

Evaluating Generative Ad Hoc Information Retrieval

SIGIR 2024 Perspective Paper

Lukas Gienapp, Harrisen Scells, **Niklas Deckers**, Janek Bevendorff, Shuai Wang, Johannes Kiesel, Shahbaz Syed, Maik Fröbe, Guido Zuccon, Benno Stein, Matthias Hagen, and Martin Potthast

Leipzig University, ScaDS.AI, University of Kassel, hessian.AI, Bauhaus-Universität Weimar, and The University of Queensland

17 July 2024



UNIVERSITÄT
LEIPZIG



U N I K A S S E L
V E R S I T Ä T



Bauhaus-Universität Weimar

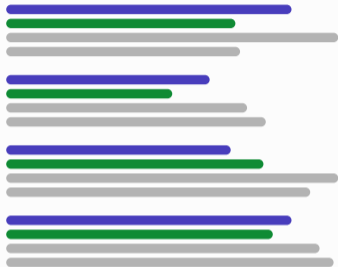
Faculty of Media



THE UNIVERSITY
OF QUEENSLAND
AUSTRALIA

Generative Information Retrieval

Query

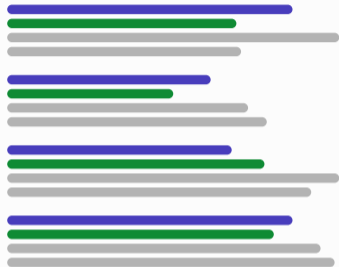


List SERP

Results are given as ranked list of links and snippets \Rightarrow Traditional IR

Generative Information Retrieval

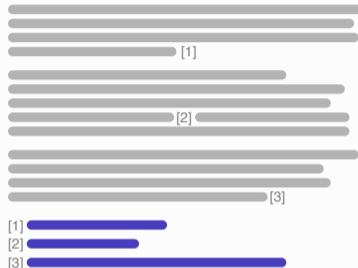
Query



List SERP

Results are given as ranked list of links and snippets \Rightarrow Traditional IR

Query



Text SERP

Results are given as single coherent response with sources \Rightarrow GenIR

The Synthetic Search Task

- Where does generative IR fit in Broder's taxonomy of search tasks (2002)?

The Synthetic Search Task

- Where does generative IR fit in Broder's taxonomy of search tasks (2002)?

Gen.	Search task	Information source	User intent	Year
1 st	informational	Document	Acquire	1995

The Synthetic Search Task

- Where does generative IR fit in Broder's taxonomy of search tasks (2002)?

Gen.	Search task	Information source	User intent	Year
1 st	informational	Document	Acquire	1995

GenIR goes beyond a single document.

The Synthetic Search Task

- Where does generative IR fit in Broder's taxonomy of search tasks (2002)?

Gen.	Search task	Information source	User intent	Year
1 st	informational	Document	Acquire	1995
2 nd	+ navigational	+ Document relations	+ Reach	1998

The Synthetic Search Task

- Where does generative IR fit in Broder's taxonomy of search tasks (2002)?

Gen.	Search task	Information source	User intent	Year
1 st	informational	Document	Acquire	1995
2 nd	+ navigational	+ Document relations	+ Reach	1998

Users of GenIR do not anticipate to reach a specific existing page.

The Synthetic Search Task

- Where does generative IR fit in Broder's taxonomy of search tasks (2002)?

Gen.	Search task	Information source	User intent	Year
1 st	informational	Document	Acquire	1995
2 nd	+ navigational	+ Document relations	+ Reach	1998
3 rd	+ transactional	+ Search verticals	+ Perform	2002

The Synthetic Search Task

- Where does generative IR fit in Broder's taxonomy of search tasks (2002)?

Gen.	Search task	Information source	User intent	Year
1 st	informational	Document	Acquire	1995
2 nd	+ navigational	+ Document relations	+ Reach	1998
3 rd	+ transactional	+ Search verticals	+ Perform	2002

In GenIR, information condensation is done on system side, not user side.

The Synthetic Search Task

- Where does generative IR fit in Broder's taxonomy of search tasks (2002)?

Gen.	Search task	Information source	User intent	Year
1 st	informational	Document	Acquire	1995
2 nd	+ navigational	+ Document relations	+ Reach	1998
3 rd	+ transactional	+ Search verticals	+ Perform	2002
4 th	+ synthetic	+ Generative models	+ Condense	2023

The Synthetic Search Task

- Where does generative IR fit in Broder's taxonomy of search tasks (2002)?
- GenIR systems go beyond existing tasks \implies 4th generation of search!

Gen.	Search task	Information source	User intent	Year
1 st	informational	Document	Acquire	1995
2 nd	+ navigational	+ Document relations	+ Reach	1998
3 rd	+ transactional	+ Search verticals	+ Perform	2002
4 th	+ synthetic	+ Generative models	+ Condense	2023

The Synthetic Search Task

- Where does generative IR fit in Broder's taxonomy of search tasks (2002)?
- GenIR systems go beyond existing tasks \implies 4th generation of search!

Gen.	Search task	Information source	User intent	Year
1 st	informational	Document	Acquire	1995
2 nd	+ navigational	+ Document relations	+ Reach	1998
3 rd	+ transactional	+ Search verticals	+ Perform	2002
4 th	+ synthetic	+ Generative models	+ Condense	2023

Synthetic Search Task

Provide a single, comprehensive, generated answer document in response to a complex information need by condensing information from multiple sources.

Evaluation of GenIR: Motivation

Problem: evaluation of Generative IR has not been systematically investigated.

Evaluation of GenIR: Motivation

Problem: evaluation of Generative IR has not been systematically investigated.

Traditional IR

- Investigated for decades
- Robust theoretical foundation
- Proven reliability of evaluation

Evaluation of GenIR: Motivation

Problem: evaluation of Generative IR has not been systematically investigated.

Traditional IR

- Investigated for decades
- Robust theoretical foundation
- Proven reliability of evaluation

Generative IR

- Recently emerged paradigm
- No theoretical foundation (yet)
- No practical experience

Evaluation of GenIR: Motivation

Problem: evaluation of Generative IR has not been systematically investigated.

Traditional IR

- Investigated for decades
- Robust theoretical foundation
- Proven reliability of evaluation

Generative IR

- Recently emerged paradigm
- No theoretical foundation (yet)
- No practical experience

Unique opportunity: transfer tried & tested methodology

Solution: Survey Traditional IR to inform Generative IR evaluation

Evaluation of GenIR: Steps

Three step process (Agosti, 2014):

1. Define evaluation objectives capturing the task
2. Derive a corresponding user model
3. Operationalize the user model for experiments

Evaluation of GenIR: Steps

Three step process (Agosti, 2014):

1. Define evaluation objectives capturing the task
2. Derive a corresponding user model
3. Operationalize the user model for experiments

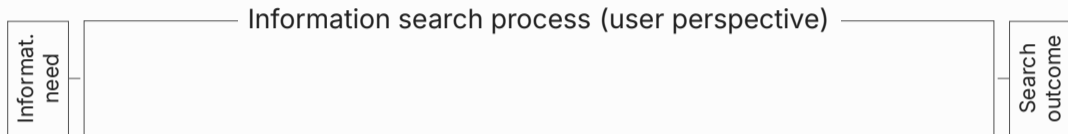
At each step, we survey existing literature from IR and related fields to apply known concepts to the new task.

Evaluation of GenIR: Objectives

- We ground our objectives in the information search process (Vakkari, 2016)

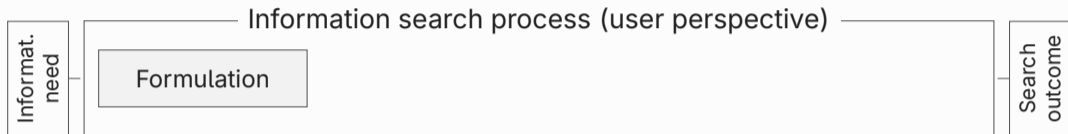
Evaluation of GenIR: Objectives

- We ground our objectives in the information search process (Vakkari, 2016)



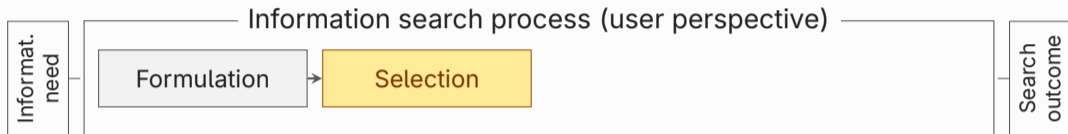
Evaluation of GenIR: Objectives

- We ground our objectives in the information search process (Vakkari, 2016)



Evaluation of GenIR: Objectives

- We ground our objectives in the information search process (Vakkari, 2016)

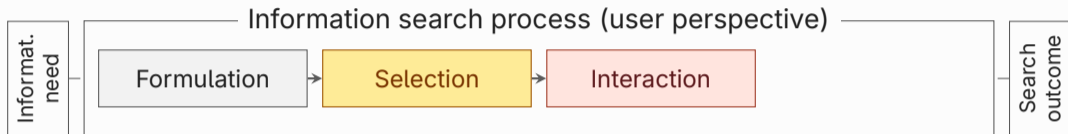


A traditional IR system user...

... selects sources from list SERP for further investigation

Evaluation of GenIR: Objectives

- We ground our objectives in the information search process (Vakkari, 2016)

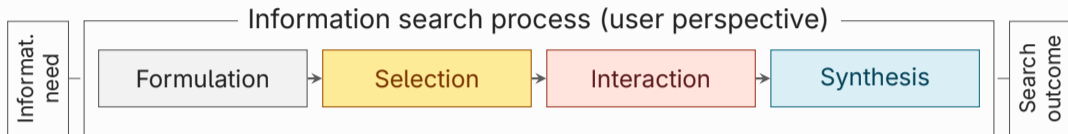


A traditional IR system user...

- ... selects sources from list SERP for further investigation
- ... analyzes selected sources for relevant information

Evaluation of GenIR: Objectives

- We ground our objectives in the information search process (Vakkari, 2016)

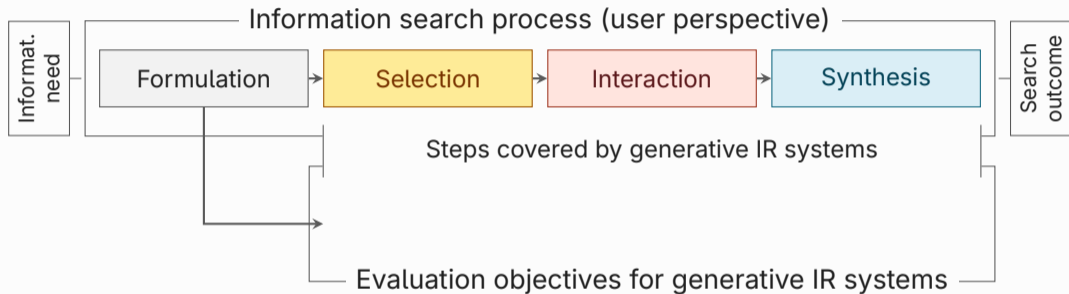


A traditional IR system user...

- ... selects sources from list SERP for further investigation
- ... analyzes selected sources for relevant information
- ... combines the information found in sources

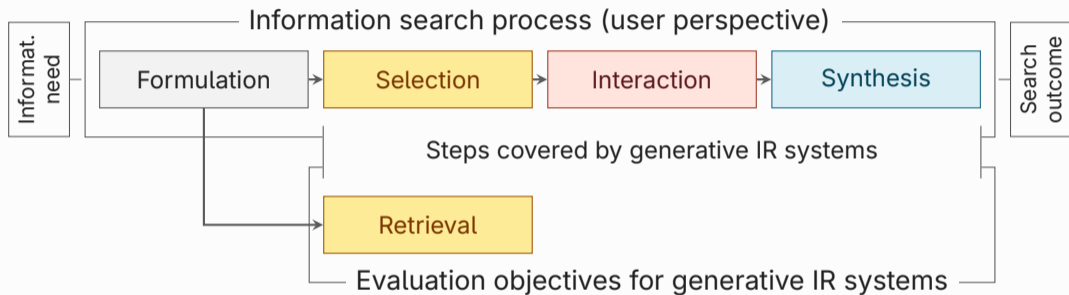
Evaluation of GenIR: Objectives

- We ground our objectives in the information search process (Vakkari, 2016)
- Generative IR systems (partly) replace steps of this process



Evaluation of GenIR: Objectives

- We ground our objectives in the information search process (Vakkari, 2016)
- Generative IR systems (partly) replace steps of this process

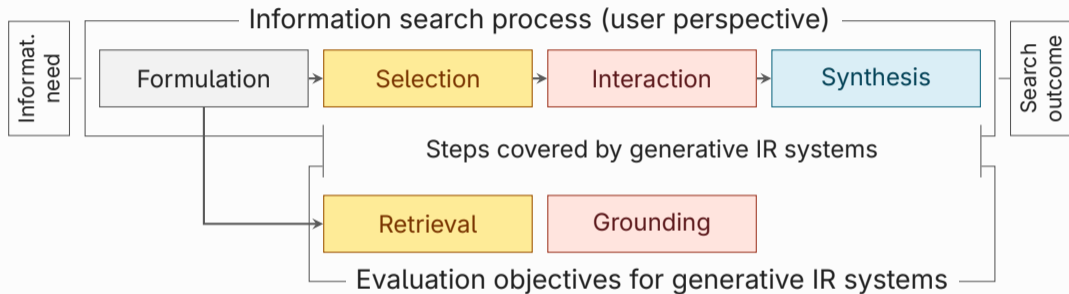


A generative IR system should...

... identify relevant, informative, correct, and diverse sources

Evaluation of GenIR: Objectives

- We ground our objectives in the information search process (Vakkari, 2016)
- Generative IR systems (partly) replace steps of this process

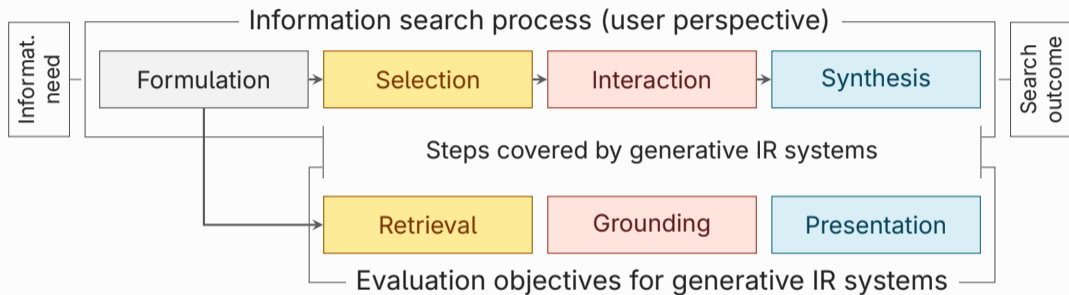


A generative IR system should...

- ... identify relevant, informative, correct, and diverse sources
- ... correlate generated output with sources correctly and consistently

Evaluation of GenIR: Objectives

- We ground our objectives in the information search process (Vakkari, 2016)
- Generative IR systems (partly) replace steps of this process



A generative IR system should...

- ... identify relevant, informative, correct, and diverse sources
- ... correlate generated output with sources correctly and consistently
- ... condense information into a concise, coherent, and accessible form

User Model for GenIR: Overview

- Offline IR evaluation experiments are grounded in a user model
- User model consists of three components (Carterette, 2011):

User Model for GenIR: Overview

- Offline IR evaluation experiments are grounded in a user model
- User model consists of three components (Carterette, 2011):

1. Utility

How valuable a result is to a user

User Model for GenIR: Overview

- Offline IR evaluation experiments are grounded in a user model
- User model consists of three components (Carterette, 2011):

1. Utility

How valuable a result is to a user

2. Browsing

How users proceed through results

User Model for GenIR: Overview

- Offline IR evaluation experiments are grounded in a user model
- User model consists of three components (Carterette, 2011):

1. **Utility** How valuable a result is to a user
2. **Browsing** How users proceed through results
3. **Accumulation** How users combine across results

User Model for GenIR: Overview

- Offline IR evaluation experiments are grounded in a user model
- User model consists of three components (Carterette, 2011):

1. Utility	How valuable a result is to a user	<u>Trad. IR example: nDCG</u> Gain function
2. Browsing	How users proceed through results	Discount function
3. Accumulation	How users combine across results	Expected total utility

User Model for GenIR: Overview

- Offline IR evaluation experiments are grounded in a user model
- User model consists of three components (Carterette, 2011):

		<u>Trad. IR example: nDCG</u>
1. Utility	How valuable a result is to a user	Gain function
2. Browsing	How users proceed through results	Discount function
3. Accumulation	How users combine across results	Expected total utility

We motivate specific instantiations of each in our paper.

User Model for GenIR: Granularity

Text SERPs can be evaluated on different levels of granularity:

User Model for GenIR: Granularity

Text SERPs can be evaluated on different levels of granularity:

Response-level evaluation

- Text SERP is annotated as a whole
- Each receives a single label
- Label can be used directly

User Model for GenIR: Granularity

Text SERPs can be evaluated on different levels of granularity:

Response-level evaluation

- Text SERP is annotated as a whole
- Each receives a single label
- Label can be used directly

Statement-level evaluation

- Text SERP is split into statements
- Aggregation over statement labels
- Needs an aggregation measure

User Model for GenIR: Granularity

Text SERPs can be evaluated on different levels of granularity:

Response-level evaluation

- Text SERP is annotated as a whole
- Each receives a single label
- Label can be used directly




Statement-level evaluation

- Text SERP is split into statements
- Aggregation over statement labels
- Needs an aggregation measure

The evaluation objectives are implemented differently depending on granularity.

User Model for GenIR: Utility

Five main utility dimensions structure how the response serves the user:

Objective	Statement Level	Response Level
 Retrieval		
 Grounding		
 Presentation		

User Model for GenIR: Utility

Five main utility dimensions structure how the response serves the user:

Objective	Statement Level	Response Level
● Retrieval	Correctness	Coverage
● Grounding		
● Presentation		

Correctness : a statement conveys information that is factual, reliable, and relevant.

Coverage : a response should provide a broad range of in-depth information.

User Model for GenIR: Utility

Five main utility dimensions structure how the response serves the user:

Objective	Statement Level	Response Level
● Retrieval	Correctness	Coverage
● Grounding	Consistency	
● Presentation		

Consistency (external): a statement should accurately convey its sources.

Consistency (internal): a response should not contain conflicting information.

User Model for GenIR: Utility

Five main utility dimensions structure how the response serves the user:

Objective	Statement Level	Response Level
● Retrieval	Correctness	Coverage
● Grounding	Consistency	
● Presentation	Clarity	Coherence

Clarity : a statement should be expressed in clear and user-accessible.

Coherence : a response should be arranged to form a coherent narrative.

User Model for GenIR: Browsing

- Browsing in traditional IR can be described by C/W/L framework
- Browsing in generative IR is formed by reading behavior of users
- Three relevant effects observed in reading comprehension studies:

User Model for GenIR: Browsing

- Browsing in traditional IR can be described by C/W/L framework
- Browsing in generative IR is formed by reading behavior of users
- Three relevant effects observed in reading comprehension studies:

Progression – users read a text sequentially from the start

User Model for GenIR: Browsing

- Browsing in traditional IR can be described by C/W/L framework
- Browsing in generative IR is formed by reading behavior of users
- Three relevant effects observed in reading comprehension studies:

Progression – users read a text sequentially from the start

Decay – attention of users diminishes throughout the text

User Model for GenIR: Browsing

- Browsing in traditional IR can be described by C/W/L framework
- Browsing in generative IR is formed by reading behavior of users
- Three relevant effects observed in reading comprehension studies:

Progression – users read a text sequentially from the start

Decay – attention of users diminishes throughout the text

Saturation – users abort reading when their information need is satisfied

User Model for GenIR: Browsing

- Browsing in traditional IR can be described by C/W/L framework
- Browsing in generative IR is formed by reading behavior of users
- Three relevant effects observed in reading comprehension studies:

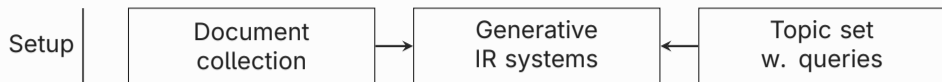
Progression – users read a text sequentially from the start

Decay – attention of users diminishes throughout the text

Saturation – users abort reading when their information need is satisfied

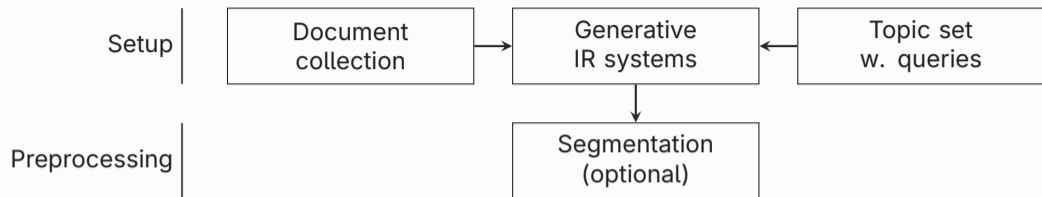
Traditional IR browsing models can thus be transferred to generative IR.

Summary: Evaluation Experiments



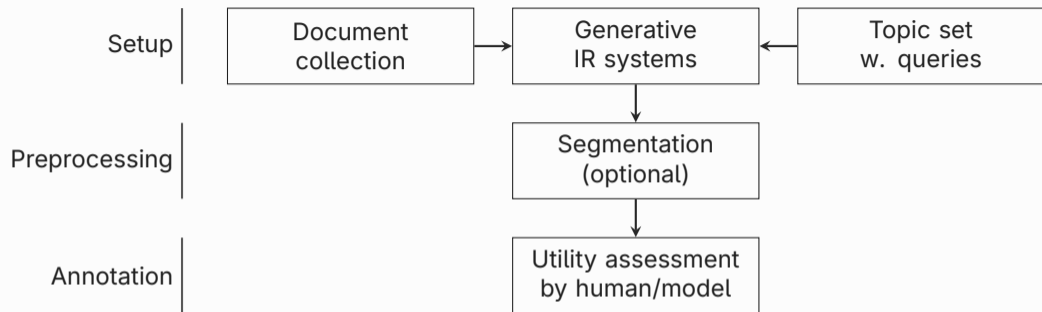
- Traditional experimental setup

Summary: Evaluation Experiments



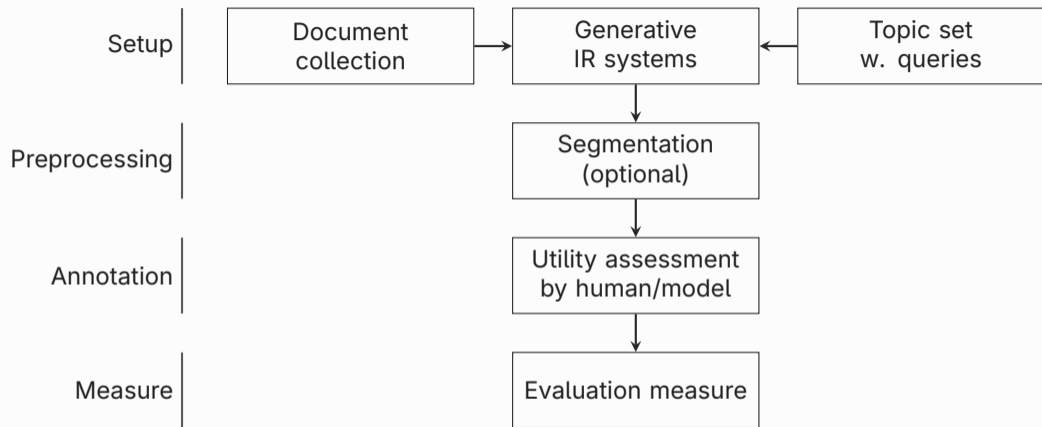
- Traditional experimental setup with added optional preprocessing step

Summary: Evaluation Experiments



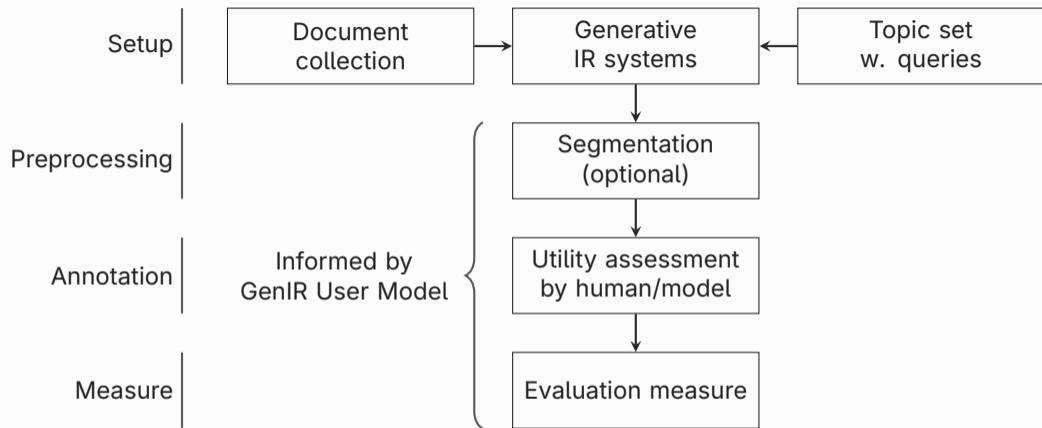
- Traditional experimental setup with added optional preprocessing step

Summary: Evaluation Experiments



- Traditional experimental setup with added optional preprocessing step

Summary: Evaluation Experiments



- Traditional experimental setup with added optional preprocessing step
- Segmentation, assessment, measures are informed by a GenIR user model

Outlook

Key Insights & Contributions

- We posit GenIR as new synthetic search task, orthogonal to traditional tasks
- We survey existing evaluation methodology and transfer it to new paradigm
- We provide systematic theoretical foundation for GenIR evaluation experiments

Outlook

Key Insights & Contributions

- We posit GenIR as new synthetic search task, orthogonal to traditional tasks
- We survey existing evaluation methodology and transfer it to new paradigm
- We provide systematic theoretical foundation for GenIR evaluation experiments

Current & Future Developments

- TREC -RAG adopts a compatible evaluation methodology (response- and statement-level, covering all three objectives, human- and model-based) and is (coincidentally) an instantiation of our framework
- Verification of proposed method through user studies is planned
- Possibilities of model-based evaluation for GenIR are planned

Outlook: User Model for GenIR: Granularity

Text SERPs can be evaluated on different levels of granularity:

Response-level evaluation

- Text SERP is annotated as a whole
- Each receives a single label
- Label can be used directly

Statement-level evaluation

- Text SERP is split into statements
- Aggregation over statement labels
- Needs an aggregation measure

Outlook: User Model for GenIR: Granularity

Text SERPs can be evaluated on different levels of granularity:

Response-level evaluation

- Text SERP is annotated as a whole
- Each receives a single label
- Label can be used directly

Statement-level evaluation

- Text SERP is split into statements
- Aggregation over statement labels
- Needs an aggregation measure

TREC RAG uses a segmentation into sentences (easy segmentation and evaluation).

Outlook: User Model for GenIR: Utility

Five main utility dimensions structure how the response serves the user:

Objective	Statement Level	Response Level
 Retrieval		
 Grounding		
 Presentation		

Outlook: User Model for GenIR: Utility

Five main utility dimensions structure how the response serves the user:

Objective	Statement Level	Response Level
● Retrieval	Correctness	Coverage
● Grounding		
● Presentation		

Correctness : TREC RAG plans to detect hallucinations through human annotation of bad nuggets; misinformation is not yet explored.

Coverage : TREC RAG checks whether all relevant nuggets are actually used.

Outlook: User Model for GenIR: Utility

Five main utility dimensions structure how the response serves the user:

Objective	Statement Level	Response Level
● Retrieval	Correctness	Coverage
● Grounding	Consistency	
● Presentation		

Consistency (external): TREC RAG checks whether statements are conveyed by the sources.

Consistency (internal): Planned.

Outlook: User Model for GenIR: Utility

Five main utility dimensions structure how the response serves the user:

Objective	Statement Level	Response Level
● Retrieval	Correctness	Coverage
● Grounding	Consistency	
● Presentation	Clarity	Coherence

Clarity : TREC RAG uses statement-level clarity metrics.

Coherence : Planned.

Outlook: User Model for GenIR: Browsing

- Browsing in traditional IR can be described by C/W/L framework
- Browsing in generative IR is formed by reading behavior of users
- Three relevant effects observed in reading comprehension studies:

Progression – users read a text sequentially from the start

Decay – attention of users diminishes throughout the text

Saturation – users abort reading when their information need is satisfied

Outlook: User Model for GenIR: Browsing

- Browsing in traditional IR can be described by C/W/L framework
- Browsing in generative IR is formed by reading behavior of users
- Three relevant effects observed in reading comprehension studies:

Progression – users read a text sequentially from the start

Decay – attention of users diminishes throughout the text

Saturation – users abort reading when their information need is satisfied

TREC RAG plans to do aggregation by averaging utility over all sentences.
Enforcing a maximal response length implies a browsing model
and punishes redundancy.