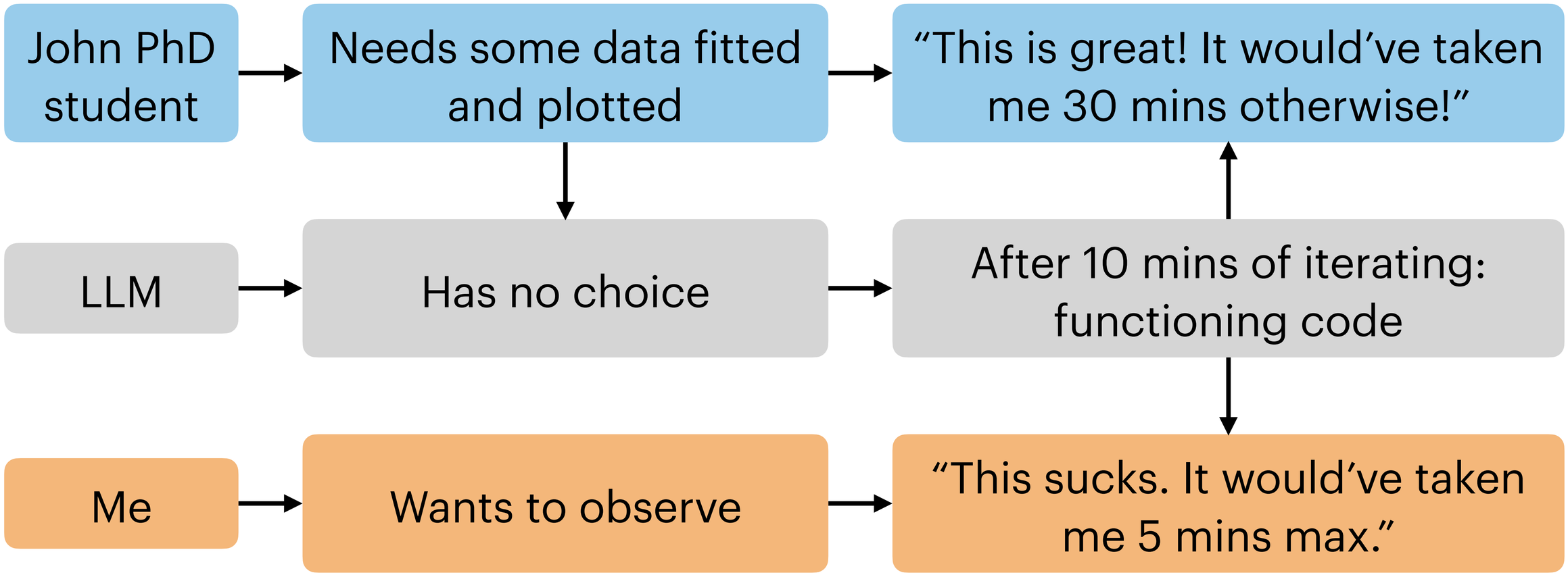
Confirmation bias in AI experimentation

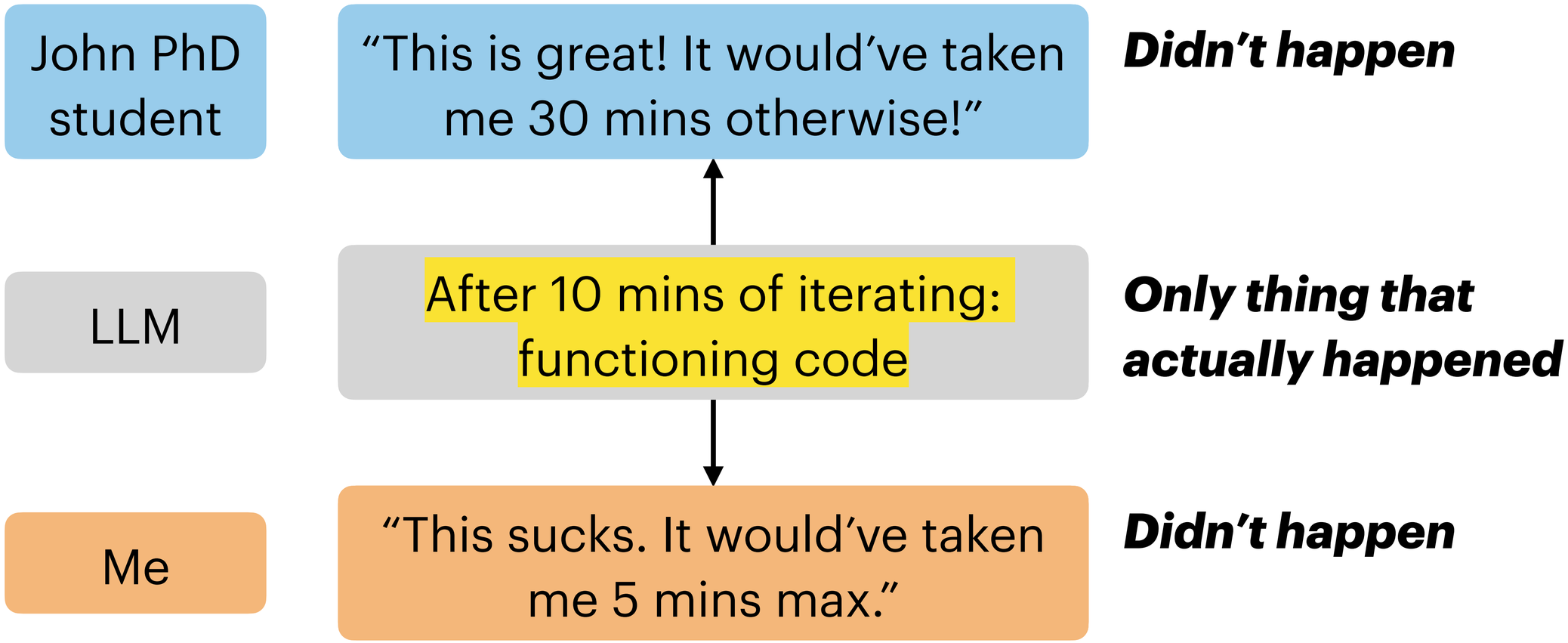
AI Discussion Forum

Renske Wierda — 28/04/2026

A real life case study



However


John PhD student
“This is great! It would’ve taken me 30 mins otherwise!”
Didn’t happen
LLM
After 10 mins of iterating: functioning code
Only thing that actually happened
Me
“This sucks. It would’ve taken me 5 mins max.”
Didn’t happen

“This is great! It would’ve taken me 30 mins otherwise!”

Both of these are examples of confirmation bias

“This sucks. It would’ve taken me 5 mins max.”

There’s a reason why medical trials have to include both the placebo and the nocebo effect. Humans are very good at convincing themselves.

LLMs are cognition hazards (blogpost)

- Chatbots specifically are known to increase positive confirmation bias
  - E.g. the Barnum effect and (more extreme) AI psychosis

- This happens because they're trained to give **plausible** output, which skews towards *agreeable* and *statistically likely*

- LLMs for coding/math have the same issues, see Schwartz's AI grad student blogpost which extensively described how Claude would just fudge numbers to create results that *looked* good, not *were* good

The consequences

Vibes-based experiments don't really hold up

> *"Well I would've been faster than the AI"*
> — John AI skeptic

> *"Well you've prompted it wrong so of course the output is bad"*
> — John AI enthusiast

(see also Wikipedia's list of cognitive biases)

Can we experiment with AI workflows in ways that account for these effects (without doing double-blind studies)?