10/29/26: FOLDS seminar, Meena Jagadeesan

Setup: *strategic classification.*

Individuals choose features $G(x)$ that the instititon then sees. The institution performs risk minimization to learn a predictor f:

$$f^* = \min_{f \in \mathcal{F}} \mathbb{E}[\ell(f(G(x)), y)].$$

Ex: LLM detection (an intervention of quality), Medicare Advantage payments [SJ: ?]

LLM detection: surprisingly, detection can *increase* LLM usage, unless detection operates on a "restricted" set of features

The paper / output is represented by $z \in \mathbb{R}^d$

The detector simply looks at the first feature z_1: $F(z) = \mathbf{1}\{z_1 > v\}$, with penalty β

Quality $Q(z) = \sum_{i=1}^{d} Q_i(z_i)$

Q_i concave, Q_1 monotone (either $\nearrow$ or $\searrow$)

Some assumptions about LLMs vs humans

Users choose α, which decides how much work to delegate to LLM vs to do themselves; production cost $\gamma^t(1-\alpha)$, how much work they do

Post-processing: to explicitly avoid detection, choose a smaller z_1, and pay gaming cost C

[SJ: So why wouldn't I just keep reparaphrasing until I get z_1 below the desired v? That's the gaming cost. Maybe in the Pangram case, the gaming cost is trivial.]

[SJ: What if I don't have perfect access to z_1, e.g., I don't know how they're evaluating me?]

[SJ: Rakesh: is it bad that the LLM is a substitute here, as opposed to a complement? My follow-up: in modern workflows, LLMs and humans collaborate, so it's not clear to me whether it's a substitute or a complement...]

Utility function combines these metrics.

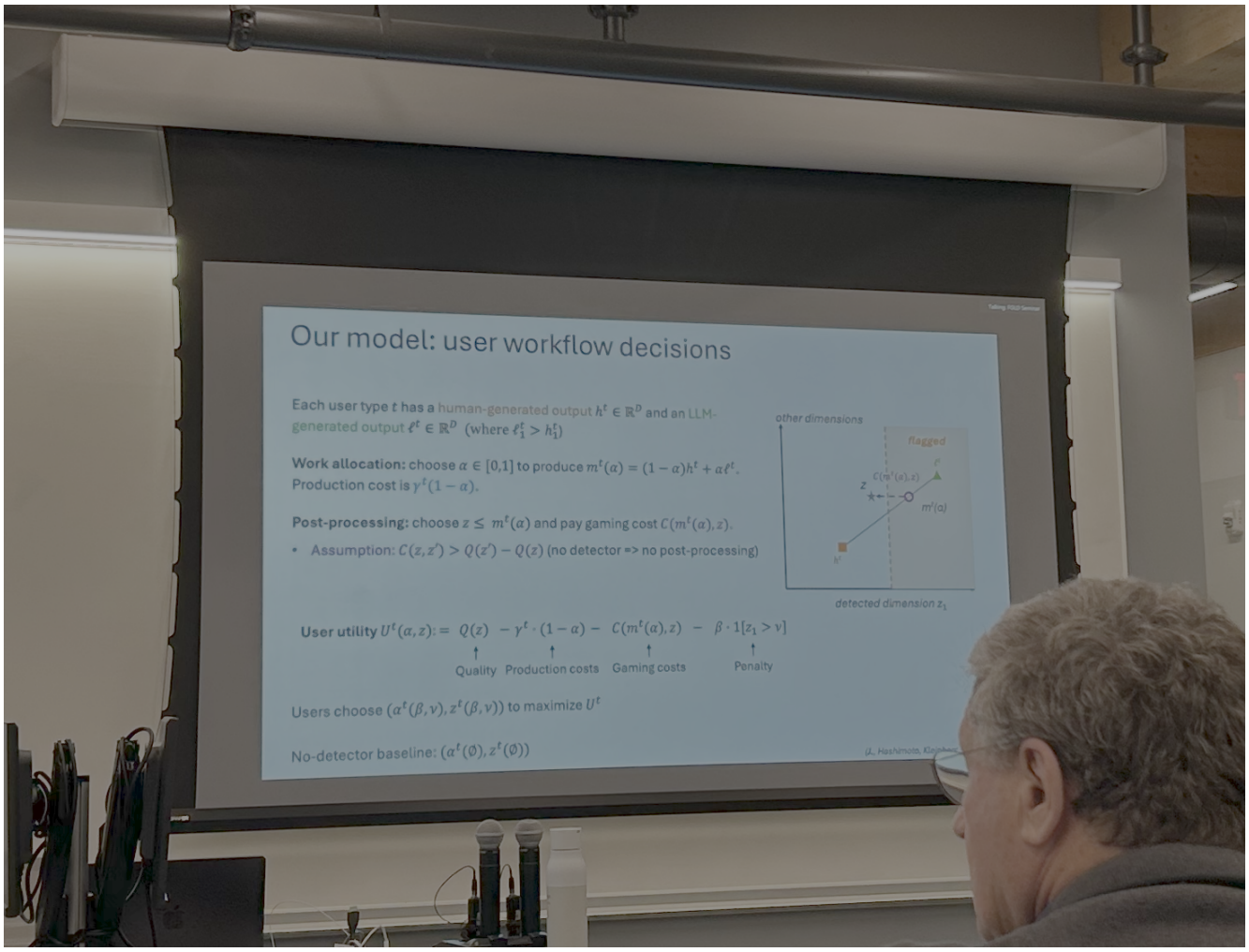


Figure 12: Setup.

(E.g., you might paraphrase or rewrite)

[SJ: How is Meena computing "quality" here? Relates to my question to the group about how to compute "expository quality" score for math papers? "rephrasy".]

Theorem 1: sometimes the detector increases . (The LLM is smarter, and it's quality-wise inconsequential and cheap to paraphrase to avoid via Pangram.)

[SJ: Surbhi asked a question but speaks too fast for me to understand.]

[SJ: Rakesh: Goodhart's law?]

Some complicated conditions on the next slide.

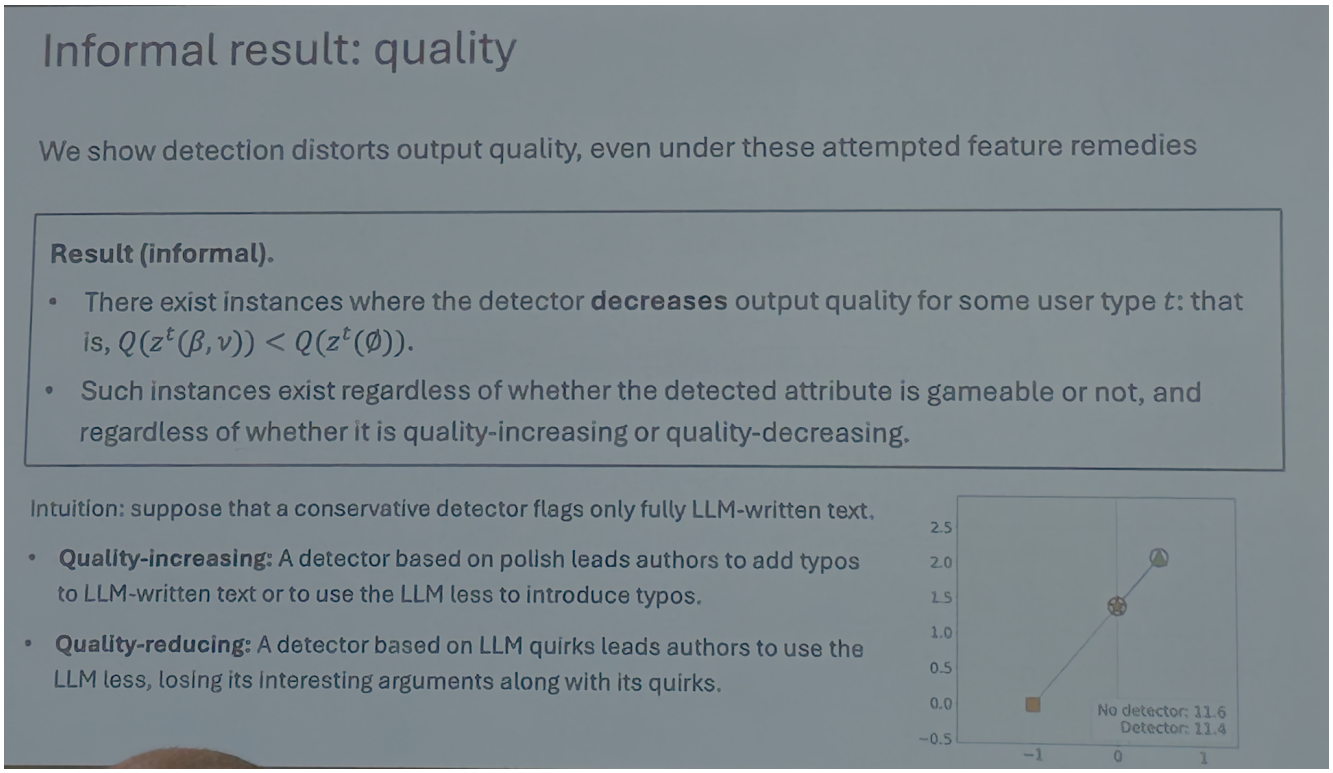


Figure 13: Enter Caption

A slide on the empirical analysis: rise-then-fall words on arXiv.

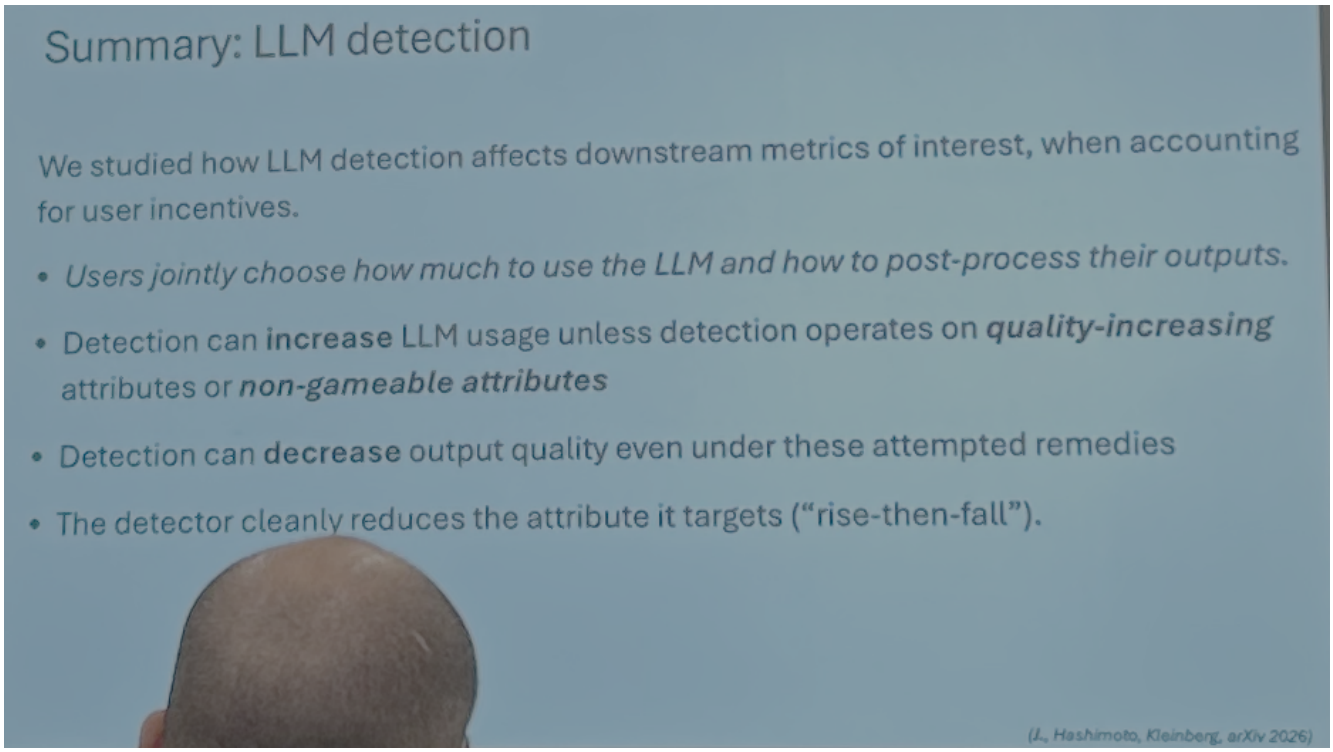


Figure 14: Summary of detection part.

Part 2: upcoding among insurers in Medicare Advantage [SJ: ?]

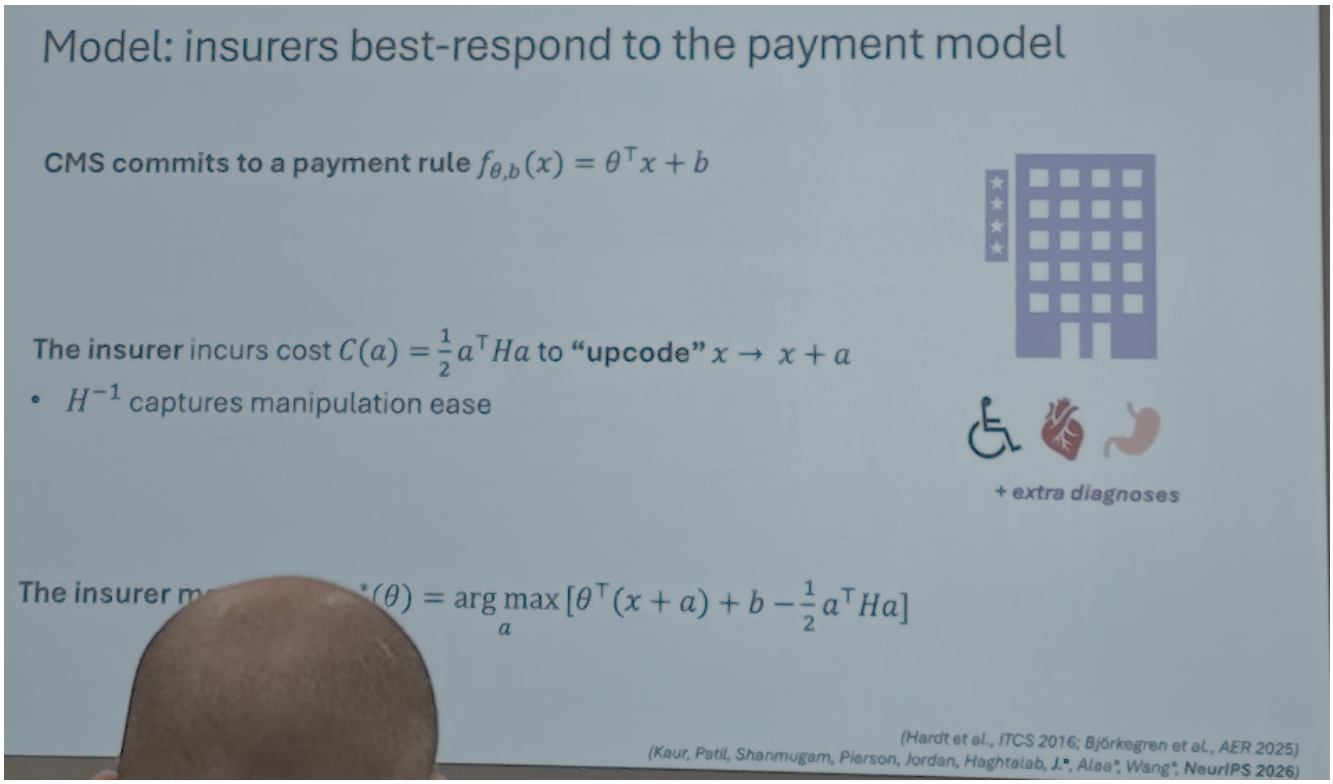


Figure 15: Upcoding intro.

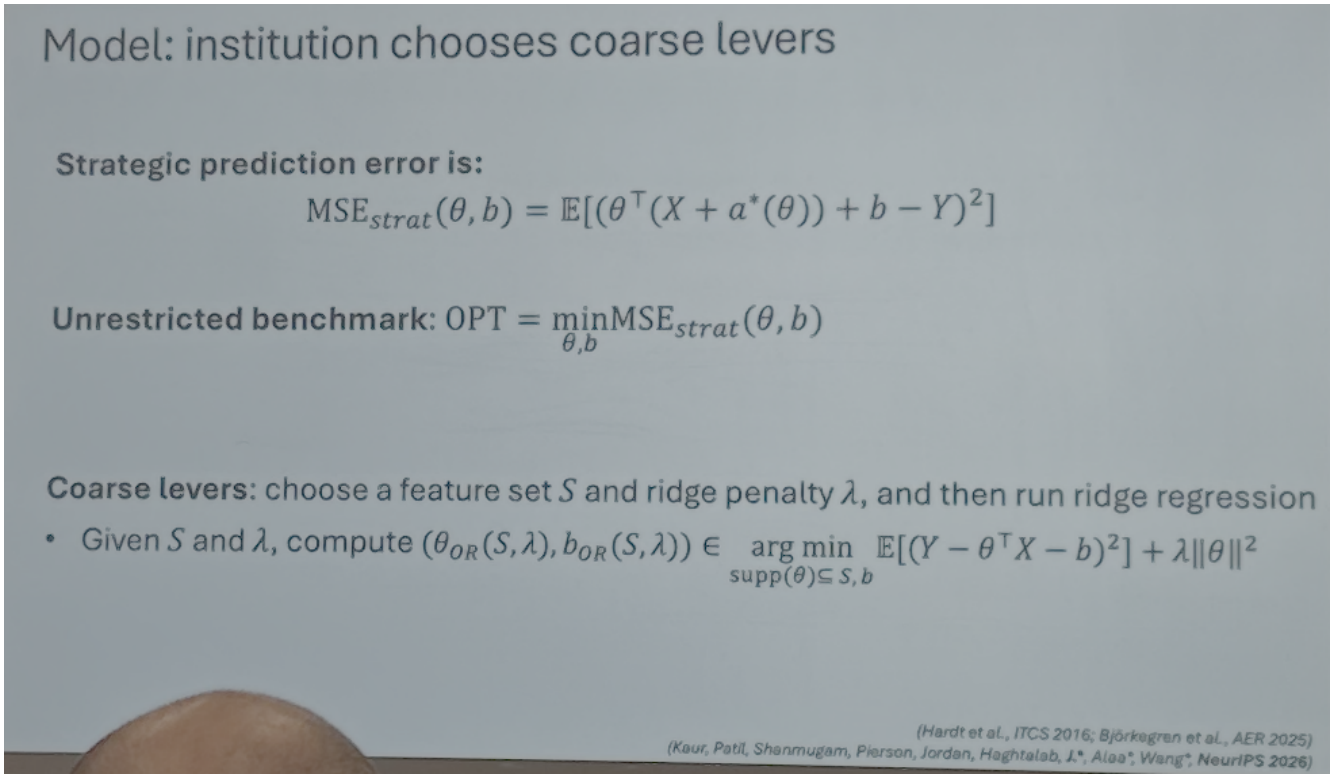


Figure 16: Institution's problem.

[SJ: Missed the punchline / ending; ask for slides.]

References