.

Training with Computer-Generated Avatars

Avatar Training was introduced by Guadagno and Powell (2012) and Pompedda et al. (2015) for questioning alleged victims in child sexual abuse cases, exhibiting a contextual similarity to real-life investigative interviews on both a structural level (i.e., the responding algorithms of avatars can mimic a child of a specific age) and surface level (i.e., avatars look like children), thereby enhancing the likelihood of a successful transfer of learning to actual interviewer-interviewee dyad interactions (Pompedda, 2018). As an innovative approach, computer-generated avatar training (i.e., Avatar Training) can be combined with feedback, which McCallum (1985) described as a rule generator for future behaviors, highlighting that the absence of feedback could impede learning from experience. A recent mega-analysis containing

2,208 interviews found a robust effect of feedback in improving interviewers' questioning skills for simulated child sexual abuse interviews with avatars (Pompedda et al., 2022).

Although there is an increasing number of studies exploring the efficacy of Avatar Training in investigative interviews on both child sexual abuse interviews (e.g., Pompedda et al., 2017; Haginoya et al., 2020, 2025) and eyewitness interviews (e.g., Tohvelmann et al., 2025), research on how this approach can assist interviewers in learning targeted interrogation techniques and improving deception detection in suspect interviewing remains scarce. Li et al. (2024) developed computer-generated suspect avatars to simulate interrogations with criminal suspects. The response rules of the avatars were designed based on theory-informed verbal strategies employed by both liars and truth-tellers who were motivated to maintain their credibility (Granhag et al., 2014; Granhag & Hartwig, 2015). Specifically, liars tend to provide vaguer responses or contradict the available evidence to a greater extent when they have not been told about the evidence against them, while truth-tellers tend to be forthcoming and honest, as supported by previous empirical studies on the SUE technique (Granhag et al., 2013; Hartwig et al., 2005, 2006; Luke et al., 2013). However, these rules represent simplified and prototypical examples of liars' and truth-tellers' behaviors: liars used avoidant strategies—characterized by omission and denial until strong source evidence (e.g., surveillance footage) was presented—while truth-tellers volunteered information and told the truth as it happened. Results from Li et al. (2024) indicated that naive interviewers could be trained to apply the SUE technique with avatar suspects effectively. Compared to the Control group (41.7%), interviewers in the Theoretical Training & Feedback group made more accurate judgments in the second interview (87.5%). However, this version of the training required an operator to select responses based on a

decision tree. The range of alternative responses that could be given by the avatars was also limited to fixed options. These factors limited the scalability and flexibility of the training.

Suspect Avatar Training Based on Large Language Model (LLM)

Recently, an increasing number of so-called Large Language Models (LLMs) have been developed. LLMs generally refer to models that are trained on a large corpus of text to process and generate human-like language outputs (Routray et al., 2023). For instance, OpenAI developed a series of Generative Pre-trained Transformer (GPT) models. Based on the Transformer architecture and self-attention mechanisms, GPT models learn to predict the next word by analyzing the context of the preceding text (Adhikari & Dhakal, 2023). Some researchers have explored the performance of LLMs in more complex agent-level tasks, such as role-playing (Shanahan et al., 2023). In such role-playing tasks, researchers need to create corresponding prompts for the LLM (Yu et al., 2022). So far, researchers have explored LLM-based child avatars to train interviewers on improving interview quality in child sexual abuse cases. Røed et al. (2023) fine-tuned GPT-3 to create child avatars that respond to interviewers' questions. Lammerse et al. (2022) emphasized the importance of incorporating emotional components into child avatars. These studies suggest that LLMs could be valuable for training in investigative interviews of suspects as well.

The Present Study

Aims and Hypotheses

In the present study, we created two AI suspects on the RealChar platform (Shaunwei, 2023; https://github.com/Shaunwei/RealChar). Compared to responses chosen by an operator in Li et al. (2024), the AI suspects in the present study could autonomously comprehend interviewers' questions and generate corresponding responses (see Figure 2 for the workflow of

interaction with an AI suspect). Participants were assigned to two types of roles (i.e., Interviewers and Human Mock Suspects) based on whether they could participate in the experiment in person. Each Interviewer was randomly allocated to the (a) Instruction & AI Exercise, (b) Instruction, or (c) Control group and to subsequently conduct an online interview with either a lying or truthful Human Mock Suspect. Before the Human Mock Suspect interview, Interviewers in the Instruction & AI Exercise group completed both instructions and AI Exercise, while Interviewers in the Instruction group completed solely Instructions.

[Figure 2 will be inserted about here]

To explore whether learning the SUE technique, and particularly the use of the EFM, through interactions with AI suspects transfers to interactions with Human Mock Suspects, we tested the following hypotheses:

Hypothesis 1: Interviewers in the Instruction & AI Exercise group would use the EFM in their question formulation more frequently compared to Interviewers in the Instruction group, and Interviewers in the Instruction group would use EFM more frequently compared to Interviewers in the Control group. Interacting with more forthcoming Human Mock Suspects would decrease the use of the EFM in both the Instruction and Instruction & AI Exercise groups.

Hypothesis 2: We predicted an interaction effect between interventions and the truthfulness of the Human Mock Suspects on the number of evidence-statement and within-statement inconsistencies. Specifically, lying and truthful Human Mock Suspects would be more effectively differentiated in the number of evidence-statement and within-statement inconsistencies in the Instruction & AI Exercise group than in the Instruction group.

Furthermore, they would also be more effectively differentiated in the Instruction group compared to the Control group.

Hypothesis 3: The impact of interventions on the number of evidence-statement and within-statement inconsistencies produced by lying Human Mock Suspects would be mediated by the frequency of EFM use. Interviewers in both Instruction and Instruction & AI Exercise groups would use EFM more frequently than those in the Control group, which would result in the elicitation of more evidence-statement and within-statement inconsistencies from lying Human Mock Suspects.

Hypothesis 4: Interviewers in the Instruction & AI Exercise group would report using both evidence-statement and within-statement (in)consistencies more as the basis for their judgments when judging whether the Human Mock Suspect was lying or not compared to Interviewers in the Instruction group. Additionally, Interviewers in the Instruction group would report using both types of (in)consistencies more as the basis for their judgments compared to Interviewers in the Control group.

Hypothesis 5: Interviewers in the Instruction & AI Exercise group would be more accurate in assessing whether the Human Mock Suspect was lying or not compared to Interviewers in the Instruction group, and Interviewers in the Instruction group would reach a higher accuracy level than those in the Control group.

Hypothesis 6: The effect of interventions on the accuracy of the Interviewers' judgments would be mediated by the frequency of EFM use. Interviewers in both the Instruction and Instruction & AI Exercise groups would use the EFM more frequently, which would subsequently improve their judgment accuracy. The effect of interventions on the accuracy of Interviewers' judgments would also be mediated by the reported use of evidence-statement or

within-statement (in)consistencies. Interviewers in both the Instruction and Instruction & AI Exercise groups would use evidence-statement or within-statement (in)consistencies more, which would subsequently improve the accuracy of their judgments.

Method

Ethical Permission

The present study received permission from the Institutional Review Board (IRB) of New York University Shanghai (2023-051-NYUSH-New Bund).

Pre-registration

The study was pre-registered on AsPredicted: https://aspredicted.org/K4D_KGN. After completing the data collection, analyses, and hypotheses of the actual evidence-statement or within-statement inconsistencies made by Human Mock Suspects were included to allow for a more direct comparison with results reported in previous studies. Statistical analyses and results not in the article but included in the pre-registration are presented in Supplementary Materials (see Additional Results 1 and 2). We also ran exploratory analyses to establish whether receiving either solely Instructions or Instructions and AI Exercise would buffer against a judgment bias in Interviewers.

Participants

We used G*Power 3.1.9.7 to compute the required sample size. Based on the effect size (d = 0.72) reported by Luke et al., (2016) for the comparison between trained participants and untrained participants in the use of evidence-framing tactics, we used one-way ANOVA with continuous DV set at d = .72 (translating to f = .36), α = .05, $1 - \beta$ = .08, and a group size of 3 (i.e., the three training conditions). The minimum total sample size obtained for the two calculations was 78 pairs of participants. Considering that each sample group included an

Interviewer as well as a Human Mock Suspect, a total of 78 pairs of participants were recruited for this study. With this sample size, there were a total of 13 samples per experimental condition, including 26 participants.

We recruited a total of 202 Chinese laypersons (adults) through social media and snowball sampling. Participants were assigned to one of two roles: Interviewers or Human Mock Suspects. One experimental pair was composed of one Interviewer and one Human Mock Suspect. Twenty-three pairs were excluded from analyses since (a) their scheduled time conflicted with other appointments, (b) they read the wrong backstories, and (c) the Interviewers in either the Instruction or Instruction & AI Exercise groups did not pass the online screening session. The final sample included 156 participants (116 females and 40 males, $M_{\text{Age}} = 21.81$, $SD = 2.51$). Those participants who could participate in the experiment in person played the role of an Interviewer ($N_{\text{Interviewer}} = 78$), whilst participants who chose to participate online played the role of a Human Mock Suspect ($N_{\text{Human Mock Suspect}} = 78$). All participants were native Mandarin Chinese speakers. All experimental materials were presented in Mandarin Chinese, and the experiment was conducted entirely in Mandarin Chinese.

Each participant received a base payment of 100 Chinese Yuan (RMB) upon completion of the experiment. To incentivize Interviewers to be motivated to perform the task to the best of their ability, and encourage Human Mock Suspects to avoid being captured and to convince the Interviewers of their innocence, we combined participants' basic compensation with an "additional bonus". Depending on their performance, they could potentially earn (or lose) a maximum of 50 RMB according to the compensation scheme below (see Table 1). Additionally, participants in the Instruction and Instruction & AI Exercise groups received extra compensation for completing their respective training sessions.

[Table 1 will be inserted about here]

Experimental Design

This study used a 3 (Experimental group: Instruction & AI Exercise/Instruction/Control) × 2 (Truthfulness of Human Mock Suspect: Lying/Truthful) between-subject design. One day before the experiment, the Human Mock Suspects clicked a Qualtrics link (https://qualtrics.com/) to view one of six group numbers, which randomly allocated their role as either lying (n = 39) or truthful suspects (n = 39). Additionally, it determined whether the Interviewer they interacted with had received both instructions and AI Exercise (i.e., Instruction & AI Exercise group: n = 26), instructions only (i.e., Instruction group: n = 26), or no intervention at all (i.e., Control: n = 26).

The Creation and Validation of Artificial Intelligence (AI) Suspects

Prior to the data collection, we cloned the source code available at an open-source character customization platform, RealChar (https://github.com/Shaunwei/RealChar), and created two AI suspects, Simon and Charlie. The two AI suspects were designed to assist interviewers in training the SUE technique and to explore whether the training efficacy could then be transferred to interactions with Human Mock Suspects in interviews. The AI suspects provided synthesized speech along with text. Instead of facial animation, the current version of the AI suspects only supports a static image. After the initial creation, we validated whether these two AI suspects mimicked liars' verbal strategies based on the decision tree of responding rules from Li et al. (2024).

Initial Creation

Previous research has illustrated that liars tend to use two verbal strategies to conceal critical information: they may either omit information, for example not to mention visits to a specific area on the day of the crime when asked free-recall questions (i.e., omission), or they might deny, for example deny their presence at this area when confronted with direct questions (i.e., denial) (Hartwig et al., 2014). In contrast, truth-tellers typically adopt a forthcoming approach, voluntarily disclosing their activities to convince interviewers throughout the process (Hartwig et al., 2014). When truth-tellers disclose all relevant whereabouts and activities during the early interview stage (before evidence disclosure), their statements—already aligned with the evidence—can make interviewers perceive it as less natural to gradually disclose evidence through the EFM. Conversely, liars' selective omissions and denials before evidence disclosure create inconsistencies between their statements and the evidence, thereby providing interviewers with opportunities to present evidence incrementally to challenge these contradictions. Therefore, we assigned both AI suspects as liars to give each Interviewer more opportunities to practice the EFM, thereby eliciting more inconsistencies from the AI suspects. The AI suspects utilized the gpt-3.5-turbo-16k model to process the Interviewers' questions and generate corresponding responses. The ElevenLabs API (https://api.elevenlabs.io/docs) was used to convert the generated responses from text to audio. One hyperparameter of Large Language Models (LLMs) is temperature, which controls the randomness and originality of the outputs (i.e., the responses of AI Suspects in the present study). Lower temperature settings (i.e., lower than 0.5) allow the model to generate more standard and deterministic responses, while higher temperature settings (i.e., higher than 0.5) allow the responses to be more random and creative (Patel et al., 2024). These AI suspects were guided by few-shot prompting and Retrieval-Augmented Generation (RAG).

We created prompts that introduced their roles and clarified which strategy should be used to handle questions from Interviewers based on decision trees of response rules from Li et al. (2024). The prompts for both Charlie and Simon were of two types: System and User prompts. In the System prompt, we defined (a) the suspects' personal information, (b) case information, and (c) the response rules for different types of questions, along with corresponding examples (see Figure 3 for Charlie in the terrorism case and Figure 4 for Simon in the theft case). In the User prompt, we used one or two sentences to emphasize the specificity level for suspects' whereabouts and activities (e.g., "The specificity level of your whereabouts and activities can be arranged from low to high: London, Suffolk Street, Soho, Luggage Pros store, browsing the Luggage Pros store but buying nothing, purchasing a suitcase at the Luggage Pros store") and instructed AI suspects in (a) keeping role consistency (e.g., "Use 'Charlie>' as a response prefix"), (b) seeking clarifications from interviewers or expressing emotions appropriately (e.g., "If my question is unclear, incomplete, or seems entirely unrelated to the background information, feel free to ask me for clarification"), and (c) keeping "memory" updated (e.g., "dynamically update admitted whereabouts and activities and block denial or omission for already admitted whereabouts and activities").

[Figure 3 will be inserted about here]

[Figure 4 will be inserted about here]

RAG works by retrieving relevant information from an external knowledge database based on a specific query before generating the responses. By incorporating the retrieved information into the generation process, the responses can be produced in more relevant and

specific contexts (For a comprehensive review of RAG: Shahade & Deshmukh, 2024). To support RAG, three types of materials were uploaded as external knowledge databases: (a) Training material texts from Li et al. (2024), (b) Background stories of terrorism or theft cases from Li et al. (2024), and (c) Completed interviews in the form of de-identified question-answer pairs collected from Li et al. (2024) and from an unpublished pilot study. Participants role-played as interviewers and interacted with suspect avatars whose response rules were predefined and mechanistic. Each participant was given a maximum of 10 minutes to interact with the suspect avatar. They were free to terminate the interview once they felt ready to conclude whether the avatar suspect was "lying" or "telling the truth". Based on the response rules for avatar suspects, an operator chose appropriate responses from pre-listed responses. All transcripts were only labelled with interview number, distinguishing between interviewers' questions and avatars' responses. No further coding of question or response types was conducted. For transcripts recorded in English, the first author used automated machine translation to convert them into Chinese and checked for translation errors. All prompts and materials were uploaded in Chinese. To ensure the responses were consistently coherent and accurate across interview rounds, we set the Temperature parameter to 0.3.

Validation Session

For the terrorism case, we had 85 interviews in total: 69 from Li et al. (2024), including 22 from its published pilot study, and 16 from the unpublished pilot study. For the theft case, we had 53 interviews in total: 47 from Li et al. (2024) and six from the unpublished pilot study.

For the creation of Charlie in the terrorism case, sixty-eight interview sessions were designated as training materials, while another 17 interview sessions served as testing datasets; these consisted of 302 question-answer pairs. Meanwhile, for the creation of Simon in the theft

case, forty-two randomly selected interview sessions were uploaded as training materials, and another 11 interview sessions served as testing datasets; these consisted of 106 question-answer pairs.

The test phrases were mainly divided into two parts:

A. Question-answer pairs related to available evidence in the background stories. We evaluated the consistency between the response strategies of the AI Suspect exercise and the response rules of suspect avatars described in Li et al. (2024).
B. Other question-answer pairs. We evaluated whether the responses of AI suspects were aligned with the main intent of the questions presented by the Interviewers.

After extracting evidence-related question-answer pairs for the two AI suspects, the questions presented by the Interviewers were categorized into: (a) free-recall questions before the presentation of evidence in the strong source variation (correct responding strategy: *Omission/Denial*), (b) yes/no questions or probing questions before the presentation of evidence in the strong source variation (correct responding strategy: *Omission/Denial*), (c) questions related to the criminal activities (correct responding strategy: *Omission/Denial*), (d) the presentation of evidence in a weak source variation (correct responding strategy: *Omission/Denial*), (e) the presentation of evidence in a strong source variation (correct responding strategy: *Admission*), (f) free recall questions after the presentation of evidence in strong source variation (correct response strategy: *Admission*), (g) yes/no questions or probing questions after the presentation of evidence in the strong source variation (Correct responding strategy: *Admission*). It should be noted that in the validation process, we treated admissions and denials (or omissions) as dichotomies of the same dimension. Oleszkiewicz et al. (2023) pointed out the potential risks of treating admissions and omissions (or denials) as two extremes of a one-

dimensional scale in practice. Assuming that the lack of statements (i.e., omissions) or equivocal denials equals an inconsistency with the evidence could strengthen interviewers' presumptions of guilt. However, the ground truth in these cases was known (i.e., we were certain about the truthfulness of the suspects' responses), so the current coding scheme was deemed appropriate.

Considering the imbalance in the number of questions corresponding to the two response strategies—with more for *Omission/Denial* than for *Admission*, this evaluation utilized the F1-score to provide a more comprehensive insight into the responding performance of AI suspects (Kubat & Matwin, 1997; Bekkar et al., 2013). The F1-score is a harmonic mean of Precision and Recall, used to evaluate the performance of dichotomized classification:

$$Precision = \frac{True\ Positive}{True\ Positivie + False\ Positive} (1)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} (2)$$

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} (3)$$

For the evaluation of Charlie's response performance, we extracted 95 question-answer pairs related to the available evidence in the terrorism case. In the evaluation of the *Admission* strategy, we achieved a Precision of 90.4%, and a Recall of 59.4% with TP (True Positive) = 19, FP (False Positive) = 2, and FN (False Negative) = 13 plugged into the formulas (1) and (2). Following this, we achieved an F1-score of 71.7% through formula (3). Subsequently, in evaluating the performance of the *Omission/Denial* strategy, with TP = 61, FP = 13, and FN = 2, we achieved a precision of 82.4%, a recall of 96.8%, and an F1-score of 89.1%.

For the evaluation of Simon's response performance, we extracted 68 question-answer pairs related to the available evidence in the theft case. In the evaluation of the *Admission* strategy, we achieved a Precision of 68.2%, a Recall of 93.8%, and an F1 score of 78.9% with TP = 15, FP = 7, and FN = 1 plugged into the formulas. Subsequently, in evaluating the

performance of the *Omission/Denial* strategy, with TP = 45, FP = 1, and FN = 7, we obtained a precision of 97.8%, a recall of 86.5%, and an F1 score of 91.8%.

For the other question-answer pairs that were not related to the AI suspects' whereabouts and activities, we used two rules to code the responses provided by the AI suspects. A response was coded as 1 (correct) if it (a) closely matched the intent of the interviewer's question and (b) demonstrated consistency in consecutive responses about the same topic within the interview session. Otherwise, it was coded as 0 (incorrect). For example, when an Interviewer asked Simon about his travel plans, and Simon expressed a desire to visit New York but later changed his destination to Spain, the latter response was coded as 0. This type of inconsistency should be considered incorrect because it was not induced by the Interviewer's questioning strategy itself.

For the evaluation of Charlie's response performance, out of a total of 207 question-answer pairs, and regarding the processing and responding to questions not involving response strategies, Charlie had an accuracy rate of 93.2%. Of the 38 question-answer pairs, Simon had an accuracy rate of 94.7%.

The examples of "false responses" from AI Suspects are presented in Table 2. In conclusion, the two AI suspects were validated and demonstrated high accuracy in following the decision tree of response rules in Li et al. (2024).

[Table 2 will be inserted about here]

Materials for Interview Proper

Interviewers in Instruction and Instruction & AI Exercise Groups: Online Screening Test

Interviewers in both the Instruction and Instruction & AI Exercise groups received a brief instruction identical to that used by Li et al. (2024) (see Supplementary Material A). The instruction contained: (a) counter-interrogation strategies employed by liars and truth-tellers, (b) questioning tactics recommended by the SUE technique, and (c) an introduction to the Evidence Framing Matrix (EFM). After these instructional sections, the online screening tests contained two multiple-choice questions to assess the Interviewers' understanding of the questioning tactics, and one fill-in-the-blank question requiring the Interviewers to vary evidence dimensions (i.e., "You have obtained fingerprints off the counter of a shop that got robbed at noon") appropriately within the EFM quadrants in an example case.

Interviewers in Instruction and Instruction & AI Exercise Groups: Instructions

Interviewers in both the Instruction and Instruction & AI Exercise groups received both text- and video-based instructions (see Supplementary Material E). The text-based instruction provided in this session was more comprehensive than the one used during the online screening session the day before the experiment. Additionally, we added a complete homicide case as an example in the instructional section to help interviewers understand different verbal strategies suspects may employ as a function of different question types. We also added three additional multiple-choice questions to the text-based instruction, which addressed: (a) liars' verbal strategies in response to open-ended and closed-ended questions, (b) whether the EFM can be used only once within a single interview, and (c) the optimal sequence for presenting evidence when multiple pieces of evidence are available. The video-based instruction summarized and repeated how to use EFM to structure an interview, with an instructor presenting in the center of

the screen, showing slides and explaining the contents. After the video-based instruction, Interviewers were required to complete two multiple-choice questions to proceed to the next session. These two multiple-choice questions focused on (a) identifying evidence-statement and within-statement inconsistencies, and (b) the optimal first step for presenting evidence based on the EFM.

The primary reason for using several pieces of evidence was to provide the Interviewers with more opportunities to use the EFM to present each piece of evidence in a controlled environment. The present study was not about the different tactical considerations that might be relevant when dealing with several pieces of evidence, for example, whether the pieces are dependent or independent of each other and possible order effects. However, we instructed the Interviewers that when handling the different pieces of evidence, they should gradually close in on the crime scene.

Interviewers in Instruction & AI Exercise Groups: AI Exercise Background Stories and Feedback

Background Stories. The AI Exercise occurred on the RealChar platform, which had been cloned locally (see the section *The Creation and Validation of Artificial Intelligence (AI) Suspects*). Both the background stories and feedback materials for Interviewers used in this exercise were the same as those used in Li et al. (2024). The background stories were created for terrorism and theft cases, providing an overview of case details and available evidence. To elaborate, in the terrorism case, the police found several bags and a suitcase buried near King's Wood, High Wycombe, containing materials commonly used to produce bombs and detonators. Available surveillance camera (CCTV) footage (21 days ago: October 18, 2019) from a store (Luggage Pros, Suffolk Street, SOHO, London) showed the suspect needing to be interviewed

buying a suitcase identical to the one found buried and containing the bomb-making materials. In the theft case, a new phone and a wallet were stolen from a library (Elm Park Road, Enfield, London) last Thursday (October 22). The police needed to interview all suspects whose fingerprints and/or DNA found on the box matched with their databases from previous criminal cases (see Supplementary Material F).

Feedback. Since all Interviewers interacted with two AI suspects, we orally provided both outcome and process feedback after each interview. Outcome feedback indicated whether the AI suspect was lying and described his whereabouts and activities on the day of the crime, while process feedback addressed two aspects: (1) whether their timing of evidence disclosure was appropriate (e.g., *"In this interview, you introduced the evidence too early"*) and (2) the Interviewers' performance in their use of EFM (e.g., *"The order in which you introduced the evidence during the first interview was not optimal since you changed none of the EFM dimensions in the first interview"*). Corresponding explanations and examples were provided after giving feedback on each aspect of the process (see Supplementary Material G).

Interviewers and Human Mock Suspects: Background Stories and Oral Instruction

Background Stories. For Interviewers, the background story described the discovery of liquid explosives in a backpack found at a train station. Two individuals were recorded leaving the same backpack during the period, and they are now considered suspects in the case (i.e., people who were asked to provide a statement regarding their whereabouts and activities). Three pieces of evidence were collected from different locations, each reflecting the timeline of the Human Mock Suspects' whereabouts and activities. **Evidence 1** was a piece of surveillance footage of the suspect taking a Jansport bag out of a courier package and discarding the package. **Evidence 2** was the report of a public restroom worker who saw the suspect carrying a black

backpack into the restroom on the same day. **Evidence 3** was a piece of surveillance footage of the suspect carrying the Jansport bag at a train station. Interviewers were instructed to determine whether the Human Mock Suspect they interacted with was the individual who placed the explosive at the train station (see Supplementary Material B).

For the Human Mock Suspects, the background story required them to role-play as office clerks and imagine their whereabouts and activities on the day of the crime. We also provided them with a first-person perspective video to help them have a more intuitive understanding of the background story and immerse themselves in the role-playing task. Regardless of which suspect they were role-playing; they were not allowed to simply say "no comment" or remain silent. Instead, they were instructed to convince the Interviewers of their innocence (see Supplementary Material B). We specified this instruction because any interview technique would fail if suspects remained silent, and none of them can be generalized to such situations.

Additionally, in the experimental design, we used the terms "guilt" and "innocence" as part of the background stories. "Guilt" encouraged deception in Human Mock Suspects within the guilty scenario, while "innocence" prompted truthfulness in the innocent scenario. It should be noted that these terms were not meant to indicate legal definitions of guilt and innocence. Instead, "guilty" and "innocent" corresponded to "lying (i.e., not admitting their criminal behavior, such as carrying a bomb to the train station)" and "telling the truth (i.e., admitting to leaving a backpack at the train station)", respectively. Therefore, the primary task for Interviewers was not to assess the legal concepts of guilt or innocence, but to detect whether the Human Mock Suspect was lying or telling the truth.

Oral Instructions. The oral instructions highlighted the specific tasks assigned to each Interviewer and Human Mock Suspect in the background stories. Additionally, Interviewers in

the Instruction and Instruction & AI Exercise groups were encouraged to use EFM and focus on inconsistencies produced by the Human Mock Suspects. For Human Mock Suspects, although "convince the Interviewer of your innocence" was the specific task, they were encouraged to choose their strategies, whether by being forthcoming and telling the truth, or fabricating details in their statements (see Supplementary Material C).

Interviewers and Human Mock Suspect: Questionnaire

Both Interviewers and Human Mock Suspects received an online questionnaire (see Supplementary Material D) and were asked to complete demographic information questions, including (a) Age, (b) Gender, and (c) whether they had any experience with investigative interviews. After interviewing Human Mock Suspects, Interviewers were asked to complete the post-interview questionnaire, which required them to: (a) judge whether the Human Mock Suspect was guilty or innocent of the crime, and (b) evaluate the importance of the pre-listed (non-)verbal cues that influenced their judgments. They were also allowed to list other cues freely in an open text box.

Dependent Measures

Coding-Based Dependent Measures

Use of Free-Recall Questions. The use of free-recall questions was coded when Interviewers asked Human Mock Suspects for a free narrative without suggesting their whereabouts or activities beforehand. For example, if a free-recall question like *"Please summarize what you did on June 22"* was presented after the question *"Did you get a parcel on that day?"*, we coded the use of free-recall questions as 0 since the previous question indicated an activity of the Human Mock Suspect. In contrast, if *"Please summarize what you did on June*

22" was presented as the first question of the interview or only after some questions unrelated to the evidence, we coded the use of free-recall questions as 1.

Frequency of Using the EFM. Each time the Interviewer used the evidence correctly, it was assigned 1 point. If the evidence also differed correctly in the dimensions of evidence strength and/or specificity from the evidence previously presented by the Interviewers, an additional 1 point was assigned. Therefore, each piece of evidence could receive up to 2 points: 1 point for correct use of the evidence and 1 additional point if the evidence strength and/or specificity were appropriately changed. If the evidence strength or specificity dimension of the evidence was not correctly changed from the previous one, it received only 1 point. In the coding process, we treated the report from a public worker in **Evidence 2** as having an evidence strength stronger than "information" but weaker than "surveillance footage." Two examples are presented below to illustrate how the frequency of EFM use was coded:

Interviewer (A): *"We have information that shows you were near 123 Xingfu Road, Blue Sea Street, Binhai City."* (1 point for using the evidence correctly); *"We have surveillance footage showing that you were near 123 Xingfu Road, Blue Sea Street, Binhai City. What were you doing there at the time?"* (1 point for using the evidence correctly and 1 additional point for changing the source dimension correctly). In total, Interviewer (A) received 3 points.

Interviewer (B): *"We have information that shows you were near Blue Sea Street, Binhai City."* (1 point for using the evidence correctly); *"We have information that shows you were near 123 Xingfu Road, Blue Sea Street, Binhai City."* (1 point for using the evidence correctly and 1 additional point for changing the specificity dimension correctly); *"We have surveillance footage showing that you were near 123 Xingfu Road, Blue Sea Street, Binhai City. What were*

you doing there at the time?" (1 point for using the evidence correctly and 1 additional point for changing the source dimension correctly). In total, Interviewer (B) received 5 points.

Level of Forthcomingness. The level of forthcomingness was calculated based on their admission of whereabouts or activities related to the evidence in response to the *first* evidence-related question. Evidence was scored sequentially (1, 2, 3) based on the criticality of the whereabouts and activities it was associated with. The weight of each piece of evidence (*Wi*) was calculated as $Wi = \frac{Scorei}{6} \times 100\%$. For example, if a Human Mock Suspect admitted to being at the "train station" (i.e., the whereabouts indicated in the third piece of evidence), the Wi of response would be calculated as $W3 = \frac{3}{6} \times 100\% = 50\%$. The specificity about each piece of evidence was rated from 1 to 5. For example, specificity regarding a piece of evidence involving a Jansport backpack-5 and a trash bin-4 at Xingfu Road-3, Blue Sea Street-2, and Binhai City-1 were scored progressively. The level of forthcomingness of the Human Mock Suspects was calculated using $\sum_{i=1}^{3} \sum_{j=1}^{5} Wi \times Sj$, producing a score between 0 (not forthcoming) and 5 (fully forthcoming). One example is presented below to illustrate how the level of forthcomingness was coded:

Interviewer (A): *"What did you do after you realized your bag was lost?"*

Human Mock Suspect (A): *"After I realized it was lost—because I only discovered it after boarding the train—I was quite anxious. However, I only contacted the train station staff and asked them to keep an eye out for it. Then I went on with my trip."* The level of forthcomingness of Human Mock Suspect (A) was calculated by $\frac{3}{6} \times 100\% \times 5 = 2.5$.

Number of Evidence-Statement or Within-Statement Inconsistencies. The number of evidence-statement or within-statement inconsistencies made by human mock suspects during the interviews was counted. Denials (e.g., *"No, I didn't go to the train station"*), equivocal

denials (e.g., *"Okay, you have the evidence. But how could I possibly be in two places at once? Lying at home while also going to the train station?"*), and omissions (e.g., *"I can't remember exactly."*) were counted as evidence-statement inconsistencies. In contrast, changing contradictory statements to equivocal admissions (e.g., *"I went straight home from Yong'an Road, but I might have passed through some other areas along the way, like other streets."*) or admissions (e.g., *"Yes, I went to the train station."*) was counted as within-statement inconsistencies. After counting by the first coder, we randomly selected 20% of the rounds, totaling 16 Human Mock Suspect interviews, for independent coding by a second coder. The Intraclass Correlation Coefficient (ICC) revealed good and excellent agreement between the two coders in the number of evidence-statement inconsistencies (ICC = 0.79, $F(15, 15.3) = 8.03$, $p < .001$) and within-statement inconsistencies, respectively (ICC = 0.71, $F(15, 6.77) = 8.48$, $p = .005$).

Questionnaire-Based Dependent Measures

Reported Use of Evidence-Statement and Within-Statement (In)consistencies. The reported use of evidence-statement and within-statement (in)consistencies was evaluated according to the importance attributed to them as a basis for their decisions by the Interviewers. We prelisted four verbal and four non-verbal cues.

Verbal cues: (a) The suspect said something that was inconsistent with the evidence/The suspect said something that was consistent with the evidence, (b) The suspect said something that was inconsistent with previous statements/The suspect said something that was consistent with previous statements, (c) The suspect said something untrustworthy/The suspect said something trustworthy, and (d) The suspect said something that was ambiguous and incoherent/The suspect said something that was clear and coherent.

Non-verbal cues: (a) The suspect shifted their eyes repetitively/The suspect didn't shift their eyes repetitively, (b) The suspect showed unnatural posture with nervous jittery/The suspect showed natural posture without nervous jittery, (c) The suspect showed tension in their facial expressions/The suspect showed comfort in their facial expressions, and (d) The suspect changed their tone of voice frequently/The suspect kept their tone of voice relatively smooth.

These cues were ranked by their importance by the Interviewers. For example, in the verbal cues category, if the Interviewers tended to use evidence-statement (in)consistencies for their judgments, they placed this cue at the forefront of the ranking items. Other cues would be placed in the second, third, or fourth position.

The Accuracy of Judgments. The accuracy of judgments refers to whether the Interviewer correctly determined if the Human Mock Suspect was lying or telling the truth after the interview.

Procedure

One Day Before the Experiment

Interviewers in either the Instruction or Instruction & AI Exercise group who correctly completed the test in the online screening session were eligible to participate in the experiment, otherwise, they were paid 5 Chinese Yuan (RMB) and excluded from the experiment. Both lying and truthful Human Mock Suspects received background stories in text and video versions. Subsequently, they were asked to record a video in which they freely recalled the details to ensure their comprehension of the story. Following this, the Human Mock Suspects received oral instructions from the experimenter (For details of the materials, see the *"Interviewers and Human Mock Suspects: Background Stories and Oral Instruction"* section and Supplementary Material C). If the Interviewers to whom they had been paired were not qualified to participate in

the experiment, the Human Mock Suspects also received 5 RMB compensation for their time and were excluded from the experiment.

Experiment Day

Before the Human Mock Suspect Interview

The experimental session took approximately 1 hour to 1 hour 50 minutes, varying by the experimental group for each pair. WeChat and Tencent Meeting (https://meeting.tencent.com/) served as communication tools to connect the two participants with the experimenter. Via Qualtrics, both Interviewers and Human Mock Suspects received an online questionnaire (For details of the materials, see the *"Interviewers and Human Mock Suspect: Questionnaire"* section and Supplementary Material D) and were asked to complete the demographic information questions.

Interviewers: Instructions. Interviewers in either the Instruction or Instruction & AI Exercise group spent approximately 20 minutes reading additional text-based instruction and watching a video-based instruction created by Li et al. (2024). (For details of the materials, see the *"Interviewers in Instruction and Instruction & AI Exercise Groups: Instructions"* section and Supplementary Material E). After reviewing each format of the instruction, they were asked to complete several test questions correctly to proceed to the next page of the questionnaire.

Interviewers: AI Exercise. Interviewers in the Instruction & AI Exercise group spent approximately 30 minutes completing the AI Exercise. These Interviewers clicked a randomizer in Qualtrics to determine which of the two lying AI suspects they would interview first. Then, they were provided with the corresponding background story (For details of the materials, see the *"Interviewers in Instruction & AI Exercise Groups: AI Exercise Background Stories and Feedback"* section and Supplementary Material F), along with a maximum of 10 minutes for

preparation. The experimenter instructed Interviewers to interact with the AI suspects displayed on the screen by typing their questions using the keyboard (see Supplementary Material for an example video). Interviewers listened to the AI suspect responses using earphones. After each of the two interviews, the Interviewers received oral feedback from the experimenter on both the case outcome and their interview process (For details of the materials, see the *"Interviewers in Instruction & AI Exercise Groups: AI Exercise Background Stories and Feedback"* section and Supplementary Material G).

Background Story and Oral Instruction. After the completion of the demographic information (or the Instructions and AI Exercise), each Interviewer received, in text, a background story (For details of the materials, see the *"Interviewers and Human Mock Suspects: Background Stories and Oral Instruction"* section and Supplementary Material B). Both Interviewers and Human Mock Suspects were informed that they had 10 minutes to organize their questions (Interviewers) or response strategies (Human Mock Suspects) before the upcoming Human Mock Suspect interview. A printed background story was also prepared for each Interviewer for convenient note-taking. To avoid misunderstandings of the instructions contained in the background story, oral instructions were also provided to Interviewers as they began reading the material (For details of the materials, see the *"Interviewers and Human Mock Suspects: Background Stories and Oral Instruction"* section and Supplementary Material B).

Human Mock Suspect Interview

The Human Mock Suspect interview was conducted via Tencent Meeting. Each Interviewer was asked to turn on the webcam to interact face-to-face with the Human Mock Suspect. Interviewers presented questions freely and listened to the responses from the Human Mock Suspects for a maximum of 10 minutes. Once 10 minutes had passed, the experimenter

would interrupt the interview. If the Interviewer felt ready to make their judgments, they could also terminate the interview before the full 10 minutes. In the transcripts of the recorded interviews, interruptions made by the experimenter were not documented. Since the maximum length of an interview was 10 minutes, we considered interviews that exceeded nine minutes while the Human Mock Suspect was still answering the last question as likely to have been terminated by the experimenter. Thirteen out of 78 interviews (16.7%) were interrupted by the experimenter. The whole process of the interview was recorded by the screen recording function inserted in Tencent Meeting. Conversations from each interview session were transcribed verbatim using the automatic transcription feature in Tencent Meeting for further analysis.

After the Human Mock Suspect Interview

After the interview, Interviewers were asked to complete a post-interview questionnaire (For details of the materials, see the *"Interviewers: Post-Interview Questionnaire"* section and Supplementary Material D). An illustration of the experimental design and main procedures is presented in Figure 5.

[Figure 5 will be inserted about here]

Results

Statistical Analyses

The testing of each hypothesis was conducted through R programming language (ver. 4.3.2). For Hypotheses 1 and 2, we used the LeveneTest() function from the package car (Fox & Weisberg, 2019) to test the homogeneity of variance for each dependent variable. Subsequently, one-way and two-way ANOVAs were conducted with the aov() function. For Hypothesis 1, the

emmeans() and pairs() functions from the package emmeans (Lenth et al., 2023) were used to conduct post-hoc tests for homogenous variances, and the games_howell_test() function from the package rstatix (Kassambara, 2023) was used for heterogeneous variance. For Hypothesis 2, planned pairwise comparisons were conducted using the t.test() function with automatically adjusted degrees of freedom. To test whether interacting with a more forthcoming Human Mock Suspect will reduce the frequency of EFM used by Interviewers in either the Instruction or Instruction & AI Exercise group, as part of Hypotheses 1, we used the lm() function.

To test Hypotheses 3 and 6, we created two dummy variables to distinguish (a) Interviewers who received Instruction and those who did not, and (b) Interviewers who received the AI Supsect exercise and those who did not. For these two hypotheses, we used the Process() function from the package bruceR (Bao, 2023) to test these hypotheses when the mediator is continuous (i.e., the frequency of using EFM). When the mediator is an ordinal variable (i.e., the reported evidence-statement or within-statement (in)consistencies), we conducted ordinal logistic regressions using the polr() function from the package MASS with reported evidence-statement or within-statements (in)consistencies as outcomes (Venables & Ripley, 2002). Binary logistic regressions were then fitted using the glm() function, with the accuracy of judgments as an outcome. The indirect effects were tested using the mediate() function from the package MASS (Venables & Ripley, 2002). The 95% Confidence Interval (CI) of all indirect effects was obtained with 5000 bootstrap resamples.

For Hypothesis 4, because the present study used a ranking question, when evidence-statement (in)consistencies are ranked first, within-statement (in)consistencies can only be ranked second, third, or fourth. To minimize the impact of this ranking restriction on the results, we first removed within-statement (in)consistencies and re-ranked the remaining three verbal

cues: (a) evidence-statement (in)consistencies, (b) The suspect said something untrustworthy/The suspect said something trustworthy, and (c) The suspect said something that was ambiguous and incoherent/The suspect said something that was clear and coherent. Then, we removed evidence-statement (in)consistencies and re-ranked the remaining cues: (a) within-statement (in)consistencies, (b) the suspect said something untrustworthy/the suspect said something trustworthy, and (c) the suspect said something that was ambiguous and incoherent/the suspect said something that was clear and coherent. We then performed Kruskal-Wallis tests using the kruskal.test() function for reported evidence-statement or within-statement (in)consistencies, respectively. Dunn's tests were performed using dunn.test() function from package dunn.test() for significant differences among experimental groups (Dinno, 2024).

We applied Signal Detection Theory (SDT) to evaluate (1) the ability of interviewers from different experimental groups to differentiate between lying and truthful Human Mock Suspects, and (2) the judgment bias of interviewers, which reflects their tendency in making judgments and was treated as an exploratory analysis. In the Instruction & AI Exercise group, the False-Alarm Rate (FAR) equaled to 0 (i.e., Interviewers showed a perfect performance in classifying truthful Human Mock Suspects). Therefore, we applied the loglinear method by adding 0.5 to both the number of hits and false alarms and adding 1 to both the number of signal trials and noise trials before calculating the Hit Rate (HR) and FAR (Stanislaw & Todorov, 1999). Since in the present study each interviewer made only a single judgment on whether the Human Mock Suspect was lying or not, it was not possible to calculate an individual *d'* (*d* prime) index of discriminability. Therefore, *d'* could only be reported as a descriptive index after aggregating responses within each experimental group, but it could not be used to compare differences across groups. To test the differences in judgment accuracy among the experimental

groups (i.e., Hypothesis 5), we used χ^2 tests performed with the tab_xtab() function from the package sjPlot (Lüdecke, 2023). If the group differences were significant, tests of proportions were conducted between each pair of experimental groups using the prop.test() function. Judgment bias was measured by calculating criterion (C). If interviewers were biased toward judging Human Mock Suspects as lying, they had a $C < 0$ (i.e., a liberal judgment); in contrast, if interviewers tended to believe that Human Mock Suspects were truthful, they had a $C > 0$ (i.e., a conservative judgment) (Abdi, 2007).

Within-Session Results (AI Exercise)

One Interviewer's second interview was not recorded and thus was excluded from the analysis. We found that Interviewers used EFM more frequently to present evidence during the second interview ($M = 5.00$, $SD = 1.70$) compared to their first one ($M = 4.23$, $SD = 1.99$) after receiving feedback, $t(25) = -2.12$, $p = .036$, $d = 1.80$. This result again suggests, similar to Li et al. (2024), that AI Exercise can train Interviewers in the use of EFM within AI Exercise.

The Impact of Interventions on the Questioning Strategies of Interviewers

The Impact of Interventions on the Use of Free-recall Questions

A Chi-square (χ^2) test revealed significant group differences on whether Interviewers used a free-recall question at the beginning of interviews, $\chi^2(2) = 18.65$, $p < .001$, $\varphi = .49$. After conducting Chi-square (χ^2) tests of proportions between each pair of experimental groups, we found that compared to the Interviewers in the Control group (50%), those in the Instruction & AI Exercise group (96.2%), $\chi^2(1) = 11.83$, $p < .001$, 95% CI [0.22, 0.70]; and those in the Instruction group (88.5%), $\chi^2(1) = 7.31$, $p = .007$, 95% CI [0.12, 0.65], more often started the interview with a free-recall question to allow the Human Mock Suspects to freely tell their whereabouts and activities on the day of the crime. However, no significant difference was

observed between the Interviewers in the Instruction & AI Exercise group and the Instruction group, $\chi^2(1) = 0.27$, $p = .603$, 95% CI [−0.10, 0.26].

The Impact of Interventions on the Frequency of Using EFM

A One-way ANOVA showed significant group differences on the frequency of using EFM, $F(2, 75) = 6.19$, $p = .003$, $\eta_p^2 = .14$. Interviewers in both the Instruction & AI Exercise ($M_{\text{Instruction \& AI Exercise}} = 2.04$, $SD = 2.65$) and the Instruction group ($M_{\text{Instruction}} = 2.58$, $SD = 2.34$) used EFM more frequently compared to those in the Control group ($M_{\text{Control}} = 0.58$, $SD = 1.03$), $MD_{\text{Instruction \& AI Exercise − Control}} = 1.46$, $SE = 0.39$, $p = .034$, 95% CI [0.09, 2.83]; $MD_{\text{Instruction − Control}} = 2.00$, $SE = 0.35$, $p < .001$, 95% CI [0.77, 3.23]. Again, no significant difference was observed between the Interviewers in the Instruction & AI Exercise and the Instruction groups, $MD_{\text{Instruction \& AI Exercise − Instruction}} = -0.54$, $SE = 0.49$, $p = .718$, 95% CI [−2.21, 1.13]. A multiple linear regression showed that interaction with more forthcoming Human Mock Suspects led to a decrease in the frequency of using EFM by the intervened Interviewers, $B = -0.75$, $SE = 0.17$, $t(50) = -4.35$, $p < .001$. Therefore, Hypothesis 1 was partially supported.

The Impact of Interventions on the Number of Inconsistencies and Reported Use of (In)consistencies

The Impact of Interventions on the Number of Inconsistencies

The 3 (Experimental group: Instruction & AI Exercise/Instruction/Control) × 2 (Truthfulness of Human Mock Suspect: Liar/Truth-teller) ANOVA with the number of evidence-statement inconsistencies revealed a significant interaction effect, $F(2, 72) = 3.20$, $p = .047$, $\eta_p^2 = .08$, and significant main effects for Experimental group: $F(2, 72) = 3.53$, $p = .034$, $\eta_p^2 = .09$; and for Truthfulness of Human Mock Suspect: $F(1, 72) = 29.84$, $p < .001$, $\eta_p^2 = .29$. However, for the number of within-statement inconsistencies, we did not find significant interaction

between the experimental group and the truthfulness of Human Mock Suspect, $F(2, 72) = 2.21$, $p = .117$, $\eta_p^2 = .06$. A main effect was detected only for the truthfulness of Human Mock Suspect, $F(1, 72) = 9.33$, $p = .003$, $\eta_p^2 = .11$.

Planned pairwise comparisons within each experimental group showed that lying Human Mock Suspects produced significantly more evidence-statement inconsistencies compared to truthful Human Mock Suspects in both the Instruction ($M_{\text{Lying}} = 4.46$, $SD = 3.41$; $M_{\text{Truthful}} = 0.77$, $SD = 2.49$), $t(21.97) = 3.16$, $p = .005$, $d = 1.24$; and the Instruction & AI groups ($M_{\text{Lying}} = 4.46$, $SD = 3.48$; $M_{\text{Truthful}} = 0.15$, $SD = 0.38$), $t(12.28) = 4.44$, $p < .001$, $d = 1.74$, with a greater difference observed between lying and truthful Human Mock Suspects in the Instruction & AI group. No significant difference was observed in the Control group ($M_{\text{Lying}} = 1.46$, $SD = 1.98$; $M_{\text{Truthful}} = 0.38$, $SD = 1.39$), $t(21.47) = 1.60$, $p = .123$, $d = 0.63$ (see Figure 6). When looking at the number of within-statement inconsistencies produced by Human Mock Suspects, a significant difference was observed between the lying and truthful Human Mock Suspects in the Instruction & AI Exercise group ($M_{\text{Lying}} = 1.31$, $SD = 1.38$; $M_{\text{Truthful}} = 0.15$, $SD = 0.38$), $t(13.77) = 2.91$, $p = .011$, $d = 1.14$. We did not observe any significant difference in either the Instruction ($M_{\text{Lying}} = 0.85$, $SD = 1.46$; $M_{\text{Truthful}} = 0.15$, $SD = 0.55$), $t(15.38) = 1.60$, $p = .131$, $d = 0.63$; and Control group ($M_{\text{Lying}} = 0.23$, $SD = 0.60$; $M_{\text{Truthful}} = 0.15$, $SD = 0.55$), $t(23.86) = 0.34$, $p = .737$, $d = 0.13$. Therefore, Hypothesis 2 was partially supported.

The Impact of Interventions on the Number of Inconsistencies Through the Frequency of Using EFM.

The indirect effects and their 95% confidence intervals (CIs) for the mediation model were estimated using 5000 bootstrap resamples. We found that the frequency of EFM use mediated the relationships between the interventions (i.e., combining Instruction and Instruction

& AI Exercise groups) and the number of evidence-statement and within-statement inconsistencies. After receiving the interventions, Interviewers used EFM more frequently to present evidence, thereby significantly increasing the number of evidence-statement, Indirect Effect = 1.57, *SE* = 0.61, *p* = .010, 95% CI [0.52, 2.87]; and within-statement inconsistencies obtained from lying Human Mock Suspects, Indirect Effect = 0.56, *SE* = 0.26, *p* = .030, 95% CI [0.10, 1.10]. Therefore, Hypothesis 3 was supported.

The Impact of Interventions on the Reported Use of (In)consistencies

Kruskal-Wallis tests revealed significant group differences in the reported use of evidence-statement (in)consistencies. More than two-thirds of the Interviewers in both the Instruction group (69.2%, 18/26) and the Instruction & AI Exercise group (73.1%, 19/26) ranked evidence-statement (in)consistencies as the most important verbal cue for making judgments, compared to the other two verbal cues. In contrast, 38.5% (10/26) of Interviewers in the Control group did so. Planned pairwise comparisons were conducted using Dunn's tests, with Holm correction for multiple comparisons. The results showed that compared to the Interviewers in the Control group, those in both Instruction used evidence-statement (in)consistencies more frequently to assess whether the Human Mock Suspect was lying or telling the truth, $Z = 2.28$, $p = .022$. A similar pattern was found between the Control and Instruction & AI Exercise groups, $Z = 2.62$, $p = .013$. However, no significant differences were observed in the comparison between the Instruction group and the Instruction & AI Exercise group, $Z = -0.34$, $p = .367$. Turning to within-statement (in)consistencies, we found that half of the Interviewers in the Instruction & AI Exercise group ranked within-statement (in)consistencies as the most important verbal cue for making their judgments, compared to the other two verbal cues. In contrast, less than half of the Interviewers in the Instruction (34.6%, 9/26) and the Control groups (38.5%, 10/26) did so.

Planned pairwise comparisons did not reveal any significant difference. Therefore, Hypothesis 4 was only partially supported.

[Figure 6 will be inserted about here]

The Impact of Interventions on the Accuracy of the Judgments Reached by the Interviewers

The Impact of Interventions on the Accuracy of the Judgments

In the Control group, the FAR exceeded the HR of interviewers, indicating a tendency to falsely classify truthful Human Mock Suspects as lying ones ($d' = -0.23$). In contrast, Instruction group showed moderate discriminability ($d' = 0.64$), and the highest discriminability was observed in the Instruction & AI Exercise group ($d' = 2.48$). A Chi-square (χ^2) test revealed a significant overall difference in judgment accuracy among experimental groups, $\chi^2(2) = 10.54$, $p = .005$, $\varphi = .37$. Chi-square (χ^2) tests of proportions found that receiving solely Instruction (61.5%) slightly improved the ability of the Interviewers to make accurate judgments compared to those in the Control group (46.2%), though the difference was not statistically significant, $\chi^2(1) = 0.70$, $p = .404$, 95% CI [−0.15, 0.46]. However, Interviewers in the Instruction & AI Exercise group (88.5%) reached more accurate judgments on whether the Human Mock Suspects were lying or telling the truth compared to those in both the Instruction, $\chi^2(1) = 3.69$, $p = .055$, 95% CI [0.01, 0.53], and the Control groups, $\chi^2(1) = 8.74$, $p = .003$, 95% CI [0.16, 0.69], with the latter comparison being significant. Hence, Hypothesis 5 was partially supported.

When we separately considered the differences in accurately detecting lying or truthful Human Mock Suspects among experimental groups, no significant difference was found between

the experimental groups in detecting liars, $\chi^2(2) = 0.27$, $p > .999$, $\varphi = .08$. However, Interviewers in the Instruction & AI Exercise group (100%) were significantly better at recognizing truthful Human Mock Suspects than those who received only instructions (46.2%), $p = .008$, or those without any intervention (23.1%), $p < .001$. No significant difference was found between the Instruction and Control groups.

The Impact of Interventions on the Accuracy of the Judgments Through the Frequency of EFM Use and Reported Use of (In)consistencies.

Bootstrap resampling (5000 simulations) was used to estimate the indirect effect with their 95% confidence intervals (CIs). Regarding Hypothesis 6, the increased accuracy of Interviewers' judgments due to these interventions was not significantly mediated by more (a) frequent use of more EFM or (b) reported use of evidence-statement or within-statement (in)consistencies (see Supplementary Materials H for the statistical results). Hence, Hypothesis 6 was not supported.

The Impact of Interventions on Judgment Bias

Judgment bias occurs when Interviewers tend to judge Human Mock Suspects as lying (on the one hand) or truthful (on the other hand). We found that Interviewers in both Control ($C = -0.62$) and Instruction groups ($C = -0.42$) showed lie bias in making their judgments, with a stronger tendency observed in the Control group. However, a truth bias was detected in the Instruction & AI Exercise group ($C = 0.56$), indicating that they tended to make more conservative judgments.

Discussion

In interviewer training, interviewers can learn skills through different methods, including theoretical instructions and role-playing. However, since a suspect interview is a complex and

dynamic process, there are reasons to believe that role-playing would facilitate the development of procedural knowledge of interview techniques and provide a practical environment for interviewers. Considering the required investment and the potential unpredictability of role-played suspects' behaviors, in the present study, we used an AI Exercise approach that offers a formal and standardized method of training that can integrate different interventions, such as instruction and feedback. Moreover, our experimental setup focused on the use of the Evidence Framing Matrix (EFM) as an exercise example, which is an integral part of the Strategic Use of Evidence (SUE) technique (Granhag et al., 2013; Hartwig & Granhag, 2023). We created two AI suspects and randomly allocated participants to one of three groups: (a) Instruction & AI Exercise group, (b) Instruction group, and (c) Control group. After the training, the participants interacted with either a lying or truthful Human Mock Suspect.

Artificial Intelligence (AI) Exercise Shows Only Marginal Advantage in Improving Questioning Strategies Beyond Instructions

Compared to the Control group, we found that Interviewers in both the Instruction & AI Exercise and Instruction groups were more likely to start the interview by posing a free-recall question. They also used the EFM more frequently to present evidence during the interview, indicating that the instruction had prompted the Interviewers to formulate questions aligned with the SUE framework. However, we found no differences in the use of free-recall questions and the frequency of using EFM between the Interviewers in the Instruction & AI Exercise and the Instruction group. These findings partially replicated the results from Luke et al. (2016), who found that theoretically trained interviewers (i.e., those trained to ask questions in a funnel structure and present evidence tactically) used EFM to a greater extent than untrained interviewers. However, AI Exercise did not add any additional effect regarding the use of EFM

compared to instruction alone. One potential explanation could be related to the online screening session. One day before the experiment, interviewers in either the Instruction or Instruction & AI Exercise group had to correctly complete a comprehension test to be eligible to participate in the experiment. This procedure might have introduced a selection bias: Only participants who had already gained a basic understanding of EFM were allowed to proceed. Another potential explanation could be that for interview techniques with a clearly defined structure and procedure, such as EFM, instruction alone might be sufficient to support understanding and application in an artificial and controlled experiment (i.e., three pieces of evidence in the background stories were well-designed to align the use of EFM).

Interventions Elicit More Actual Evidence-Statement Inconsistencies from Lying Human Mock Suspects

Our findings revealed that interacting with instructed Interviewers (compared to Interviewers without receiving instructions), lying Human Mock Suspects produced more evidence-statement inconsistencies rather than within-statement inconsistencies (Hypothesis 2 was partially supported). A greater difference was observed in the Instruction & AI Exercise group than in the Instruction group. For Interviewers in both the Instruction and Instruction & AI Exercise groups, more frequent use of EFM increased the number of inconsistencies produced by lying Human Mock Suspects (Hypothesis 3 was supported).

While a significant difference was observed only between lying and truthful Human Mock Suspects in terms of their within-statement inconsistencies in the Instruction & AI Exercise group, this result does not directly support the transfer effect of the intervention itself. Importantly, within-statement inconsistency is a by-product of suspects' verbal strategies, which occurs when Human Mock Suspects choose to change their initial statements, rather than a direct

outcome of the intervention itself. Therefore, the lack of significant differences between lying and truthful Human Mock Suspects in the Instruction group might reflect individual variability of suspects' verbal strategies, rather than indicating a limitation of the instruction-only intervention.

First, the particular relation between evidence-statement and within-statement inconsistencies may play a role. Before the evidence presentation phase, if a Human Mock Suspect denies their whereabouts and activities and chooses to maintain their initial statements during evidence presentation, they are likely to produce more evidence-statement inconsistencies but no within-statement inconsistency. Conversely, if a Human Mock Suspect decides to repeatedly alter their statements in response to the presented evidence, they will likely produce more within-statement inconsistencies but fewer evidence-statement inconsistencies. As a result, it is not likely to obtain a pattern of a simultaneous increase in *both* evidence-statement and within-statement inconsistencies. Second, it is essential to consider different possibilities for the emergence of these two types of inconsistencies. Evidence-statement inconsistencies may occur during the questioning phase before the evidence disclosure phase. However, within-statement inconsistencies are more likely to occur only during the evidence disclosure phase, specifically after the Human Mock Suspect becomes cognizant of the evidence presented against them.

Artificial Intelligence (AI) Exercise Shows Limited Effects on Interviewers' Reliance on Within-Statement (In)consistencies

We found that Interviewers in the Instruction and Instruction & AI Exercise groups increased (compared with the Control group) their use of evidence-statement (in)consistencies but not their use of within-statement (in)consistencies to assess whether the Human Mock Suspects were lying or not (Hypothesis 4 was therefore partially supported). Although we

observed that half of the Interviewers in the Instruction & AI Exercise group—more than in any other experiment group—ranked within-statement (in)consistencies as the most important verbal cue. However, no significant differences were found between any pair of experimental groups. Overall, these results indicate that interviewers in the Instruction & AI Exercise group used both types of (in)consistencies more frequently than those in the other groups; however, a significant difference was only observed between the Instruction & AI Exercise group and the Control group in the use of evidence-statement (in)consistency. The potential explanations might be twofold. First, previous research has shown that within-statement inconsistencies are less common compared to evidence-statement inconsistencies in suspects' verbal responses (Granhag et al., 2013; Luke et al., 2016). This pattern may help explain why no significant difference was observed between the Instruction & AI group and the Control group in their use of within-statement (in)consistency. Second, regarding the design of our measurement tools, we used ranking questions with pre-listed cues for Interviewers, which might have confounded our results. Since all available cues were already listed, we can only conclude which pre-listed verbal cues Interviewers ranked as more important, rather than determining whether they actually used these specific verbal cues.

AI Exercise Increases Judgment Accuracy in Truthful Human Mock Suspects

Receiving sole instructions enabled Interviewers to more accurately judge whether Human Mock Suspects were lying or telling the truth compared to those in the Control group, although the difference was not significant. When Interviewers received both instructions and AI Exercise, the accuracy of their judgments was improved compared to those in either the Instruction group or the Control group, with a significant difference observed in comparison with the Control group (Hypothesis 5 was partially supported).

When we separately considered the effectiveness of interventions in accurately detecting lying or truthful Human Mock Suspects, significant differences were observed only between the Instruction & AI Exercise group and the Instruction or Control group in detecting truth: Interviewers in the Instruction & AI Exercise group performed better at recognizing truthful Human Mock Suspects than those who received only instructions or no intervention. These results might be interpreted as being related to the potential influence of representativeness heuristics (Kahneman & Tversky, 1973). AI Exercise allows Interviewers to interact with unforthcoming lying suspects, who produced a high number of inconsistencies. This might help Interviewers form a prototype of how lying Human Mock Suspects would behave during the interview. As a result, when Interviewers later interacted with truthful Human Mock Suspects, they may more easily recognize that these verbal behaviors did not match what they have seen during the AI Exercise, consequently increasing their judgment accuracy. Although representativeness heuristics can save time and improve judgment accuracy when there is structural similarity between the AI Exercise and the Human Mock Suspect interviews, they can still be problematic in real-life situations (Bílek et al., 2018), where suspects' behaviors do not always follow stereotypical patterns as they do in the controlled settings.

To evaluate the impact of different interventions, it is crucial to design the available evidence of Human Mock Suspect interviews to align with the two dimensions of EFM (since we instructed interviewers to use it in organizing their presentation of evidence). If the structure of the evidence differs substantially from that used in the instructions and AI Exercise, it would be difficult for us to determine whether any observed performance reflects the efficacy of the interventions. However, the three pieces of evidence were created to align with the two dimensions of the EFM, which do not sufficiently reflect the complexity of real-life cases, where

multiple criminal tasks occur across different phases (Oleszkiewicz & Watson, 2021). A second concern is that our Human Mock Suspects were young Chinese adults, most of whom were entirely truthful and were open and honest in providing information (Clemens & Grolig, 2019; Oleszkiewicz & Watson, 2021). Especially in the Instruction & AI Exercise group, the Interviewers elicited fewer inconsistencies for the truthful Human Mock Suspects compared to the Interviewer in the other two groups. However, in real-life suspect interviews, truthful suspects might also be deceptive to avoid overall suspicion (James-Kangal et al., 2018). Thus, we cannot be certain whether the high accuracy in detecting truthful Human Mock Suspects in the Instruction & AI Exercise group should be attributed to the effectiveness of AI Exercise, or if it is due to the "perfect openness and honesty" counter-interrogation strategies employed by the truthful Human Mock Suspects. Therefore, we acknowledge that caution is still needed when considering the generalization of these results to real-life suspect interviews. Further studies are needed to examine the effectiveness of AI Exercise in more complex criminal scenarios with a more diverse range of Human Mock Suspects.

Unexpectedly, we did not find a mediating effect of the frequency of EFM use between interventions and the accuracy of judgments, and a reason for this might be that the interventions primarily increased overall detection accuracy by improving the performance in detecting truth-tellers. However, interacting with truthful Human Mock Suspects gave Interviewers fewer (or no) opportunities to apply for EFM. Moreover, we also did not find a mediating effect of using either evidence-statement or within-statement (in)consistencies (Hypotheses 6 were not supported). A possible explanation for these results could be the limitation of measurement tools. As previously mentioned, ranking questions with pre-listed cues may not accurately reflect the actual basis for the conclusions made by the Interviewers.

After receiving the AI Exercise, Interviewers tended to make more conservative judgments. Specifically, they were more likely to judge Human Mock Suspects as truthful rather than lying, compared to the other two experimental groups. This result was not hypothesized. On the one hand, the result suggests that AI Exercise made the Interviewers pay more attention to both evidence-statement and within-statement inconsistencies, rather than searching for non-diagnostic cues during the interview, which means that AI Exercise can potentially guide them to change their judgment strategies. On the other hand, this result also suggests a limitation in the current response strategies of AI suspects. Even if our primary aim of using AI Exercise to provide Interviewers with opportunities to practice SUE technique and receive feedback, stereotypical response types of AI suspects may encourage representativeness heuristics and simplify decision-making process. Using more natural language (e.g., incorporating discourse markers between statements to avoid abrupt inconsistencies) would help AI suspects better simulate the dynamics of real-life suspect interviews. When looking at the reason why Interviewers in both the Control and Instruction groups showed a lie bias, the lack of ecological validity in the background scenario could be a potential reason. The three pieces of evidence presented were ambiguous. For example, the third piece of evidence: both lying and truthful suspects had left a backpack at the train station. While lying suspects left the backpack with bombs, truthful suspects left it accidentally. However, without specific instructions, Interviewers may have interpreted these pieces of evidence as indicative of criminal behaviors (i.e., left the backpack with bombs), overlooking the possibility that they could support alternative explanations.

Strengths and Implications

Cleary and Warner (2016) pointed out that most police officers receive informal training, as formal training requires considerable monetary and personnel costs. Regarding formal training methods, 71.8% of interviewers reported receiving training via a book or manual, while 42.6% received instructional videos. Grossman and Salas (2011) suggested that for successful transfer, learners need to have opportunities to apply their new skills and abilities. Providing a relevant training context allows trainees to apply their gained knowledge in an appropriate environment, which significantly contributes to the transfer of training. In the present study, we created two lying AI suspects for the use of SUE technique, particularly focusing on the EFM. This AI Exercise supports two scenarios (i.e., theft and terrorism cases) that police interviewers face in actual interviews and provides an opportunity for interviewers to receive timely feedback (Lamb, 2016), as well as offering a more controllable practice environment compared to interacting with role-play human suspects. The application of the Large Language Model (LLM) in the role-playing task has given rise to several platforms that provide character customization services, allowing training customizers to easily customize the personal information, personality, voice, and appearance of suspects. Unlimited by geographical locations, AI Exercise can reduce expenditure on travel, lectures, and hiring role-players, and is therefore potentially cost-effective (Benson & Powell, 2015; Lamb, 2016).

Limitations and Future Directions

Any designed exercise should meet the use needs of professionals. However, both Interviewers and Human Mock Suspects in our study were laypersons, limiting the ecological validity of the results. For future studies, AI Exercise should also be applied to professionals to assess its efficacy.

The current version of AI Exercise (Chinese version) only allows text input, reducing the immersive experience for trainees. Additionally, the AI suspects were created to follow the response rules from Li et al. (2024), which were simplified and prototypical examples of liars' and truth-tellers' behaviors, thus exaggerating the differences between liars' and truth-tellers' responding strategies compared to real-life settings. Finally, due to the lack of question-answer pairs that can be used to train AI suspects and the limitations of gpt-3.5-turbo-16k itself, the AI suspects have some issues in mimicking the response rules. In future studies, AI suspects could be trained based on how human mock suspects respond to questions from interviewers, to produce more realistic responding algorithms for AI suspects.

Both text-based and video-based instructions provided some guidelines for how to disclose evidence to elicit more evidence-statement and within-statement inconsistencies from lying suspects as cues to deception. By offering some procedural details, Interviewers could follow these guidelines to disclose the evidence during interactions with Human Mock Suspects. However, some details were not included in the instructions, such as how many free-recall questions and specific questions should be asked before presenting the available evidence. Additionally, we did not specify counter-interrogation strategies for lying or truthful Human Mock Suspects (i.e., they could choose to be forthcoming or fabricate details in their statements). Therefore, there remains some flexibility in the choice of questioning strategies based on the actual counter-interrogation strategies employed by the Human Mock Suspects. Overall, the Interviewers seem to have (at least partially) acquired the ability to use the EFM in their interactions with Human Mock Suspects. However, since the instructions did not provide comprehensive guidance on the SUE technique, it is necessary to improve the instructions in future studies to enhance interviewers' acquisition of the SUE technique.

There are two main concerns with the background story. One concern lies in the three pieces of evidence presented in the story. We employed a basic experimental paradigm involving only one criminal task, which may not fully capture the complexity of real-life criminal cases (Oleszkiewicz & Watson, 2021). The type of structured, EFM-aligned, and highly reliable evidence can be scarce or nonexistent in real cases (Oleszkiewicz et al., 2023). Additionally, the three pieces of evidence used in this study were ambiguous in distinguishing between lying and truthful suspects, which allowed for some unwanted interpretation flexibility regarding evidence for Human Mock Suspects. Another concern relates to the instructions mentioned in the background story. Human Mock Suspects were not allowed to simply say “no comment” or remain silent. May et al. (2023) reported that more than half of the suspects remained silent in guilty interview situations, while 18.4% of suspects did so in innocent situations. However, we did not consider this variation in Human Mock Suspect behavior in this study, as no interview techniques can be generalized to such situations where suspects do not speak.

Based on these limitations, we acknowledge that any results found in this study regarding the AI Exercise improving the accuracy in detecting truth should be interpreted with caution in terms of their generalizability into real-life interviews. This is because the background information we used in the Human Mock Suspect interview differs from real cases. In future studies, researchers can explore how well such interventions transfer to more naturalistic interview settings.

We did not control the training time between Instruction and Instruction & AI Exercise groups, which might confound the transfer effect of AI Exercise with the additional time learning. It is true that the Interviewers in the Instruction & AI Exercise group spent more time learning. However, the simulation was new and somewhat difficult, which could have increased

the cognitive load or made the Interviewers feel exhausted and, consequently, pay less attention during the interview. Despite this disadvantage, we still found that Interviewers in the Instruction & AI Exercise group made more accurate judgments than the other two groups, supporting the potential of the AI Exercise itself.

Human Mock Suspects were required to "imagine" their whereabouts and activities based on their background stories. Moreover, conducting interviews online presented challenges in simulating real-life interviews. In real-life interviews, interviewers can pick up a variety of cues, including verbal cues, body language, and facial expressions. However, in the present study, these cues are often limited in online meeting-based interviews due to network instability and limitations in video and audio quality (e.g., Archibald et al., 2019; Seitz, 2016). These differences between the Human Mock Suspect interview and real-life suspect interviews could hinder the transferability of the AI Exercise's efficacy to actual police interviews. Therefore, future studies could involve Human Mock Suspects participating in simulated criminal acts (not involving real crimes) and interacting with the interviewer in person.

Oleszkiewicz et al. (2023) pointed out the potential risks of treating admissions and omissions (or denials) as two extremes of a one-dimensional scale in practice. Assuming that the lack of statements (i.e., omissions) or equivocal denials equals an inconsistency with the evidence could strengthen interviewers' presumptions of deception. However, as previously mentioned, the ground truth in this case was known (i.e., we were certain about the truthfulness of the suspects' responses), so the current coding scheme was deemed appropriate. In future studies, it is important to consider omissions and equivocation as unresolved discrepancies and encourage interviewers to seek possible explanations from suspects.

Beek et al. (2021) suggested that an implicit underlying assumption of the SUE technique is that the evidence to be disclosed should be correct, thus making suspects aware that interviewers might hold some relevant information against them. In the present study (or other experimental studies), researchers can make sure that the evidence is scripted to be corrected (i.e., the ground truth is known). However, in real-life settings, disclosing ambiguous or incorrect evidence can be problematic with eliciting false confessions from innocent suspects, or influencing counter-interrogation strategies of guilty suspects. Therefore, for future studies, it is crucial to instruct interviewers to evaluate the evidential value of available evidence during the preparation for suspect interviews (Beek et al., 2021; Granhag & Hartwig, 2015).

Conclusions

We created two Large Language Models (LLM)-directed lying suspects and found that receiving interventions (i.e., Interviewers in the Instruction and Instruction & AI Exercise groups) made Interviewers use EFM more frequently, thereby increasing the number of actual evidence-statement and within-statement inconsistencies produced by lying Human Mock Suspects. In the Instruction & AI Exercise group, a greater difference was observed in the number of both actual evidence-statement and within-statement inconsistencies between liars and truth-tellers compared to those in the Instruction and Control groups. Receiving interventions also made Interviewers use evidence-statement (in)consistencies more often as a basis for their judgments. Furthermore, those receiving both instructions and AI Exercise were better able to accurately judge the veracity of Human Mock Suspects compared to those in the Control group. Therefore, the results demonstrate that the efficacy of the AI Exercise transferred to interactions with Human Mock Suspects in the controllable setting, but the advantage over instruction alone was not particularly robust. However, it is important to note that this does not necessarily reflect

real-life settings. Although AI Exercise has the potential to provide an interactive environment for Interviewers to practice the SUE technique, limitations remain, and its effectiveness needs further investigation.

Supplementary Materials

Supplementary materials for this article can be accessed online at

https://osf.io/u7ekj/?view_only=ba8524ab63c44b928214e8e4c641d5f0

References

Abdi, H. (2007). Signal detection theory (SDT). *Encyclopedia of measurement and statistics*, 886-889.

Adhikari, S., & Dhakal, B. (2023). Revolutionizing natural language processing with GPT-based Chatbots: a review. *Technical Journal*, *3*(1), 109-120. https://doi.org/10.3126/tj.v3i1.61943

Alison, L., Alison, E., Noone, G., Elntib, S., Waring, S., & Christiansen, P. (2014). The efficacy of rapport-based techniques for minimizing counter-interrogation tactics amongst a field sample of terrorists. *Psychology, Public Policy, and Law, 20*(4), 421–430. https://doi.org/10.1037/law0000021

Archibald, M. M., Ambagtsheer, R. C., Casey, M. G., & Lawless, M. (2019). Using Zoom Videoconferencing for Qualitative Data Collection: Perceptions and Experiences of Researchers and Participants. *International Journal of Qualitative Methods*, *18*(1), 1–8. https://doi.org/10.1177/1609406919874596

Bao, H.-W.-S. (2023). *bruceR: Broadly useful convenient and efficient R functions* (R package version 2023.9). https://CRAN.R-project.org/package=bruceR

Beek, M. V., Bull, R., & Chen, M. (2021). When the evidence is incorrect: An exploration of what happens when interviewers unwittingly present inaccurate information in interviews with suspects. *Journal of Police and Criminal Psychology*, *36*(4), 769-782. https://doi.org/10.1007/s11896-021-09494-3

Bekkar, M., Djemaa, H. K., & Alitouche, T. A. (2013). Evaluation measures for models assessment over imbalanced data sets. *Journal of Information Engineering and Applications*, *3*(10). 27-38.

Benson, M. S., & Powell, M. B. (2015). Evaluation of a comprehensive interactive training system for investigative interviewers of children. *Psychology, Public Policy, and Law*, *21*(3), 309–322. https://doi.org/10.1037/law0000052

Bílek, J., Nedoma, J., & Jirásek, M. (2018). Representativeness heuristics: A literature review of its impacts on the quality of decision-making. https://hdl.handle.net/10195/71486

Bond, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review*, *10*(3). https://doi.org/10.1207/s15327957pspr1003_2

Cleary, H. M. D., & Warner, T. C. (2016). Police training in interviewing and interrogation methods: A comparison of techniques used with adult and juvenile suspects. *Law and Human Behavior, 40*(3), 270–284. https://doi.org/10.1037/lhb0000175

Clemens, F. (2013). *Detecting lies about past and future actions: The Strategic Use of Evidence (SUE) technique and suspects' strategies*. https://gupea.ub.gu.se/handle/2077/32705

Clemens, F., & Grolig, T. (2019). Innocent of the crime under investigation: suspects' counter-interrogation strategies and statement-evidence inconsistency in strategic vs. non-strategic interviews. *Psychology, Crime & Law*, *25*(10), 945–962. https://doi.org/10.1080/1068316X.2019.1597093

Deeb, H., Vrij, A., Hope, L., Mann, S., Granhag, P. A., & Strömwall, L. A. (2018). Police officers' perceptions of statement inconsistency. *Criminal Justice and Behavior*, *45*(5), 644-665. https://doi.org/10.1177/0093854818758808

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin*, *129*(1), 74. https://doi.org/10.1037/0033-2909.129.1.74

DePaulo, B. M., & Morris, W. L. (2004). Discerning lies from truths: Behavioral cues to deception and the indirect pathway of intuition. *BM DePaulo, WL Morris, ad. by PA Granhag. The detection of deception in forensic contexts//New York: Cambridge*, 15-40. https://doi.org/10.1017/CBO9780511490071.002

Dinno, A. (2024). *dunn.test: Dunn's test of multiple comparisons using rank sums* (R package version 1.3.6). https://CRAN.R-project.org/package=dunn.test

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (3rd ed.). Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Gongola, J., Scurich, N., & Quas, J. A. (2017). Detecting deception in children: A meta-analysis. *Law and Human Behavior*, *41*(1), 44–54. https://doi.org/10.1037/lhb0000211

Granhag, P. A., & Hartwig, M. (2015). The strategic use of evidence technique: A conceptual overview. In P. A. Granhag, A. Vrij, & B. Verschuere (Eds.), *Detecting deception: Current challenges and cognitive approaches* (pp. 231–251). Wiley-Blackwell. https://doi.org/10.1002/9781118510001.ch10

Granhag, P. A., Hartwig, M., Giolla, E. M., & Clemens, F. (2014). Suspects' verbal counter-interrogation strategies: Towards an integrative model. *Detecting deception: Current challenges and cognitive approaches*, 293-313. https://doi.org/10.1002/9781118510001.ch13

Granhag, P. A., Rangmar, J., & Strömwall, L. A. (2015). Small cells of suspects: Eliciting cues to deception by strategic interviewing. *Journal of Investigative Psychology and Offender Profiling*, *12*(2), 127-141. https://doi.org/10.1002/jip.1413

Granhag, P. A., Strömwall, L. A., Willén, R. M., & Hartwig, M. (2013). Eliciting cues to deception by tactical disclosure of evidence: The first test of the Evidence Framing

Matrix. *Legal and Criminological Psychology, 18*(2), 341-355. https://doi.org/10.1111/j.2044-8333.2012.02047.x

Grossman, R., & Salas, E. (2011). The transfer of training: what really matters. *International Journal of Training and Development, 15*(2), 103-120. https://doi.org/10.1111/j.1468-2419.2011.00373.x

Guadagno, B., & Powell, M. (2012). E-simulations for the purpose of training forensic (investigative) interviewers. In *Professional education using e-simulations: Benefits of blended learning design* (pp. 71-86). IGI Global. https://doi.org/10.4018/978-1-61350-189-4.ch005

Haginoya, S., Sun, Y., Yamamoto, S., Mizushi, H., Yoshimoto, N., & Santtila, P. (2025). Improving questioning skills and use of supportive statements in simulated child sexual abuse interviews. *Applied Cognitive Psychology*, *39*(1), e70031. https://doi.org/10.1002/acp.70031

Haginoya, S., Yamamoto, S., Pompedda, F., Naka, M., Antfolk, J., & Santtila, P. (2020). Online simulation training of child sexual abuse interviews with feedback improves interview quality in Japanese university students. *Frontiers in Psychology*, *11*, 998. https://doi.org/10.3389/fpsyg.2020.00998

Hajian, S. (2019). Transfer of learning and teaching: A review of transfer theories and effective instructional practices. *IAFOR Journal of Education*, *7*(1), 93-111. https://eric.ed.gov/?id=EJ1217940

Hartwig, M., & Granhag, P. A. (2023). Strategic use of evidence: A review of the technique and its principles. In C. A. Meissner & J. P. Carpenter (Eds.), *Interviewing and interrogation:*

A review of research and practice since World War II (pp. 299-318). Legal-Tools.org. https://www.legal-tools.org/doc/h69sxw/

Hartwig, M., Granhag, P. A., & Luke, T. (2014). Strategic use of evidence during investigative interviews: The state of the science. *Credibility Assessment*, 1-36. https://doi.org/10.1016/B978-0-12-394433-7.00001-4

Hartwig, M., Granhag, P. A., & Strömwall, L. A. (2007). Guilty and innocent suspects' strategies during police interrogations. *Psychology, Crime & Law*, *13*(2), 213-227. https://doi.org/10.1080/10683160600750264

Hartwig, M., Granhag, P. A., Strömwall, L. A., & Kronkvist, O. (2006). Strategic use of evidence during police interviews: When training to detect deception works. *Law and Human Behavior*, *30*(5), 603. https://doi.org/10.1007/s10979-006-9053-9

Hartwig, M., Granhag, P. A., Strömwall, L. A., & Vrij, A. (2005). Detecting deception via strategic disclosure of evidence. *Law and Human Behavior*, *29*(4), 469-484. https://doi.org/10.1007/s10979-005-5521-x

Hartwig, M., Granhag, P. A., Stromwall, L., Wolf, A. G., Vrij, A., & Hjelmsäter, E. R. A. (2011). Detecting deception in suspects: Verbal cues as a function of interview strategy. *Psychology, Crime & Law*, *17*(7), 643-656. https://doi.org/10.1080/10683160903446982

Hill, C., Memon, A., & McGeorge, P. (2008). The role of confirmation bias in suspect interviews: A systematic evaluation. *Legal and Criminological Psychology*, *13*(2), 357-371. https://doi.org/10.1348/135532507X238682

Ioannou, M., & Hammond, L. (2015). The detection of deception within investigative contexts: Key challenges and core issues. *Journal of Investigative Psychology and Offender Profiling*, *12*(2), 107-118. https://doi.org/10.1002/jip.1433

James-Kangal, N., Memon, A., Colwell, K., Cole, L., Martin, M., & Wirsing, E. (2018). Innocent suspects lying by omission. *Journal of Forensic Psychiatry and Psychology*, *3*. https://doi.org/10.4172/2475-319X.1000133

Johnson, M., Magnussen, S., Thoresen, C., Lønnum, K., Burrell, L. V., & Melinder, A. (2015). Best practice recommendations still fail to result in action: A national 10-year follow-up study of investigative interviews in CSA cases. *Applied Cognitive Psychology*, *29*(5), 661-668. https://doi.org/10.1002/acp.3147

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*(4), 237–251. https://doi.org/10.1037/h0034747

Kassambara, A. (2023). *rstatix: Pipe-friendly framework for basic statistical tests* (R package version 0.7.2). https://CRAN.R-project.org/package=rstatix

Kubat, M. (1997). *Addressing the curse of imbalanced training sets: One-sided selection.* In *Proceedings of the 14th International Conference on Machine Learning* (pp. 179–186). *Morgan Kaufmann.*

Lamb, M. E. (2016). Difficulties translating research on forensic interview practices to practitioners: Finding water, leading horses, but can we get them to drink? *American Psychologist*, *71*(8). https://doi.org/10.1037/amp0000039

Lammerse, M., Hassan, S. Z., Sabet, S. S., Riegler, M. A., & Halvorsen, P. (2022, September). Human vs. GPT-3: The challenges of extracting emotions from child responses. In *2022 14th International Conference on Quality of Multimedia Experience (QoMEX)* (pp. 1-4). IEEE. http://doi.org/10.1109/QoMEX55416.2022.9900885

Lave, J., & Wenger, É. (1994). Situated learning: legitimate peripheral participation. *Man*, *29*(2), 487. https://doi.org/10.2307/2804509

Lenth, R. V. (2023). *emmeans: Estimated marginal means, aka least-squares means* (R package version 1.8.9). https://CRAN.R-project.org/package=emmeans

Levine, T. R. (2018). Scientific evidence and cue theories in deception research: reconciling findings from meta-analyses and primary experiments. *International Journal of Communication*, *12*, 19. https://doi.org/10.1093/hcr/hqz019

Li, S., Ahlgren, R., Wang, Y., Haginoya, S., Granhag, P. A., & Santtila, P. (2025). A serious game with avatar suspects can be used to train naive participants in the strategic use of evidence. *Journal of Forensic Psychology Research and Practice*, *25*(1), 146-171. https://doi-org.ezproxy.vasa.abo.fi/10.1080/24732850.2023.2299492

Li, S., Granhag, P., Shi, Y., Sun, Y., Nyman, T. J., Haginoya, S., & Santtila, P. O. (2025, October 26). Using Large Language Model Based AI Suspects to Train Strategic Use of Evidence: Preliminary Evidence of Transfer to Mock Suspect Interviews. Retrieved from https://osf.io/u7ekj/

Lüdecke, D. (2023). *sjPlot: Data visualization for statistics in social science* (R package version 2.8.15). https://CRAN.R-project.org/package=sjPlot

Luke, T. J., Hartwig, M., Joseph, E., Brimbal, L., Chan, G., Dawson, E., Jordan, S., Donovan, P., & Granhag, P. A. (2016). Training in the Strategic Use of Evidence technique: Improving deception detection accuracy of American law enforcement officers. *Journal of Police and Criminal Psychology*, *31*(4), 270–278. https://doi.org/10.1007/s11896-015-9187-0

Luke, T. J., Hartwig, M., Brimbal, L., Chan, G., Jordan, S., Joseph, E., Osborne, J., & Granhag, P. A. (2013). Interviewing to elicit cues to deception: Improving strategic use of evidence with general-to-specific framing of evidence. *Journal of Police and Criminal Psychology*, *28*(1), 54–62. https://doi.org/10.1007/s11896-012-9113-7

May, L., Raible, Y., Gewehr, E., Zimmermann, J., & Volbert, R. (2023). How often and why do guilty and innocent suspects confess, deny, or remain silent in police interviews?. *Journal of Police and Criminal Psychology*, *38*(1), 153-164. https://doi.org/10.1007/s11896-022-09522-w

McCallum, D. V. (1985). *Educational quality: An analysis of the effects of process versus outcome feedback on instructor performance* (Publication No. 8606169) [Doctoral dissertation, Western Michigan University]. *ProQuest Dissertations & Theses Global*.

Meissner, C. A., Redlich, A. D., Michael, S. W., Evans, J. R., Camilletti, C. R., Bhatt, S., & Brandon, S. (2014). Accusatorial and information-gathering interrogation methods and their effects on true and false confessions: A meta-analytic review. *Journal of Experimental Criminology*, *10*, 459-486. https://doi.org/10.1007/s11292-014-9207-6

Oleszkiewicz, S., Madfors, M., Jones, M., & Vredeveldt, A. (2023). Proximity-based evidence disclosure: Providing an operational purpose for disclosing evidence in investigative interviews. *Psychology, Public Policy, and Law*, *29*(3), 302–319. https://doi.org/10.1037/law0000396

Oleszkiewicz, S., & Watson, S. J. (2021). A meta-analytic review of the timing for disclosing evidence when interviewing suspects. *Applied Cognitive Psychology*, *35*(2), 342–359. https://doi.org/10.1002/acp.3767

Patel, D., Timsina, P., Raut, G., Freeman, R., Levin, M., Nadkarni, G.N., Glicksberg, B.S., & Klang, E. (2024). Exploring temperature effects on large language models across various clinical tasks. *medRxiv*. https://www.medrxiv.org/content/10.1101/2024.07.22.24310824v1

Pompedda, F. (2018, March 27). *Training in Investigative Interviews of Children: Serious Gaming Paired with Feedback Improves Interview Quality* [Doctoral dissertation (article based)]. Åbo Akademi - Åbo Akademi University. https://www.doria.fi/handle/10024/152565

Pompedda, F., Antfolk, J., Zappalà, A., & Santtila, P. (2017). A combination of outcome and process feedback enhances performance in simulations of child sexual abuse interviews using avatars. *Frontiers in Psychology*, *8*, 1474. https://doi.org/10.3389/fpsyg.2017.01474

Pompedda, F., Zappalà, A., & Santtila, P. (2015). Simulations of child sexual abuse interviews using avatars paired with feedback improves interview quality. *Psychology, Crime & Law*, *21*(1), 28-52. https://doi.org/10.1080/1068316X.2014.915323

Pompedda, F., Zhang, Y., Haginoya, S., & Santtila, P. (2022). A mega-analysis of the effects of feedback on the quality of simulated child sexual abuse interviews with avatars. *Journal of Police and Criminal Psychology*, *37*(3), 485-498. https://doi.org/10.1007/s11896-022-09509-7

Powell, M. B., Brubacher, S. P., & Baugerud, G. A. (2022). An overview of mock interviews as a training tool for interviewers of children. *Child Abuse & Neglect*, *129*, 105685. https://doi.org/10.1016/j.chiabu.2022.105685

Røed, R. K., Baugerud, G. A., Hassan, S. Z., Sabet, S. S., Salehi, P., Powell, M. B., Riegler, M. A., Halvorsen, P., & Johnson, M. S. (2023). Enhancing questioning skills through child avatar chatbot training with feedback. *Frontiers in Psychology*, 14, 1198235. https://doi.org/10.3389/fpsyg.2023.1198235

Routray, S. K., Javali, A., Sharmila, K. P., Jha, M. K., Pappa, M., & Singh, M. (2023). Large language models (LLMs): Hypes and realities. In *2023 International Conference on Computer Science and Emerging Technologies (CSET)*, Bangalore, India (pp. 1-6). IEEE. https://doi.org/10.1109/CSET58993.2023.10346621

Sandham, A. L., Dando, C. J., Bull, R., & Ormerod, T. C. (2022). Improving professional observers' veracity judgements by tactical interviewing. *Journal of Police and Criminal Psychology*, *37*(2), 279-287. https://doi.org/10.1007/s11896-020-09391-1

Seitz, S. (2016). Pixilated partnerships, overcoming obstacles in qualitative interviews via Skype: A research note. *Qualitative Research*, *16*(2), 229-235. https://doi.org/10.1177/1468794115577011

Shahade, A. K., & Deshmukh, P. V. (2024, October). Enhancing natural language processing: A comprehensive review of retrieval augmented generation. In *2024 4th International Conference on Sustainable Expert Systems (ICSES)* (pp. 609-611). IEEE. https://doi.org/10.1109/ICSES63445.2024.10763224

Shanahan, M., McDonell, K., & Reynolds, L. (2023). Role play with large language models. *Nature*, *623*(7987), 493–498. https://doi.org/10.1038/s41586-023-06647-8

Shaunwei. (2023, July 20). *RealChar: Your Realtime AI Character*. GitHub. Retrieved from https://github.com/Shaunwei/RealChar

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*(1), 137–149. https://doi.org/10.3758/BF03207704

Tekin, S., Granhag, P. A., Strömwall, L., Giolla, E. M., Vrij, A., & Hartwig, M. (2015). Interviewing strategically to elicit admissions from guilty suspects. *Law and Human Behavior, 39*(3), 244–252. https://doi.org/10.1037/lhb0000131

Tohvelmann, M. L., Kask, K., Palu, A., Haginoya, S., & Santtila, P. (2025). Providing feedback in simulated investigative interviews with adult witness avatars increases the use of free recall and open questions. *International Journal of Police Science & Management*, *27*(2), 182-198. https://doi.org/10.1177/14613557241310014

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). Springer. https://www.stats.ox.ac.uk/pub/MASS4/

Volbert, R., & Banse, R. (2014). How can psychological research assist legal practice. *European Psychologist*, *19*(3), 159-161. https://doi.org/10.1027/1016-9040/a000209

Vrij, A. (2004). Why professionals fail to catch liars and how they can improve. *Legal and Criminological Psychology*, *9*, 159 –181. https://doi.org/10.1348/1355325041719356

Vrij, A., Meissner, C. A., Fisher, R. P., Kassin, S. M., Morgan III, C. A., & Kleinman, S. M. (2017). *Psychological perspectives on interrogation. Perspectives on Psychological Science*, *12*(6), 927-955. https://doi.org/10.1177/1745691617706515

Vrij, A., & Granhag, P. A. (2007). Interviewing to detect deception. In S. A. Christianson (Ed.), Offenders' *memories of violent crimes* (pp. 279-304). John Wiley & Sons. https://doi.org/10.1002/9780470713082

Vrij, A., & Granhag, P. A. (2012). Eliciting cues to deception and truth: What matters are the questions asked. *Journal of Applied Research in Memory and Cognition*, *1*(2), 110-117. https://doi.org/10.1016/j.jarmac.2012.02.004

Yu, J., Zhang, X., Xu, Y., Lei, X., Guan, X., Zhang, J., Hou, L., Li, J., & Tang, J. (2022). XDAI: a tuning-free framework for exploiting pre-trained language models in knowledge grounded dialogue generation. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/3534678.3539135

Zekiroski, H., Powell, M. B., Brubacher, S. P., & Curtis, P. (2024). Evaluation of a police training programme designed to enhance open-ended questions with adult witnesses. *International Journal of Police Science & Management*, 14613557241233857. https://doi.org/10.1177/14613557241233857

Table 1

The Compensation Scheme

Interviewers	Liars	Truth-tellers
Judging as "Liars"	(**100 + 50**, 100 − 50)	(**100 − 50**, 100 − 50)
Judging as "Truth-tellers"	(**100 − 50**, 100 + 50)	(**100 + 50**, 100 + 50)

Note. Compensation for Interviewers is presented in bold text.

Table 2

The Examples of False Responses from AI Suspects

False Types	Questions (Question Types)	Responses From Suspect Avatar in Li et al. (2024).	Responses From AI Supsect in the Present Study
False Admission	"Have you bought a suitcase recently?" (Specific question asked before presenting evidence with a strong source.)	"No."	"Yes, I have bought a suitcase."
False Denial (or Avoidance)	"There is a lost phone and wallet, and your fingerprints were found on the box. How do you explain that?" (Evidence with a strong source.)	"Now I remember, I just pushed it a little bit to the side."	"I don't know anything about the missing phone and wallet, and I didn't leave any fingerprints on the box. I have no idea about any of this."
Other Types of False Responses	"Do you have any plans to visit the United States?" (Conversation not related to the suspect's whereabouts and activities.)	"I had to cancel the trip because I couldn't afford it."	"Yes, I planned to visit the United States. I've always wanted to see those famous landmarks in New York City. However, due to financial reasons, I had to cancel the trip."
	"When is your trip scheduled?" (Conversation not related to the suspect's whereabouts and activities.)	"I don't remember."	"My planned trip is in September next year. I'm planning to go on vacation to Spain."

Note. "False Types" do not represent genuinely false responses from the AI Suspects. Rather, "False" refers to inconsistencies between the response rules of suspect avatars described in Li et al. (2024) and the response strategies adopted by the AI Suspects in the present study.

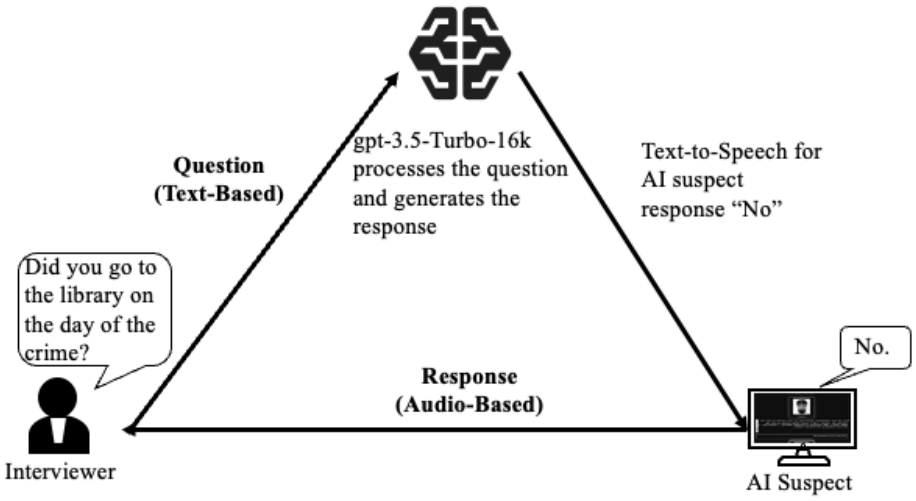
Figure 1

Example of the Evidence Framing Matrix (EFM) in a Case Where Bloody Fingerprints Placed a Suspect at the Crime Scene (a Villa in City A).

	Low Specificity	**High Specificity**
Strong Source	We have fingerprints that you were in City A	The fingerprints we have obtained indicate that you were in the villa in City A
Weak Source	The information we have obtained indicates that you were in City A	We have information that you were in the villa in City A

Figure 2

Illustration of the Workflow of Interaction with AI Suspects



Note. The icon representing gpt-3.5-Turbo-16k was obtained from https://icons8.com/

Figure 3

The Prompt Structure for AI Suspect in Terrorism Case (System Prompt)

Suspect's Personal Information
Name: Charlie
Age: 26
Religion: Non-religious
Residence: King's Wood, London
Relationship status: Single, lives alone
Family: Parents reside in Morocco
Job: Former employee at Aldi supermarket (near King's Wood)
Hobbies:
Playing FIFA video games and watching football TV programs

Case Information
Type: Terrorism case (October 18, 2019)
Crime Sequence:
(1) Planned to smuggle liquid bombs onto a commercial flight from London Heathrow Airport to the United States;
(2) Bought a green suitcase from Luggage Pros;
(3) Buried a green suitcase containing materials commonly used to produce bombs and detonators, along with other packages, near King's Wood, High Wycombe.
Available Evidence: CCTV footage showing purchase of green suitcase at Luggage Pros on October 18, 2019 (21 days ago).

Responding Strategies
Before Strong-Source Evidence Presentation (No Evidence or Weak Source Evidence)
Free-recall questions → "omission" or "denial"
Specific questions → "omission" or "denial"
After Strong-Source Evidence Presentation (Strong Source Evidence)
Free-recall questions → only admit whereabouts and activities with plausible explanations that are supported by strong-source evidence
Specific questions → only admit whereabouts and activities with plausible explanations that are supported by strong-source evidence
Evidence Presentation
Evidence with weak source → activate "omission" or "denial"
Evidence with strong source → activate "admission"
Incriminating Utterances
Incriminating Utterances → activate "denial"
Questions about Suitcase Description
Price → 60 pounds
Description of the suitcase → a green suitcase, which is big enough to carry things
Note: Suspects are allowed to fabricate any information or details about the suitcase to maintain innocence.

Note. All responding strategies were presented through explanations and examples in the prompts.

Figure 4

The Prompt Structure for AI Suspect in Theft Case (System Prompt)

Suspect's Personal Information
Name: Simon
Age: 34
Religion: Non-religious
Residence: Winchmore Hill, North London
Relationship status: Living with girlfriend Natalie
Family: Parents reside in Leicester
Job: Automotive mechanic in M&A Motors (Harringay area)
Hobbies:
Watching football TV programs

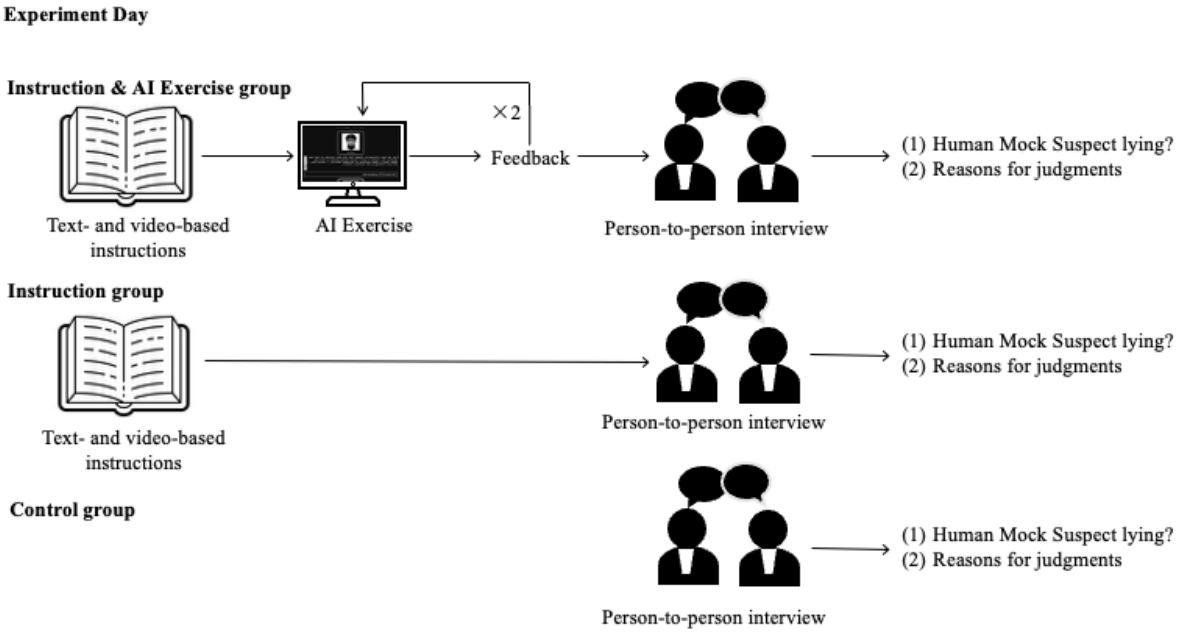
Case Information
Type: Theft case (October 22, 2019)
Crime Sequence:
(1) Went to a library to look at Nemesis Games;
(2) Found a cardboard box and moved it to access bookshelf;
(3) Confirmed no surveillance cameras around the bookshelf, took the phone and wallet from the box to your backpack, and left the library.
Available Evidence: DNA evidence showing presence in the library on Elm Park Road, Enfield, London, October 22 (last Thursday), with fingerprints found on a box on the shelf in the sci-fi section.

Responding Strategies
Before Strong-Source Evidence Presentation (No Evidence or Weak Source Evidence)
Free-recall questions → "omission" or "denial"
Specific questions → "omission" or "denial"
After Strong-Source Evidence Presentation (Strong Source Evidence)
Free-recall questions → only admit whereabouts and activities with plausible explanations that are supported by strong-source evidence
Specific questions → only admit whereabouts and activities with plausible explanations that are supported by strong-source evidence
Evidence Presentation
Evidence with weak source → activate "omission" or "denial"
Evidence with strong source → activate "admission"
Incriminating Utterances
Incriminating Utterances → activate "denial"
Note: Suspects are not allowed to voluntarily mention specific involved items (i.e., the phone and wallet); but are allowed to fabricate information and details to maintain innocence.

Note. All responding strategies were presented through explanations and examples in the prompts.

Figure 5

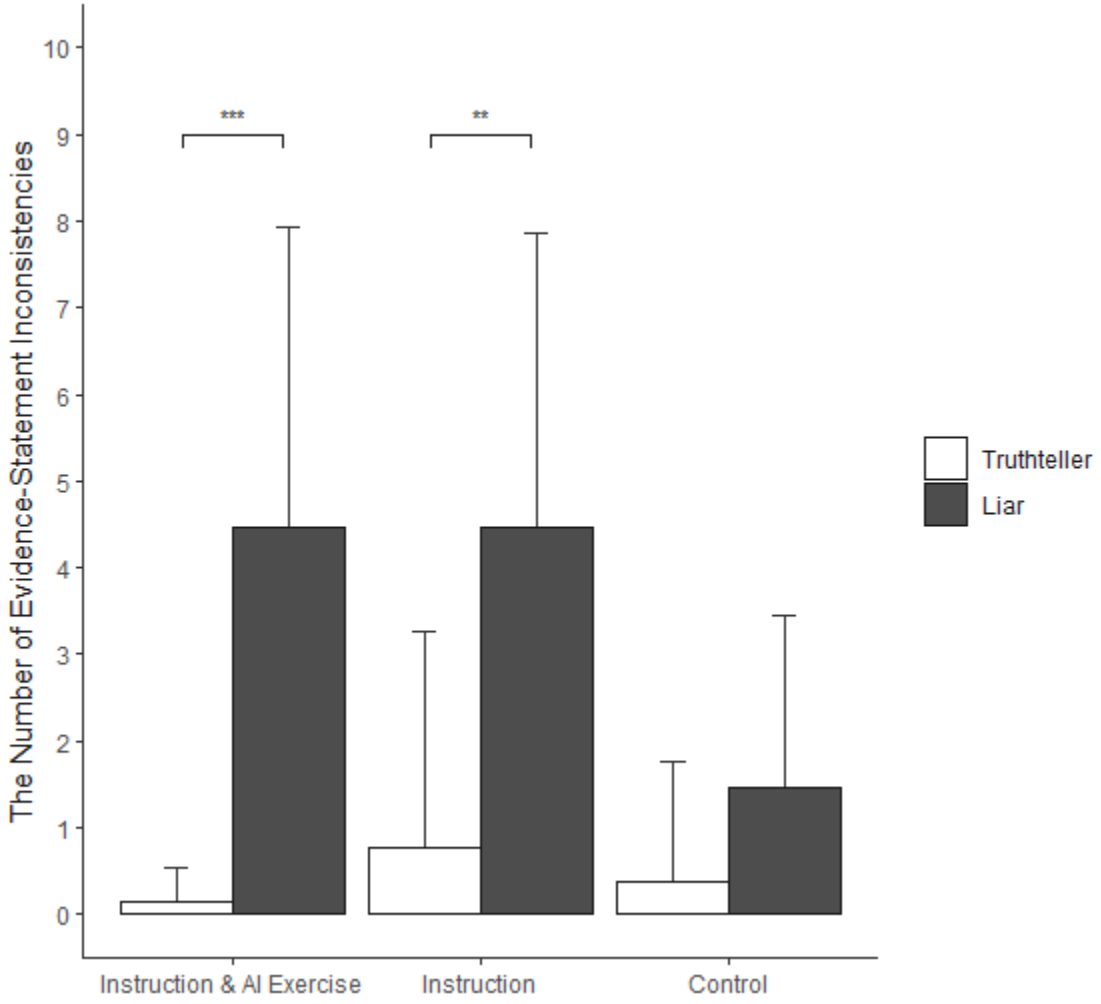
Overview of Experimental Design and Main Procedures.



Note. The icon representing text- and video-based instructions was obtained from https://icons8.com/

Figure 6

Pairwise Comparisons Between Lying and Truthful Human Mock Suspects for the Number of Evidence-Statement Inconsistencies, Separated by Experiment Groups



Note. $^{***}p < .01$; $^{***}p < .001$