Next Steps

Finalize the test-set plots and tables, then:

1. Measure inter-LLM consistency.
2. Measure inter-human-rater consistency.
3. Measure human/LLM correlation.
4. Add a simple LLM prompting baseline.
5. Quantify the gain from sampling N times rather than 1 time.
6. Update the test set with improved gold years once multiple annotations can be merged.
7. Develop baselines: simplest baseline with a single prompt.
8. Develop baselines: include entity extraction, entity type, and search query.
9. Develop baselines: add a search and update step.
10. Develop baselines: update step combines N LLM outputs.
11. Define comparison metrics:
12. Metric: exact-year accuracy.
13. Metric: no-leak accuracy (predicted year $\geq$ gold).
14. Metric: mean absolute error in years.
15. Metric: signed error bias (mean predicted minus gold year).
16. Metric: coverage of 95% intervals (if intervals are produced).
17. Metric: interval width (median and mean, if intervals are produced).
18. Metric: fraction of samples with any grounded source link.
19. Metric: average number of entities extracted per sample.
20. Metric: average number of sources per entity.
21. Metric: search cost per sample (queries and total tokens).
22. Metric: runtime per sample (end-to-end latency).
23. Refine the task definition.
24. Add a literature review.
25. Update the protocol to explicitly include the LLM steps.