

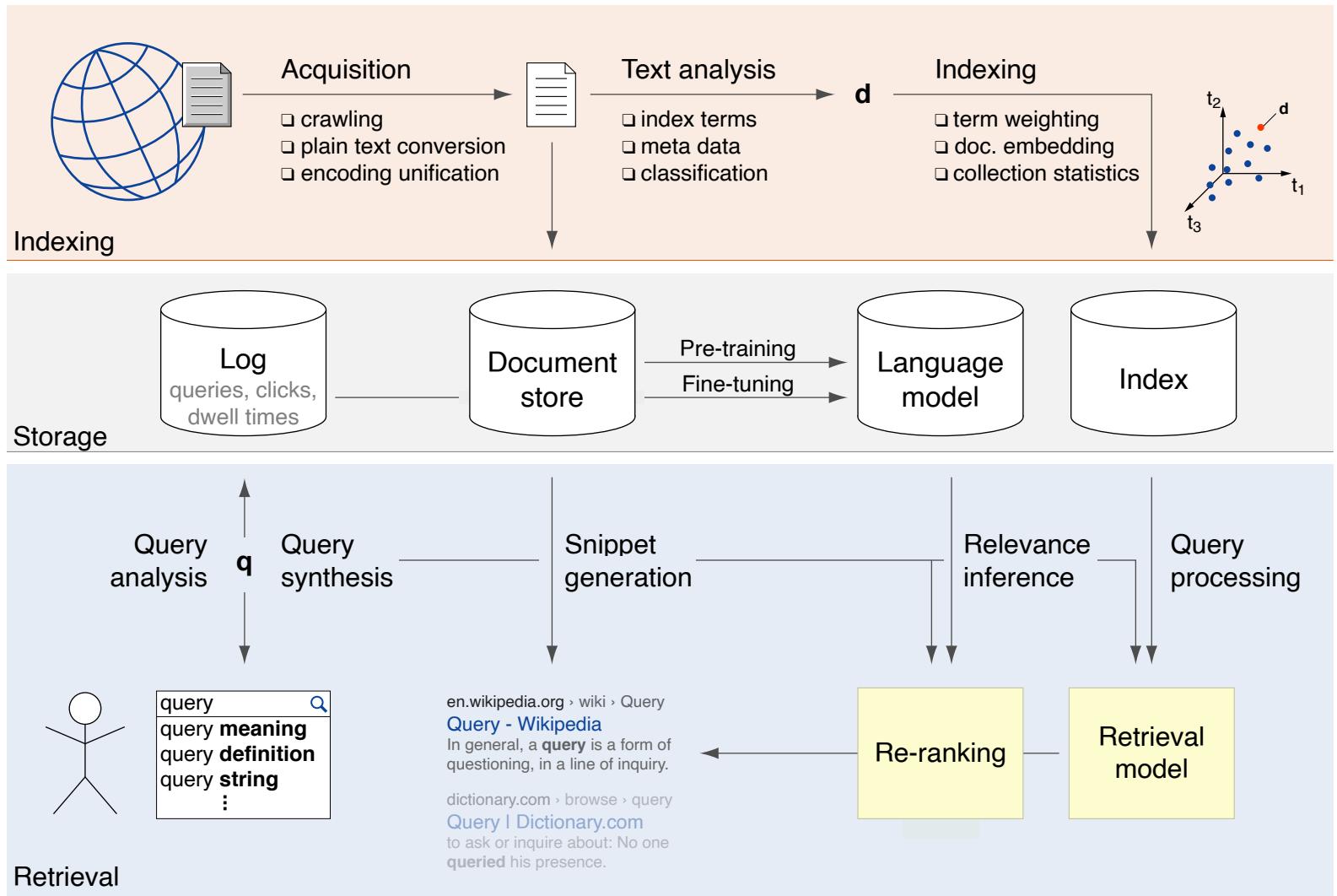
Generative Information Retrieval

Overview

- ❑ What is Generative IR
- ❑ Methods for Generative IR
- ❑ Evaluation of Generative IR
- ❑ Multi-Modal Generative IR
- ❑ Open Challenges

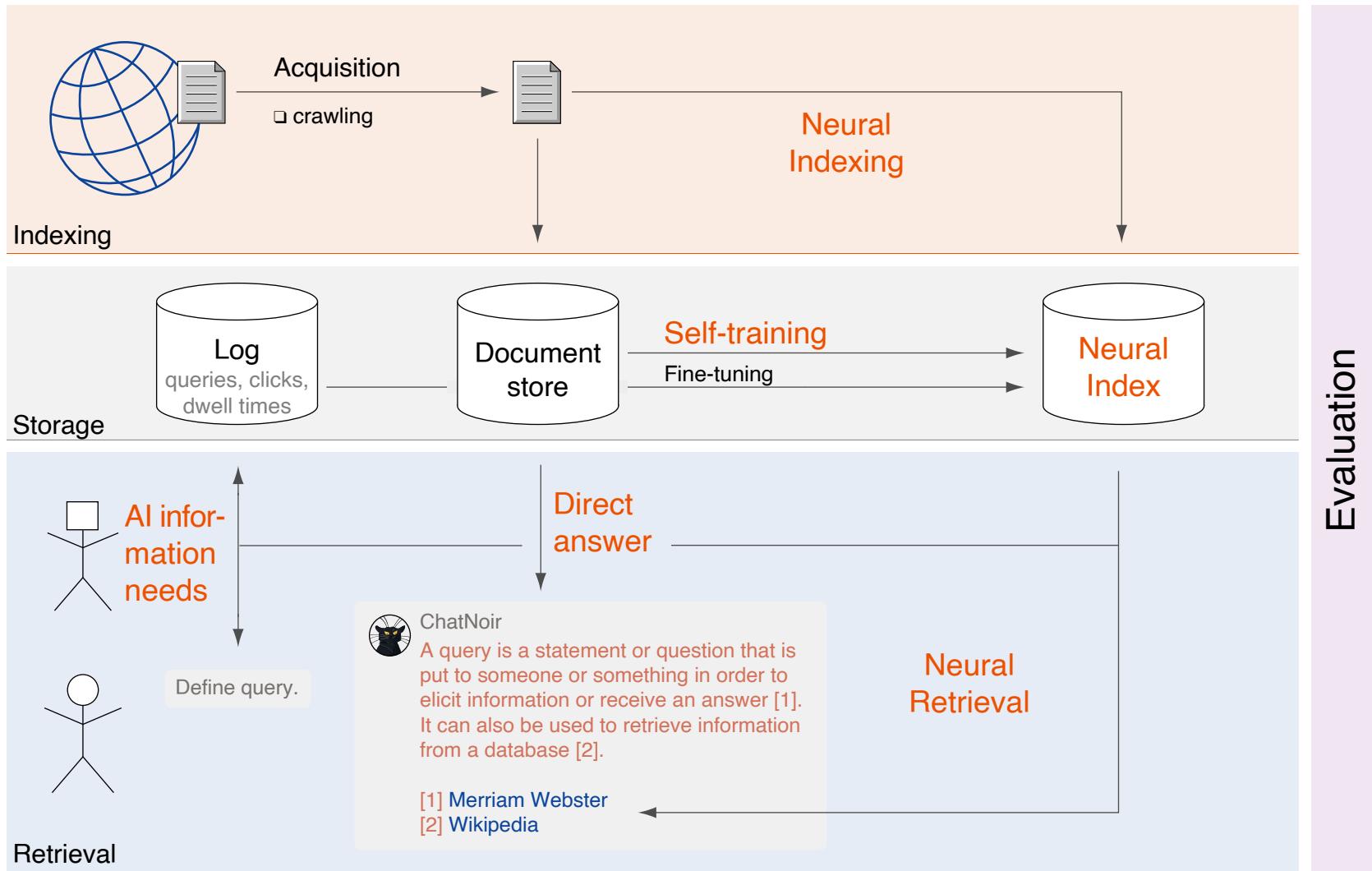
What is Generative IR

Overview



What is Generative IR

Overview



What is Generative IR

Overview



ChatNoir

Bonjour! I'm ChatNoir, your friendly cat
assistant. How can I help you today?

How can a computer's intelligence be tested?



ChatNoir

To test computer intelligence, you can give it problems to solve and evaluate how it performs. You can also give it tasks that require it to think outside the box, such as creating its own algorithms or coming up with innovative solutions to problems. Additionally, you can give it tests that measure its knowledge and understanding of a given subject.

This is outlined in [1], which also covers other definitions and tests of machine intelligence proposed, such as [2] which provides an overview of these definitions and tests. Finally, [3] is a standard test for machine intelligence.

www.wired.com/.../pass-turing-ai-test/

[1] **How to Pass the Turing Artificial Intelligence Test | WIRED**

ClueWeb22 · Crawled Aug 2022

How to Pass the Turing Artificial Intelligence Test | WIRED WIRED Staff Science Jun 19, 2012 2:00 PM How to Pass the Turing Artificial Intelligence Test Are you human or a machine? Prove it, by passing the Turing Test — a test of a machine's ability to exhibit intelligent behavior.

www.shivonzilis.com/

[2] **Machine Intelligence - Shivon Zilis**

ClueWeb22 · Crawled Aug 2022

Teams need to figure out how to test newfound capabilities, and applications need to change so they offer more than a system of record; they also need to coach employees and learn from the data they enter.

www.geeksforgeeks.org/turing-test-artificial-intelligence/

[3] **Turing Test in Artificial Intelligence - GeeksforGeeks**

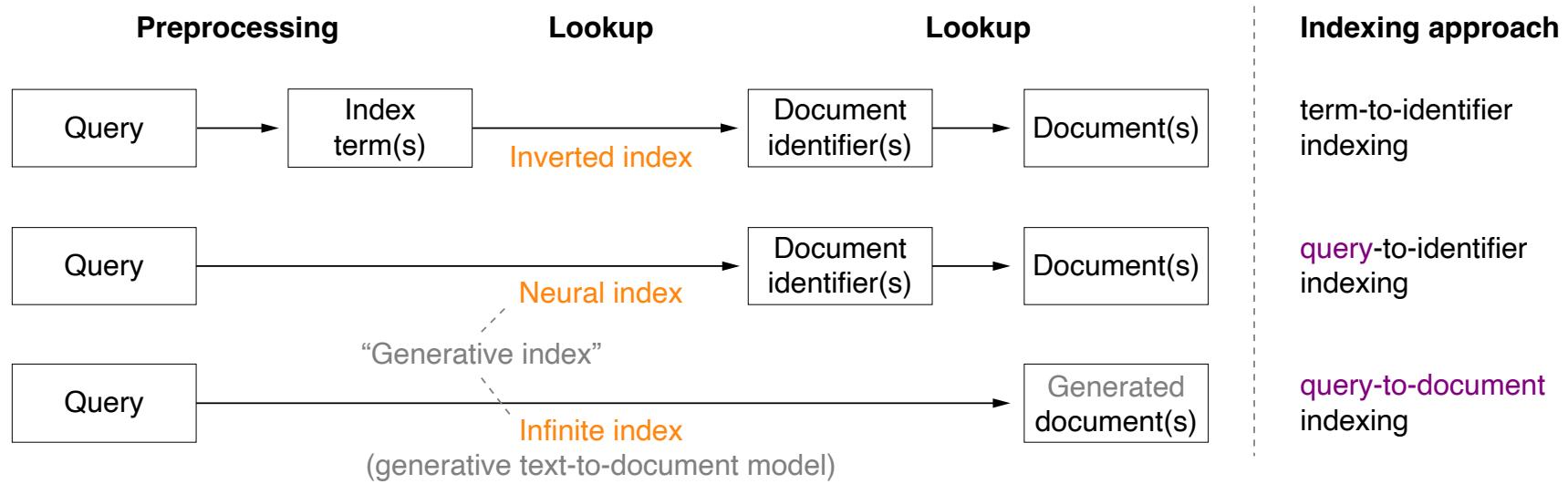
ClueWeb22 · Crawled Aug 2022

If a machine engages in a conversation with a human how to process the data it has been demonstrated by a machine, He has proposed the following skills of the test as follows: The turning judges the conversational skills of humans.

Show more...

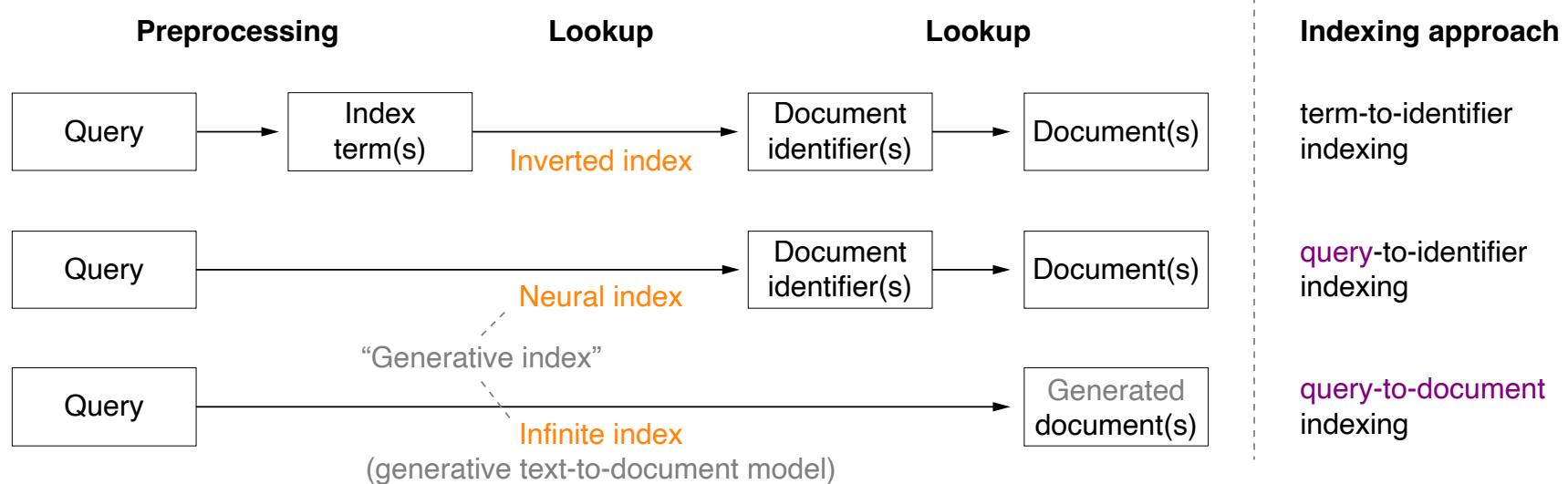
What is Generative IR

Overview



What is Generative IR

Overview



Two types of ‘Generative IR’:

- query-to-identifier:
 - Given a query, generate the document identifiers of relevant documents.
- query-to-document:
 - Given a query, generate the text of a relevant document.

What is Generative IR

query-to-identifier

Idea:

- IR systems typically have two phases: indexing and retrieval.
- Instead of storing index separately, store index inside the model parameters.
- Train model to associate documents with document identifiers.
- At inference time, autoregressively generate document identifiers.

What is Generative IR

query-to-identifier

Idea:

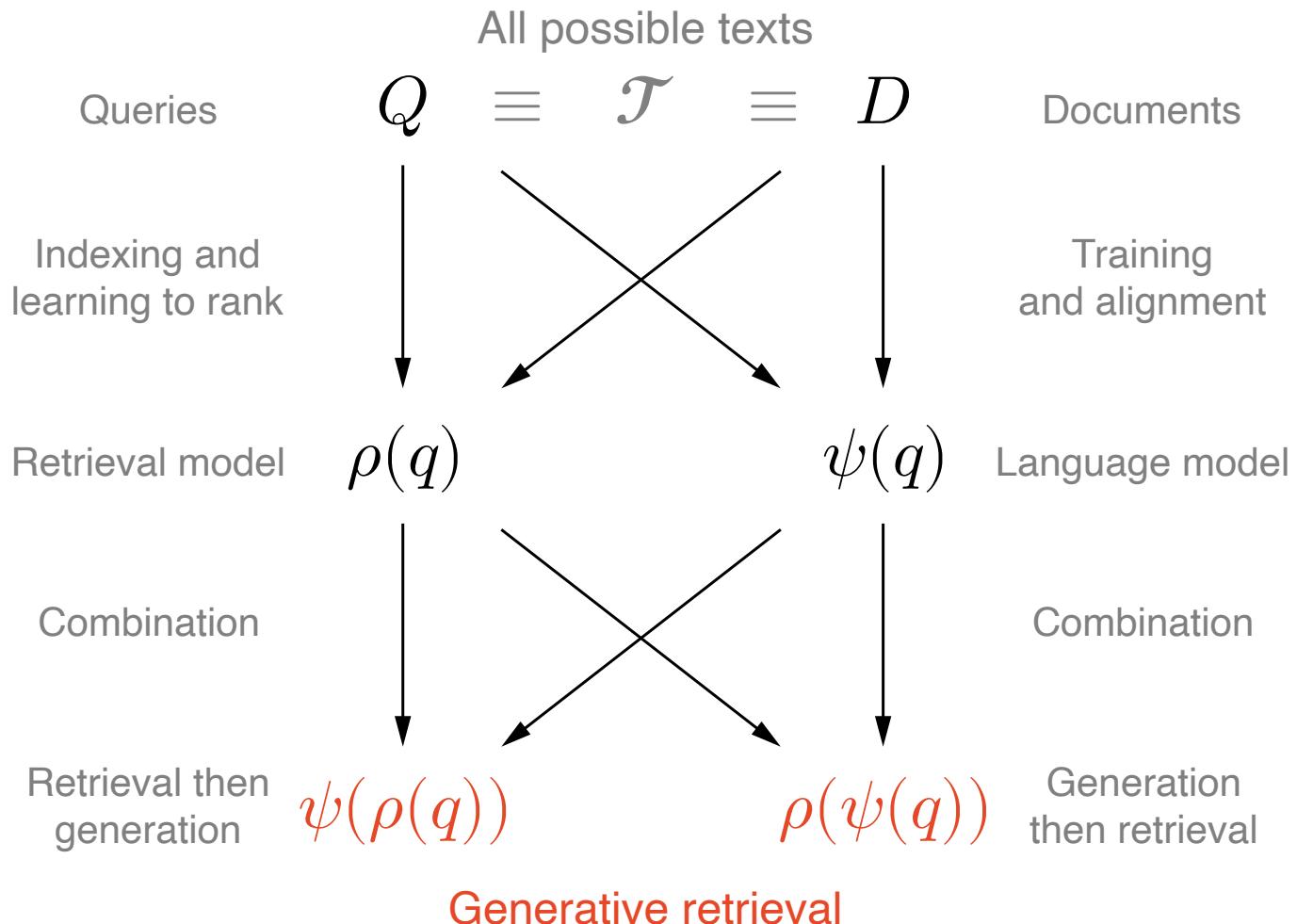
- IR systems typically have two phases: indexing and retrieval.
- Instead of storing index separately, store index inside the model parameters.
- Train model to associate documents with document identifiers.
- At inference time, autoregressively generate document identifiers.

Several architectures:

- Differentiable Search Index (DSI) [[Tay et al. 2022](#)]
 - Overfit a encoder-decoder transformer to predict identifiers given a query.
 - Improvements to DSI: DSI++ [[Vaibhav et al. 2022](#)], DSI-QG [[Zhuang et al. 2023](#)]
- Generating Substrings as Document Identifiers (SEAL) [[Bevilacqua et al. 2022](#)]
 - Constrain decoding and use generated n-gram to match documents.
 - Improvements to SEAL: MINDER [[Li et al. 2023](#)], LTRGR [[Li et al. 2023](#)]
- Neural Corpus Indexer (NCI) [[Wang et al. 2022](#)]
 - Encode semantics into document identifiers to help with decoding.

What is Generative IR

query-to-document [Gienapp et al. 2023]



What is Generative IR

query-to-document in the Wild

H best pizza place in leipzig

best pizza place in leipzig

PaLM2

Regenerate draft ⚡ 🔍

best pizza place in leipzig

Generating answers for you...

Certainly! Here are some of the best pizza places in Leipzig:

1. L'Osteria Leipzig: This Italian restaurant is located in the city center and serves delicious pizza and pasta dishes. Their pizza is so large that it can easily be shared with friends or family members. They also have fresh salads, antipasti, and a weekly menu with various pasta creations. The restaurant has a unique atmosphere and a terrace that attracts guests in the summer. Families are welcome too! [1](#)
2. La Strada Ristorante & Pizzeria: This restaurant is known for its awesome pizza and very nice service. It's definitely worth a visit! [2](#)
3. 60 Seconds to Napoli Leipzig: This restaurant serves creative Naples-style pizzas. Their pizza is rated 10/10 and their beer is rated 2/10 [3](#)
4. Rizzi: This restaurant is located in the city center and serves good Italian food with big portions [4](#)
5. Pizza Hut Leipzig, HBF: This restaurant is located in Promenaden im Hauptbahnhof and serves American-style pizza [5](#)

I hope this helps you find the perfect pizza place! 

Learn more [🔗](#)

1  [losteria.net](#) 2  [tripadvisor.com](#)
3  [tripadvisor.com](#) 4  [rizzi-leipzig.de](#)
5  [pizzahut.de](#) 6  [bing.com](#)
7  [60secondstonapoli.de](#) 8  [tripadvisor.de](#)
9  [tripadvisor.co.uk](#)

1 of 4

best pizza place in leipzig

Google Maps

Here are some highly-rated pizza places in Leipzig:

- Pizzeria Napoli Leipzig at Jahnallee 27, 04109 Leipzig, Germany has a 4.3 star rating on Google Maps.
- AMOR at Münzgasse 3, 04107 Leipzig, Germany has a 4.2 star rating on Google Maps.
- Osteria L'Area at Jahnallee 8, 04109 Leipzig, Germany has a 4.1 star rating on Google Maps.
- il CORSO fast food at Burgstraße 12, 04109 Leipzig, Germany has a 4.6 star rating on Google Maps.
- La Strada Pizzaria & Eiscafe at Kochstraße 1b, 04275 Leipzig, Germany has a 4.6 star rating on Google Maps.

Google Maps

Places



A Google Map of Leipzig, Germany, centered on the city center. Several red location pins are placed on the map, indicating the locations of highly-rated pizza places. The pins are scattered across the central business district, with labels like 'Leipzig', 'Museum der Bildenden Künste Leipzig', 'GRASSI Museum Angewandte K.', 'Felsenkeller Leipzig', 'Clara-Zetkin-Park', and 'Panometer Leipzig' visible. The map also shows major roads like 'Erlaustrasse', 'Albert-Schweitzer-Straße', and 'Hochstraße'. The surrounding neighborhoods are labeled as 'ALT-LINDENAU', 'ZENTRUM-OST', 'REUDN', 'SUDVORSTADT', and 'ZENTRUM-SÜDOST'.

Based on the search results, here are some of the best pizza places in Leipzig:

1. **Fra Diavolo:** This restaurant offers good Italian food with big portions. It has received positive reviews.
2. **Da Salvo:** Known for its authentic Italian cuisine and great pizza, Da Salvo is a popular choice among locals and visitors [1](#).
3. **Pekar:** Pekar is highly regarded as one of the best places to enjoy pizza in Leipzig. It is known for its awesome pizzas.
4. **Andria - Leipzig:** With a menu that includes Italian and pizza dishes, Andria is a recommended Italian restaurant in Leipzig. It is known for its good food and has received positive reviews [1](#).
5. **Allegretto:** Allegretto is a favorite among families and is known for its delicious Italian cuisine, including pizza.
6. **WORLD OF PIZZA Leipzig-City:** This restaurant offers decent pizza and has a friendly atmosphere. It has received positive reviews [1](#).
7. **60 Seconds to napoli Leipzig:** Known for its creative Naples-style pizzas, this restaurant is highly rated for its pizza. It also offers online ordering.
8. **Sbaroo:** Sbaroo is a cozy Italian restaurant that serves delicious pizza. It has received positive reviews and offers online ordering.
9. **L'Osteria:** Although not mentioned in the search snippets, L'Osteria is a popular pizza place in Leipzig according to another source.
10. **Rizzi:** Another highly recommended pizza place in Leipzig is Rizzi. It is known for its tasty pizzas.

Please note that these are just a few of the best pizza places in Leipzig, and there may be other great options available as well. It's always a good idea to check reviews and ratings before making a decision. Enjoy your pizza!

tripadvisor

What is Generative IR

query-to-document in the Wild

Who would win in a fight, Godzilla or King Kong?

✓ Searching for: Godzilla vs King Kong fight
✓ Generating answers for you...

The epic battle between Godzilla and King Kong has been a topic of debate for decades. Both monsters have their unique strengths and weaknesses. Godzilla is a giant radioactive lizard with immense strength, atomic breath, and regenerative abilities. King Kong, on the other hand, is a giant ape with incredible agility, intelligence, and strength.

In the 2021 movie "Godzilla vs. Kong," the two titans engage in a fierce battle for dominance. The fight is intense and lasts for several rounds, with both monsters exchanging blows and using their unique abilities to gain the upper hand. I won't spoil the ending for you, but I can tell you that it's a thrilling and action-packed movie that's worth watching. [1](#) [2](#)

As for who would win in a fight, it's hard to say. Both monsters are incredibly powerful and have their unique strengths and weaknesses. It would depend on the circumstances of the fight and the environment in which it takes place. However, one thing is for sure: it would be an epic battle that would go down in history!

Learn more [1 youtube.com](#) [2 youtube.com](#) [3 screenrant.com](#) [4 screenrant.com](#) [5 imbd.com](#)

2 of 4

Who would win in a fight, Godzilla or King Kong?

View other drafts

PaLM2

In a direct confrontation, Godzilla would likely emerge victorious over King Kong. Godzilla possesses several advantages over King Kong, including:

- Superior Size and Strength: Godzilla is significantly larger and stronger than King Kong. This size difference gives Godzilla a significant advantage in terms of raw power and the ability to inflict damage.
- Radiation Beam: Godzilla's signature atomic breath is a powerful weapon that can vaporize opponents or even cause severe burns. This ability gives Godzilla a ranged attack that King Kong lacks.
- Armored Scales: Godzilla's scales are incredibly thick and durable, making them resistant to King Kong's physical attacks. This armor can protect Godzilla from significant damage.
- Radiation Absorption: Godzilla can absorb radiation to enhance his power and resilience. This ability allows Godzilla to recover from injuries and regain strength more quickly than King Kong.

However, King Kong is not without his own strengths. He is agile and resourceful, able to use his environment to his advantage and improvise weapons. He is also intelligent and cunning, capable of planning and strategizing.

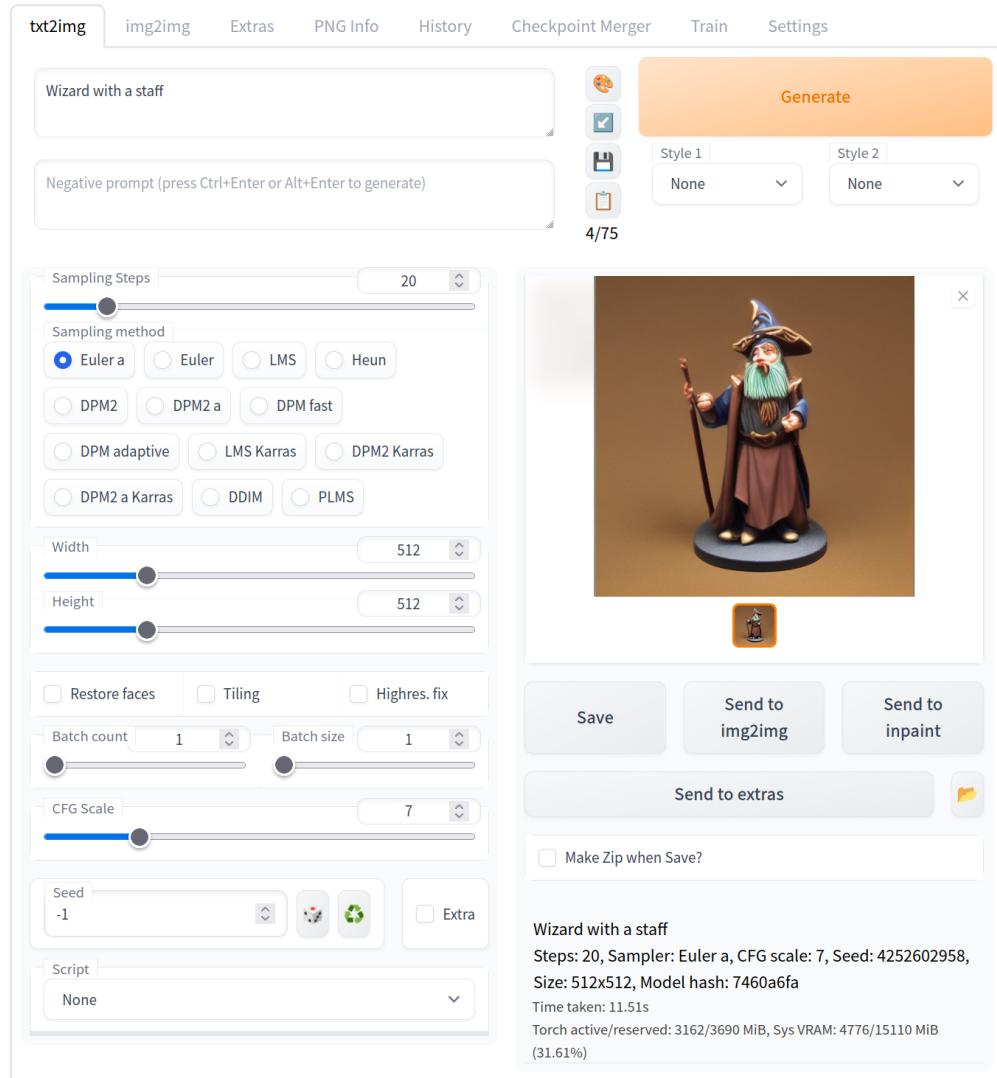
In a hypothetical fight, the outcome would depend on a variety of factors, such as the location of the battle, the availability of resources, and the tactics employed by each creature. However, based on their inherent abilities and attributes, Godzilla would likely hold the upper hand in a direct confrontation.

Check completed • Understand the results

theringer [1](#) screenrant [2](#) reddit [3](#)

What is Generative IR

query-to-document isn't just for text [Deckers et al. 2023]



Methods for Generative IR

Overview

query-to-identifier

- DSI [[Tay et al. 2022](#)]
- SEAL [[Bevilacqua et al. 2022](#)]

query-to-document

- Fusion-in-Decoder (FiD) [[Izacard and Grave 2021](#)]

Methods for Generative IR

DSI [Tay et al. 2022]

How to index?

- Predict $\text{doc_tokens} \rightarrow \text{doc_id}$. Mimics the task at inference time but cannot perform pre-training using documents using masked language modelling.
- Predict $\text{doc_id} \rightarrow \text{doc_tokens}$. Equivalent to training decoder model conditioned on document identifier.
- Bidirectional prediction. Prepend a special token to allow the model to know the prediction task.
- Masked language modelling. Document identifier and document text are concatenated and randomly masked for the model to predict which token fits.

Methods for Generative IR

DSI [Tay et al. 2022]

How to index?

- Predict $\text{doc_tokens} \rightarrow \text{doc_id}$. Mimics the task at inference time but cannot perform pre-training using documents using masked language modelling.
- Predict $\text{doc_id} \rightarrow \text{doc_tokens}$. Equivalent to training decoder model conditioned on document identifier.
- Bidirectional prediction. Prepend a special token to allow the model to know the prediction task.
- Masked language modelling. Document identifier and document text are concatenated and randomly masked for the model to predict which token fits.

What to index?

- ‘Direct indexing’. Perform no pre-processing, truncate documents to fit into input size of transformer model.

Methods for Generative IR

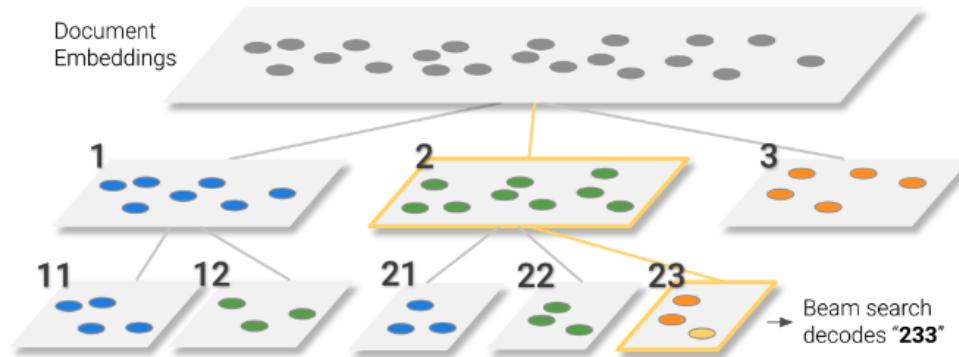
DSI [Tay et al. 2022]

How to represent document identifiers?

- Identifier should capture some semantic information about document.
- Identifier should be structured for efficient decoding.

Hierarchically cluster document embeddings to induce a trie:

- Offline, cluster documents into k clusters, assign identifier based on cluster.
- Leaves are identifiers, balanced so all leaves are same distance from root.
- Trie is traversed at inference time to decode valid document identifiers.



Methods for Generative IR

DSI [Tay et al. 2022]

Related research:

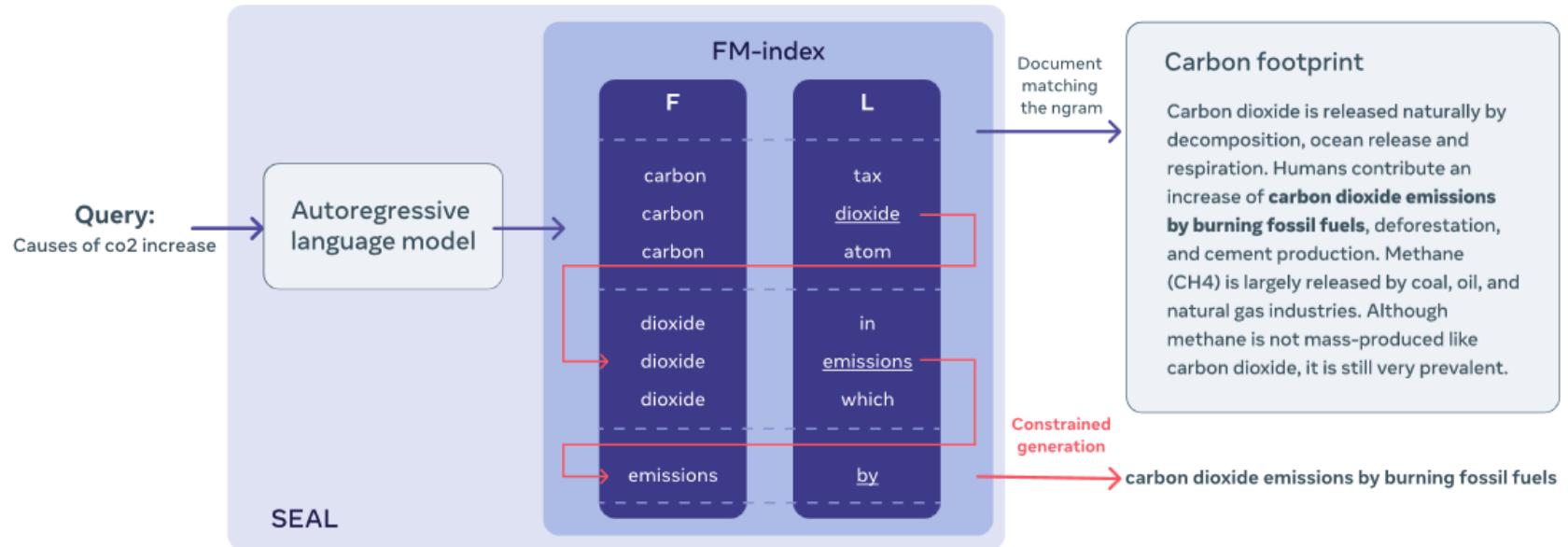
- DSI++ [[Vaibhav et al. 2022](#)]: Provide a way to incrementally index new documents into the model without it forgetting how to rank previous queries.
- DSI-QG [[Zhuang et al. 2023](#)]: Provide explicit query signals at training, rather than the implicit signal as in DSI.

Methods for Generative IR

SEAL [Bevilacqua et al. 2022]

Method:

- Create an ‘FM’ n-gram index offline.
- Use decoder to generate multiple n-grams, constrained by FM index.
- Use n-grams generated by decoder to match documents.



Methods for Generative IR

SEAL [Bevilacqua et al. 2022]

Related research:

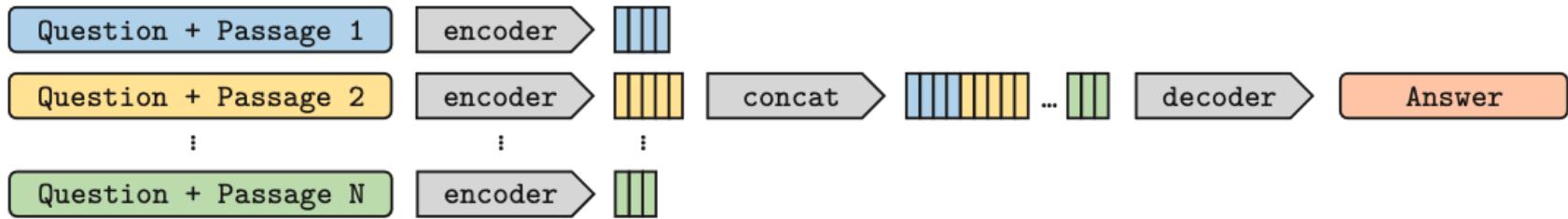
- MINDER [[Li et al. 2023](#)]: Generate n-gram identifiers for multiple ‘views’ of documents, e.g., title, substring, or query.
- LTRGR [[Li et al. 2023](#)]: After generating n-gram identifiers, use the same decoder model to score documents, trained in a LTR setting.

Methods for Generative IR

FiD [Izacard and Grave 2021]

Given a query, answer the query directly with a generated document.

- Use retrieval model to select k documents as context.
- Use (encoder-)decoder model to generate document using context.

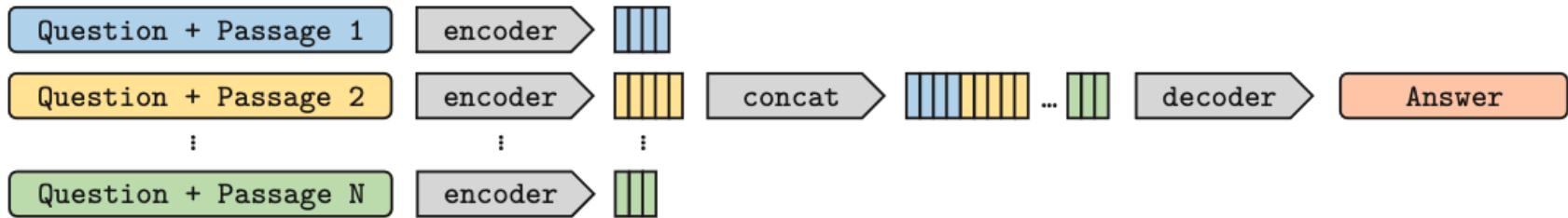


Methods for Generative IR

FiD [Izacard and Grave 2021]

Given a query, answer the query directly with a generated document.

- Use retrieval model to select k documents as context.
- Use (encoder-)decoder model to generate document using context.



Model only provides an answer, does not ground the response with references.

Methods for Generative IR

FID [[Izacard and Grave 2021](#)]

Related research:

- FID-Light [[Hofstätter et al. 2023](#)]: Applies compression to make FID faster and provides a way to inject references to documents in the output.

Evaluation of Generative IR

Overview

Two usage patterns can be discerned for generative models in IR:

- Query-to-Identifier
 - Uses a generative model to produce a ranking of materialized documents
 - Evaluation can commence under the traditional Cranfield paradigm
- Query-to-Document
 - Uses a generative model to produce a new document in-situ
 - Since the new document has no existing identifier, we cannot use Cranfield-style evaluation → **What now?**

A new evaluation setup is needed! This encompasses a...

- ... Task definition → what is the synthetical search task?
- ... User model → how does a user read generated text?
- ... Evaluation criteria → how can generated text quality be measured?

Evaluation of Generative IR

Synthesetical Search Task

- ❑ Before: informational/navigational search tasks
 - User is presented with a list of possibly relevant content
 - They have to read and evaluate each whether answers their question
- ❑ Now: synthesitical search task
 - User is presented with synthesized text response
 - Generative model decides which parts from sources are relevant
- ❑ What is new?
 - Condensation of relevant information from multiple sources
 - Text SERP instead of list SERP
- ❑ What is it used for?
 - opinionated information needs (“Invest in renewable energy?”)
 - decision-making ones (“Should I get life insurance?”).

Evaluation of Generative IR

Example

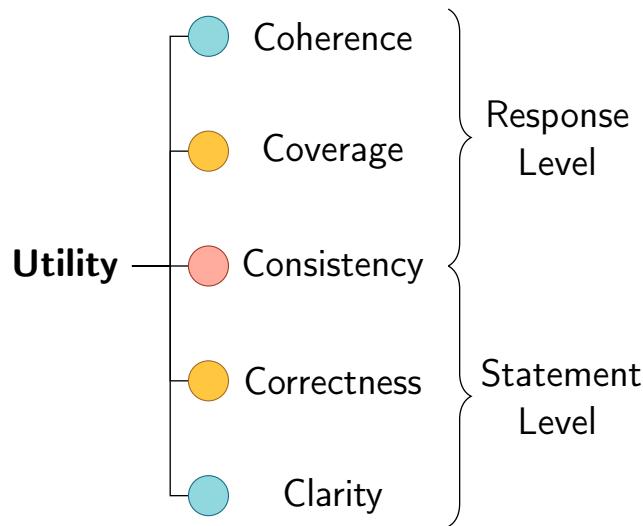
- **Query:** investing renewable energy
- **Good response:** “The global trend towards clean energy makes it a lucrative market. Solar and wind are particularly promising, offering a diverse range of investment opportunities. Studies affirm the viability of such investments, emphasizing both environmental sustainability and financial gains. Yet, the volatility of the market might not be a good fit for low risk investing.”
- **Bad response:** “Renewable energy, is like, good for the environment. Some reports say it's going up in the markets, but who really checks those, right? Might consider it for investing, the information is all there, check it out.”

Clear quality difference between the responses, but how to quantify it?

Evaluation of Generative IR

Evaluation Criteria & Utility

- Users have different expectations towards a generated text than a ranked list
 - **Retrieval aspect**: generated text should address the information need
 - **Grounding aspect**: generated text should be verifiable from sources
 - **Presentation aspect**: generated text should be coherent and readable
- We can judge these aspects of utility either for the whole response ('response-level') or at individual parts of the response ('statement level')



Evaluation of Generative IR

Utility Dimensions

- ❑ **Coherence:** the manner in which the response is structured and presented
 - structured well-formed narrative (Logical Coherence)
 - uniform style of speech (Stylistic Coherence)
- ❑ **Coverage:** cumulative extent to which presented information is relevant
 - the response covers a breadth of diverse information (Broad Coverage)
 - the response provides in-depth detailed information (Deep Coverage)
- ❑ **Consistency:** the agreement between statements and their sources
 - consistency between statements of the response (Internal Consistency)
 - consistency between a statement and its source (External Consistency)
- ❑ **Correctness:** degree to which the information provided is factually correct
 - the verifiability of a statement (Factual Correctness)
 - the relevance of a statement to the user (Topical Correctness)
- ❑ **Clarity:** the information is presented in an accessible manner
 - written in an easily readable way (Language Clarity)
 - clear communication of salient information (Content Clarity)

Evaluation of Generative IR

Example

- **Query:** investing renewable energy
- **Good response:** “The global trend towards clean energy makes it a lucrative market. Solar and wind are particularly promising, offering a diverse range of investment opportunities. Studies affirm the viability of such investments, emphasizing both environmental sustainability and financial gains. Yet, the volatility of the market might not be a good fit for low risk investing.”
 - easily readable and to the point (high clarity and coherence)
 - backs up claims with specific examples, and covers multiple view on the topic (high coverage and consistency)
 - correctness through external sources
- **Bad response:** “Renewable energy, is like, good for the environment. Some reports say it's going up in the markets, but who really checks those, right? Might consider it for investing, the information is all there, check it out.”
 - embellished and colloquial language (low clarity and coherence)
 - only refers to external sources without summarizing (lacking consistency)
 - no specific answer to the query (low coverage and correctness)

Evaluation of Generative IR

User Model for Generative IR

Traditional User Model

- User browses documents in a ranked list sequentially
- User stops browsing once their information need is fulfilled
- User is less likely to browse documents at lower ranks

Generative User Model

- User reads statements in the generated text sequentially
- User stops reading once their information need is fulfilled
- User reads with decaying attention over the span of the text

→ The behaviour of the user is equivalent for browsing and reading alike!

- We can use discount-based metrics that operationalize the traditional user model for the generative user model as well (e.g., nDCG)
- But: we need to segment the generated text into statements, and assign a utility score to each statement → statement-level utility!

Evaluation of Generative IR

Annotating Utility

Problem: How can we annotate utility if the text is newly generated?

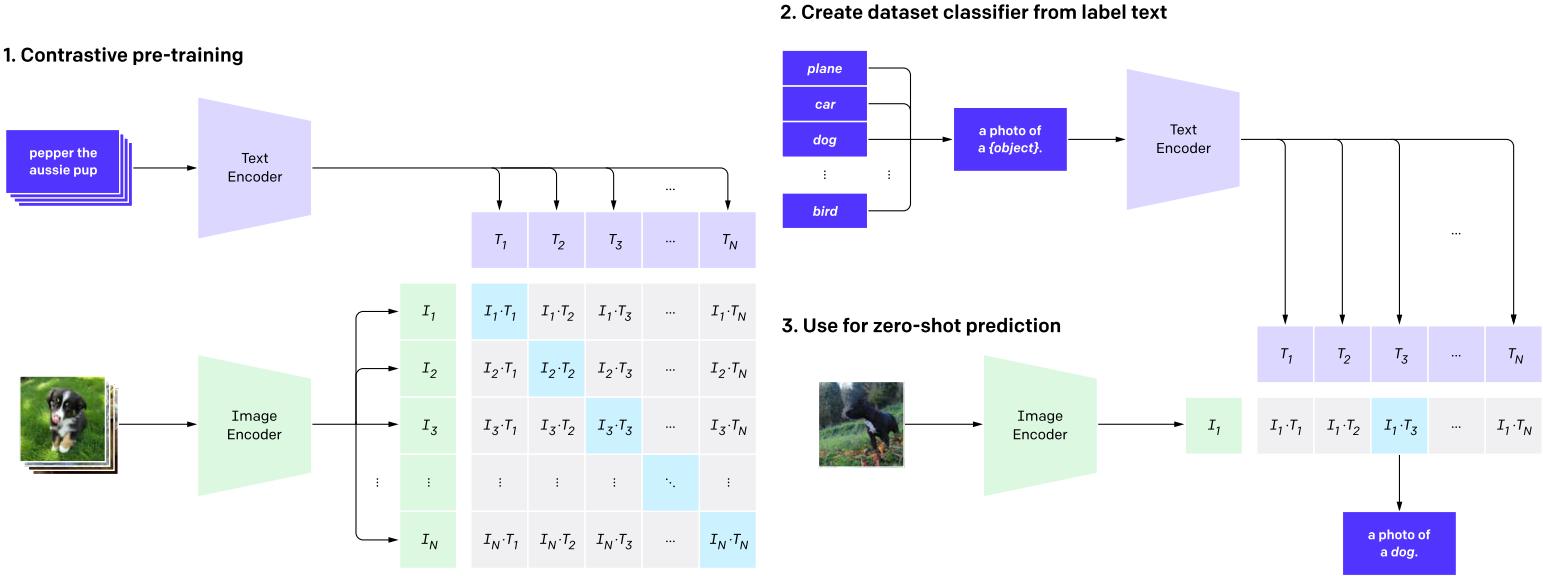
- Reference-free assessment
 - direct assessment of every new generated text
 - **Pro:** reliable given expert annotators
 - **Con:** very expensive to collect, not repeatable
- Reference-based assessment
 - transfer of pre-existing judgments from other texts
 - **Pro:** reliability hinges on transfer
 - **Con:** efficient and repeatable experiments

Multi-Modal IR

Overview

Focus of this course has been on text, search can be applied to any media

- query-to-document is represented as ‘infinite index’ [\[Deckers et al. 2023\]](#)
- Decoder models could be perceived as ‘infinite indexes’ over text.
- Future of IR may be retrieval components on the top of ‘infinite indexes’.
- For multi-modal search, many methods use CLIP [\[Radford et al. 2021\]](#)



Open Challenges

Overview

query-to-identifier:

- How to scale to large collections? All documents must be contained within model parameters.
- How to make these models more efficient at inference time? Decoding is expensive, requiring forward pass for each document.

query-to-document:

- How to properly evaluate the output from models? How to scale the evaluation or do meaningful offline evaluation?
- When should the output be grounded? How to evaluate grounding?
- How to reduce or prevent hallucination?