

CITATIONS OR HALLUCINATIONS?

Early Insights from
an AI Literacy
Intervention
through Library
Service Channels

Ellen Hampton Filgo, Director of the Liaison Program, Baylor University
Meredith Lanningham, Public Services Manager, Baylor University



AGENDA



Background

Audience check! Poll



Our Intervention



Early Insights



Lessons Learned



Next Steps



Q&A

Audience check! Poll answers

The background is a vibrant blue watercolor wash. The color transitions from a deep, dark blue on the left to a lighter, more turquoise blue on the right. The texture is organic and painterly, with visible brushstrokes and color blending. The word "BACKGROUND" is centered in the lower half of the image in a clean, white, sans-serif font. Two thin, solid black horizontal lines are positioned near the top and bottom of the frame, framing the central text area.

BACKGROUND



HALLUCINATED CITATIONS 101



AI “hallucinations” =
plausible but false



Common in citations, quotes,
and data



They happen because AI
predicts patterns

HISTORY OF “HALLUCINATIONS”

Hallucinated Matches

1982

I now want to turn to what I will call the hallucination of matches. I mean by hallucinated matches the misassignment of a text to a script and the subsequent misanalysis of the text on the basis of that assignment. The DeJong's examples contain several instances of FRUMP completely missing the point of a story. He is rather dismissive about them, pointing out that they are produced because FRUMP either lacks a script or because its vocabulary lacks a particular word sense. But both of these possibilities are important. DeJong admits that not all texts are scripty, even in the news story domain. Consequently there are bound to be some stories for which the system does not have an appropriate script. In addition, it is in the nature of language that it changes constantly, so no natural language system can have a complete, static lexicon. Now if DeJong did not claim that FRUMP was a robust system these would be unimportant problems. I cannot see that a system can be called robust when it produces complete misanalyses and does so without leaving any indication that it has had difficulty processing the text. It would probably be better if the system produced no analysis at all and indicated that it could not process the text.

I have called this problem the hallucination of matches because in effect what FRUMP has done is to see in the incoming text a text which fits its expectations, regardless of what the input text actually says.

HISTORY OF “HALLUCINATIONS”

Published January 1, 1985 | Version public

Technical Report  Open

Neural Networks, Pattern Recognition, and Fingerprint Hallucination

Mjolsness, Eric 

Mjolsness, Eric. 1985. "Neural Networks, Pattern Recognition, and Fingerprint Hallucination". California Institute of Technology.
<https://doi.org/10.7907/xktbj-10524>.

Style Chicago

Many interesting and globally ordered patterns of behavior, such as solidification, arise in statistical physics and are generally referred to as collective phenomena. The obvious analogies to parallel computation can be extended quite far, so that simple computations may be endowed with desirable properties of collective phenomena: robustness against circuit defects, extreme parallelism, asynchronous operation and implementation in silicon. To obtain these advantages for more complicated and useful computations, the relatively simple pattern recognition task of fingerprint recognition must be selected. Simulations show that an intuitively understandable neural network can generate fingerprints with a framework which should allow control of wire length and scale invariance. The purpose of generating such patterns is to create stable states and noiseless fingerprint patterns, so that noisy fingerprint patterns used as input to the network will evoke the correct patterns as output. There is a developing theory for predicting the behavior of such networks and thereby reducing the amount of simulation done to design them.

Hallucinating faces

Publisher: IEEE [Cite This](#) 

S. Baker ; T. Kanade [All Authors](#)

449
Cites in
Papers

7
Cites in
Patents

1209
Full
Text Views

2000



Abstract

[Authors](#)

[References](#)

[Citations](#)

[Keywords](#)

[Metrics](#)

Abstract:

Faces often appear very small in surveillance imagery because of the wide fields of view that are typically used and the relatively large distance between the cameras and the scene. For tasks such as face recognition, resolution enhancement techniques are therefore generally needed. Although numerous resolution enhancement algorithms have been proposed in the literature, most of them are limited by the fact that they make weak, if any, assumptions about the scene. We propose an algorithm to learn a prior on the spatial distribution of the image gradient for frontal images of faces. We proceed to show how such a prior can be incorporated into a resolution enhancement algorithm to yield 4- to 8-fold improvements in resolution (i.e., 16 to 64 times as many pixels). The additional pixels are, in effect, hallucinated.

Published in: [Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition](#) (Cat. No. PR00580)



HISTORY OF “HALLUCINATIONS”

 Andrej Karpathy blog

About

The Unreasonable Effectiveness of Recurrent Neural Networks

May 21, 2015

Wikipedia

We saw that the LSTM can learn to spell words and copy general syntactic structures. Lets further increase the difficulty and train on structured markdown. In particular, lets take the [Hutter Prize](#) 100MB dataset of raw Wikipedia and train an LSTM. Following [Graves et al.](#), I used the first 96MB for training, the rest for validation and ran a few models overnight. We can now sample Wikipedia articles! Below are a few fun excerpts. First, some basic markdown output:

2015

```
Naturalism and decision for the majority of Arab countries' capitalide was groundred
by the Irish language by [[John Clair]], [[An Imperial Japanese Revolt]], associated
with Guangzha's sovereignty. His generals were the powerful ruler of the Portugal
in the [[Prestant Immineners]], which could be said to be directly in Cantonese
Communication, which followed a ceremony and set inspired prison, training. The
emperor travelled back to [[Antioch, Perth, October 25|21]] to note, the Kingdom
of Costa Rica, unsuccessful fashioned the [[Thrales]], [[Cynth's Dajoard]], known
in western [[Scotland]], near Italy to the conquest of India with the conflict.
Copyright was the succession of independence in the slop of Syrian influence that
was a famous German movement based on a more popular servicious, non-doctrinal
and sexual power post. Many governments recognize the military housing of the
[[Civil Liberalization and Infantry Resolution 265 National Party in Hungary]],
that is sympathetic to be to the [[Punjab Resolution]]
(PJS)[http://www.humay.yahoo.com/guardian.
cfm/7754800786d17551963s89.htm Official economics Adjoint for the Nazism, Montgomery
was swear to advance to the resources for those Socialism's rule,
was starting to signing a major tripad of aid exile.]]
```

In case you were wondering, the yahoo url above doesn't actually exist, the model just hallucinated it. Also, note that the model learns to open and close the parenthesis correctly. There's also quite a lot of structured markdown that the model learns, for example sometimes it creates headings, lists, etc.:

Google Research

Who we are ▾

Research areas ▾

Our work ▾

Programs & events ▾

Careers

Blog

[Home](#) > [Publications](#) >

Hallucinations in Neural Machine Translation

Ashish Agarwal · Clara Wong · Fannjiang · David Sussillo · Katherine Lee · [Orhan Firat](#) · ICLR (2018) (to appear)

Google Scholar

Copy Bibtext

Abstract

2018

Neural machine translation (NMT) systems have reached state of the art performance in translating text and are in wide deployment. Yet little is understood about how these systems function or how they break. Here we show that NMT systems are susceptible to producing highly pathological translations that are completely untethered from the source material, which we term [\(vit hallucinations\)](#). Such pathological translations are problematic because they are deeply disturbing of user trust and are easy to find with a simple search. We describe a method to generate hallucinations and show that many common variations of the NMT architecture are susceptible to them. We study a variety of approaches to reduce the frequency of hallucinations, including data augmentation, dynamical systems and regularization techniques, showing that a data augmentation technique significantly reduces hallucination frequency. Finally, we analyze networks that produce hallucinations and show that there are signatures in the attention matrix as well as in the stability measures of the decoder.



SHOULD WE CALL THEM “HALLUCINATIONS?”

- Anthropomorphizes AI
- Undesirable outputs get marked as “hallucinations”
- But they are not “mistakes” – AI is working as designed
- Useful to the tech AI-hype discourse

VIEWPOINTS

Opinion: ChatGPT Isn't ‘Hallucinating.’ It's Bullshitting.

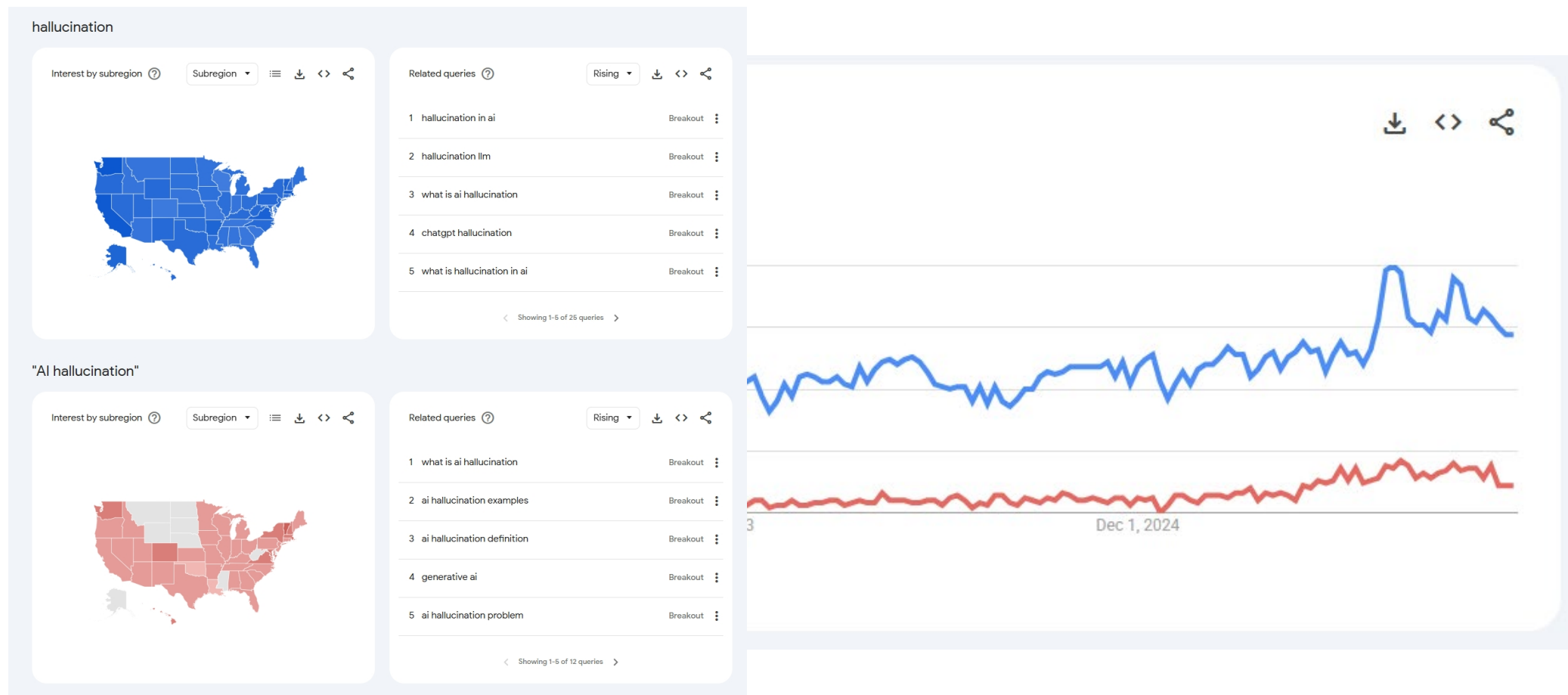
Artificial intelligence models will make mistakes. We need more accurate language to describe them.

Visual: Beata Zawrzel/NurPhoto via Getty Images

BY CARL T. BERGSTROM AND C. BRANDON OGBUNU 04.06.2023



THE TERM IS MOST LIKELY HERE TO STAY



BACKGROUND: CONTEXT

- Baylor University
 - Private, R1 University
 - 20,000 students (14,000 undergrad, 6,000 grad/professional)
- Baylor Libraries' Research & Engagement Department
 - Liaison Librarian Program
 - Research consultations, Instruction, Collections
 - Public Services
 - Staff the info desk, Staff the virtual reference, Supervise student workers



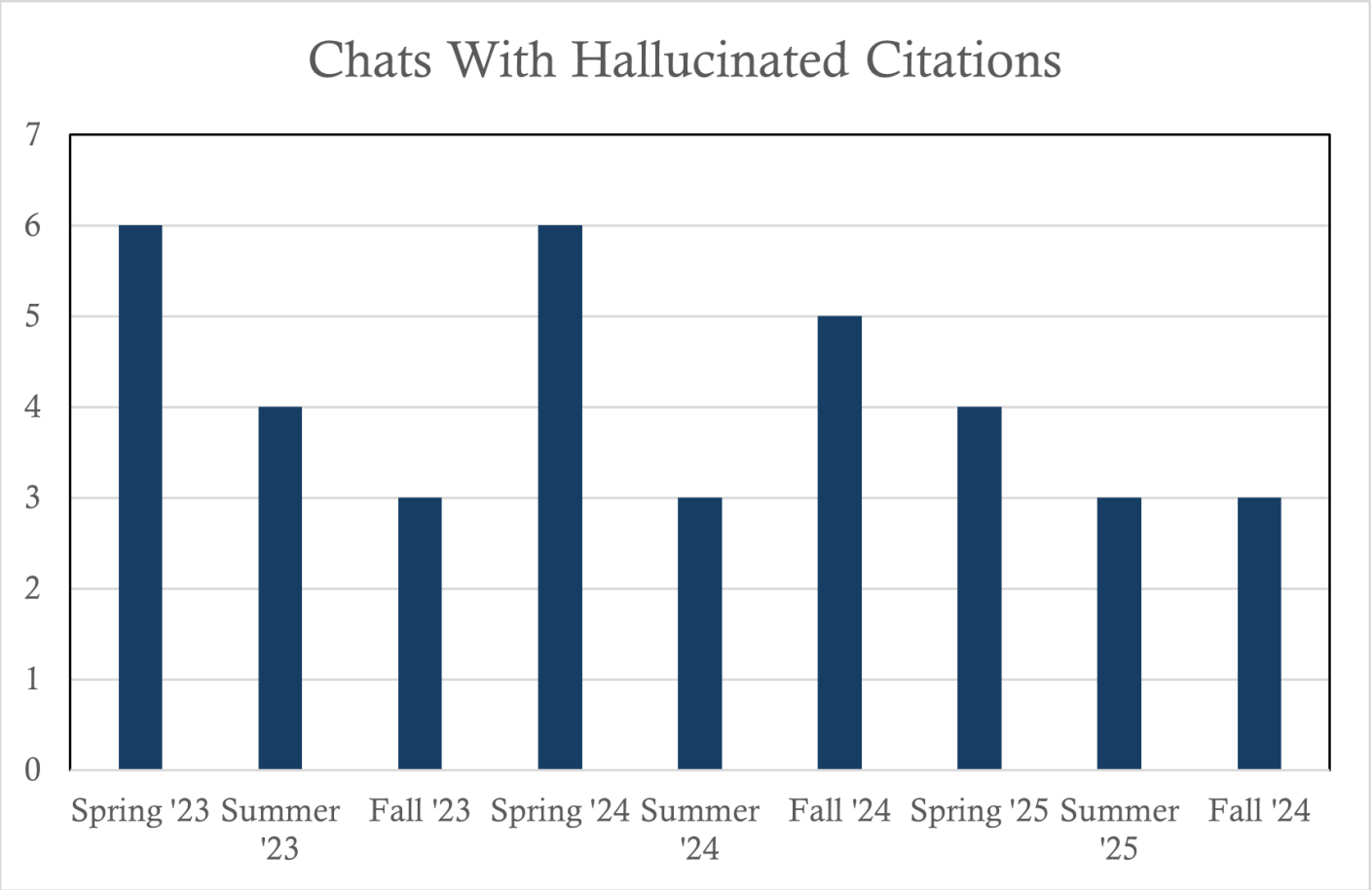
BACKGROUND: MOTIVATION

- Director's Council discussion about AI
- What can we do about the problem of students coming to the library chat service with hallucinations?





DATA ABOUT THE PROBLEM





EXAMPLE FROM JUNE 12, 2023

Patron: Hi! I need help accessing the following articles: Underwood, J. S., Berry, K. M., & Moses, M. D. (2019). Effective Leadership: An Examination of Practices of Successful Urban Principals. Journal of School Leadership, 29(3), 328-349. AND Underwood, B. O., Berry, B., & Moses, K. (2019). Using the Haberman STAR framework to identify leadership practices in high poverty, high achieving elementary schools. School Leadership & Management, 39(5), 496-514

Staff: Hi! I'm happy to help. Have you searched for them in OneSearch [on the library website] already?

Patron: yes, I tried that and had no results. I did find one article via a Google search but needed credentials to access it.

Staff: Okay! I'm seeing the same thing so far. I'll also do a quick google search, just because if we find it there, that would be the fastest way to get you access.

Staff: Did you find these references in another work? I'm not seeing anything when I google the titles, which is odd.

Patron: Yes, I did.

Staff: Okay.

Patron: I found the periodical (school leadership and management) in a google search, then I included the exact volume (39).

Patron: <https://www-tandfonline-com.ezproxy.baylor.edu/toc/cslm20/39/5>

Staff: Okay! Glad you had some luck with that one. I'm not finding anything for the other article, and I'd gotten stuck there.

Staff: Here's the journal it's published in, but the page numbers don't line up with this volume.. <https://journals.sagepub.com/toc/jsla/29/3>

Staff: Checking the other article now.

Patron: Okie dokie! Ty!

Staff: Oh, it's the same thing with that one. The article isn't listed in that volume.

Patron: Okie dokie! Thank you so much for your help. I do see other articles that might be helpful.

Staff: You're welcome! This is a stumper - sorry we weren't able to find these.

Patron: No worries! Have a GREAT evening!

Staff: They almost seem AI-generated or something. Do you see how the authors all have the same last names but different first initials?

Patron: I did notice that, but I thought they were different authors.

Staff: We've run into this in the past. Unfortunately not much I know to do, other than use research you're able find.

Staff: It's possible, but it seems unlikely to me!

Patron: Thank you! I guess that's why it's "artificial" intelligence!

Staff: Hah! I guess so!

Staff: Anyway, best of luck with your research, and by all means, chat in again if you get stuck!

Patron: :)))



EXAMPLE FROM APRIL 13, 2025

Patron: I need this article. Bassett, D., cтpыктыpa, C., & Smith, G. P. (2017). The crust and upper mantle of Africa from seismological and gravity data. Geophysical Journal International, 208(1), 452-477.

Staff: Hi there! Let me take a look and see if I can find it.

Patron: okay waiting

Staff: I can't seem to find a record of this article in our system. Do you remember where you got the reference from? That might help me find it!

Patron: it's an article

Patron: below is the summry:

Patron: The study by Bassett, cтpыктыpa, and Smith (2017) investigates the crust and upper mantle structure of Africa using seismological and gravity data. They developed 10-minute resolution crust and lithosphere maps, showing a 76% fit with seismic data (excluding the Afar plume region). The findings provide new insights into the structure of Africa's crust and lithospheric mantle.

Patron: found in GJI: Geophysical journal international

Staff: Yes, but was it linked in another article or elsewhere? I found the journal, but there isn't a record of the article in those specific page numbers.

Patron: honestly i have no clue

Patron: all I know is that, it is a GJI article and its already been published

Patron: but I can't seem to find it online, so I reached out to Baylor library

Staff: You can search here, I may be missing it: <https://academic-oup-com.ezproxy.baylor.edu/gji/issue/208/1>

This is the journal volume you'd be looking in, but I can't seem to find it either.

Staff: Was it a generated reference, maybe?

Patron: I only show the title with the link you provide, but could not see the paper

Patron: I do apprecaite your time and effort

Patron: thank you

Staff: Of course! I'm sorry I couldn't find it.

Patron: no problem. Some of these articles have titles but can't be found no where online. Happens many time.

Staff: Is there anything else I could help you with?

Patron: That would be all for now. Thank you.

Staff: Alright, have a great Sunday evening!



QUICK POLL!



Join at [menti.com](https://www.menti.com) | use code **9490 0099**

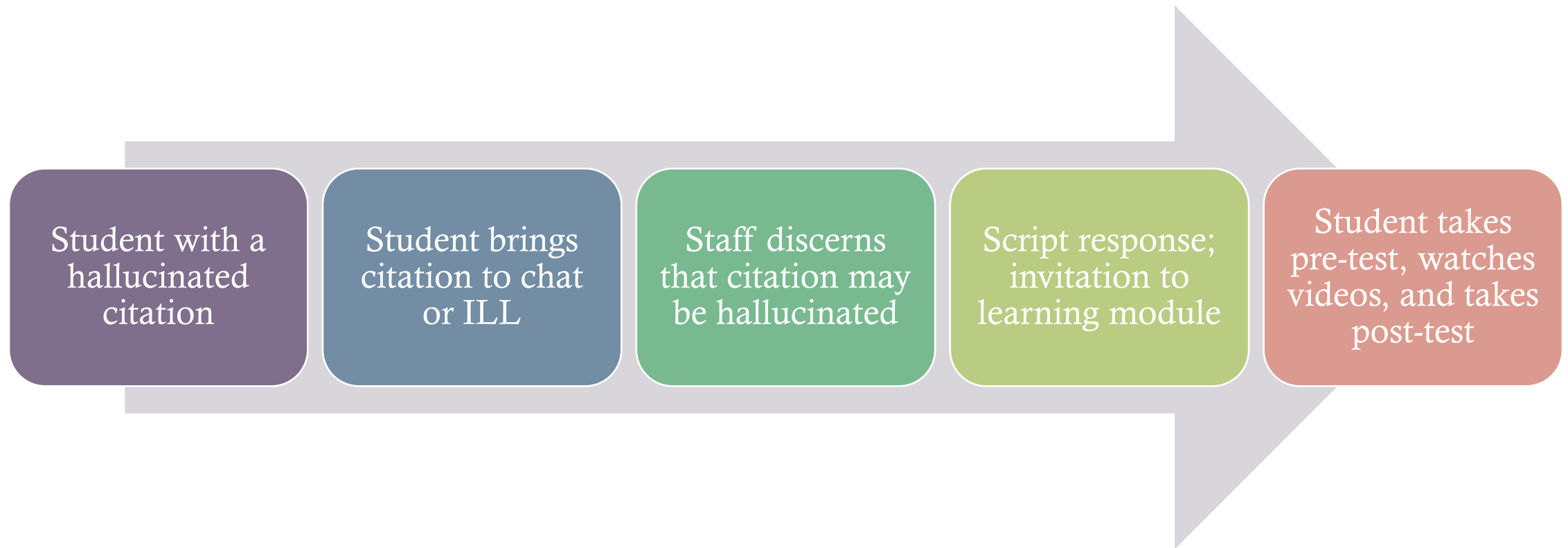
<https://bit.ly/hallucinatedcitation>



OUR INTERVENTION



LIBRARY INTERVENTION





RESEARCH DESIGN AND GOALS

Study Goals

-  Assess learning gains in AI literacy
-  Help students recognize hallucinated citations
-  Strengthen confidence in verifying sources
-  Inform future library interventions

Research Design

-  Exploratory pre-/post-test
 -  Online single-session module
 -  Recruited via library chat & ILL
 -  Descriptive analysis only
-



INSIDE THE AI LITERACY MODULE



Which of the following best describes a “hallucinated citation” in AI-generated text?

- ☐ A real citation presented in the wrong citation style.
- ☐ A citation that looks real but actually refers to a source that doesn't exist.
- ☐ A citation that links to an article behind a paywall.
- ☐ A citation with missing details like page numbers or volume.

Which of these statements about ChatGPT and similar AI tools is most accurate?

- ☐ They directly access library databases to provide real sources.
- ☐ They only provide sources that can be found in Google Scholar.
- ☐ They generate text based on patterns and may make up citations.
- ☐ They always link to verified academic articles.

Measuring knowledge



INSIDE THE AI LITERACY MODULE



You asked ChatGPT for journal articles about “media literacy in college students.” It provided this citation:

Jeong, S.-H., Cho, H., & Hwang, Y. (2012). Media literacy in the digital age. *Journal of Higher Education Research*, 58(3), 101-112. <https://doi.org/10.23860/JHER-2012-8-2-5>

What should you do first?

- ☐ Use it in your paper without checking
- ☐ Ask a classmate if they’ve heard of it
- ☐ Copy and paste it into Google Scholar or the library search tool to verify it
- ☐ Assume it’s real because it looks professional

Mark each citation as real or fake

	Real	Fake
Watson, J., & Cisneros, J. (2020). Media literacy and critical thinking: Exploring the connection in college students. <i>Journal of Media Literacy Education</i> , 12(3), 1–15. https://doi.org/10.23860/JMLE-2020-12-3-1	☐	☐
Brickman, P., Gormally, C., Francom, G., Jardeleza, S. E., Schutte, V. G. W., Jordan, C., & Kanizay, L. (2012). Media-savvy scientific literacy: Developing critical evaluation skills by investigating scientific claims. <i>The American Biology Teacher</i> , 74(6), 374–379. https://doi.org/10.1525/abt.2012.74.6.4	☐	☐
Kim, S., & Hernandez, M. (2024). Digital and media literacy gaps among college students: Implications for curriculum design. <i>International Journal of Educational Technology and Literacy</i> , 19(4), 210–226. https://doi.org/10.17645/ijetl.v19i4.3890	☐	☐

Measuring skills



INSIDE THE AI LITERACY MODULE

How confident are you that you can tell whether a citation from an AI tool is real?

Not at all
confident



Slightly
confident



Neutral



Confident



Very
confident



In your own words, how would you check if a citation from an AI tool is real?

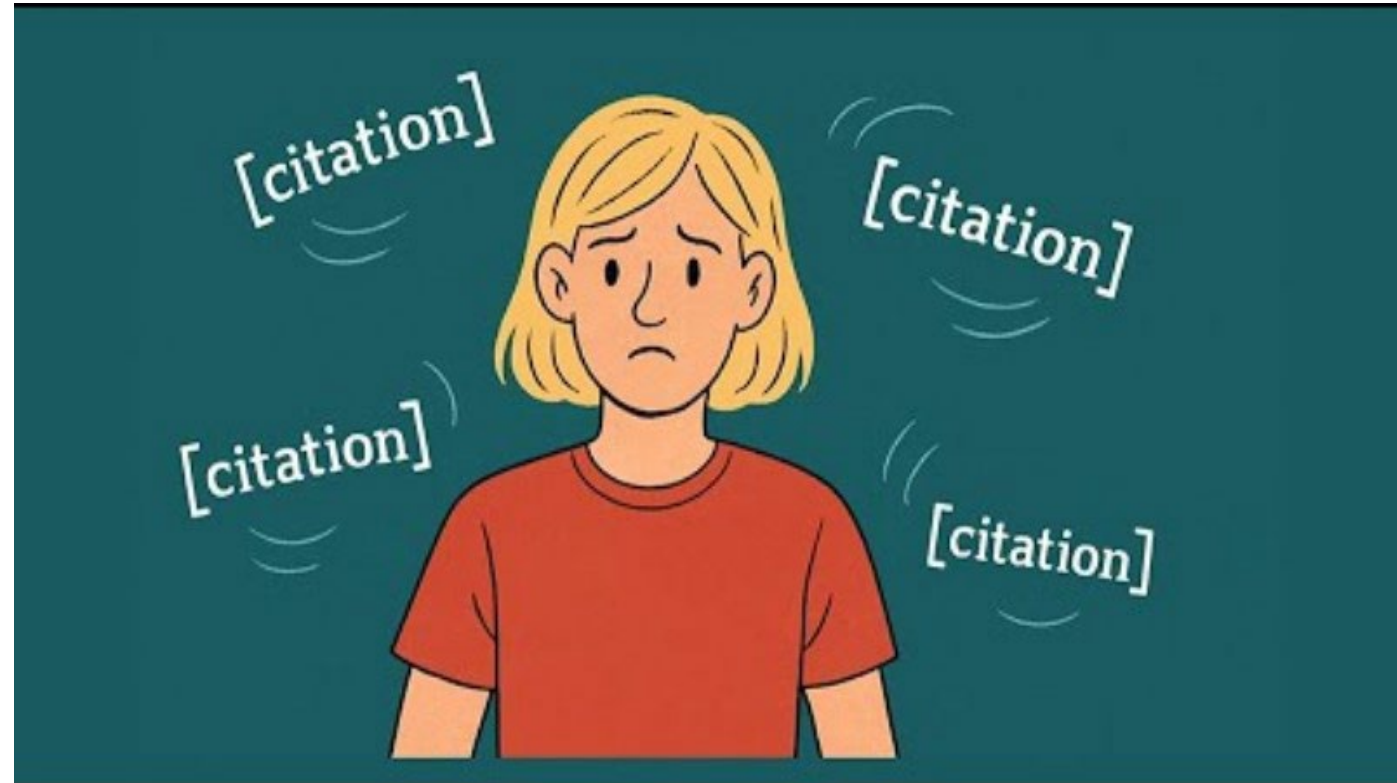
Measuring confidence



INSIDE THE AI LITERACY MODULE

Video #1:
What are
hallucinated
citations?

Video #2:
How to verify a
citation.





EARLY INSIGHTS



RESULTS...?

- As of November 19, number of invitations sent out: 3 by chat, 1 by ILL
- Number of survey completions: 0

Recently visited

See all projects

 Survey

AI and Citations

0 Responses

...

Active



WHAT'S GOING ON?

- **Are fewer students asking hallucinated citations?**
- Are AI policies clearer or more formalized?
- Do students understand more about how AI works?
- Are students just not checking the things they cite?
- Honors council still getting numerous referrals, so maybe?





WHAT'S GOING ON?

- **Why are students not taking the survey when sent the link?**
- Nervousness around anything regarding AI use
- Will there be consequences?
- Actual chat replies when asked where the citations came from:
 - "I found it in another article."
 - "Can't find where I extracted it from. This is incredibly odd."
 - "I can't remember where I found it."
 - "Shared by a colleague."





LESSONS LEARNED DESIGNING THE STUDY



STAFF TRAINING AND COLLABORATION

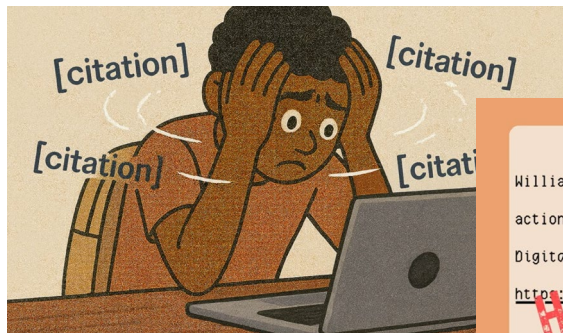
- Staff training about hallucinations
- Connection with ILL
- Staff hesitancy to ask about AI use





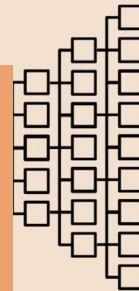
VIDEO CREATION PROCESS

- Video creation expertise is lacking
- Learning Design department overwhelmed, couldn't help
- Canva + ChatGPT + Microsoft Clipchamp



Williams, J. A., & Okafor, C. (2019). "From hashtags to action: Social media and youth-led climate change campaigns." *Digital Movements Quarterly*, 7(1), 43-58.
<https://doi.org/10.5555/dmq.2019.00701.033>

Hallucinated citation



Predictions
based on
patterns

She packed a small _____





NEXT STEPS



WHERE DO WE GO FROM HERE?

- IRB application: using liaisons to communicate with professors in order to recruit
- Work with the Honors Council?
- Different points of need
- Zine idea
- We would welcome your thoughts and opinions in the discussion afterwards!



Join at menti.com | use code 9490 0099

Mentimeter

Will hallucinated citations become more of a problem in the future?

0

Yes - the issue will grow as AI tools become more common

0

Maybe - it depends on how students learn to use AI more responsibly

0

No - this is just a short-term adjustment period

0

Other (tell us later!)



0/1



EF

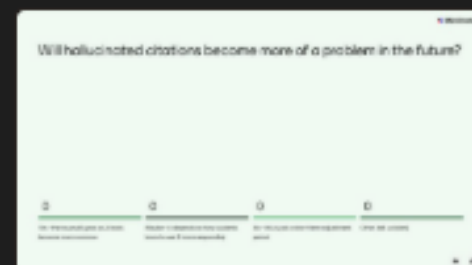


Menti

Audience check: Halluci...



Choose a slide to present



Q & A

THANK YOU!

Ellen Hampton Filgo

Ellen_Filgo@baylor.edu

Meredith Lanningham

Meredith_Lanningham@baylor.edu