Scaling Groundtruth Data Creation with AI

Let's brainstorm practical ways to scale data creation and sustain efforts virtually.

How can AI help us scale groundtruth data creation effectively?

Can't know yet :) needs experimentation in practice

it depends

It can identify the sources of ground truth data.

Through data gathering, verifiability and traceability. It can provide numbers available in the literature.

scaling of the ground truth data

We create enough good quality ground truth data for AI to learn.

I am not sure it can contribute to that.

The structure of the LCA ground truth development process may benefit from human refinement before its ready for AI acceleration.

How can AI help us scale groundtruth data creation effectively?

Use verifiable data

Potentially, with a sufficiently rigorous framework in place

Optimize based on directed human in the loop scenarios

Feeding many LCAs into AI, training AI to think/act like an expert/professor in the field of LCA

By using AI to generate diverse, realistic LCA scenarios while humans curate and validate the modeling decisions that matter.

web scrapping and collection of publicly available data

Train and AI with LCA knowledge

By summarizing many sources with inconsistent format

How can AI help us scale groundtruth data creation effectively?

Create a master list of rules

ChatGPT App that captures LCA questions/responses and asks for feedback on the response

Peer-reviewed publications!

What virtual methods or tools could keep our data creation efforts going remotely?

Zoom

Federal LCA commons

GitHub?

Amazon Prime

Have a simple user interface app?

Curated library of AI agents or other tools

Create something like ecoquery. It worked well for ecoinvent.

Shared scenario libraries, collaborative QA review boards (e.g., Notion/Miro), and version-controlled Q&A datasets where experts asynchronously refine modeling decisions rather than just answers.

What virtual methods or tools could keep our data creation efforts going remotely?

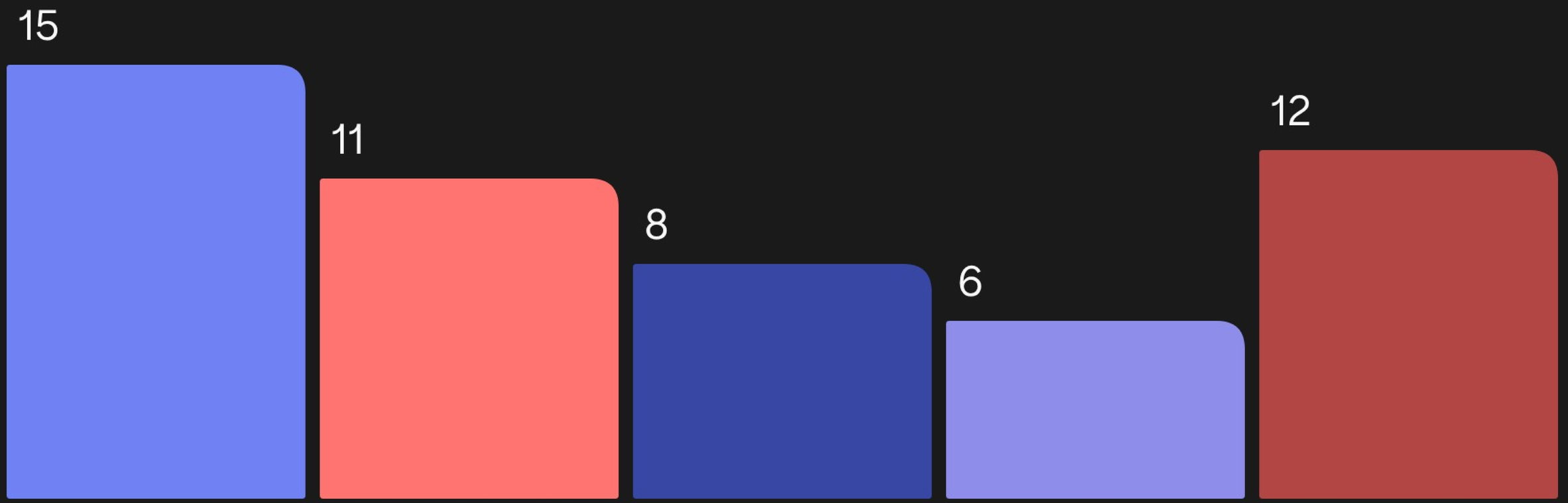
GH repo of prompts + example flows

a vibe-coded app which can reduce the effort to do this

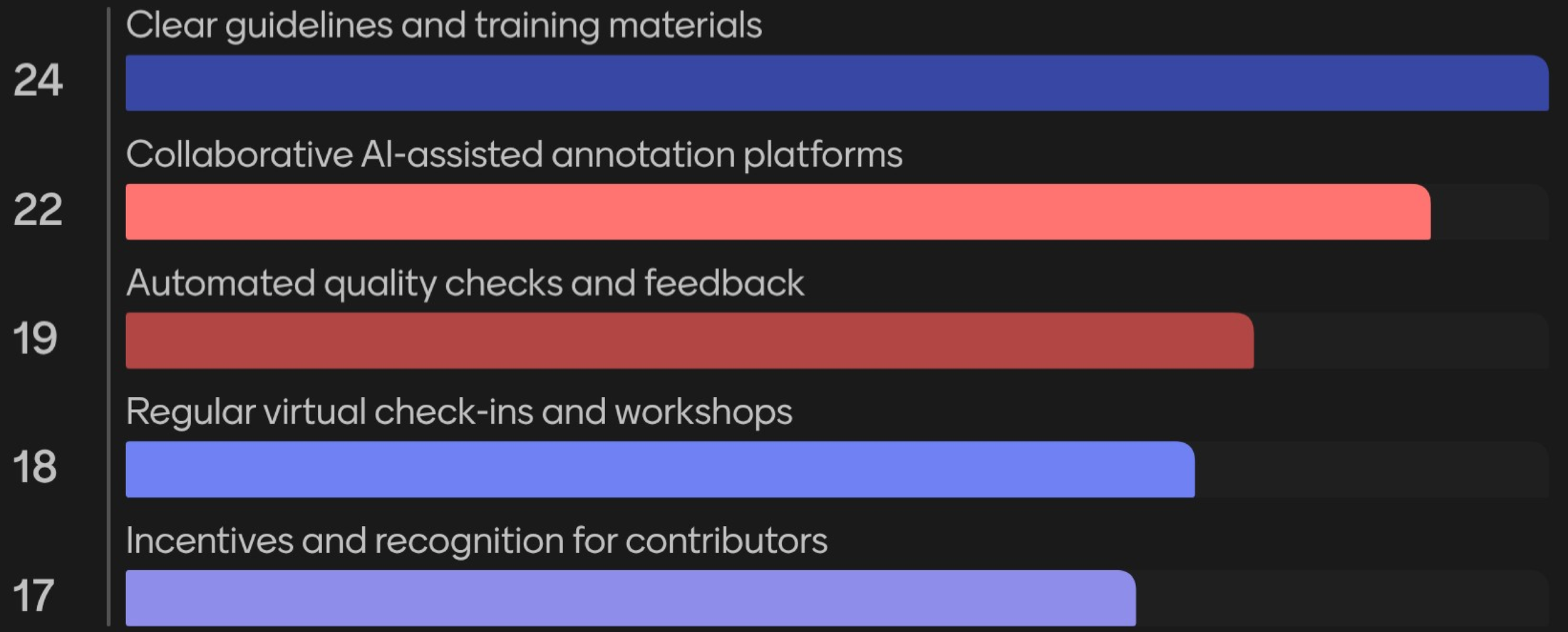
Hackathon!

Notion

Which of these challenges might impact scaling groundtruth data creation with AI?



Prioritize these strategies to improve virtual collaboration for data creation.



In one word, what is the biggest opportunity AI brings to data creation?

What support or resources do you need to help scale this work virtually?

Training

Prompt generation and validation

Guidance

Customer trust!

Open source data

Community, conversation, training, guidance

Shared scenario library, modeling-focused rubric, async expert review platform.

Share experience of use cases

What support or resources do you need to help scale this work virtually?

- Training
- Communication
- Collaborative work, more workshop
- Accessibility for the AI illiterate , but domain experts
- Data sharing platform
- Public Q&A forums
- Ai training and tools