Proposed Solutions

- Just look for formulaic prose, or “As a large language model…” ☺
- Metadata ← trivial to remove from text!
- Giant database of completions ← privacy?
- Discriminator models, like GPTZero or DetectGPT or Ghostbuster ← too many false positives?
- **Watermarking:** inserting a statistical signal into the LLM’s choice of tokens

Brief History of LLM Watermarking

My moment of terror (~July 2022): Led me to propose a watermarking scheme based on the "Gumbel Softmax Rule" and pseudorandom functions, prove theorems, give talks about it.

Kirchenbauer et al. (Jan. 2023): Modify the LLM probabilities using a pseudorandom function of n-grams

Christ, Gunn, Zamir (June 2023): Watermarked output that's cryptographically indistinguishable from normal LLM output—rediscovered most of my proposal and more

Kuditipudi et al. (July 2023): Watermarking using "one-time pad" rather than pseudorandom function