How good is it?

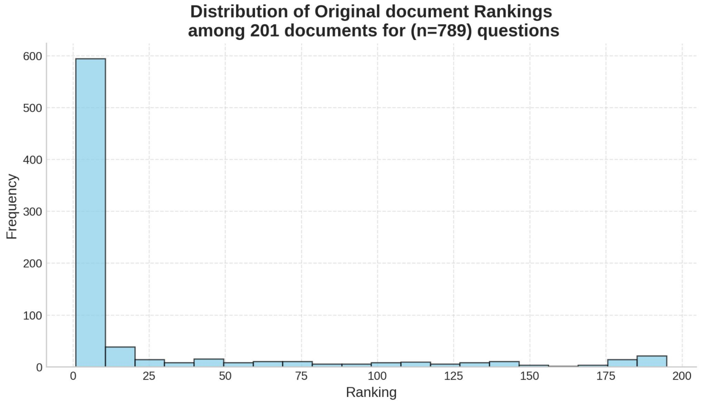
JC

Quality control, test performance, test evolutions
JC got the work started with generating test data:

- For each TWiki, ask an LLM to list X questions that can be answered by the TWiki & response

Invert for testing:

- **Search**: Ask question and get the rank of the TWiki among the documents (see plot)
  - **With current setup, correct doc returned ⅔ of time**
- **Assistant**: Ask question and see if answer matches (cosine sim) the “known” answer



Small test: Ability to find correct TWiki using 789 questions from 201 TWikis