Red Teaming

Red teaming is treating a system with benevolent malice.

The goal is to come up with and test ways to make the system fail.

We can think of it as cross-examining the machine.

Why do lawyers need to do this?

Red teaming is a posture and a mindset.

We need to avoid confirmation bias when engaging with legal tech.

Legal AI couldn't be better at trapping us in confirmation bias.

What do you learn?

It’s a minesweep. You get a map of dangers.

You also get a map of where the system appears resilient.

Red Teaming & Experimental Design

Both are structured ways to try to prove yourself wrong.

We're not training to be computer scientists.

We're training to critically engage with legal technology.

Beyond hallucination

What are the errors or vulnerabilities we want to learn about?

Breakdown of performance based on size of task or prompt

Interaction between hard-coded and generative AI systems produces new errors

Distortion from previous prompts and conversations

Misapplying instructions

Inability to judge uncertainty / false certainty

Misapplying terms of art

Inability to handle hypotheticals

Misapplication of guardrails

Misapplication of legal reasoning

Ability to differentiate between dicta and holding

Handling nuance

Spewing lots of garbage

Failure to identify user error

Not asking follow-up questions - jumpt to conclusions and start drafting

Getting stuck in loops without a way out

Semantic drift outside of appropriate legal issues