Teacher's Key

How Humans and AI Detect Lies: What the Research Shows (Level C1)

Vocabulary Matching

1 veracity = **N**
2 intuition = **A**
3 stereotype = **H**
4 susceptible = **L**
5 pervasive = **G**
6 heuristic = **M**
7 corroborate = **I**
8 discriminate = **E**
9 countermeasure = **D**
10 fabricate = **O**
11 outperform = **J**
12 impair = **K**
13 plausible = **B**
14 caveat = **C**
15 generalize = **F**

Fill in the Blank

1 veracity
2 intuition
3 stereotype
4 susceptible
5 pervasive
6 heuristic
7 corroborated
8 discriminate
9 fabricate
10 outperform
11 impair
12 plausible
13 caveat
14 generalize

True or False

1 **True.** Real-world evidence is harder to collect.
2 **False.** Many behaviors showed no discernible link, or only a weak one.
3 **True.** The average was 54 percent correct.
4 **False.** Intuition performed at chance; relying on detail worked better.
5 **True.** It found accuracy well above chance but far from perfect.
6 **False.** They learn statistical differences between honest and invented texts.
7 **True.** The researchers collected 1,640 statements about intentions.
8 **False.** Only 31.76 percent asked for it.
9 **True.** It also treats polygraph-type tools at borders as high-risk.
10 **False.** It says this has not been shown in the studies reviewed.

Language Focus: Comparatives with Degree Modifiers

Example answer: The model was **considerably more accurate than** the human judges.

Comprehension Questions: Model Answers

1 Because most studies define better as accuracy on staged tasks in which some participants were told to lie, so the results may not describe real cases, and the two kinds of judge are often tested on different tasks.

2 People rely on behavioral cues such as gaze aversion, but DePaulo and colleagues found many cues had no discernible or only a weak link to deception. Bond and DePaulo found 54 percent average accuracy: 61 percent for truths and 47 percent for lies, a pattern that suggests a truth bias, which makes people miss lies.

3 Using a simple rule such as level of detail, especially detail that can be corroborated, let judges distinguish lies from truths where intuition performed at chance. The polygraph measures physiology and performed far better than chance but far from perfectly; the review could not confirm it survives countermeasures.

4 They analyze language and learn statistical differences between honest statements and statements that participants were asked to fabricate. They do not detect lies as such because they pick up patterns in a dataset, not lying itself.

5 Humans: near chance or below (46.47 percent in one experiment); algorithms: 69 percent (Kleinberg and Verschuere), 66.86 percent (von Schenk et al.), and about 71 to 83 percent for a fine-tuned language model on familiar statements. When humans could overrule the Kleinberg and Verschuere model, accuracy fell to chance, so in that study human input impaired performance.

6 Any three of: accuracy fell to about 49 to 56 percent on unseen statement types; datasets are low-stakes, invented stories; real lies are often embedded in true accounts; base rates produce many false accusations; a 2025 preprint by Barkett and colleagues reported that several advanced language models judged deceptive statements poorly.

7 If few people lie, even an accurate test flags many honest people, so most of those flagged may be truthful. In the National Research Council's hypothetical screening of 10,000 employees with 10 spies, a polygraph set to catch eight spies would also fail about 1,600 innocent people, so almost everyone who failed would be innocent.

8 It treats polygraph-type tools in migration and border control as high-risk and bans emotion inference in workplaces and schools, which suggests regulators see these tools as risky and of uncertain validity.

9 Open answer. Firm: algorithms beat intuitive judgment on narrow text-based lab tasks; no method has been shown reliable enough to justify accusing someone on its own. Not firm: that machines are better with real suspects, real stakes, or unfamiliar liars; the reading says this has not been shown in the studies reviewed.

10 Not Given. The reading does not mention police officers or their training.

11 No. The writer says the evidence supports only a limited claim: machines are better in specific settings, and their performance with real suspects, real stakes, and unfamiliar liars has not been shown.

12 To let the reader see both the promise and the limits, so that the conclusion is limited rather than a claim that AI is simply better.

Notes on the Evidence

- Sources checked (September 2026). Bond and DePaulo 2006, Personality and Social Psychology Review (abstract: 206 documents, 24,483 judges, 54 percent, 47 percent of lies, 61 percent of truths, more accurate with audible than visible lies). DePaulo et al. 2003, Psychological Bulletin (abstract: 158 cues). Global Deception Research Team 2006 (abstract: 75 countries; gaze aversion is the dominant stereotype). Verschuere et al. 2023, Nature Human Behaviour (abstract: nine studies; intuition at chance; relying on detailedness discriminated lies from truths). National Research Council, The Polygraph and Lie Detection (report released October 2002, book dated 2003): summary chapter and the Academies' 2002 news release. Hauch et al. 2015 (abstract: 44 studies). Kleinberg and Verschuere 2021, Acta Psychologica (abstract rechecked: corpus of truthful and deceptive answers about intentions, n = 1,640; machine learning 69 percent; hybrid-overrule at chance; truth bias proposed as the explanation). Von Schenk et al., TSE working paper 23-1448 (June 2023; abstract rechecked: classifier 66.86 percent, human 46.47 percent, N = 2,040, 31.76 percent requests), later published in iScience (2024), whose record at TSE lists a revision dated 10 June 2026. Loconte et al. 2023, Scientific Reports (Scenario 1 accuracy 71 to 83 percent; unseen statement type about 49 to 56 percent). Barkett, Long, and Thakur 2025, arXiv 2506.21561 (abstract: eight LLMs, 4,800 judgments; o4-mini, GPT-4.1, and R1 good at truth accuracy, poor at deception accuracy).
- Evidence limits. The 54 percent figure is an average across many studies and hides variation. The 46.47 percent human accuracy in the von Schenk experiment is specific to that task; a Max Planck project page rounds the figures to 67 and 50 percent, so check the published paper before quoting exact decimals. The Kleinberg and Verschuere abstract gives n = 1,640 without saying whether this counts statements or participants; the reading says statements. Verschuere et al. is disputed: a comment by Zickfeld and Elbaek reports a modest gain in Study 8 (58.93 against 54.26 percent). The 67 to 83 percent range in paragraph 7 combines three studies with different tasks and applies only to familiar statement types. The Barkett preprint tested general-purpose models by prompting, not fine-tuned detectors. No study reviewed compared algorithms with trained interviewers in real cases; the search was not exhaustive.
- Regulation. Annex III point 7 of the EU AI Act lists polygraph-type tools used by authorities in migration, asylum, and border control as high-risk. Article 5(1)(f) bans emotion inference in workplaces and education, with medical and safety exceptions, and has applied since 2 February 2025. The Digital Omnibus on AI, Regulation (EU) 2026/1744 (Parliament vote 16 June 2026, Council approval 29 June 2026, Official Journal 24 July 2026, in force 27 July 2026, per secondary legal sources), fixed 2 December 2027 as the application date for Annex III high-risk obligations, replacing 2 August 2026. The Official Journal text itself was not opened. Not legal advice.
- Revision (September 2026). Corrected the National Research Council claim: the report did not estimate the outcome of a real government screening; it gave a hypothetical of 10,000 employees with 10 spies, in which a test set to catch eight spies would also fail about 1,600 innocent people. Named the unattributed 2025 preprint (Barkett and colleagues) and the fine-tuned model study (Loconte and colleagues, Italy). Hedged the Kleinberg and Verschuere finding to that study and attributed the truth-bias explanation to its authors. Paragraph 7 now reports 67 to 83 percent, consistent with paragraph 5 (the earlier 67 to 69 percent left out the language-model result), and drops currently. Updated the AI Act wording after the Omnibus was adopted and published. Replaced credibility with intuition in the vocabulary because credibility also fitted the veracity fill-in blank. Relabeled the Language Focus as comparatives with degree modifiers and replaced two examples that did not show the structure. Rewrote the Yes/No/Not Given item, whose old answer (No) was arguably Not Given. Hindi and Arabic translations have not been checked by a native speaker.