Here's what everyone does today: find a base model, write some clever system prompts, maybe add a few examples, then fine-tune if performance isn't good enough. Rinse and repeat.

But this approach suffers from what researchers call "brevity bias"—the tendency to compress knowledge into short, generic instructions that lose critical domain-specific details.

Even worse? When you try to iteratively improve prompts by having an LLM rewrite them, you often get "context collapse"—where the model compresses accumulated knowledge into increasingly shorter summaries until performance tanks.

Stanford documented one striking example: a context that contained 18,282 tokens and achieved 66.7% accuracy suddenly collapsed to just 122 tokens at the next update, with accuracy dropping to 57.1%—worse than the 63.7% baseline.