

Optimizing and Evaluating Enterprise RAG

A Content Design Perspective

Sarah Packowski AI ContentOps Architect, IBM

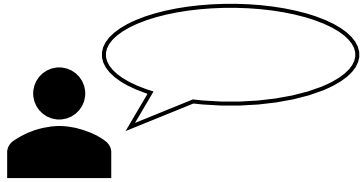
Inge Halilovic Content Strategist, IBM

Jenifer Schlotfeldt Content Experience Architect, IBM

Trish Smith Content Designer, IBM

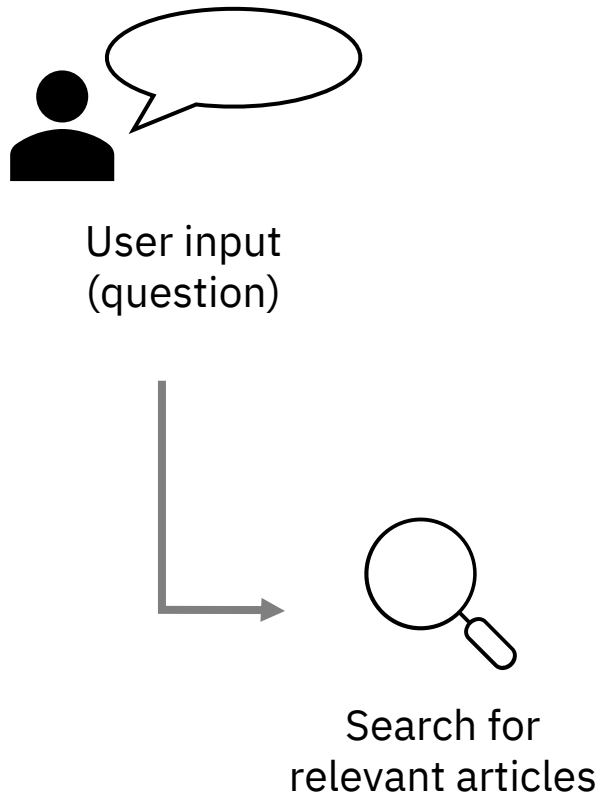
1. Retrieval-augmented generation (RAG)
2. Content design
3. Modular, model-agnostic RAG solution
4. Optimizing knowledge base content for RAG
5. Evaluating RAG results

Retrieval-augmented generation (RAG)

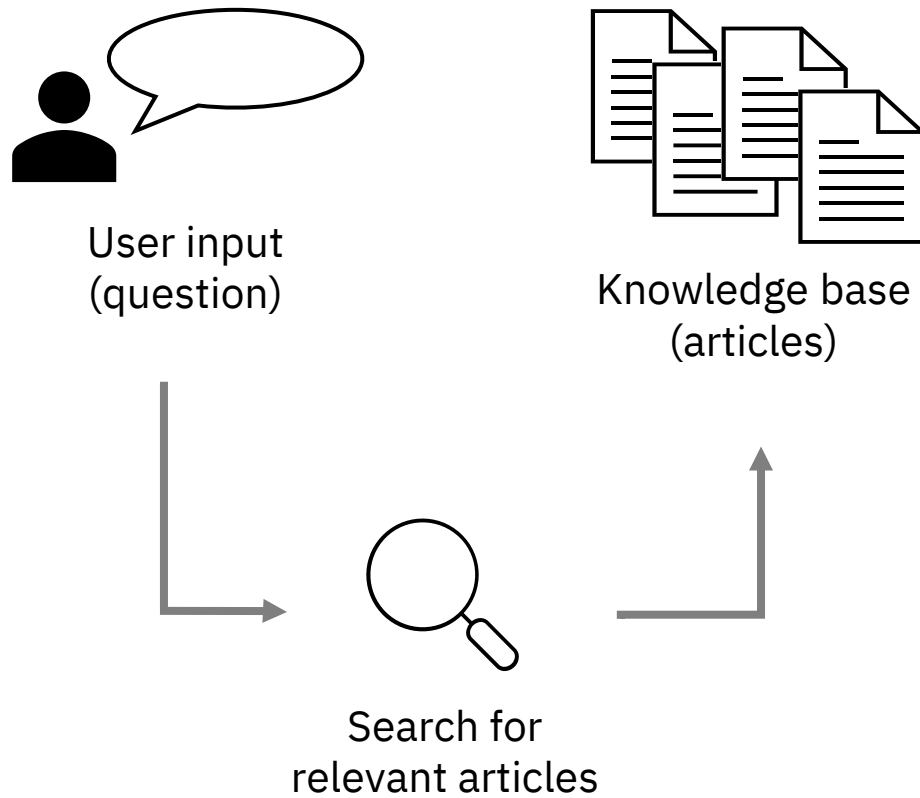


User input
(question)

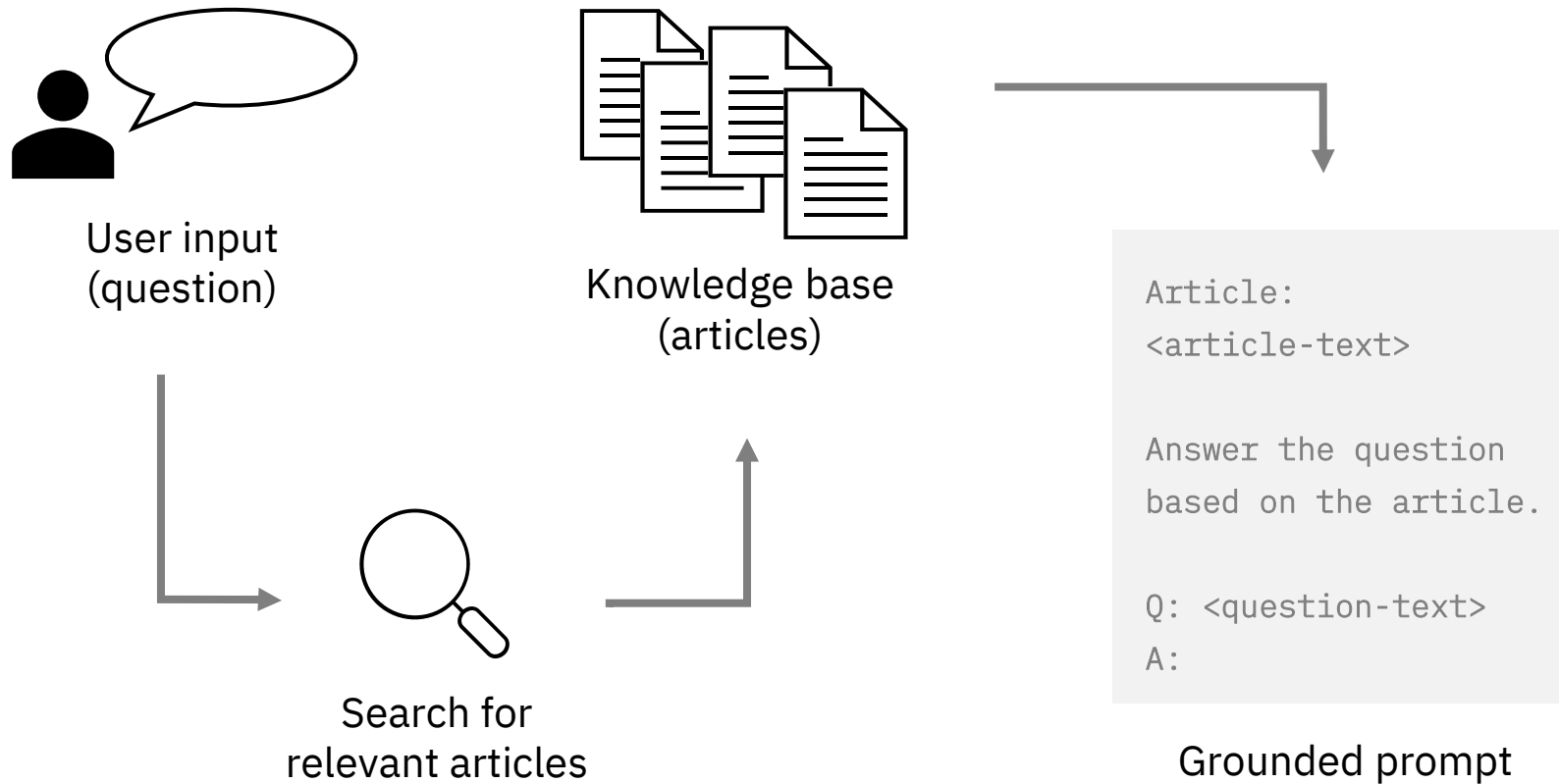
Retrieval-augmented generation (RAG)



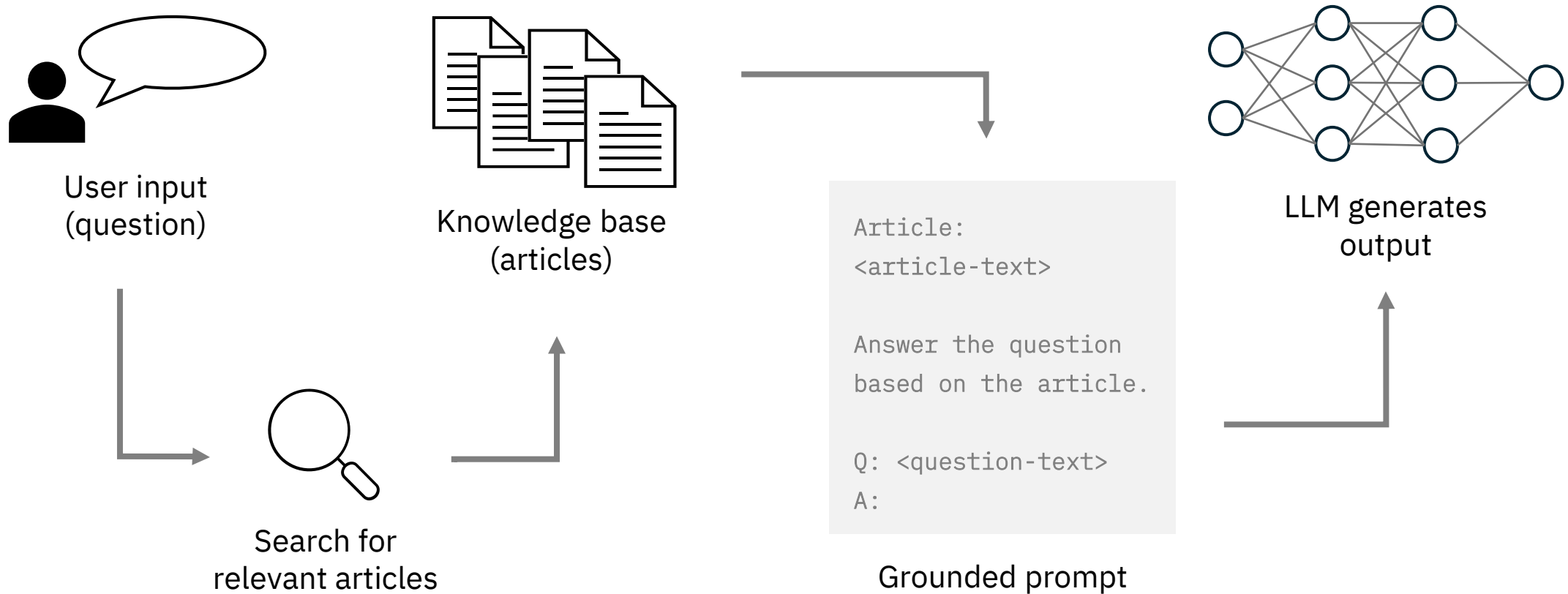
Retrieval-augmented generation (RAG)



Retrieval-augmented generation (RAG)



Retrieval-augmented generation (RAG)



Content design

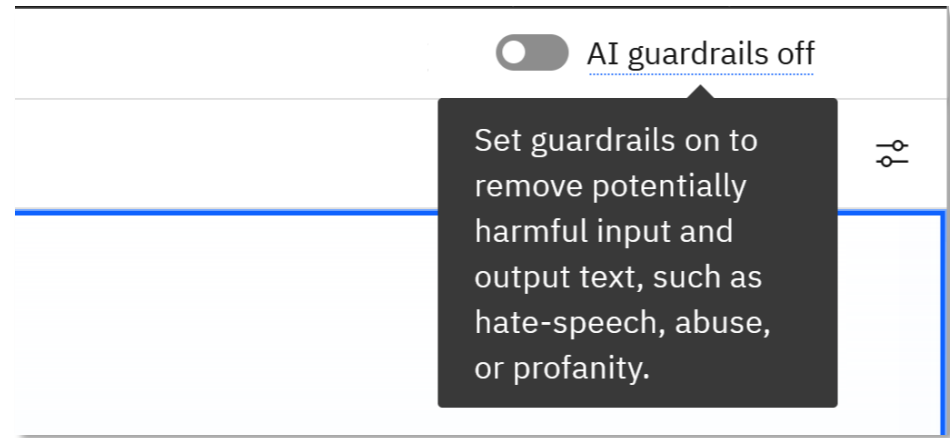
Create:

- Interface text
- Error messages
- Documentation
- Samples
- Tutorials
- Videos

Content design

Create:

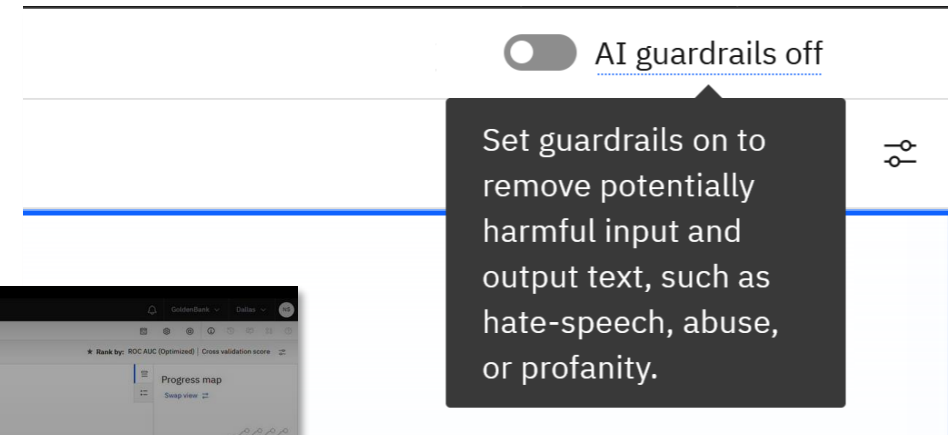
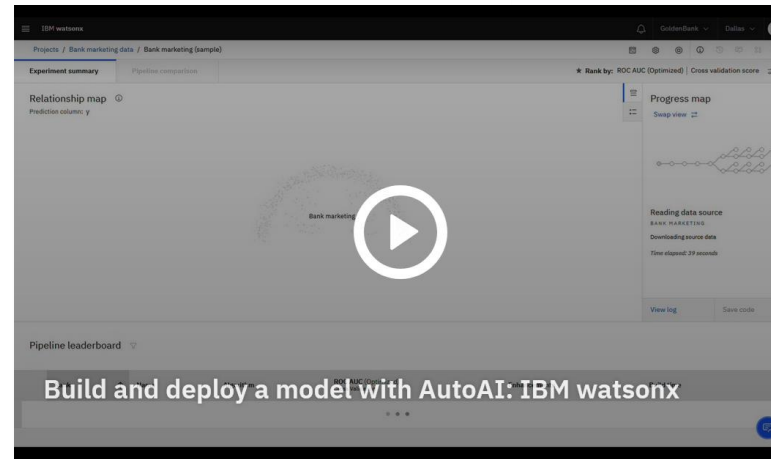
- Interface text
- Error messages
- Documentation
- Samples
- Tutorials
- Videos



Content design

Create:

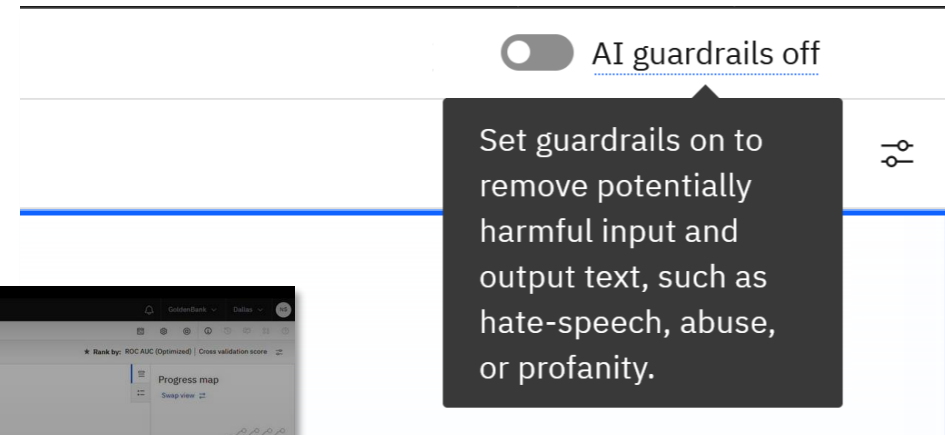
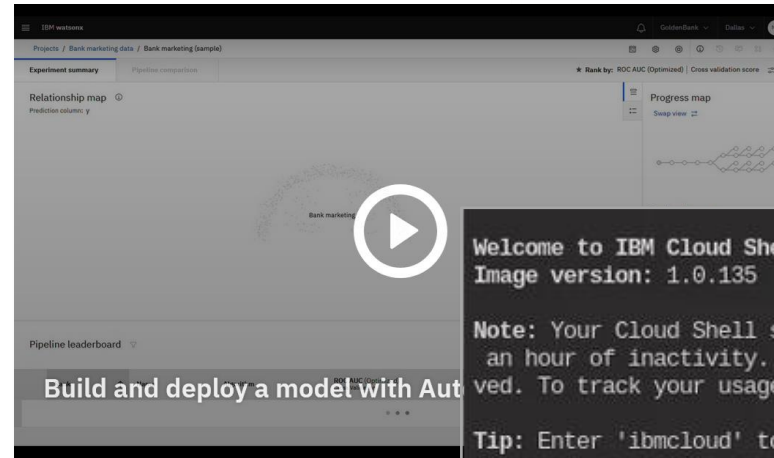
- Interface text
- Error messages
- Documentation
- Samples
- Tutorials
- Videos



Content design

Create:

- Interface text
- Error messages
- Documentation
- Samples
- Tutorials
- Videos

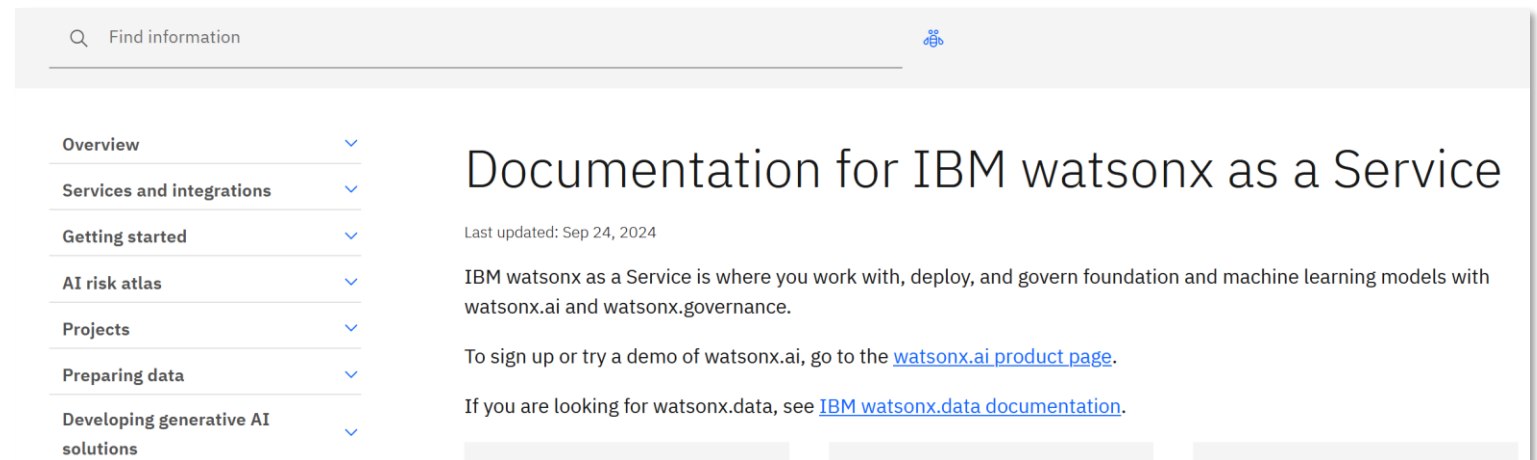
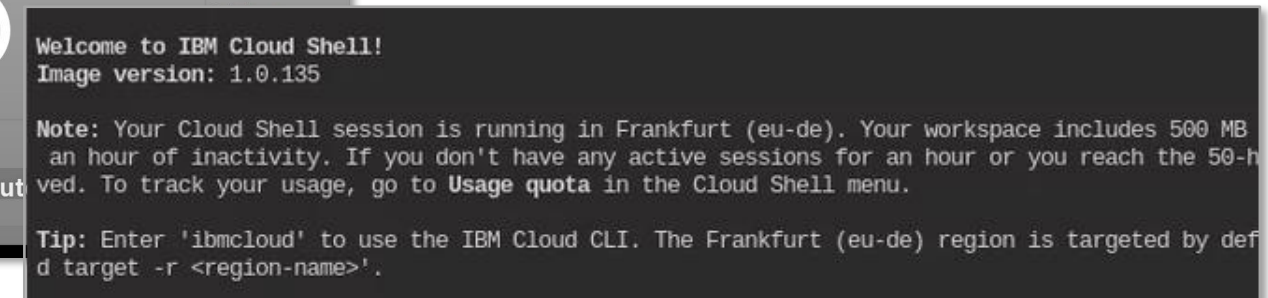
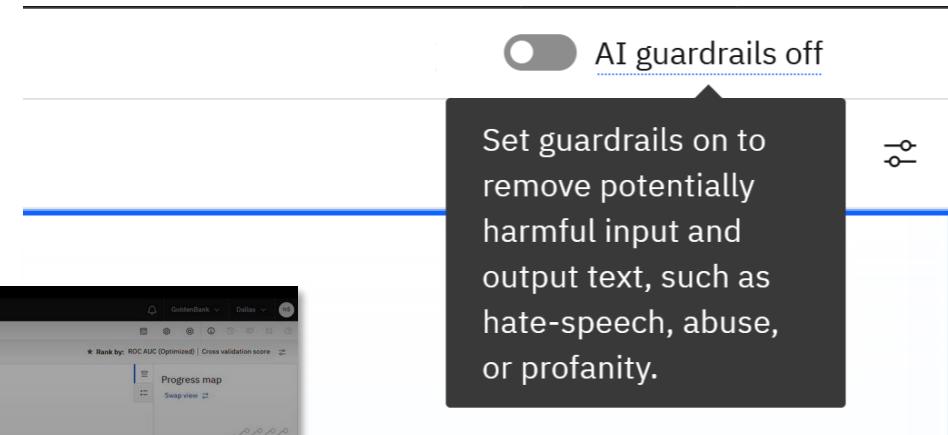
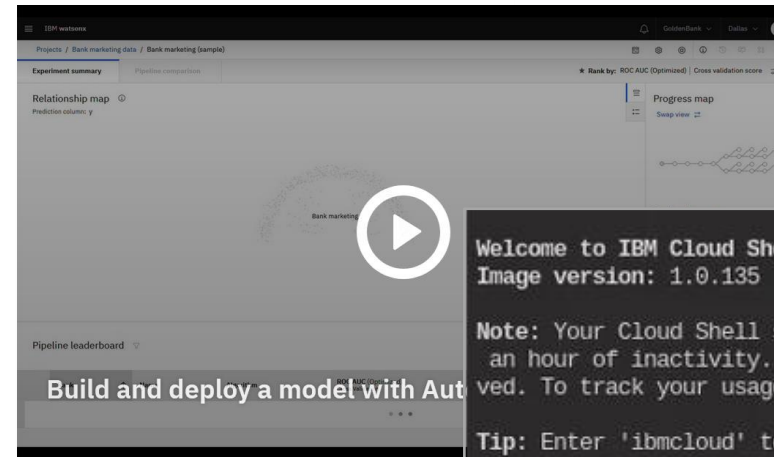


```
Welcome to IBM Cloud Shell!  
Image version: 1.0.135  
  
Note: Your Cloud Shell session is running in Frankfurt (eu-de). Your workspace includes 500 MB  
an hour of inactivity. If you don't have any active sessions for an hour or you reach the 50-h  
ved. To track your usage, go to Usage quota in the Cloud Shell menu.  
  
Tip: Enter 'ibmcloud' to use the IBM Cloud CLI. The Frankfurt (eu-de) region is targeted by def  
d target -r <region-name>'.
```

Content design

Create:

- Interface text
- Error messages
- Documentation
- Samples
- Tutorials
- Videos



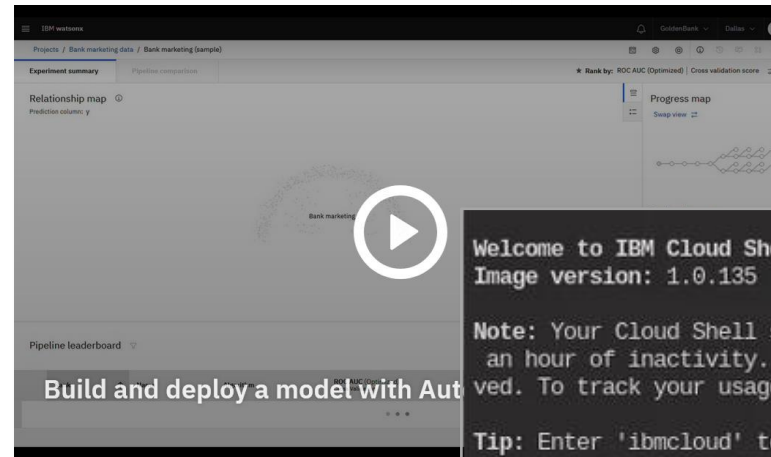
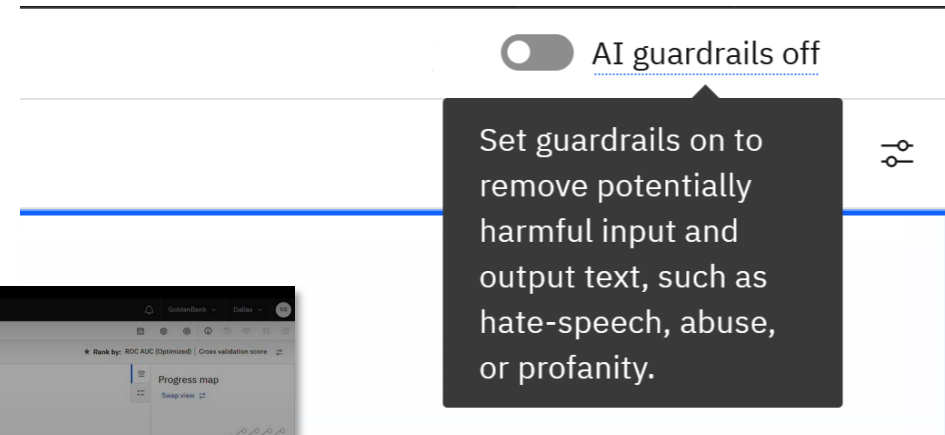
Content design

Create:

- Interface text
- Error messages
- Documentation
- Samples
- Tutorials
- Videos

Processes:

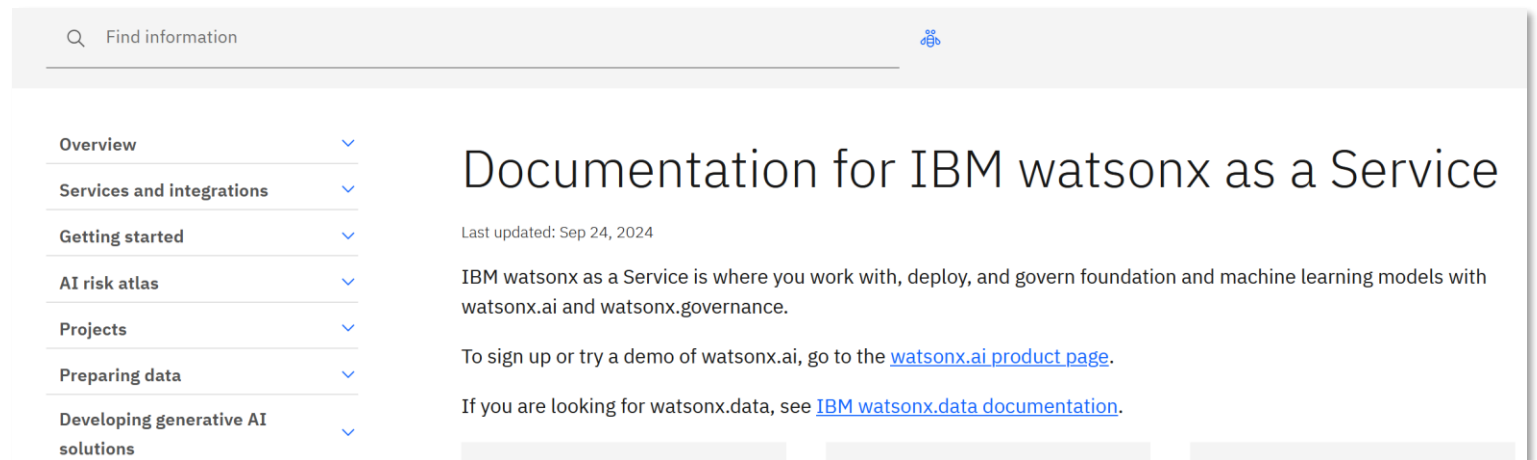
- Content strategy
- Writing style guidelines
- Terminology database
- Automated editing tools
- Accessibility requirements
- Globalization



Welcome to IBM Cloud Shell!
Image version: 1.0.135

Note: Your Cloud Shell session is running in Frankfurt (eu-de). Your workspace includes 500 MB of storage and 2 CPU cores. Your session will timeout after 1 hour of inactivity. If you don't have any active sessions for an hour or you reach the 50-hour limit, your session will be terminated. To track your usage, go to **Usage quota** in the Cloud Shell menu.

Tip: Enter 'ibmcloud' to use the IBM Cloud CLI. The Frankfurt (eu-de) region is targeted by default. To change the region, use the command 'ibmcloud config --set target -r <region-name>'.



Content design

Language is our
whole thing!



Entertainment

Lindsay Lohan bitten by snake
on holiday in Thailand

<https://www.standard.co.uk/showbiz/celebrity-news/lindsay-lohan-bitten-by-snake-on-holiday-in-thailand-a3727976.html>

Content design

Language is our
whole thing!

*Image shared in a team Slack channel,
earlier this week, with comments like:
“Look at that misplaced modifier!” and
“Why would a snake go on holiday?”*



Entertainment

Lindsay Lohan bitten by snake on holiday in Thailand

<https://www.standard.co.uk/showbiz/celebrity-news/lindsay-lohan-bitten-by-snake-on-holiday-in-thailand-a3727976.html>


Example RAG user experience

[Docs](#) / [Analyzing data and working with ...](#) / [Foundation models](#)

Q

What is decoding?

×



[Foundation model parameters: decoding and stopping criteria](#)

Last updated: Jul 06, 2023

You can specify parameters to control how the model generates output in response to your prompt. **Decoding** is the process a model uses to choose the tokens in the generated output. Greedy **decoding** selects the token

[Avoiding undesirable output](#)

Last updated: Jul 05, 2023

To reduce the risk of off-topic hallucination with use cases that don't




Generative AI search and answer

These answers are generated by a large language model in watsonx.ai based on content from the product documentation. [Learn more](#)

Q: What is decoding?

A: Decoding is the process a model uses to choose tokens in generated output.

Source links:
[Foundation model parameters: decoding and stopping criteria](#)

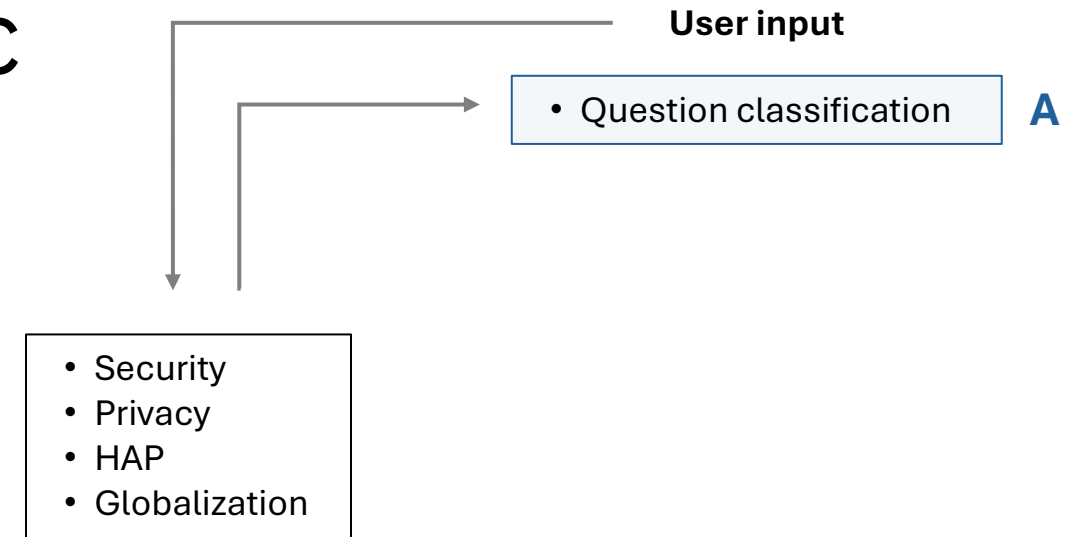


Optimizing and Evaluating Enterprise RAG: A Content Design Perspective

https://github.com/spackows/ICAAI-2024_RAG-CD

Modular, model-agnostic RAG implementation

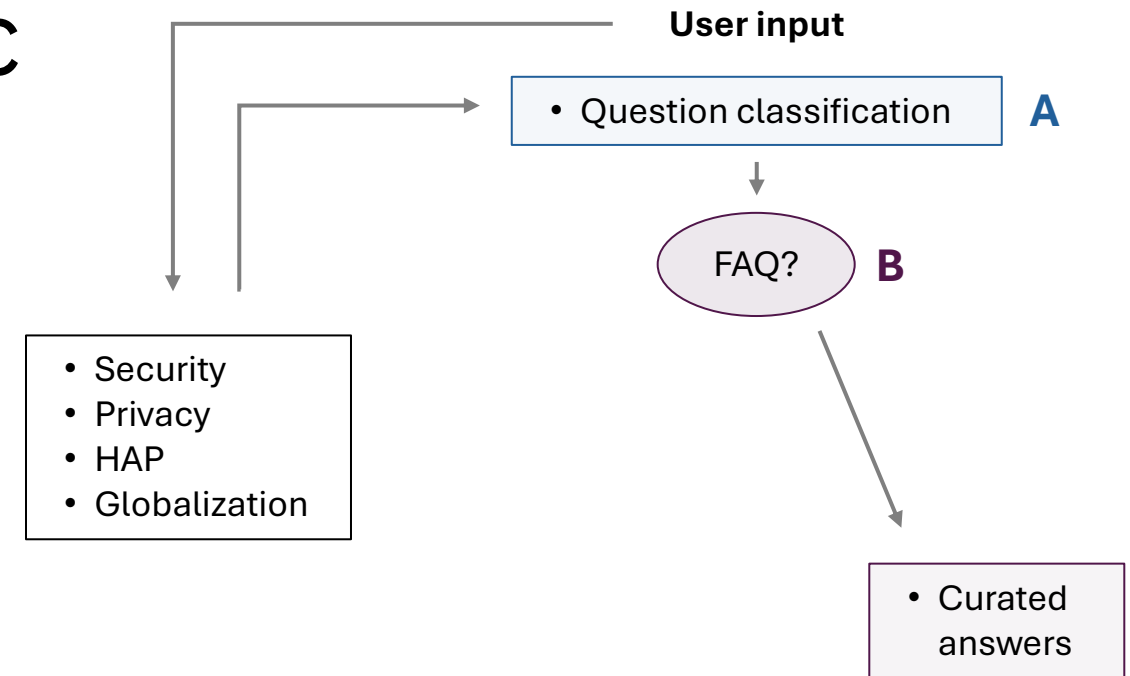
A: Only valid, unharmed, questions, in English move on



Modular, model-agnostic RAG implementation

A: Only valid, unharmed, questions, in English move on

B: Return curated answers for frequently asked questions

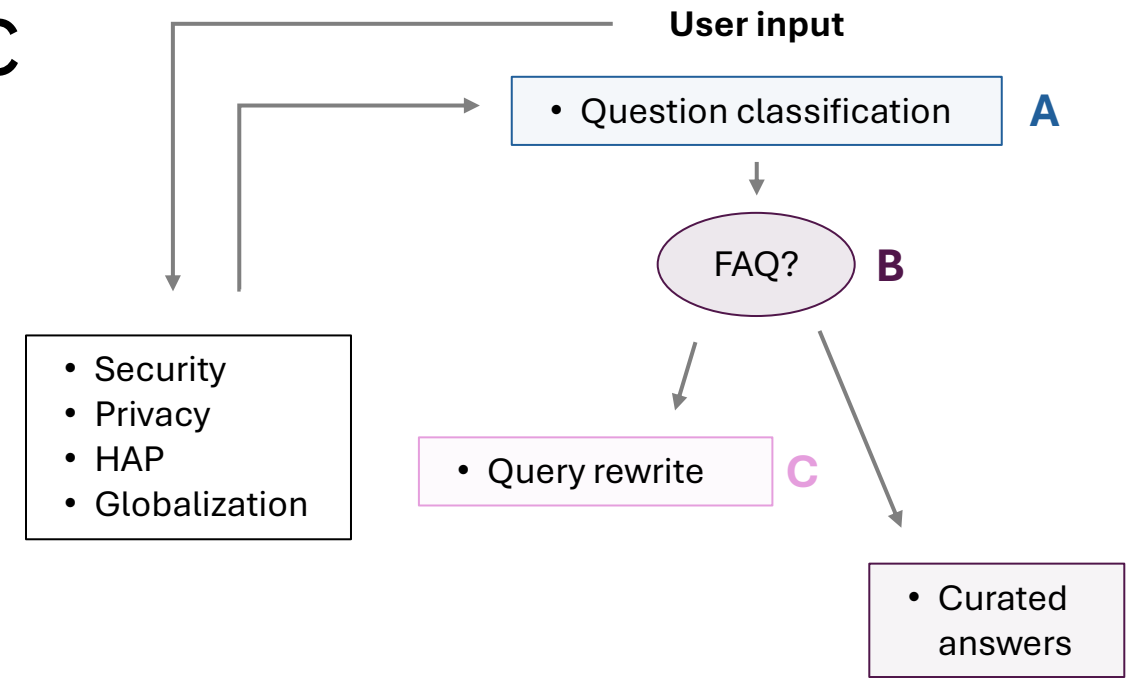


Modular, model-agnostic RAG implementation

A: Only valid, unharmful, questions, in English move on

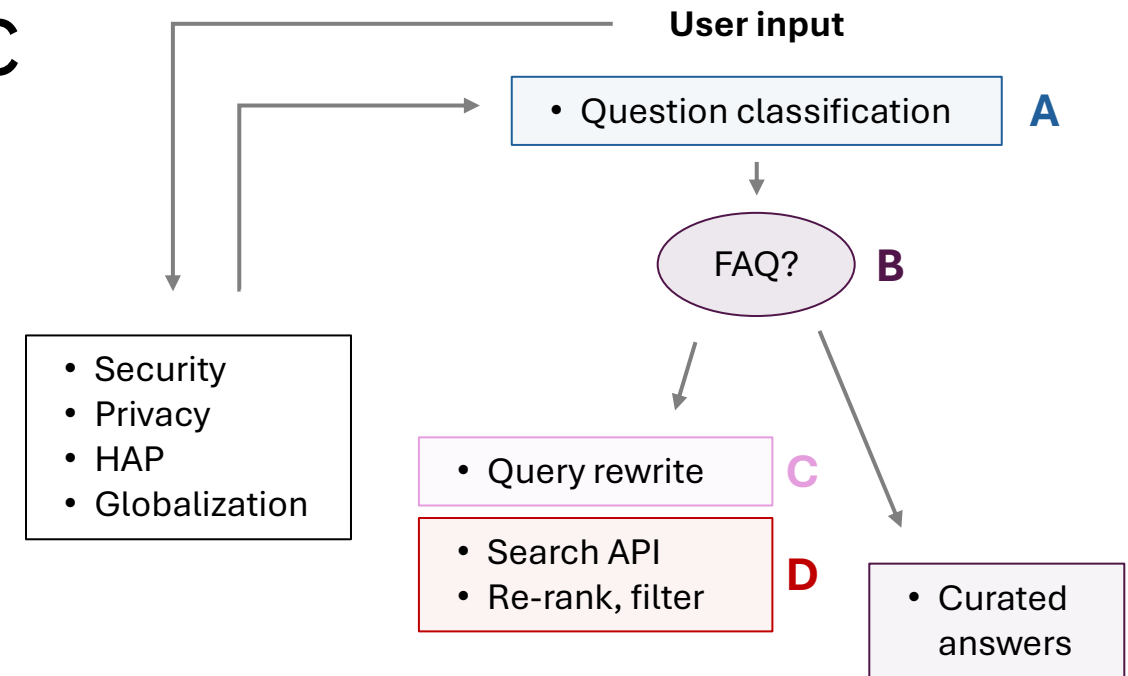
B: Return curated answers for frequently asked questions

C: Clarify ambiguous, typo-filled questions



Modular, model-agnostic RAG implementation

- A**: Only valid, unharmed, questions, in English move on
- B**: Return curated answers for frequently asked questions
- C**: Clarify ambiguous, typo-filled questions
- D**: Use whatever search works*

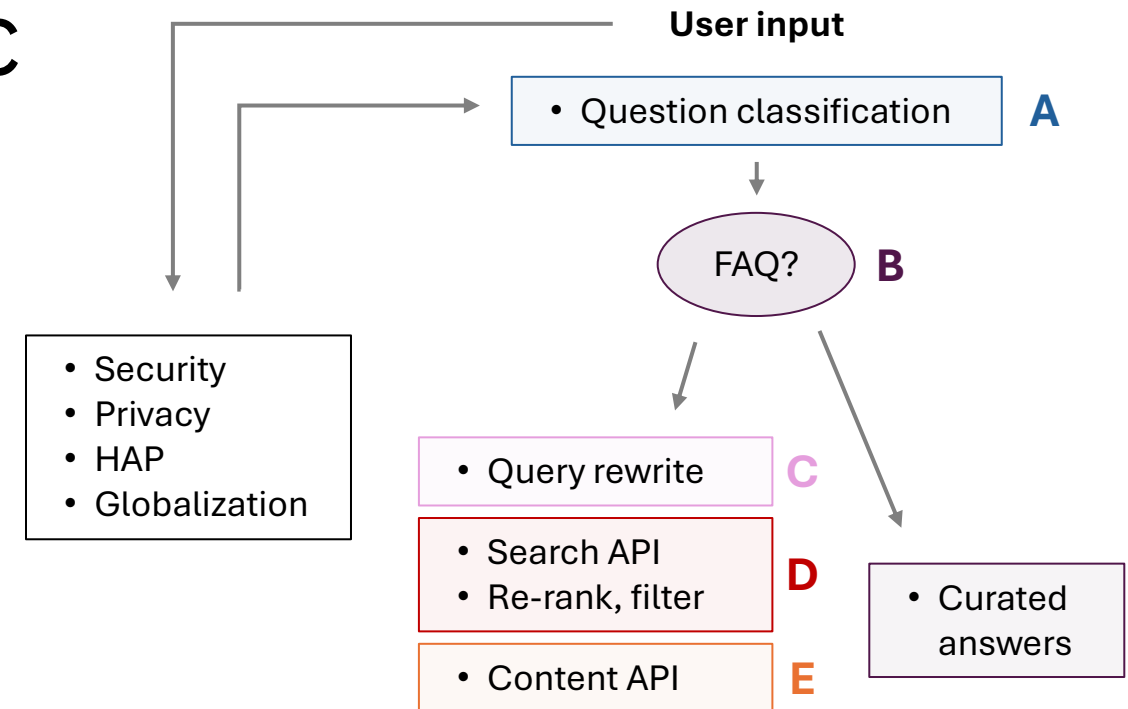


*Articles are short, self-contained, and optimized for search, so traditional search methods work as well as or better than vector databases.

Modular, model-agnostic RAG implementation

- A**: Only valid, unharmed, questions, in English move on
- B**: Return curated answers for frequently asked questions
- C**: Clarify ambiguous, typo-filled questions
- D**: Use whatever search works*
- E**: Extract *complete text* of relevant articles

*Articles are short, self-contained, and optimized for search, so traditional search methods work as well as or better than vector databases.



Modular, model-agnostic RAG implementation

A: Only valid, unharmful, questions, in English move on

B: Return curated answers for frequently asked questions

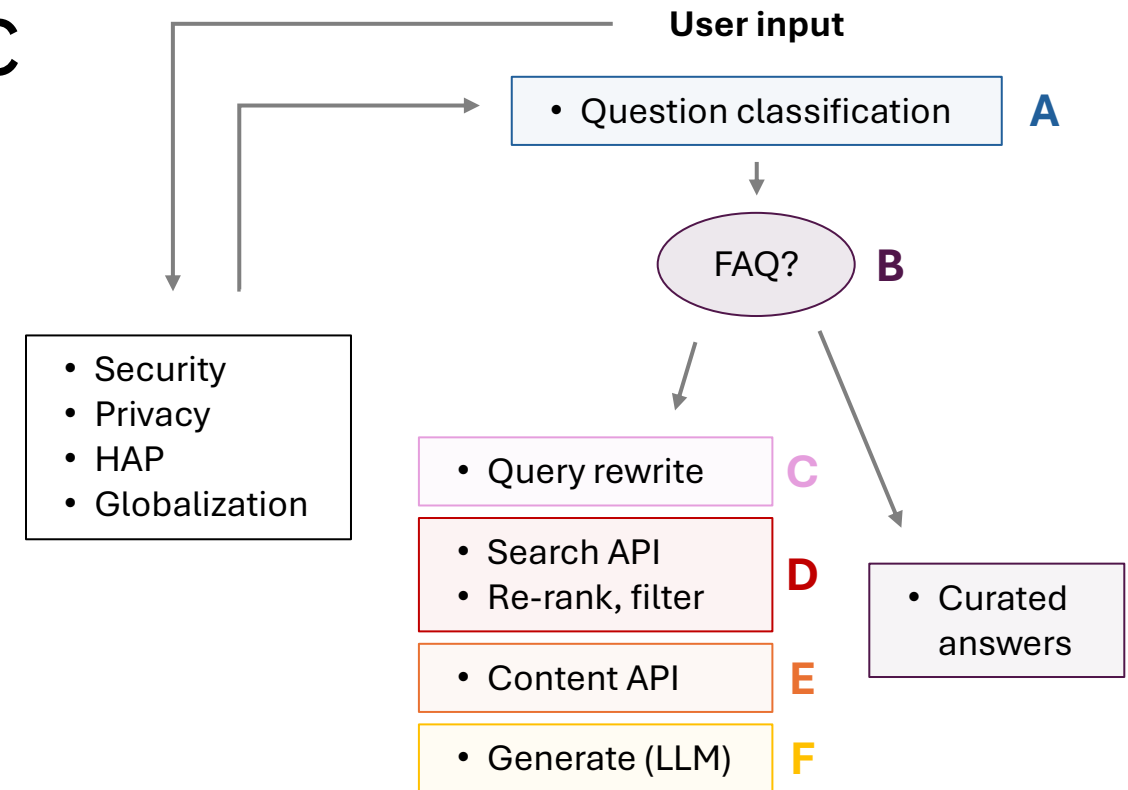
C: Clarify ambiguous, typo-filled questions

D: Use whatever search works*

E: Extract *complete text* of relevant articles

F: LLM has a simple job: answer based on articles

*Articles are short, self-contained, and optimized for search, so traditional search methods work as well as or better than vector databases.



Modular, model-agnostic RAG implementation

A: Only valid, unharmed, questions, in English move on

B: Return curated answers for frequently asked questions

C: Clarify ambiguous, typo-filled questions

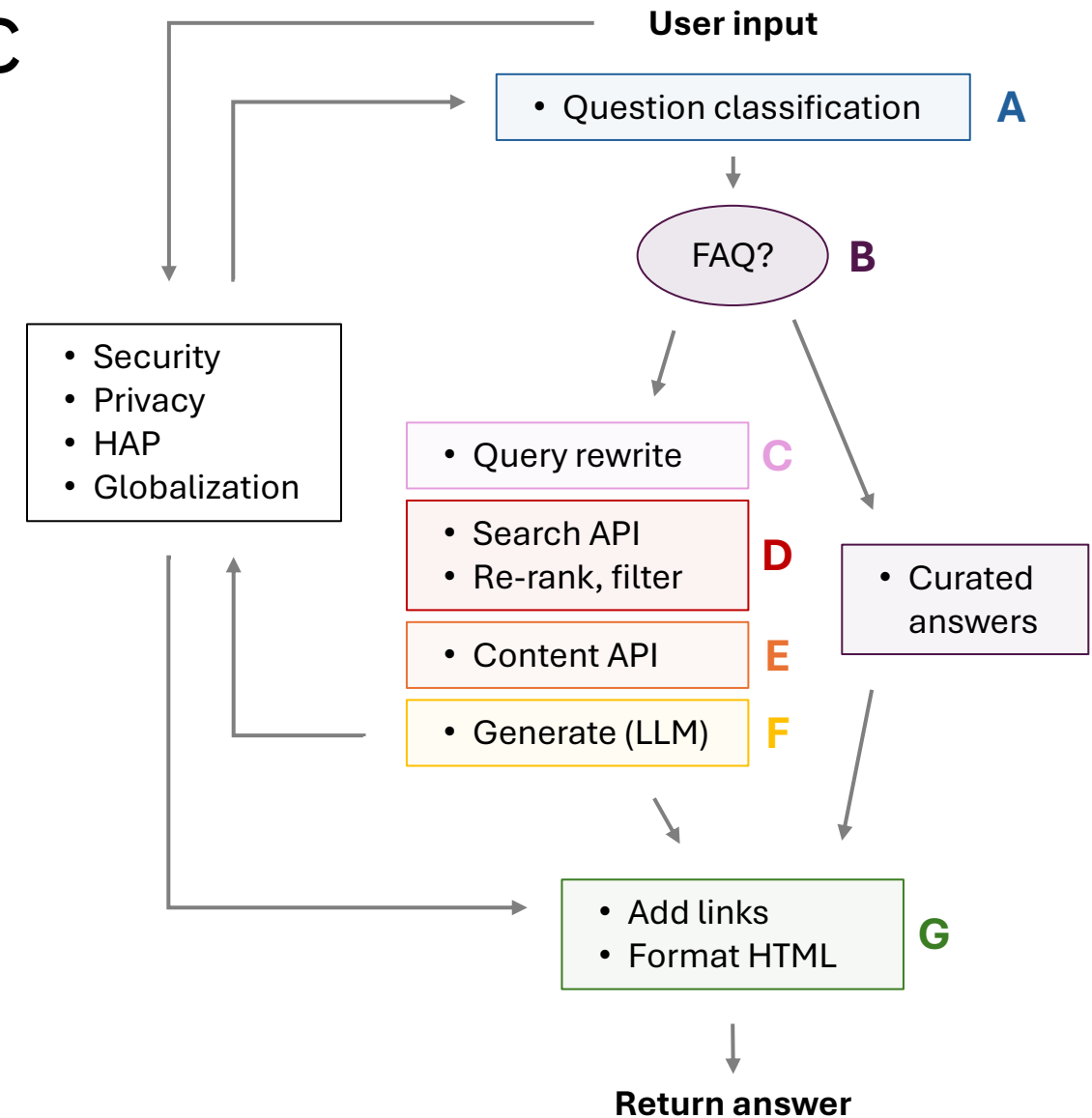
D: Use whatever search works*

E: Extract *complete text* of relevant articles

F: LLM has a simple job: answer based on articles

G: Return unharmed answer to user's language

*Articles are short, self-contained, and optimized for search, so traditional search methods work as well as or better than vector databases.



Modular, model-agnostic RAG implementation

A: Only valid, unharmed, questions, in English move on

B: Return curated answers for frequently asked questions

C: Clarify ambiguous, typo-filled questions

D: Use whatever search works*

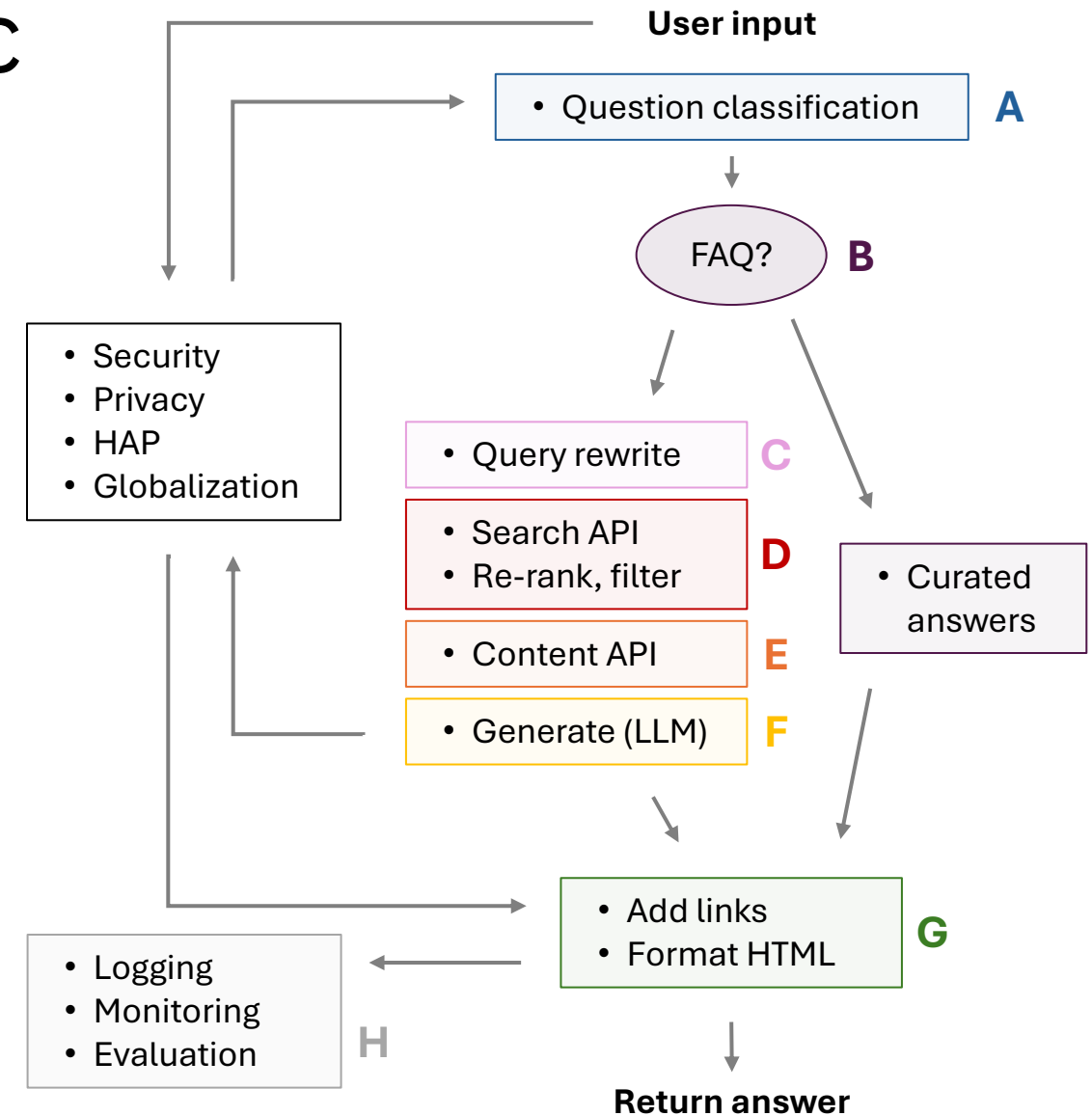
E: Extract *complete text* of relevant articles

F: LLM has a simple job: answer based on articles

G: Return unharmed answer to user's language

H: Monitor in Slack, capture results for evaluation

*Articles are short, self-contained, and optimized for search, so traditional search methods work as well as or better than vector databases.



Optimizing knowledge base content for RAG

- Topic-based writing
- SEO
- Meaningful anchor text
- Image descriptions
- Intentional table design
- Consistency
- Concise writing style
- Writing for translation
- Avoid idioms
- Include synonyms
- Consider accessibility
- Use active voice
- Avoid multiple negatives
- Lead-in sentence for lists
- List parallelism
- Simplify complex procedures
- Frame optional, condition steps
- Place modifiers carefully
- Inanimate object possessive
- Computerization
- Avoid user blame
- May vs. might vs. allow
- That vs. which
- With vs. use vs. together
- Nouns vs. names
- Avoid and/or

Evaluating RAG results

RAG eval

ADMINNERTAGSEVALCHARTS

Total: 110☐ Resolution

[2023-07-07 08:02:46 (UTC)] qa_xFBpKFYJqwnI

What is Decision Optimization?

Custom data:

- lang: fr
- question_class: what-is

Ignore

IBM Decision Optimization gives you access to IBM's industry-leading solution engines for mathematical programming and constraint programming.

- [Decision Optimization](#)
- [Decision Optimization experiments](#)

Valid question ?☒ Yes ☐ No ☐ ?

Correct class ?☒ Yes ☐ No ☐ ?

Article exists ?☒ Yes ☐ No ☐ ?

Search success ?☒ Top ☐ Top 3 ☐ Fail ☐ ?

Good answer ?☒ Yes ☐ Partly ☐ No ☐ ?

[2023-07-07 15:40:13 (UTC)] qa_ZpXhbUn5XcDI

how do I train a LLM?

Custom data:

- lang: en
- question_class: what-is

Ignore

No answer was found in the documentation

- [IBM Db2 for i connection](#)
- [IBM Db2 for z/OS connection](#)
- [IBM Data Virtualization Manager for z/OS connection](#)
- [Operations](#)

Valid question ?☒ Yes ☐ No ☐ ?

Correct class ?☐ Yes ☒ No ☐ ?

Article exists ?☐ Yes ☒ No ☐ ?

Search success ?☐ Top ☐ Top 3 ☒ Fail ☐ ?

Good answer ?☐ Yes ☐ Partly ☒ No ☐ ?

Evaluating RAG results

What about fully automated evaluation?

Novel questions:

- Question-article relevance
- Article-answer faithfulness
- Question-answer relevance
- Fact comparison
- LLM as answer judge
- User feedback

Evaluating RAG results

What about fully automated evaluation?

Novel questions:

- Question-article relevance
- Article-answer faithfulness
- Question-answer relevance
- Fact comparison
- LLM as answer judge
- User feedback

What if the
article is wrong?

Evaluating RAG results

What about fully automated evaluation?

Novel questions:

- Question-article relevance
- Article-answer faithfulness
- Question-answer relevance
- Fact comparison
- LLM as answer judge
- User feedback

What if the
article is wrong?

Faithful to the correct
part of the article?

Evaluating RAG results

What about fully automated evaluation?

Novel questions:

- Question-article relevance
- Article-answer faithfulness
- Question-answer relevance
- Fact comparison
- LLM as answer judge
- User feedback

What if the article is wrong?

Faithful to the correct part of the article?

Users don't want to do QA for you

Evaluating RAG results

What about fully automated evaluation?

Novel questions:

- Question-article relevance
- Article-answer faithfulness
- Question-answer relevance
- Fact comparison
- LLM as answer judge
- User feedback

What if the article is wrong?

Faithful to the correct part of the article?

Users don't want to do QA for you

With known answers:

- String similarity
- Semantic similarity
- LLM as judge

Evaluating RAG results

What about fully automated evaluation?

Novel questions:

- Question-article relevance
- Article-answer faithfulness
- Question-answer relevance
- Fact comparison
- LLM as answer judge
- User feedback

What if the article is wrong?

Faithful to the correct part of the article?

Users don't want to do QA for you

With known answers:

- String similarity
- Semantic similarity
- LLM as judge

What if there's more than one way to write the answer?

Evaluating RAG results

What about fully automated evaluation?

Novel questions:

- Question-article relevance
- Article-answer faithfulness
- Question-answer relevance
- Fact comparison
- LLM as answer judge
- User feedback

What if the article is wrong?

Faithful to the correct part of the article?

Users don't want to do QA for you

With known answers:

- String similarity
- Semantic similarity
- LLM as judge

What if there's more than one way to write the answer?

What if there's more than one right answer?

Evaluating RAG results

When search returns poor results, people just search again.

But with RAG solutions “answering” questions, users have higher expectations.

Air Canada ordered to pay customer who was misled by airline's chatbot

Company claimed its chatbot ‘was responsible for its own actions’ when giving wrong information about bereavement fare



<https://www.theguardian.com/world/2024/feb/16/air-canada-chatbot-lawsuit>

Sample code, discussion

https://github.com/spackows/ICAAI-2024_RAG-CD

She wore a lovely hat
on her head, which
was much too large



My neighbor
was walking her
dog in heels

<https://www.scribendi.com>