Hybrid Synthetic-Human Test Collections for Enhanced Retrieval Evaluation

Susitra Gnanasambhandham

CPSC 8470 | Spring 2025

BACKGROUND

In traditional IR, evaluating search systems requires manually labeled relevance judgments

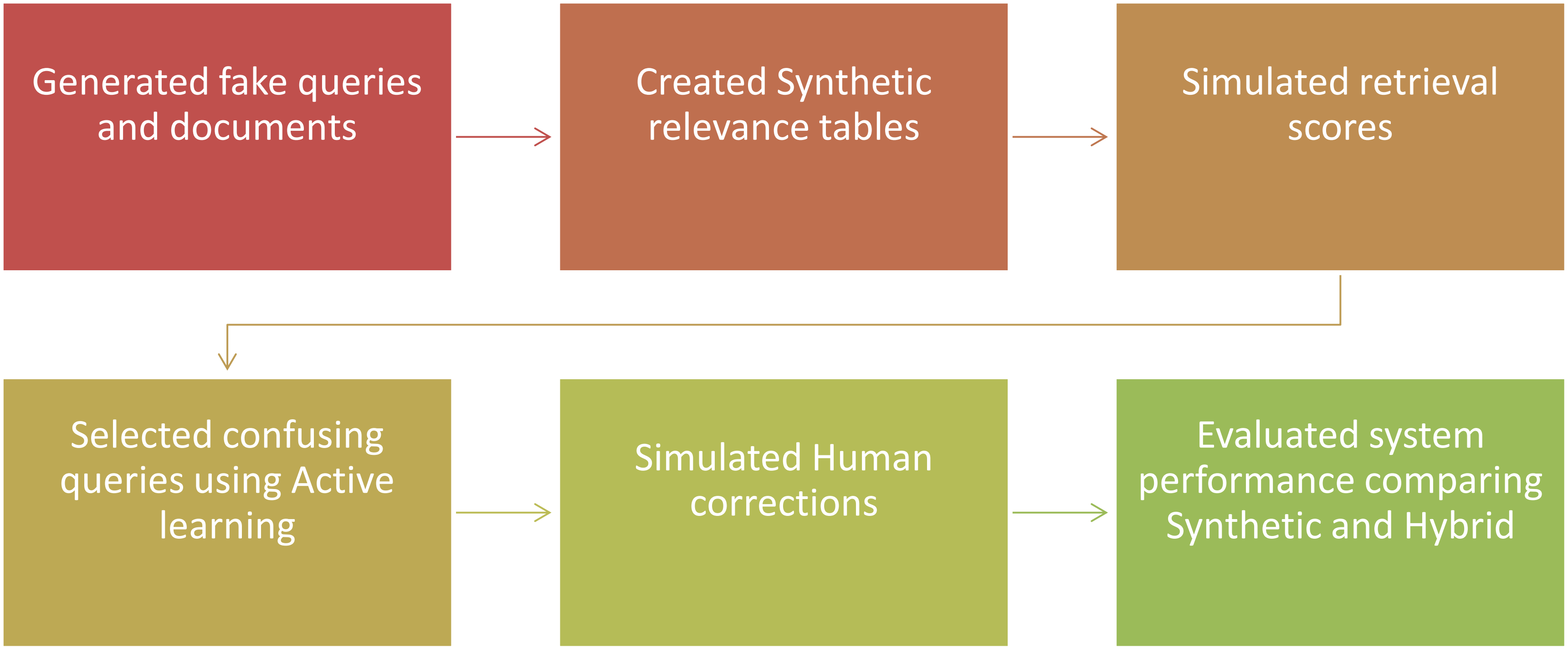
Synthetic labels often miss the subtle aspects of human relevance

Finding a balance between scalability and realism is crucial.

PROJECT OBJECTIVES

- Create a Hybrid Human-Synthetic test collection
- Use of Synthetic labels for speed and fix most confusing part with human-like correction
- Target and correct the "most confusing" queries using active learning.

METHODOLOGY OVERVIEW



Generated fake queries and documents

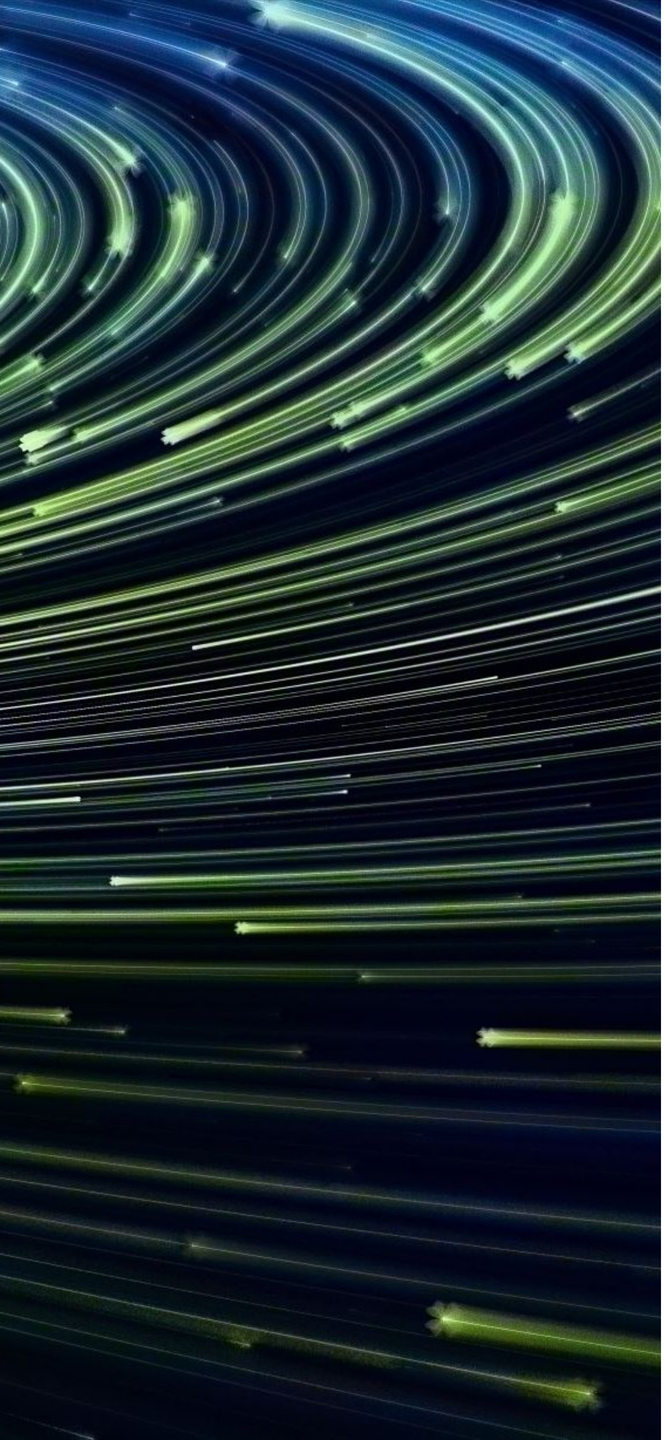
Created Synthetic relevance tables

Simulated retrieval scores

Selected confusing queries using Active learning

Simulated Human corrections

Evaluated system performance comparing Synthetic and Hybrid



ACTIVE LEARNING

- Calculated average retrieval scores for each query.
- Queries with scores close to 0.5 = most uncertain.
- Selected top 20% confusing queries for human correction.
- Improved sample selection based on system uncertainty.

HUMAN CORRECTION

Corrected top 20 documents manually for selected queries.

Human judgments were given priority over synthetic ones.

Minimal manual effort ensured scalability.

EVALUATION METRICS

- nDCG@10: Evaluates the ranking quality.
- Precision@10: Measures the accuracy of top-retrieved results.
- Used a smarter threshold of 0.4 for predictions.
- Applied double weighting for human-corrected queries during evaluation.

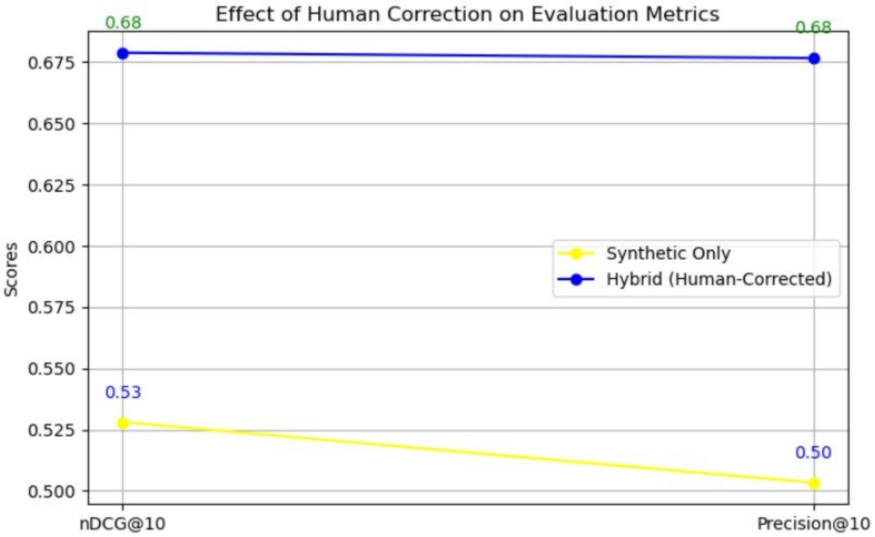
RESULTS TABLE

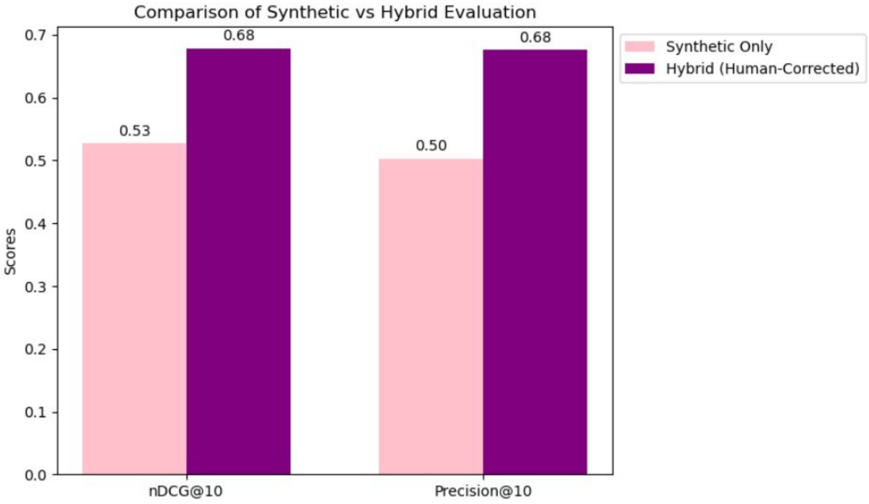
Synthetic-only evaluation		Hybrid (Human corrected) evaluation	
nDCG@10	Precision@10	nDCG@10	Precision@10
0.5280	0.5033	0.6789	0.6767

Clear and significant improvement with minimal human effort.

GRAPHS - VISUAL RESULTS

Strong visual comparison between Synthetic and Hybrid methods.





Line Graph: Human correction visibly boosts nDCG@10 and Precision@10.

Bar Chart: Strong visual comparison between Synthetic and Hybrid methods.

CONCLUSION

- Hybrid collections significantly enhance evaluation realism.
- Small-scale human intervention leads to major reliability gains.

FUTURE WORK

Implement smarter active learning strategies.

Explore using Large Language Models for scalable relevance simulation.

THANK YOU!

Happy to take any questions!