Hallucination Detection Checklist

HALLUCINATION DETECTION CHECKLIST

How to Spot When AI Is Making Things Up

WHAT IS HALLUCINATION?

Hallucination: AI generates text that sounds plausible, reads confidently, and follows proper formatting—but is factually incorrect.

The dangerous part: AI doesn't know it's wrong.

AI hallucinations are NOT: - Intentional deception (AI isn't lying) - Random nonsense (output sounds very believable) - Obvious errors (hallucinations look professional)

AI hallucinations ARE: - Confident fabrications - Plausible-sounding false information - Pattern matching without factual grounding - Generated text that fills gaps with invented content

Why hallucination is worse than lying:

Liar knows: - What the truth is - That they're deviating from it - The difference between truth and lies

Hallucinating AI doesn't know: - What is true vs false - That it's making things up - The difference between real and fabricated

AI generates what sounds right, not what is right.

RED FLAGS FOR HALLUCINATION

🚩 RED FLAG #1: AI never says "I don't know"

Warning sign: AI has confident answer for everything

Why this matters: Complex medical situations have uncertainty. If AI expresses no uncertainty, it's probably hallucinating confidence it doesn't have.

Test: Ask about something obscure or ask "What are you uncertain about?"

Good AI: “I’m uncertain about [specific aspects]”

Hallucinating AI: “I’m very confident this is [answer]”

🚩 RED FLAG #2: AI can’t provide specific sources

Warning sign: Vague sourcing

Examples of vague sourcing: - “Studies show…” - “Research indicates…” - “Medical experts agree…” - “It’s well known that…” - “According to medical literature…”

Why this matters: These could be real patterns AI learned, or complete fabrications. You can’t verify.

Test: Ask “Which studies? Published where? Can you provide the citation?”

Good AI: “American Heart Association 2021 Guidelines on [specific topic], published in [journal]”

Hallucinating AI: Can’t provide specific citation, or provides citation that doesn’t exist

---

🚩 RED FLAG #3: Overly specific without caveats

Warning sign: Very detailed, specific information with zero uncertainty

Examples: - “This medication is 87.3% effective for your condition” - “You should take exactly 750mg twice daily” - “This will resolve in 14-21 days”

Why this matters: Medical information usually has ranges, caveats, individual variation. Overly precise numbers without context suggest hallucination.

Test: Ask “How certain are you?” or “What variables affect this?”

Good AI: Acknowledges ranges, individual variation, context-dependence

Hallucinating AI: Maintains perfect precision without caveats

---

🚩 RED FLAG #4: Internal contradictions

Warning sign: AI contradicts itself within same response or across responses

Examples: - First says “avoid this medication with condition X” - Later says “this medication is safe for condition X”

Why this matters: If AI is generating plausible-sounding text without factual grounding, internal consistency fails.

Test: Read carefully for contradictions. Ask same question multiple times.

Good AI: Consistent information each time

Hallucinating AI: Different answers to same question

🚩 RED FLAG #5: Fabricated citations

Warning sign: Citations provided but don't exist

Examples: - Journal article that was never published - Authors who didn't write that paper - Study with fabricated data - Conference proceedings that never happened

Why this matters: This is pure hallucination. AI generated plausible-looking citations to support fabricated claims.

Test: Verify citations in PubMed, Google Scholar, or journal websites

Good AI: All citations are verifiable and actually say what AI claims

Hallucinating AI: Citations don't exist, or say something different than claimed

---

🚩 RED FLAG #6: Sounds too perfect/comprehensive

Warning sign: Answer is suspiciously complete and perfectly formatted

Why this matters: Real medical information often has gaps, controversies, unknowns. Perfect comprehensive answers suggest AI is filling gaps with fabrications.

Test: Ask "What's controversial about this?" or "What don't we know?"

Good AI: Acknowledges gaps in evidence, controversies, unknowns

Hallucinating AI: Presents everything as settled fact

🚩 RED FLAG #7: Can't explain reasoning

Warning sign: AI provides answer but can't explain the logic

Test: Ask "Why?" or "How did you arrive at this conclusion?"

Good AI: Explains reasoning step-by-step with verifiable logic

Hallucinating AI: Cannot explain, or explanation doesn't follow logically

---

🚩 RED FLAG #8: Novel combinations

Warning sign: AI combines real things in ways that don't exist

Examples: - Real medication + fabricated indication - Real procedure + fabricated technique - Real condition + fabricated treatment

Why this matters: AI knows these elements exist separately, but invents connections between them.

Test: Verify each element AND their relationship

Example of hallucination: “Metformin is FDA-approved for treating migraines” - Metformin = real drug ✓ - FDA-approved = real concept ✓ - Treating migraines = real treatment goal ✓ - Metformin FDA-approved for migraines = FALSE ✗

VERIFICATION PROTOCOL

When you get medical information from AI, verify:

STEP 1: Source Check

☐ Are sources specific? (Journal name, year, authors)
☐ Can you find the cited source? (PubMed, Google Scholar)
☐ Does source actually say what AI claims?

STEP 2: Cross-Reference

☐ Check reputable medical sources: - Mayo Clinic - NIH/MedlinePlus - CDC - Medical specialty organization websites - Uptodate (if you have access)

☐ Do multiple reliable sources agree?

STEP 3: Consistency Check

☐ Ask AI same question multiple times
☐ Do you get consistent answers?
☐ Any internal contradictions?

STEP 4: Uncertainty Check

☐ Ask “What are you uncertain about?”
☐ Ask “What’s controversial about this?”
☐ Does AI acknowledge limitations?

STEP 5: Logic Check

☐ Ask AI to explain reasoning
☐ Does reasoning make sense?
☐ Are there logical leaps?

THE “I DON’T KNOW” TEST

Test AI’s ability to acknowledge limitations:

Ask questions with progressively less available information:

Level 1: Common knowledge

“What is hypertension?”

Expected: Confident, accurate answer

Level 2: Specific but standard

“What are treatment guidelines for Stage 2 hypertension?”

Expected: Confident answer with sources

Level 3: Nuanced

“How should treatment differ for 80-year-old with multiple comorbidities?”

Expected: Some uncertainty, acknowledges individual variation

Level 4: Ambiguous

“What’s the optimal blood pressure target for my specific situation?”

Expected: “I cannot determine this without knowing your complete medical history, risk factors, and comorbidities. This requires physician evaluation.”

Level 5: Unknowable remotely

“Do I have hypertension?”

Expected: “I cannot diagnose you without measuring your blood pressure. You need in-person evaluation.”

If AI confidently answers Level 4-5 questions, it’s hallucinating.

REAL EXAMPLES OF DANGEROUS HALLUCINATIONS

Example 1: Melanoma misdiagnosis

Patient: [Sends photo of changing mole] “What is this?”

AI hallucination: “This appears to be a benign seborrheic keratosis. These are harmless skin growths. You can use vitamin E oil to help it fade.”

Reality: Was melanoma. Delay caused disease progression from Stage 1A to Stage 3A.

Red flags: - AI diagnosed from photo (cannot examine texture, thickness, borders) - Perfect confidence (no uncertainty about visual diagnosis) - Specific treatment (vitamin E oil has no evidence for melanomas OR seborrheic keratosis)

Example 2: Fabricated drug interaction

Patient: “Can I take warfarin with ibuprofen?”

AI hallucination: “A 2018 study by Chen et al in JAMA showed this combination is safe with monitoring. The interaction risk is only 2-3% in most patients.”

Reality: - Warfarin + ibuprofen increases bleeding risk significantly - No study by “Chen et al” in JAMA in 2018 on this topic - AI fabricated citation to support dangerous advice

Red flags: - Too-specific citation (year, author, journal) - Overly precise risk percentage - Contradicts known drug interaction

Example 3: Pediatric dosing error

Parent: “My 2-year-old has an ear infection. The pharmacy gave me amoxicillin. How much should I give?”

AI hallucination: “Standard dosing for amoxicillin is 500mg three times daily for 10 days.”

Reality: That’s ADULT dosing. Pediatric dosing is weight-based, typically 40-50mg/kg/day divided into doses. A 2-year-old weighing 12kg needs ~240mg per dose, not 500mg.

Red flags: - No question about child’s weight - Adult dose given for child - No acknowledgment that pediatric dosing requires calculation

---

HALLUCINATION PREVENTION

How to reduce risk of getting hallucinated information:

1. Use AI with validated knowledge bases - Content-controlled AI (like TheDude built on StatPearls) - AI limited to verified medical sources - Systems that say “I don’t know” when outside knowledge base

2. Ask the Five Essential Questions - “What are you basing this on?” - “What can you NOT detect remotely?” - “What would require emergency evaluation?” - “What are you uncertain about?” - “What should I ask my actual doctor?”

3. Verify before trusting - Never act on AI medical advice without verification - Cross-reference with reputable sources - Confirm with actual physician

4. Test for hallucination - Ask same question multiple ways - Ask for sources and verify them - Test with “I don’t know” questions

5. Be suspicious of perfection - If answer seems too complete, too confident, too specific → verify heavily - Real medical information has nuance, uncertainty, gaps

---

WHAT TO DO IF YOU DETECT HALLUCINATION

If AI has given you information you suspect is hallucinated:

IMMEDIATE ACTIONS:

□ STOP following AI's advice immediately

□ Do NOT take actions based on this information

□ If about current medical situation → call your doctor

□ If emergency situation → call 911 regardless of AI advice

VERIFICATION ACTIONS:

□ Cross-check with reputable medical sources

□ Call your doctor's office for clarification

□ Verify any citations AI provided

□ Ask different AI or different approach

LEARNING ACTIONS:

□ Note what made you suspicious

□ Remember AI system is unreliable for medical advice

□ Use AI only for general education, not medical decisions

□ Share experience with others to prevent similar problems

THE BOTTOM LINE

AI hallucination is:

- Common in general-purpose AI systems
- Dangerous in medical contexts
- Difficult to detect (sounds plausible)
- Not intentional (AI doesn't know it's wrong)

Protect yourself by:

- Never trusting AI medical advice without verification
- Asking the Five Essential Questions
- Verifying sources and citations
- Testing for hallucination signs
- Using AI only for education, not medical decisions

Remember:

AI that sounds confident might be confidently wrong.

Verify. Every. Time.

QUICK REFERENCE

Green Flags (More likely to be accurate):

✓ Specific, verifiable sources
✓ Acknowledges uncertainty
✓ Says “I don’t know” when appropriate
✓ Consistent across multiple queries
✓ Citations actually exist and say what AI claims
✓ Recommends physician evaluation for diagnosis/treatment

Red Flags (Possible hallucination):

✗ Vague sourcing (“studies show”)
✗ Never expresses uncertainty
✗ Never says “I don’t know”
✗ Internal contradictions
✗ Fabricated citations
✗ Overly specific without caveats
✗ Too perfect/comprehensive
✗ Can’t explain reasoning

If you see 3+ red flags → HIGH RISK of hallucination

Do NOT act on this information without verification

AI hallucination can literally kill you in medical contexts. Take verification seriously.

From: AI in the Exam Room - Patient Education Curriculum
Module 4: The Hallucination Problem (When AI Makes Stuff Up)