Handling Hallucinations

Fact-Checking AI

ITAG Skillnet AI Advantage

90-Minute Lunch & Learn

Learn to identify, prevent, and verify AI-generated content

Today's Agenda

10 min	What Are AI Hallucinations?
15 min	Why Hallucinations Happen
15 min	The VERIFY Framework
20 min	Hands-On: Spotting & Fixing Hallucinations
15 min	Prevention Strategies
10 min	Tools & Best Practices
5 min	Q&A and Wrap-Up

Learning Objectives

Understand

- What AI hallucinations are and why they occur
- The different types of hallucinations
- Risk factors that increase hallucination likelihood

Apply

- The VERIFY framework for fact-checking
- Prevention techniques in your prompts
- Tools to validate AI-generated content

By the end: You'll have practical skills to confidently use AI while maintaining accuracy and trust.

What is an AI Hallucination?

When AI generates content that is:

Factually Incorrect

States something false as if it were true

Fabricated

Invents sources, quotes, or data that don't exist

Confidently Wrong

Presents errors with high certainty

Key Insight: AI doesn't "know" it's wrong. It generates plausible-sounding text based on patterns, not understanding.

Real-World Hallucination Examples

⚠ Fabricated Citation

"According to a 2023 study published in the Journal of Business Analytics by Dr. Sarah Mitchell..."

The journal, study, and author don't exist

⚠ Invented Statistics

"Research shows that 73% of companies using AI report a 45% increase in productivity."

These specific numbers were fabricated

⚠ False Historical Facts

"The Treaty of Dublin, signed in 1847, established..."

No such treaty exists

⚠ Misattributed Quotes

"As Einstein famously said, 'The definition of insanity is doing the same thing...'"

Einstein never said this

The Scale of the Problem

3-27%
Hallucination rate in LLMs (varies by task type)

40%+
Of legal citations in one study were fabricated

$0
Cost for AI to sound confident about wrong answers

Notable Case: In 2023, lawyers were sanctioned after submitting a brief with AI-generated fake case citations. The AI had invented 6 non-existent court cases.

Why Do Hallucinations Happen?

How LLMs Actually Work

- Pattern matching - not knowledge retrieval
- Next-token prediction - optimized for plausibility
- No fact database - can't verify claims
- Training data gaps - incomplete world knowledge

Key Insight

LLMs are designed to generate *plausible* text, not *accurate* text.

They don't distinguish between:

- What they "know" vs. what they're generating
- Verified facts vs. plausible-sounding content

Types of Hallucinations

Intrinsic Hallucinations

Contradicts the source/input provided

Input: "The meeting is on Tuesday"

AI Output: "I've scheduled the meeting for Wednesday"

Extrinsic Hallucinations

Adds information that cannot be verified

Input: "Summarize this article"

AI Output: "The article also mentions..." (content not in article)

Hallucination Type	Risk Level	Detection Difficulty
Fabricated citations/sources	High	Medium - requires verification
Invented statistics	High	Hard - sounds authoritative
Outdated information	Medium	Easy - check dates
Subtle factual errors	Medium	Hard - requires expertise

High-Risk Domains for Hallucinations

Critical Risk

- Medical advice
- Legal citations
- Financial data
- Safety procedures

High Risk

- Historical facts
- Scientific claims
- Technical specifications
- Current events

Lower Risk

- Creative writing
- Brainstorming
- Code structure
- Format templates

Rule of Thumb: The more specific and verifiable the claim, the higher the risk of hallucination. General patterns are safer than specific facts.

The VERIFY Framework

A systematic approach to fact-checking AI output

V
Validate Sources - Check if cited sources actually exist
E
Examine Specifics - Scrutinize numbers, dates, and names
R
Research Claims - Cross-reference with authoritative sources
I
Identify Uncertainty - Look for hedging language or confidence markers
F
Flag Suspicious Content - Mark claims that need expert review
Y
Yield to Expertise - Consult domain experts for critical decisions

V - Validate Sources

What to Check

- Do cited papers/articles actually exist?
- Are the authors real people?
- Do the URLs work?
- Is the publication real?
- Does the quote appear in the source?

Quick Validation Tools

- **Google Scholar** - Academic papers
- **CrossRef** - DOI verification
- **Wayback Machine** - Archived pages
- **LinkedIn** - Verify authors exist
- **Direct URL check** - Visit the source

Quick Exercise

AI Output: "According to Johnson & Smith (2022) in the Harvard Business Review..."

Action: Search HBR for this article. Does it exist? Are these real authors?

E - Examine Specifics

Red Flags

- Very precise percentages (73.4%)
- Round numbers that seem too perfect
- Specific dates for vague events
- Exact quotes from famous people
- Detailed statistics without sources

Verification Steps

- Ask: "Where does this number come from?"
- Search for the exact statistic online
- Check if the date/timeline is plausible
- Verify quotes via quote databases
- Cross-reference with official sources

Pro Tip: Ask the AI "What is the source for this statistic?" - If it can't provide one or provides a fake one, treat the data as unverified.

R - Research Claims

The 2-Source Rule

For any important factual claim, verify with at least 2 independent, authoritative sources before using in professional work.

Tier 1: Primary

- Official company sites
- Government databases
- Peer-reviewed journals
- Original research

Tier 2: Secondary

- Major news outlets
- Industry publications
- Expert analysis
- Verified Wikipedia

Tier 3: Caution

- Blog posts
- Social media
- Forums
- Unverified claims

I - Identify Uncertainty

AI confidence ≠ accuracy. Learn to spot uncertainty markers.

Good Signs (AI acknowledges limits)

- "I believe..." / "I think..."
- "Based on my training data..."
- "I don't have access to current..."
- "You should verify this..."
- "Approximately..." / "Around..."

Warning Signs (Overconfidence)

- Extremely specific claims
- No hedging language
- Precise numbers without sources
- Definitive statements on evolving topics
- "Studies show..." (which studies?)

Prompt Tip: Add "Express your confidence level and flag any claims you're uncertain about" to your prompts.

F & Y - Flag and Yield

F - Flag Suspicious Content

Create a systematic review process:

- Mark all statistics with [VERIFY]
- Highlight quotes for validation
- Note claims outside AI's training date
- Tag domain-specific assertions

```
"Market grew 45% [VERIFY] according to Gartner [CHECK SOURCE]"
```

Y - Yield to Expertise

Know when to involve experts:

- **Legal claims** → Legal team
- **Medical info** → Healthcare pros
- **Financial data** → Finance team
- **Technical specs** → Engineers

> AI assists, humans verify. Final responsibility is always human.

Hands-On Exercise

Spot the Hallucinations

You'll review AI-generated content and identify potential hallucinations using the VERIFY framework.

What You'll Do

- Read AI-generated business content
- Identify suspicious claims
- Apply VERIFY steps
- Rewrite with proper verification

Tools Available

- Google Search
- Google Scholar
- The interactive demo
- Your critical thinking!

Exercise: Review This AI Output

AI-Generated Market Analysis

"The global AI market is projected to reach $1.8 trillion by 2030, according to a 2023 McKinsey report titled 'AI: The Next Digital Frontier.' This represents a compound annual growth rate of 42.6%.

Dr. Amanda Chen, Chief AI Researcher at Stanford's Institute for Human-Centered AI, stated in her keynote at the 2023 World Economic Forum: 'Organizations that fail to adopt AI within the next 18 months will face a 60% higher risk of market displacement.'

A recent Deloitte survey found that 78.3% of Fortune 500 companies have already implemented generative AI tools, with an average productivity improvement of 34%."

Your Task: How many potential hallucinations can you spot? Apply the VERIFY framework.

Exercise Analysis

Potential Issues Found

- **V:** McKinsey report title needs verification
- **E:** 42.6% CAGR - suspiciously precise
- **E:** "78.3%" - overly specific percentage
- **V:** Is Dr. Amanda Chen real? Verify LinkedIn
- **R:** WEF keynote - can we find the speech?
- **E:** "60% higher risk" - source needed
- **V:** Deloitte survey - does it exist?

Verification Results

- McKinsey report exists but title is different
- Market projections vary by source
- Dr. Chen - unable to verify exact quote
- Deloitte survey exists but numbers differ
- "60% risk" claim - no source found

Result: 5+ claims need correction or removal

Prevention Strategy #1: Better Prompts

Reduce hallucinations through prompt engineering

Hallucination-Prone Prompt

```
"Tell me about the benefits of AI in healthcare with statistics."
```

Problems: Vague, invites fabrication of stats

Hallucination-Resistant Prompt

```
"Describe potential benefits of AI in healthcare. Do NOT include specific statistics unless you can cite the exact source. If you're uncertain about a claim, say so."
```

Better: Sets boundaries, requests honesty

Prevention Strategy #2: Constrain the Output

Anti-Hallucination Prompt Additions

Add to your prompts:

"Important guidelines:
- Only include facts you are highly confident about
- If you cite a source, it must be a real, verifiable source
- Mark any uncertain claims with [NEEDS VERIFICATION]
- Say 'I don't know' rather than guessing
- Don't invent statistics - use qualitative descriptions instead
- If asked about events after your training date, acknowledge limitations"

Key Insight: LLMs follow instructions. Explicitly telling them NOT to fabricate significantly reduces hallucinations.

Prevention Strategy #3: Ground with Context

RAG - Retrieval Augmented Generation

The Concept

Instead of relying on AI's training data, provide verified source material in the prompt.

- Paste relevant documents
- Include verified data
- Reference specific sources
- Constrain to provided info

Grounded Prompt Example:

```
"Based ONLY on the following quarterly report data,
summarize our Q3 performance:

[Paste actual Q3 data here]

Do not add any information not contained in the above data.
If asked about something not covered, say 'This information
is not in the provided data.'"
```

Prevention Strategy #4: Chain-of-Thought

Ask AI to show its reasoning - makes errors more visible

Without Chain-of-Thought:
"What's the ROI of implementing CRM software?"

AI Response:
"CRM implementation typically yields a 245% ROI within the first year."

 No reasoning shown, hard to verify

With Chain-of-Thought:
"What factors determine CRM ROI? Think through the calculation step by step, noting any assumptions."

AI Response:
"ROI depends on: 1) Implementation cost, 2) Training costs, 3) Time savings, 4) Sales improvement... My assumptions are... These are estimates because..."

✔ Reasoning visible, assumptions clear

Prevention Strategy #5: Self-Critique

Ask the AI to check its own work

Two-Step Approach:

Step 1 - Generate:
"Write a summary of recent AI developments in healthcare."

Step 2 - Critique:
"Review your previous response. Identify any claims that:
1. Might be outdated (my knowledge cutoff is [date])
2. Include specific statistics that should be verified
3. Reference sources that need confirmation
4. Make definitive claims that should be hedged

List each concern and suggest how to address it."

Why it works: The same pattern-matching that creates hallucinations can also detect them when specifically prompted to look.

Tools for Verification

Source Verification

- **Google Scholar** - Academic papers
- **Semantic Scholar** - AI-powered search
- **CrossRef** - DOI lookup
- **Snopes** - Fact-checking
- **PolitiFact** - Claims verification

AI Detection Tools

- **Originality.ai**
- **GPTZero**
- **Copyleaks**
- **Turnitin**
- *Note: Detection isn't verification*

Built-in Features

- **Bing Chat** - Live citations
- **Perplexity** - Source links
- **Google Bard** - "Google it" button
- **ChatGPT** - Browse mode
- *Still verify independently!*

Building a Review Workflow

1. **Generate** - Create content with anti-hallucination prompts
2. **Flag** - Mark all verifiable claims (stats, quotes, sources)
3. **Verify** - Check each flagged item against authoritative sources
4. **Revise** - Correct errors, remove unverifiable claims
5. **Review** - Expert check for domain-specific content
6. **Publish** - Only after human sign-off

Best Practices Summary

Do

- Always verify statistics and sources
- Use anti-hallucination prompts
- Ground AI with verified context
- Apply the VERIFY framework
- Have experts review critical content
- Maintain healthy skepticism
- Document your verification process

Don't

- Trust AI citations without checking
- Use AI statistics in professional work unverified
- Assume confidence = accuracy
- Skip verification for "small" details
- Publish AI content without review
- Ignore your domain expertise
- Rely solely on AI for critical decisions

The Trust Equation

AI Output + Human Verification = Trustworthy Content

Speed	+	Accuracy
AI generates drafts quickly		Humans verify and validate

The Goal: Use AI to accelerate work while maintaining the accuracy standards your work requires.

Interactive Demo

Hallucination Detector Tool

Try our interactive fact-checking tool to practice identifying and verifying AI-generated claims.

- Paste AI-generated content
- Auto-detect potential hallucinations
- Get verification suggestions
- Generate corrected versions

Access the Demo

Navigate to the **Demo** section in the course materials

Or use any AI tool with the anti-hallucination prompts we've learned

Take-Home Exercise

Practice Assignment

1. **Generate** - Ask AI to write a 200-word summary about a topic in your field
2. **Analyze** - Apply the VERIFY framework to identify potential issues
3. **Verify** - Fact-check at least 3 specific claims
4. **Revise** - Create a corrected version with proper sourcing
5. **Reflect** - Document what you learned about AI limitations

Bonus Challenge: Try the same prompt with different AI models (ChatGPT, Claude, Gemini) and compare the hallucination patterns.

Key Takeaways

Always Verify

Trust but verify - especially statistics and sources

Prevent

Use anti-hallucination prompts

VERIFY

Apply the framework systematically

Human + AI

AI assists, humans decide

"The goal isn't to avoid AI - it's to use it wisely."

Resources

Further Reading

- Stanford HAI - AI Index Report
- MIT Technology Review - AI Coverage
- OpenAI Documentation on Limitations
- Anthropic's Research on Honesty

Verification Tools

- scholar.google.com
- snopes.com
- crossref.org
- perplexity.ai (with citations)

Stay Updated: AI capabilities change rapidly. What hallucinates today may improve tomorrow - but verification remains essential.

Questions & Discussion

What challenges have you encountered with AI accuracy?

Share

Your hallucination experiences

Ask

About specific use cases

Discuss

Industry-specific concerns

Thank You!

Handling Hallucinations: Fact-Checking AI

ITAG Skillnet AI Advantage

Remember the VERIFY Framework

Validate - **E**xamine - **R**esearch - **I**dentify - **F**lag - **Y**ield

Access course materials and the interactive demo through the course portal