EXHIBIT TWENTY-ONE

Insights from Books → Web

Data team offsite Oct 28, 2025

Topics

- Vellum
  - Understanding, prioritization, experiments
  - (Xiaolan covering curation!)
- AI Generated Web Content filtering

Look at appendix and links in slide deck for details!

Takeaways

- Book ordering + metadata is complicated
- Plot your data more
  - (Corollary of "read your data")
  - Sampled semantic clusters → spot non-obvious things
    - (used lots image/video side, not as much text so far)
- Let's curate more subsets of web
  - May make it easier to filter for badness if we don't need to worry about removing good
    - Ex. classifiers confusing fiction for spam

Vellum

Why book scanning?

- We have decided that we will no longer use Unsafe books datasets
  - …can't even run experiments on the old data anymore 🙃
- Big impact from removing Unsafe datasets
  - 6.4 million documents, 880B raw tokens; [redacted] of DT 4.0 weighted datamix
  - Regressions - [redacted]x on MMLU/GPQA
  - Books important for long context, knowledge, etc
- *Bartz v. Anthropic*
  - Training on books from piracy sites: Not okay
  - Buying a book, cutting off the spine, scanning the pages: Fine