Small data

Jon & Wendy

https://www.pelican-crossing.net/
https://www.cst.cam.ac.uk/people/jac22
Spot the bird
Signal to noise can only get worse

- LLMs and their miscontents.
- Flu Trends.
- Model collapse.
The threats

- Regulators: data protection, privacy, antitrust, competition.
- Here be monsters.
- Washington Post Google C4 dataset checker:
Alternative Futures


B. Individualization. Personalized small language models that can run on a device in a person's pocket. Something like that possible now with image generators.

C. Combine the two above? Could individualized LMs be federated to collaborate on stuff? Share the query processing but not the data?

D. Paradigm shift. Build a new platform with a new paradigm and business model. (Ref Tim Wu, *The Attention Merchants*)
What's the damage?

• Short-term damage
  - Writers - cut to gig workers as unscrupulous corporations try to take advantage
  - Actors - become stock library items (Hollywood),
  - Coders - as writers

• Long-term appears likely to be little change.
  • Recent articles advise return to Stack Overflow as LLM-generated code not reliable
  • But distraction for the technology industry from current problems.
  • How many of us fell for it?
Conclusions & End Notes

• Nothing to see here...
• ...wasn’t it all just a distraction from what?

• Jill Fain Lehman,
• "Statistics means never having to say you're sorry."
See also

- >>. In a lawsuit filed in California last month, the writers Sarah Silverman, Richard Kadrey, and Christopher Golden allege that Meta violated copyright laws by using their books to train LLaMA, a large language model similar to OpenAI’s GPT-4—an algorithm that can generate text by mimicking the word patterns it finds in sample texts. But neither the lawsuit itself nor the commentary surrounding it has offered a look under the hood: We have not previously known for certain whether LLaMA was trained on Silverman’s, Kadrey’s, or Golden’s books, or any others, for that matter.>>