

AI in the Real World

Data, Power, & Society - Day 07

Detecting Generative AI

CMSE 492 (aka CMSE 101) • Fall 2026

Prof. Danny Caballero

Week 2 work is graded

- Check if you met the standard for your assignments (Case Studies, Individual Report, Forum Post).
- You will have until **Wednesday, Sept 23rd at 11:59pm** to turn back in this work.

Week 2 Case Studies Feedback

Great diversity of cases!

Likely reasons you missed the standard

- Narrative didn't dive into the details of the case (*look over the scaffolding questions*)
 - You might not be able to answer them all, that's ok.
- References that supported your case were present, but not inline to show where the evidence came from.
- Missed some elements like the summary and context of the case.

If you have questions, just ask and we can help you parse the feedback.

Picking Up From Monday

We didn't get through everything with the dice on Monday. Before we move on, let's finish it.

1. Discuss Narrative 3 (two dice) — where we left off
2. Simulate the two-dice roll with code
3. Use that to make sense of Anthropic's watermarking technique
4. Then start your own Practices case for this week's Case Study

Narrative 3 – Reminder

The forecaster stood before the map and declared that today would be **[word 1]**. The chance of rain, she explained, was **[word 2]**. Looking out the window, everything outside appeared **[word 3]**. The crowd's mood shifted to **[word 4]** as they braced for what was shaping up to be a **[word 5]** afternoon.

Narrative 3 – Reminder

Sum	Ways to roll it	Word 1	Word 2	Word 3	Word 4	Word 5
2	1/36	apocalyptic	certain	terrifying	panic	disastrous
3	2/36	catastrophic	highly likely	ominous	dread	dreadful
4	3/36	severe	probable	gray	worry	difficult
5	4/36	unsettled	possible	hazy	unease	uneventful
6	5/36	cloudy	uncertain	overcast	indifference	routine
7	6/36	mild	fifty-fifty	ordinary	calm	ordinary
8	5/36	pleasant	unlikely	bright	relief	pleasant
9	4/36	sunny	doubtful	clear	cheer	enjoyable
10	3/36	gorgeous	improbable	radiant	joy	wonderful
11	2/36	perfect	remote	dazzling	excitement	magical
12	1/36	miraculous	impossible	otherworldly	euphoria	legendary

Discussion: Narrative 3

Let's discuss the reflection questions from Monday:

1. Compare your 20 rolls with your group mates. Are your distributions identical? Should they be?
2. Add all your counts together and sketch the distribution of your group's rolls.
3. How is this different from the "weighted" tables in Narratives 1 and 2? There, *we* decided which word was more likely. Here, who or what decided it?
4. What might happen to the narrative if the number of dice rolls was increased to 1,000? Or 10,000? Notice which words fall near the middle of any distribution compared to those near the ends (e.g., 6, 7, 8 vs. 2 and 12).

One Thing This Activity Gets Wrong on Purpose

Dice tables are a **lookup** and a transformer is not.

The rolling shows you what the model does *with* the probabilities, which is the last step of a long process. The transformer is how those probabilities get **computed in the first place**, using everything in the context at once.

That's the piece the dice model cannot show you.

Simulating the Dice, Many Times

Question 4 asked what happens at 1,000 or 10,000 rolls. Let's find out with code instead of by hand.

`scripts/two-dice-simulation.py` rolls two fair dice over and over and reports the observed distribution at checkpoints, next to the theoretical one ($1/36$, $2/36$, $3/36$, ...) — as text in the terminal, and as a bar chart that redraws live.

```
python3 scripts/two-dice-simulation.py --checkpoints 20 200 2000 20000
```

Watch what happens to the gap between the **observed** bars and the **expected** line as the count grows.

<https://github.com/msu-cerl/cmse101/blob/main/scripts/two-dice-simulation.py>

What If the Dice Were Secretly Weighted?

Now run it with a hidden tilt so that one sum is deliberately favored:

```
python3 scripts/two-dice-simulation.py --checkpoints 20000 \  
    --bias-sum 7 --bias-strength 2.5
```

- At 20 rolls, could you tell the dice were rigged?
- At 20,000, the tilt is unmistakable — even though **any single roll** still looks like an ordinary roll of two dice.
- Notice: this "roll" isn't two physical dice anymore. It's a probability table that we wrote down and nudged.

```
--bias-sum - sets the output we are attempting to bias the dice towards
```

```
--bias-strength - sets the amount of bias (oversampling) that we are doing
```

<https://github.com/msu-cerl/cmse101/blob/main/scripts/two-dice-simulation.py>

Activity: Watermarking AI-Produced Text

Our dice activities produced toy models of LLMs — but the text an LLM produces is random, *contextually* random.

The [European Union has passed a regulation](#) requiring generative AI outputs to be detectable as such.

Anthropic has developed a [watermarking technique](#) that biases the selection of the next word using the prior context (i.e., what was written before) — the same move as the rigged dice, just on tokens instead of sums.

Go to deClaude.org/watermarking and work through the simulations.

Reflection Questions for Watermarking

1. How does an LLM typically produce the next word in a sentence?
2. How is Claude changing that with a "secret key"? What is it changing about the selection of the next word?
3. How does this secret key approach allow generated text to be detected?
4. What does this biasing of text generation using a secret key say about how these GenAI companies think about the task of writing?

Starting Your Practices Case

This week you add **Practices** and **Quantitative Literacy** to your DTPA analysis. From the Week 3 materials:

- **Use** — how does the tool actually get used? Who prompts it, who receives the output, what did that person do before, and where (if anywhere) does someone check the result?
- **Labor and resources** — who annotated, moderated, or wrote the material the system depends on, under what conditions? What does it consume to train and to run, and who lives next to that?
- **Constraint and critique** — critique is someone saying it's wrong; constraint is a change in what someone is now permitted to do. Name at least one of each where they exist, and say whether the critique produced the constraint.

Full framework and "what meeting the standard looks like": [Week 3 materials](#)

Find a Case and Start Your Analysis

Research a specific, documented use of Generative AI in some industry. "AI in healthcare" is not a case; "a hospital system that deployed an AI scribe to draft clinical notes" is.

Places to start looking:

Where	What you'll find
AI Hallucination Cases Database	2,000+ court decisions, filterable to Michigan
AI Incident Database	Structured, cross-domain, sourced
404 Media	Ongoing reporting on AI in workplaces, schools, libraries
The Data Workers' Inquiry	Worker-authored accounts; <i>content warning</i>

Use today and Friday to find your case and start drafting Use / Labor & Resources / Constraint & Critique with your group.

Reminders

Exit Tickets, D2L, and the Course Website

- Don't forget to complete today's Exit Ticket by 5pm (on D2L)
- Case Study 3 due 11:59pm Sunday
- Individual and Metacognitive Reflection 3 due 11:59pm Sunday
- Spend time before Sunday (~90 minutes) going over Week 4's preparatory materials
 - Forum Post due 11:59pm **Sunday**

Your instructors have given detailed feedback on Week 2's assignments.