Simulation of Interactive Information Retrieval

A Guided Tour

March 22, 2026 | Seattle, WA, USA

Who we are

— **Saber Zerhoudi,** Postdoctoral Researcher
University of Passau, Germany

— **Adam Roegiest,** VP of Research and Technology
Zuva, Canada

— **Johanne Trippas,** Vice-Chancellor's Senior Research Fellow
RMIT University, Australia

Outline

01

Foundations of Simulation

02

Advanced Simulation

03

Evaluating Simulation

12:30-01:30 PM
Lunch

01:30-03:00 PM
DC / Tutorials

03:00-03:30 PM
Coffee Break

03:30-05:00 PM
DC / Tutorials

06:00-08:00 PM
Welcome Reception

CHIIR 2026

01

Foundations of Simulation

Simulation is more common **Than You Think**

You have already performed simulation.

Every metric embeds a user model

When you compute Precision@10, you assume a user who reads exactly 10 documents from the top. *That is a behavioural model.*

The model is implicit, not absent

Metrics are rarely framed as simulation, but every one encodes assumptions about how users interact with a ranked list.

The only question is fidelity

You are not choosing whether to simulate. You are choosing how realistic your simulation needs to be.

The question is not "should I simulate?",
it is "is my simulation realistic enough?"

Metrics as Simulation

Recall / Precision / ...@k

Always examines exactly k documents, top to bottom, no skipping

Limitation: Behaviour is fully predetermined, no decisions

Reciprocal Rank

Scans from rank 1, stops at the first relevant document found

Limitation: Assumes a single-target, navigational search task

Average Precision

A population of R users, each stops after finding a different number of relevant documents

Limitation: Still rule-based, no user adapts to what they see

Rank-Biased Precision

Models utility gained by searcher.

User processes list top-to-bottom with persistence (probability) P

Example: Exploratory user: p=0.90 vs Lookup user: p=0.50

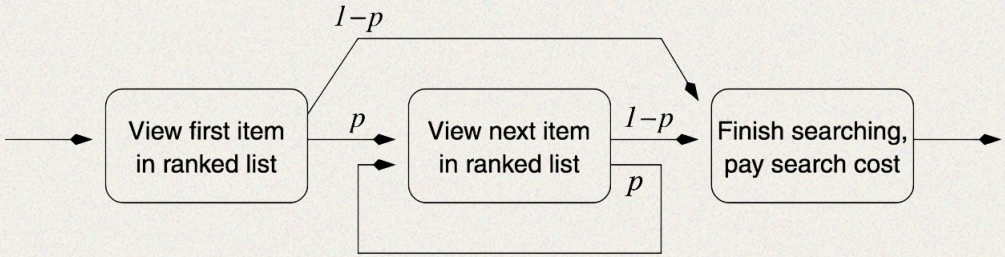


Figure 1. The user model assumed by rank-biased precision.[1]

[1] Taken from: Rank-Biased Precision for Measurement of Retrieval Effectiveness, *Moffat, Zobel, TOIS 2008*

Why simulation?

Traditional Evaluation

Costly. *User studies require recruitment, compensation, and lab time*

Hard to Scale. *Difficult to test across many system variants or user populations*

Not Reproducible. *Inherent lack of exact replication across studies*

Not Comparable. *Hard to compare multiple interactive systems fairly*

≠

What Simulation Offers

Reproducible. *Run the same experiment identically across labs and time*

Scalable. *Test thousands of user scenarios at minimal cost*

Controlled. *Isolate variables and test hypothetical perturbations*

Dynamic. *Model non-static user behaviour that changes with context*

IR metrics focus on rules-based models**,**
but real user behaviour is dynamic and context-dependent

Complex Searcher Model (Maxwell et al.)[2]

A search session is a sequence of decisions, each one can be modelled independently.

Modular

Each decision point (query formulation, snippet evaluation, click, stopping) is a separate, replaceable component.

Observable

Every decision can use information available at that moment: the query, the snippet text, documents already seen, time elapsed.

Evaluable

The quality of each decision can be measured against ground truth, enabling fine-grained simulation diagnostics.

If you can formalize **your problem, you can** simulate **it**

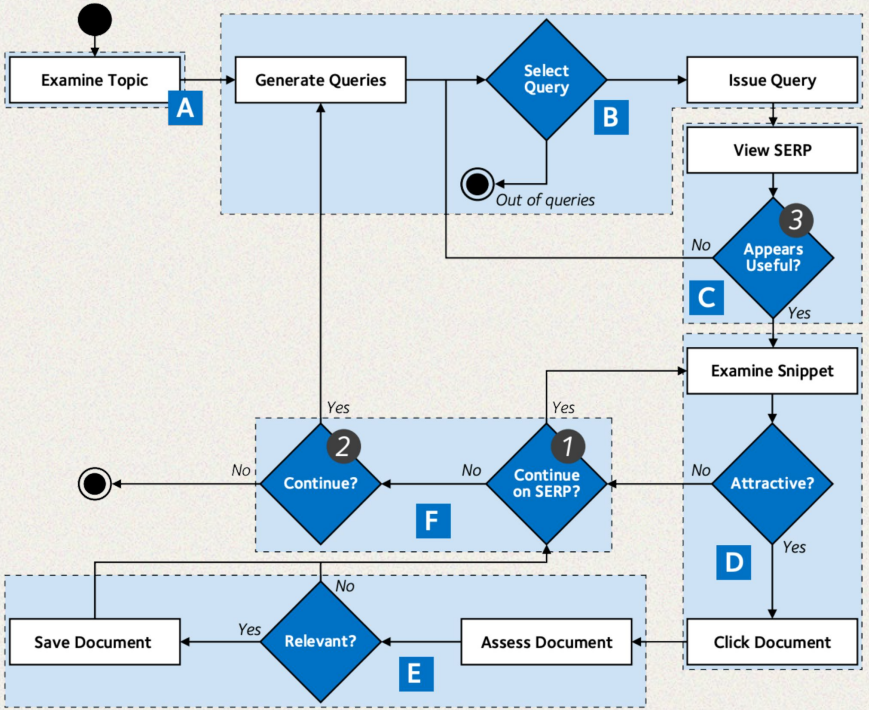
[2] Searching and Stopping: An Analysis of Stopping Rules and Strategies, *Maxwell et al., CIKM 2015*

Complex Searcher Model (Maxwell et al.)[2]

Topic: "Recent advances in retrieval-augmented generation (RAG)"

A **Examine Topic.** *Researcher identifies key concepts: RAG architecture, retrieval component, generation quality, recent work*

B **Generate & Select Query.** *Formulates query: "RAG retrieval augmented generation survey"*



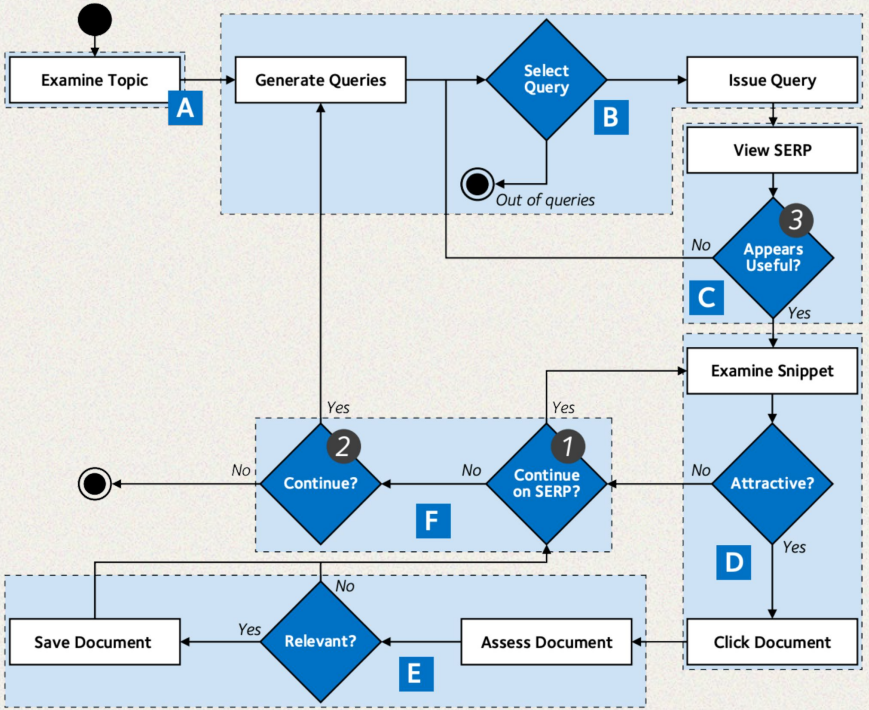
[2] Searching and Stopping: An Analysis of Stopping Rules and Strategies, *Maxwell et al., CIKM 2015*

Complex Searcher Model (Maxwell et al.)[2]

Topic: "Recent advances in retrieval-augmented generation (RAG)"

C **View SERP & Examine Snippets.** *SERP returns 10 results. Result 1 is a Semantic Scholar survey; Result 4 is an arXiv paper on "adaptive retrieval."*

D **Attractiveness Decision.** *Result 4 snippet mentions "query-dependent retrieval gating" and "hallucination reduction" → Novel angle → Click*



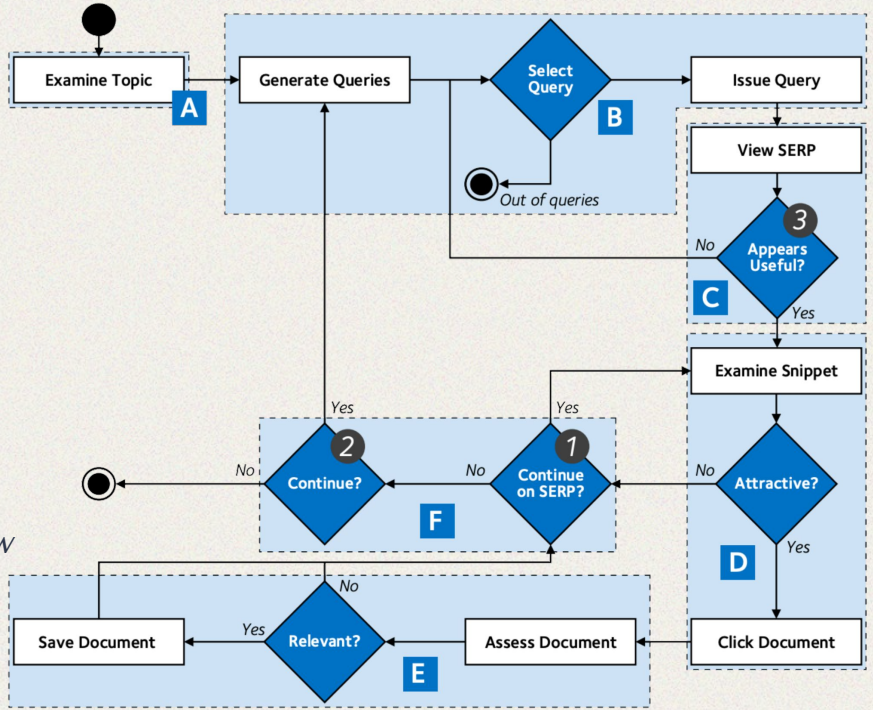
[2] Searching and Stopping: An Analysis of Stopping Rules and Strategies, *Maxwell et al., CIKM 2015*

Complex Searcher Model (Maxwell et al.)[2]

Topic: "Recent advances in retrieval-augmented generation (RAG)"

E **Assess Document.** *Paper introduces gating mechanism that skips retrieval when the LLM is confident. Relevant but narrow → Save for reference section.*

F **Continue on SERP?** *Back to SERP. Result 7 covers evaluation benchmarks for RAG → new sub-problem. Decides: Issue refined query.*



[2] Searching and Stopping: An Analysis of Stopping Rules and Strategies, *Maxwell et al., CIKM 2015*

Simulation Example

```
QUERY   600  10 extinction wildlife
SERP    600  15 EXAMINE_SERP
SNIPPET 600  18 SNIPPET_NOT_RELEVANT APW19...0801
SNIPPET 600  21 SNIPPET_NOT_RELEVANT APW19...1129
SNIPPET 600  24 SNIPPET_NOT_RELEVANT APW19...0405
SNIPPET 600  27 SNIPPET_RELEVANT APW19...1290
DOC     600  47 EXAMINING_DOCUMENT APW19...1290
SNIPPET 600  50 SNIPPET_NOT_RELEVANT APW19...0561
QUERY   600  60 extinction species wildlife
SERP    600  65 EXAMINE_SERP
SNIPPET 600  68 SNIPPET_RELEVANT APW19...0801
DOC     600  88 EXAMINING_DOCUMENT APW19...0801
MARK    600  91 CONSIDERED_RELEVANT APW19...0801
SNIPPET 600  94 SNIPPET_NOT_RELEVANT APW19...1129
SNIPPET 600  97 SNIPPET_RELEVANT APW19...0561
DOC     600 117 EXAMINING_DOCUMENT APW19...0561
MARK    600 120 CONSIDERED_RELEVANT APW19...0561
SNIPPET 600 123 SNIPPET_NOT_RELEVANT APW19...1668
SNIPPET 600 126 SNIPPET_NOT_RELEVANT APW19...0166
QUERY   600 136 extinction prevent wildlife
SERP    600 141 EXAMINE_SERP
SNIPPET 600 144 SNIPPET_NOT_RELEVANT APW19...0801
SNIPPET 600 147 SNIPPET_RELEVANT APW19...1129
DOC     600 167 EXAMINING_DOCUMENT APW19...1129
MARK    600 170 CONSIDERED_RELEVANT APW19...1129
SNIPPET 600 173 SNIPPET_NOT_RELEVANT APW19...0405
SNIPPET 600 176 SNIPPET_NOT_RELEVANT APW19...1290
SNIPPET 600 179 SNIPPET_NOT_RELEVANT APW19...0561
...
```

```
QUERY   600  10 extinction wildlife
SERP    600  15 EXAMINE_SERP
SNIPPET 600  18 SNIPPET_NOT_RELEVANT APW19...0861
SNIPPET 600  21 SNIPPET_NOT_RELEVANT APW19...1129
SNIPPET 600  24 SNIPPET_NOT_RELEVANT APW19...0405
SNIPPET 600  27 SNIPPET_RELEVANT APW19...1290
DOC     600  47 EXAMINING_DOCUMENT APW19...1290
SNIPPET 600  50 SNIPPET_NOT_RELEVANT APW19...0561
SNIPPET 600  53 SNIPPET_NOT_RELEVANT APW19...1434
SNIPPET 600  56 SNIPPET_RELEVANT APW19...0030
DOC     600  76 EXAMINING_DOCUMENT APW19...0030
QUERY   600  86 wildlife extinction in the
                philippines
SERP    600  91 EXAMINE_SERP
SNIPPET 600  94 SNIPPET_RELEVANT APW19...0801
DOC     600 114 EXAMINING_DOCUMENT APW19...0801
MARK    600 117 CONSIDERED_RELEVANT APW19...0801
SNIPPET 600 120 SNIPPET_NOT_RELEVANT APW19...1129
SNIPPET 600 123 SNIPPET_RELEVANT APW19...0561
DOC     600 143 EXAMINING_DOCUMENT APW19...0561
MARK    600 146 CONSIDERED_RELEVANT APW19...0561
SNIPPET 600 149 SNIPPET_NOT_RELEVANT APW19...1668
SNIPPET 600 152 SNIPPET_NOT_RELEVANT APW19...0166
SNIPPET 600 155 SNIPPET_NOT_RELEVANT APW19...0986
SNIPPET 600 158 SNIPPET_RELEVANT APW19...0738
DOC     600 178 EXAMINING_DOCUMENT APW19...0738
MARK    600 181 CONSIDERED_RELEVANT APW19...0738
SNIPPET 600 184 SNIPPET_NOT_RELEVANT APW19...0566
...
```

```
QUERY   600  10 extinction wildlife
SERP    600  15 EXAMINE_SERP
SNIPPET 600  18 SNIPPET_NOT_RELEVANT APW19...0801
SNIPPET 600  21 SNIPPET_NOT_RELEVANT APW19...1129
QUERY   600  26 extinction species wildlife
SERP    600  31 EXAMINE_SERP
SNIPPET 600  34 SNIPPET_NOT_RELEVANT APW19...0801
SNIPPET 600  37 SNIPPET_NOT_RELEVANT APW19...1129
SNIPPET 600  40 SNIPPET_NOT_RELEVANT APW19...0561
SNIPPET 600  43 SNIPPET_RELEVANT APW19...1668
DOC     600  63 EXAMINING_DOCUMENT APW19...1668
SNIPPET 600  66 SNIPPET_NOT_RELEVANT APW19...0166
QUERY   600  76 extinction animals wildlife
SERP    600  81 EXAMINE_SERP
SNIPPET 600  84 SNIPPET_NOT_RELEVANT APW19...0801
SNIPPET 600  87 SNIPPET_RELEVANT APW19...1129
DOC     600 107 EXAMINING_DOCUMENT APW19...1129
MARK    600 110 CONSIDERED_RELEVANT APW19...1129
QUERY   600 120 extinction prevent wildlife
SERP    600 125 EXAMINE_SERP
QUERY   600 130 extinction spotted wildlife
SERP    600 135 EXAMINE_SERP
SNIPPET 600 138 SNIPPET_RELEVANT APW19...0801
DOC     600 158 EXAMINING_DOCUMENT APW19...0801
MARK    600 161 CONSIDERED_RELEVANT APW19...0801
SNIPPET 600 164 SNIPPET_NOT_RELEVANT APW19...1129
SNIPPET 600 167 SNIPPET_RELEVANT APW19...0405
DOC     600 187 EXAMINING_DOCUMENT APW19...0405
...
```

Beyond Search

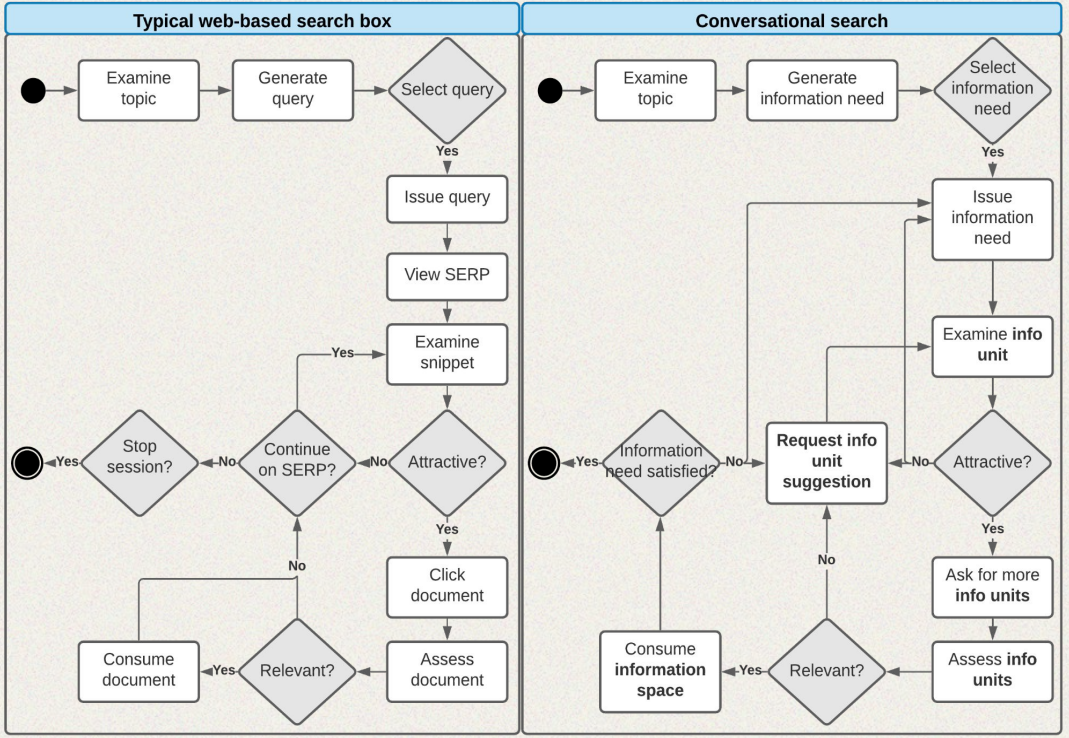
Different modalities require different simulation strategies.

Web search. Queries, SERPs, clicks, documents

Conversational search. Utterances, info units, follow-ups

The underlying decision structure is similar (express need, evaluate response, decide to continue) but the surface interactions differ enough that each requires its own simulation approach.

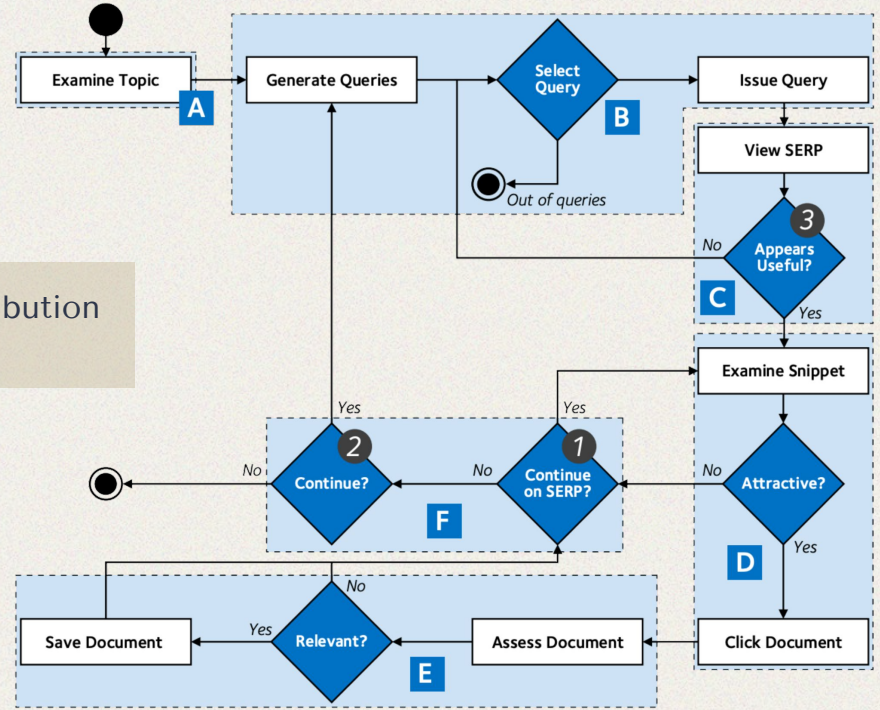
This is why flexible frameworks **matter.**



From Rules **to** Probabilities

Each decision point can be driven by a probability distribution learned from data, not a fixed rule.

From "what rule should we use?" **to** "what distribution best describes real user behaviour?"



Markov Models & Click Models

Markov Chains for User Modeling

Model user behaviour as transitions between discrete states.

Each state represents an action (e.g., querying, scanning, clicking, stopping).

Next state depends only on current state ("memoryless" property).

Transition probabilities can be learned from real user interaction logs.

Click Models

Cascade Model. *User scans results top to bottom, clicks the first attractive result, then stops examining.*

User Browsing Model (UBM). *Models position bias: click probability depends on both relevance and distance from last click.*

Dynamic Bayesian Network (DBN). *Adds hidden relevance and satisfaction variables. User may click, assess, then continue or stop based on satisfaction.*

Learn parameters from data (click logs),
then use the model to generate synthetic user interactions

Foundations: What We've Covered

Metrics are simulations.

Every IR metric encodes a fixed rule about user behaviour *(Precision@k, RR, AP, RBP)* all simulate different users.

The CSM decomposes search into decisions.

Query formulation, snippet evaluation, clicking, relevance judgement, stopping: each modelled independently.

Different modalities need different models.

Web search and conversational search share structure but differ in surface interactions.

Decisions can be probabilistic.

Markov chains and click models replace fixed rules with distributions learned from real interaction data.

Next: Your turn to think about simulation design.

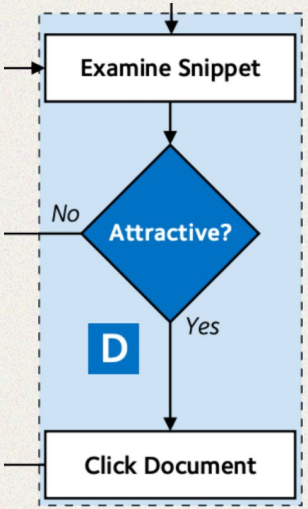
Activity Time

D Snippet Attractiveness

In groups (no more than 3) or solo:
How would you simulate a user deciding whether a snippet is attractive enough to click?

Consider: rules, probabilities, text similarity, position bias, LLMs, or something else entirely.

Take 5-10 minutes, then report back.



02

Advanced Simulation

To Framework *or* **Not To** Framework

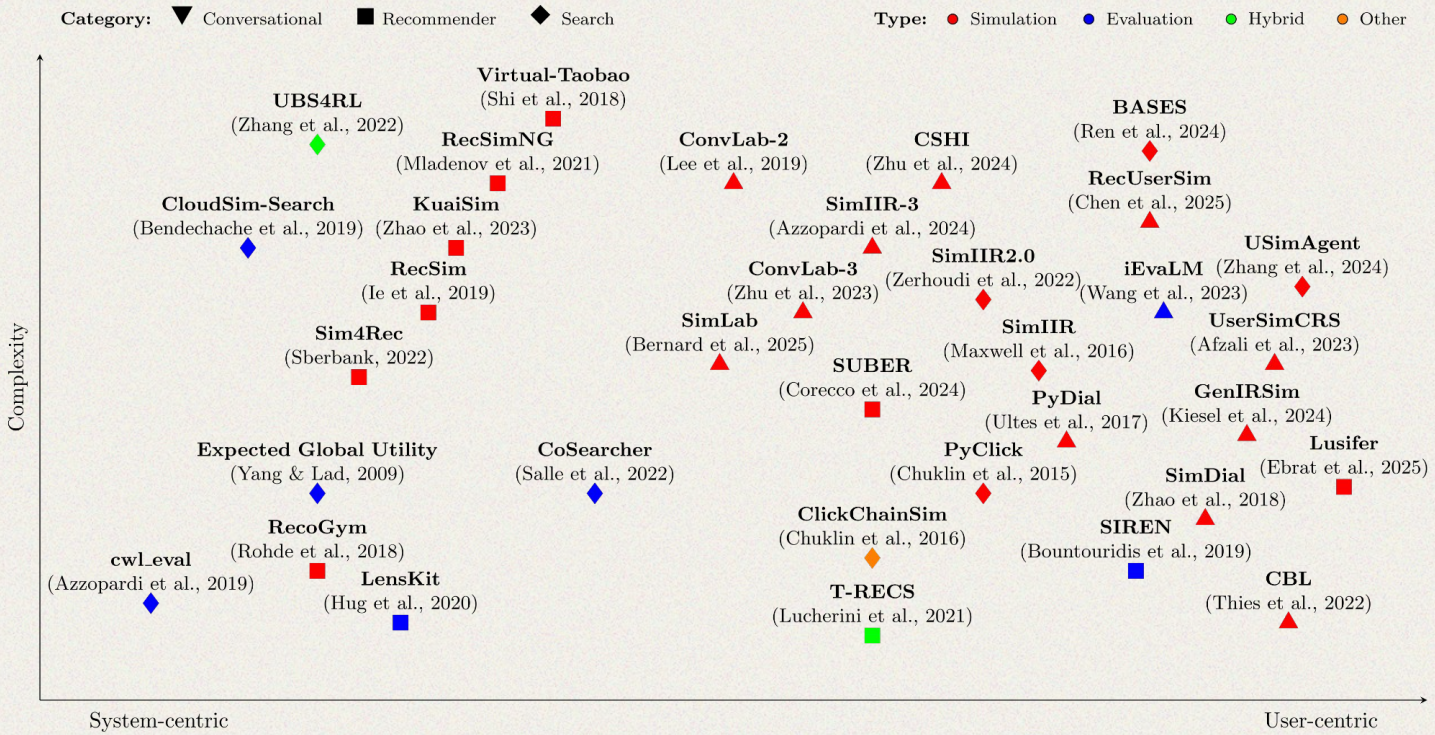
With more complex simulation**, you may find it easier to use an existing framework rather than having to manually encode a search process.**

The **downside** is that you have to buy-in into that framework and implement the choices that it wants in the way it wants.

The **upside** is that you get *(usually)* more plug-and-play of existing techniques and replication/reproducibility.

The tough part is deciding which framework to use.

To Framework *or* **Not To** Framework



[3] Theory and Toolkits for User Simulation in the Era of Generative AI: User Modeling, Synthetic Data Generation, and System Evaluation, *Balog et al., SIGIR 2025*

SimIIR 3 (Azzopardi et al.)[4]

SimIIR is a highly configurable simulation framework **with the primary goal of making simulation easy to reproduce.**

Originally focused on the **Complex Searcher Model.**

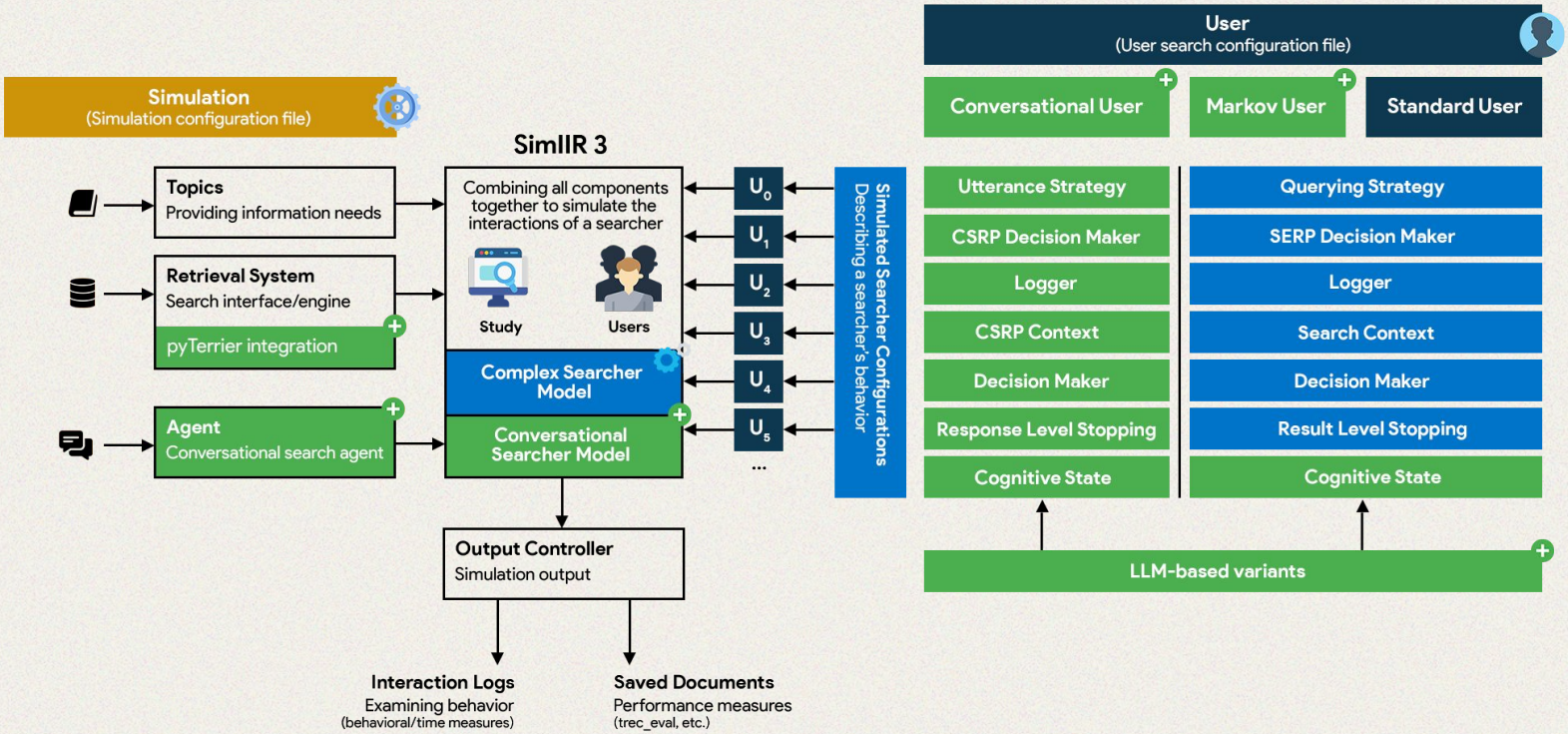
This version unifies two existing versions of the SimIIR framework.

Adds support for conversational search and generative models.

Plus some other quality of life improvements (e.g., PyTerrier support).

[4] SimIIR 3: A framework for the simulation of interactive and conversational information retrieval, *Azzopardi et al., SIGIR-AP 2024*

SimIIR 3 (Azzopardi et al.)[4]



[4] SimIIR 3: A framework for the simulation of interactive and conversational information retrieval, *Azzopardi et al., SIGIR-AP 2024*

SimIIR 3 (Azzopardi et al.)[4]

The framework comes with a set of different simulation components **pre-built.**

Many are derived from existing IR literature *(e.g., time constraints, language models).*

Can be extended by writing new Python classes to add new simulation functionality.

Each decision point has access to all the history of the current simulated user/topic.

Configuration is done via a (mostly) human-readable XML file.

```
query_generator:
  - BaseGenerator
  - BiTermGenerator
  - PredeterminedQueryGenerator
  - SingleReversedTriInterleavedGenerator
  - TriTermGenerator
  - [...]
serp_impression:
  - BaseSerpImpression
  - PerfectSerpImpression
  - SimpleSerpImpression
  - StochasticSerpImpression
decision_maker:
  - BaseDecisionMaker
  - DifferenceDecisionMaker
  - FixedDepthDecisionMaker
  - IftBasedDecisionMaker
```

[4] SimIIR 3: A framework for the simulation of interactive and conversational information retrieval, *Azzopardi et al., SIGIR-AP 2024*

Expedited Repo Walkthrough

Boy, that sucked didn’t it?

Code = Pain & Low-Level = Pain

We know many are interested in simulation but **have very little interest in the low-level code.**

We created **IIRSim** to help address these issues.

IIRSim wraps **SimIIR 3** *(and potentially other frameworks)* with a GUI.

Allows experimentation, testing, and understanding of various simulation components.

```python
import os
import sys
import gc
import logging
from utils.progress_indicator import ProgressIndicator
from sims.search_user import SimulatedUser
from sims.conversational_search_user import SimulatedConversationalUser
from utils.config_readers.simulation_config_reader import SimulationConfigReader


def main(config_filename):
    """
    The main simulation!
    For every configuration permutation, create a Simulated user object, and run the simulation (the while loop).
    Then save, report, and repeat ad naseum.
    """
    logging.basicConfig(filename='sim.log',level=logging.DEBUG)
    config_reader = SimulationConfigReader(config_filename)

    for configuration in config_reader:
        if configuration.user.type == 'ConversationalSearchUser':
            user = SimulatedConversationalUser(configuration)
        else:
            user = SimulatedUser(configuration)

        progress = ProgressIndicator(configuration)
        configuration.output.display_config()

        while not configuration.user.logger.is_finished():
            #progress.update()  # Update the progress indicator in the terminal.
            user.decide_action()

        configuration.output.display_report()
        #print "complete."
        configuration.output.save()
        gc.collect()
```

IIRSim Demonstration

Activity Time

Each task focuses on a different aspect of simulating interactive information retrieval.

Start with Task 1 and work your way up!

03

Evaluating Simulation

Why evaluate simulations?

If we want to use simulations to tell us something about humans, we need to make sure it behaves like a human.

Some evaluation is easier because it's rules-based (e.g., average reading speed, time constraints).

Much is just justifying the rationale for why certain rules are appropriate.

Evaluation becomes trickier when there are more possible outcomes.

Why evaluate **simulations?**

[5] Evaluating simulated user interaction and search behaviour, *Zerhoudi et al., ECIR 2022*
[6] Information Retrieval Evaluation as Search Simulation: A General Formal Framework for IR Evaluation, *Zhang et al., ICTIR 2017*
[7] Modeling Expected Utility of Multi-Session Information Distillation, *Yang et al., ICTIR 2009*
[8] A probability ranking principle for interactive information retrieval, *Fuhr, 2008*
[9] Discounted Cumulated Gain Based Evaluation of Multiple-Query IR Sessions, *Järvelin et al., ECIR 2008*
[10] Evaluating Multi-Query Sessions, *Kanoulas et al., SIGIR 2011*

Validating with Human Data

Requires human analogues to simulation data
(e.g., a user study, search logs).

Can examine descriptive statistics:

- If simulation stops at the same time.
- If simulation issues the same queries.
- If simulation marks the same documents as relevant.

Issues arise in that the simulation is generic but comparing to individuals is closer to personalization.

Requires you to articulate what you're trying to accomplish.

Evaluating Traces

Behavioural Fidelity

Compare simulated vs. real interaction traces step by step.

Examine sequences of actions: queries, clicks, dwell time, stopping points.

Statistical tests assess whether distributions of simulated and real traces differ significantly.

Emerging User-Oriented Metrics. *Session Difficulty Score, User Frustration Rate, Struggle Recovery Prediction*

Cognitive Fidelity

Beyond actions: does the simulated session show realistic cognitive trajectories?

Inferred states (confusion, satisfaction, scent-following) provide a new evaluation dimension.

Cognitive-enhanced models arguably improve session outcome prediction over behavioural baselines.

A good **simulation should be both** behaviourally **and** cognitively **faithful**

Bias in Simulation

Simulations provide a reflection of human behaviour.
How to simulate the decisions in the search process says much more about the researcher than the simulation itself.

Large Language Models / Generative Models increase the likelihood of unintentional bias creeping into the simulation:

- Different relevance assessing behaviours than humans.
- Different query generation behaviours than humans.

Care must be taken to ensure the simulation represents the population of interest faithfully.

All simulations are wrong, some are useful

Thank you! | Questions?

szerhoudi@acm.org | adam@roegiest.com | j.trippas@rmit.edu.au