Writers are needed more than ever!

Sarah Packowski
AI ContentOps Architect,
IBM

RAG and agentic projects need content professionals

Growing In Content 2025

Agenda

1. RAG and agentic solutions
2. The role of content
3. Content strategy
4. Evaluating, improving results

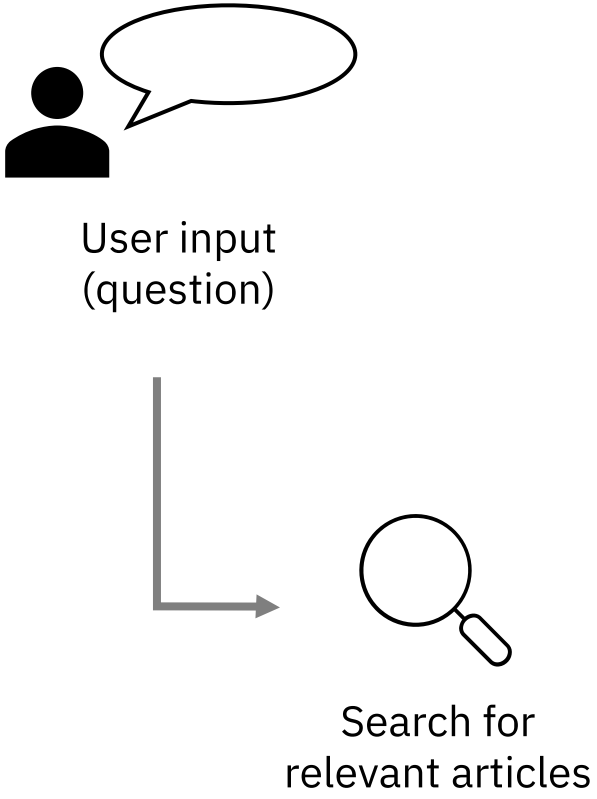
Agenda

1. RAG and agentic solutions
2. The role of content
3. Content strategy
4. Evaluating, improving results

Retrieval-augmented generation (RAG)

User input
(question)

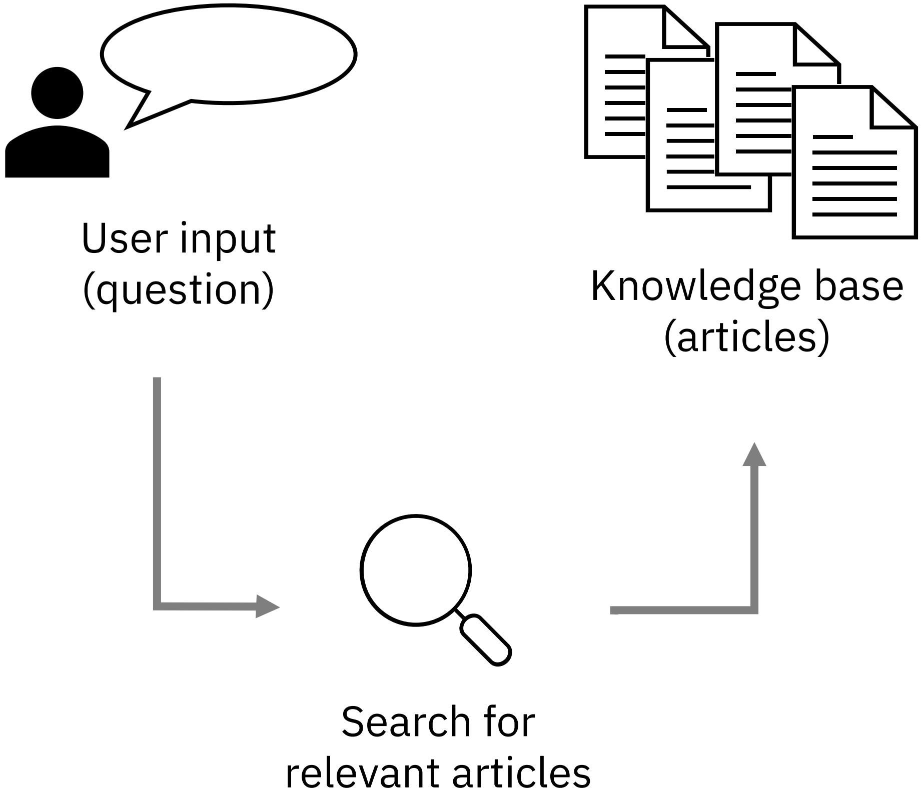
Retrieval-augmented generation (RAG)



User input
(question)

Search for
relevant articles

Retrieval-augmented generation (RAG)

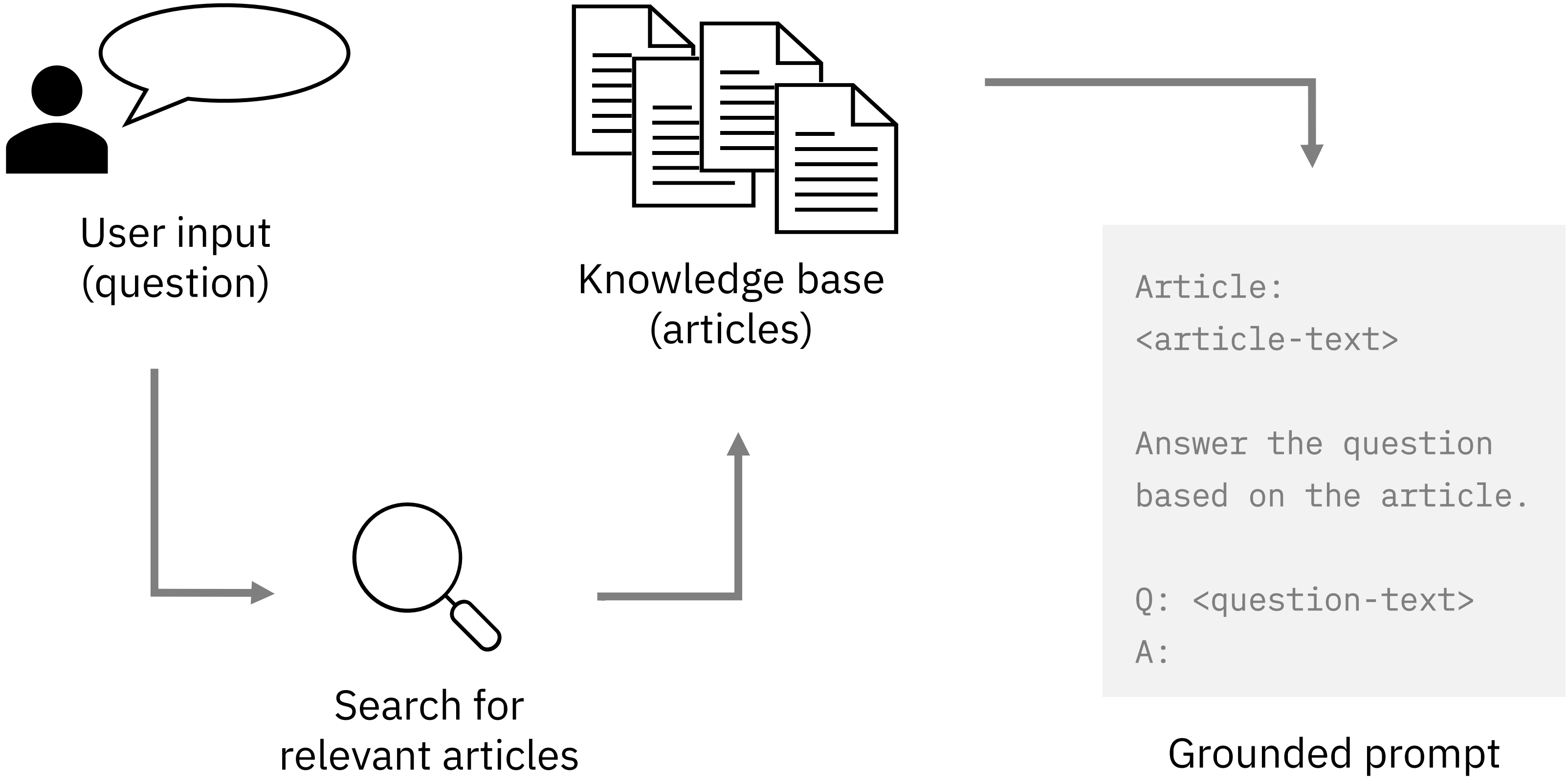


User input
(question)

Knowledge base
(articles)

Search for
relevant articles

Retrieval-augmented generation (RAG)



User input (question)

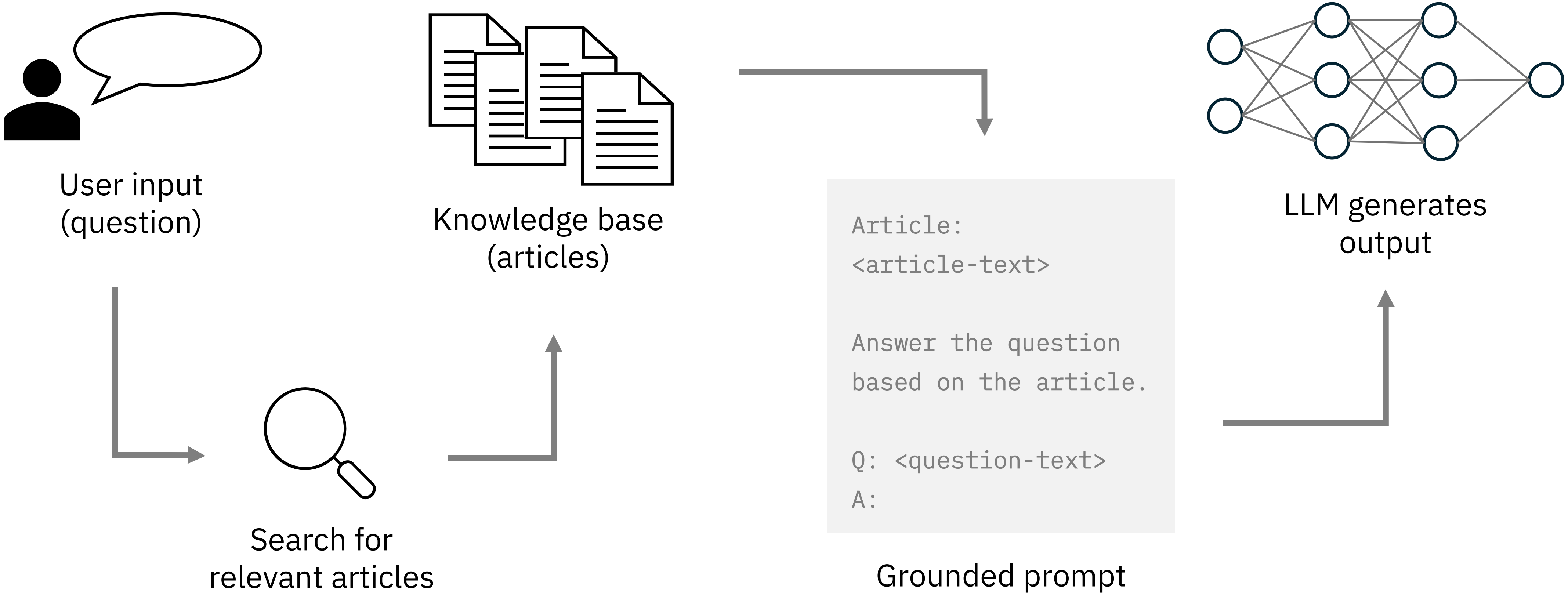
Knowledge base (articles)

Search for relevant articles

```
Article:
<article-text>

Answer the question
based on the article.

Q: <question-text>
A:
```

Grounded prompt

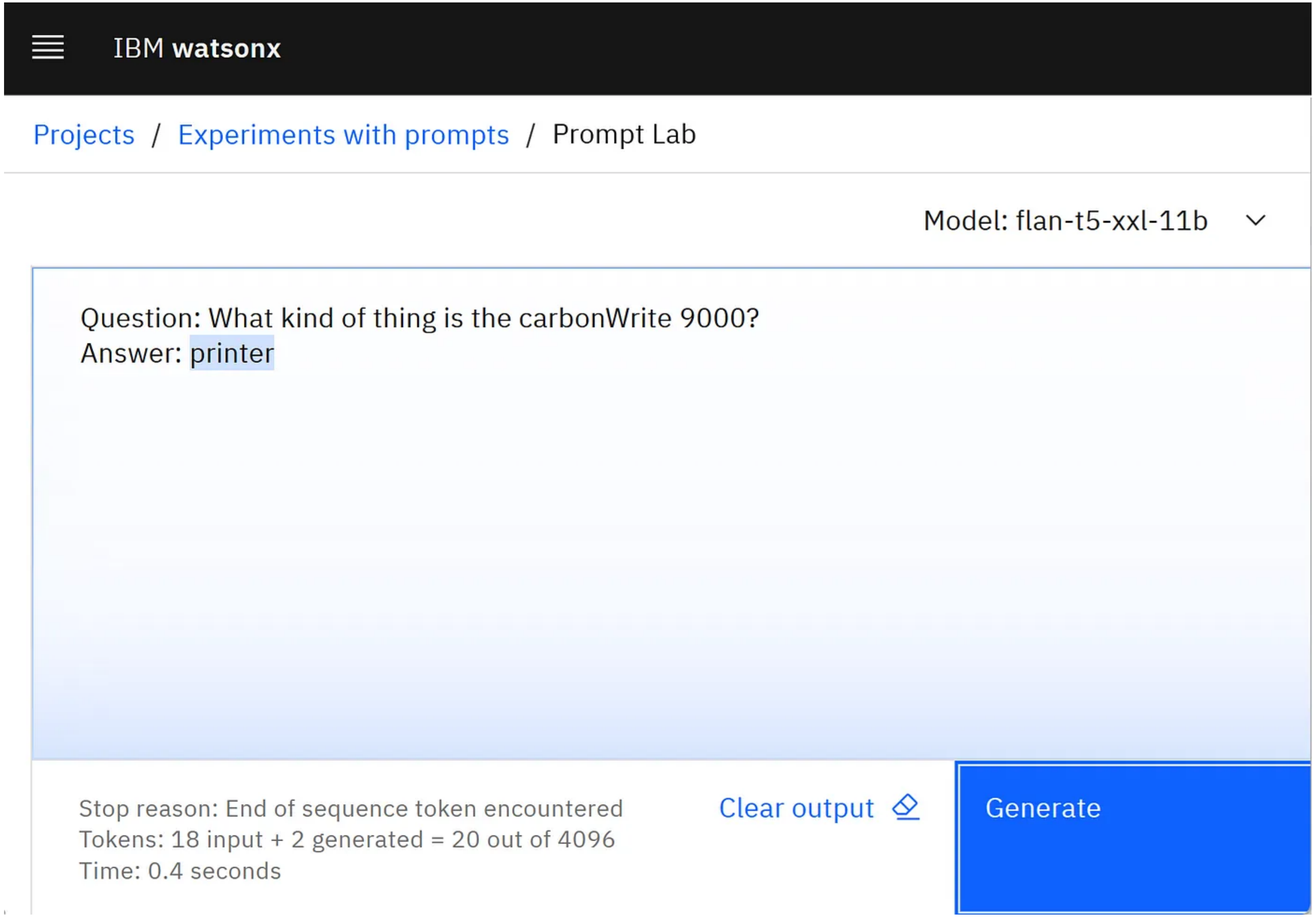
Retrieval-augmented generation (RAG)



```
Article:
<article-text>

Answer the question
based on the article.

Q: <question-text>
A:
```

Example

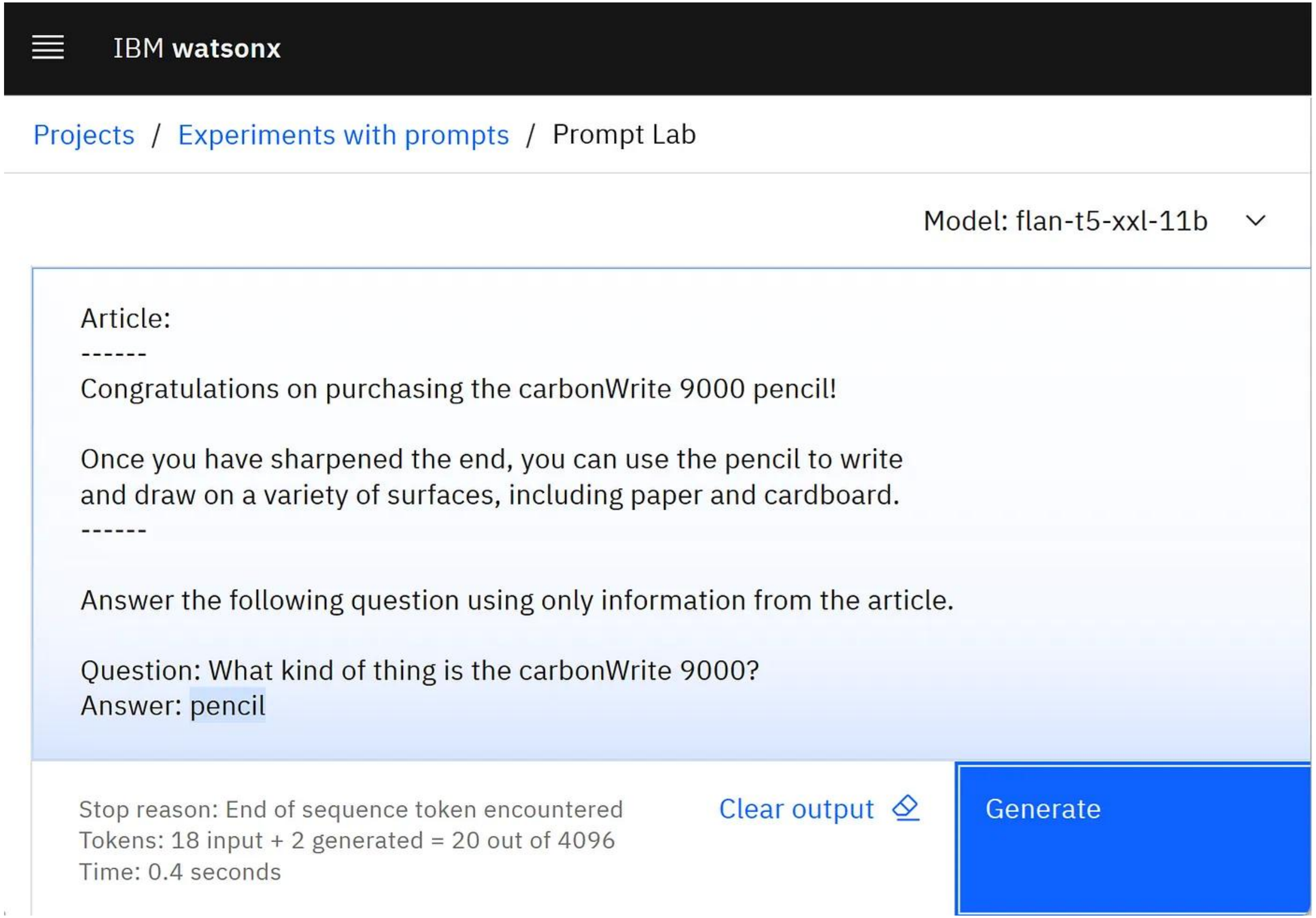
You have invented a new hand-held writing implement, called the *carbonWrite 9000*



Question: What kind of thing is the carbonWrite 9000?
Answer: printer

Stop reason: End of sequence token encountered
Tokens: 18 input + 2 generated = 20 out of 4096
Time: 0.4 seconds

watsonx.ai



Projects / Experiments with prompts / Prompt Lab

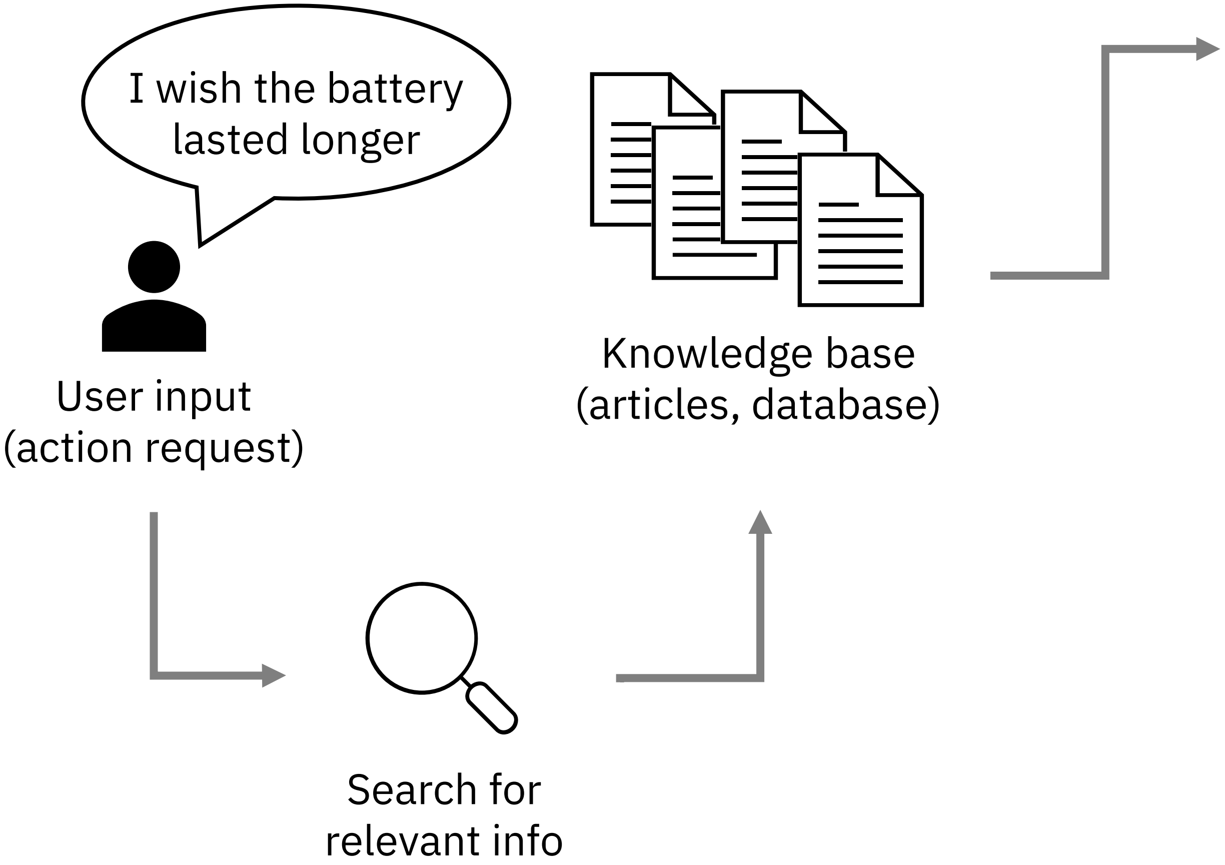
Model: flan-t5-xxl-11b

```
Article:
------
Congratulations on purchasing the carbonWrite 9000 pencil!

Once you have sharpened the end, you can use the pencil to write
and draw on a variety of surfaces, including paper and cardboard.
------

Answer the following question using only information from the article.

Question: What kind of thing is the carbonWrite 9000?
Answer: pencil
```

```
Stop reason: End of sequence token encountered
Tokens: 18 input + 2 generated = 20 out of 4096
Time: 0.4 seconds
```

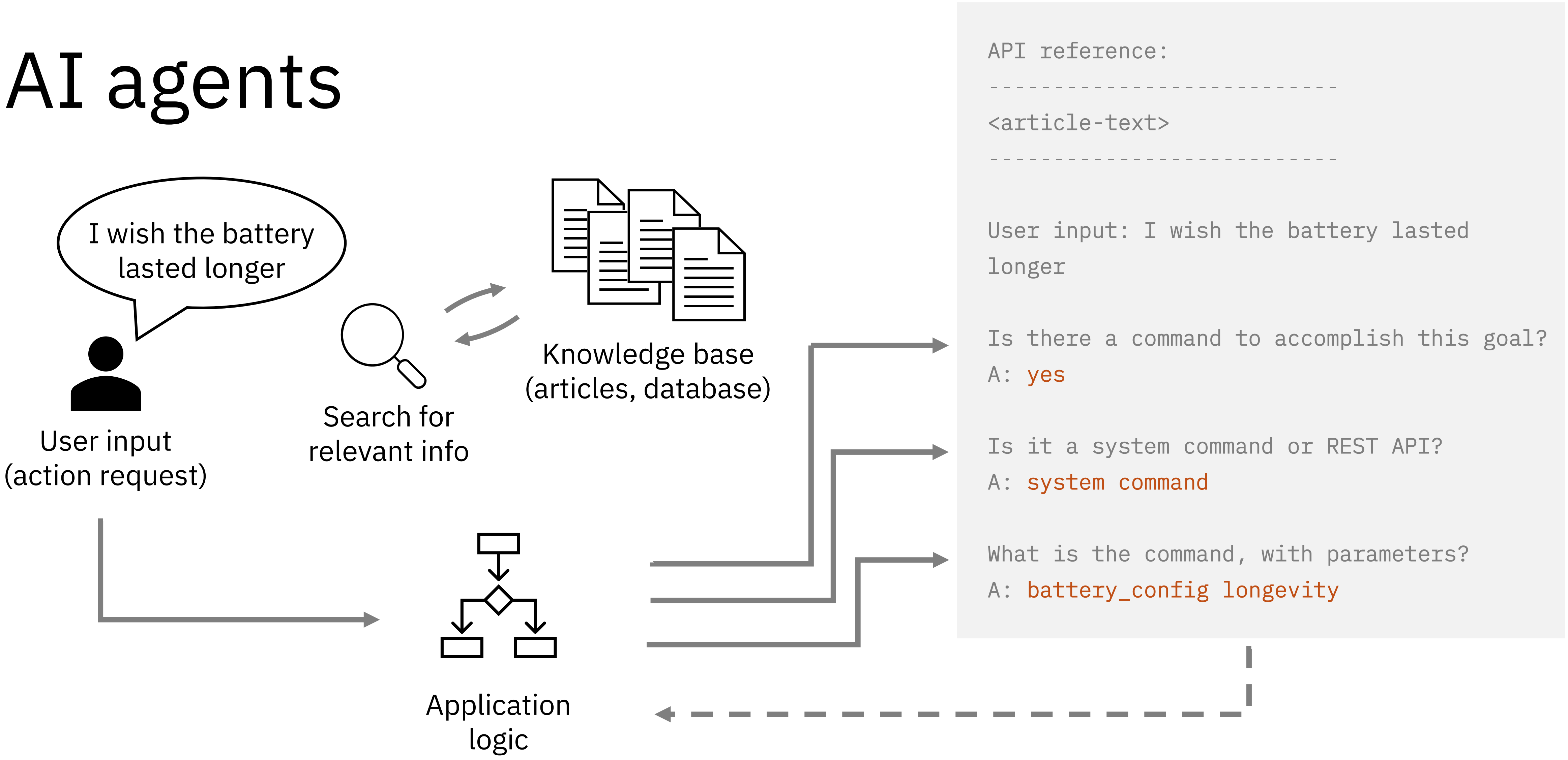
watsonx.ai

AI agents


I wish the battery lasted longer
User input (action request)
Knowledge base (articles, database)
Search for relevant info

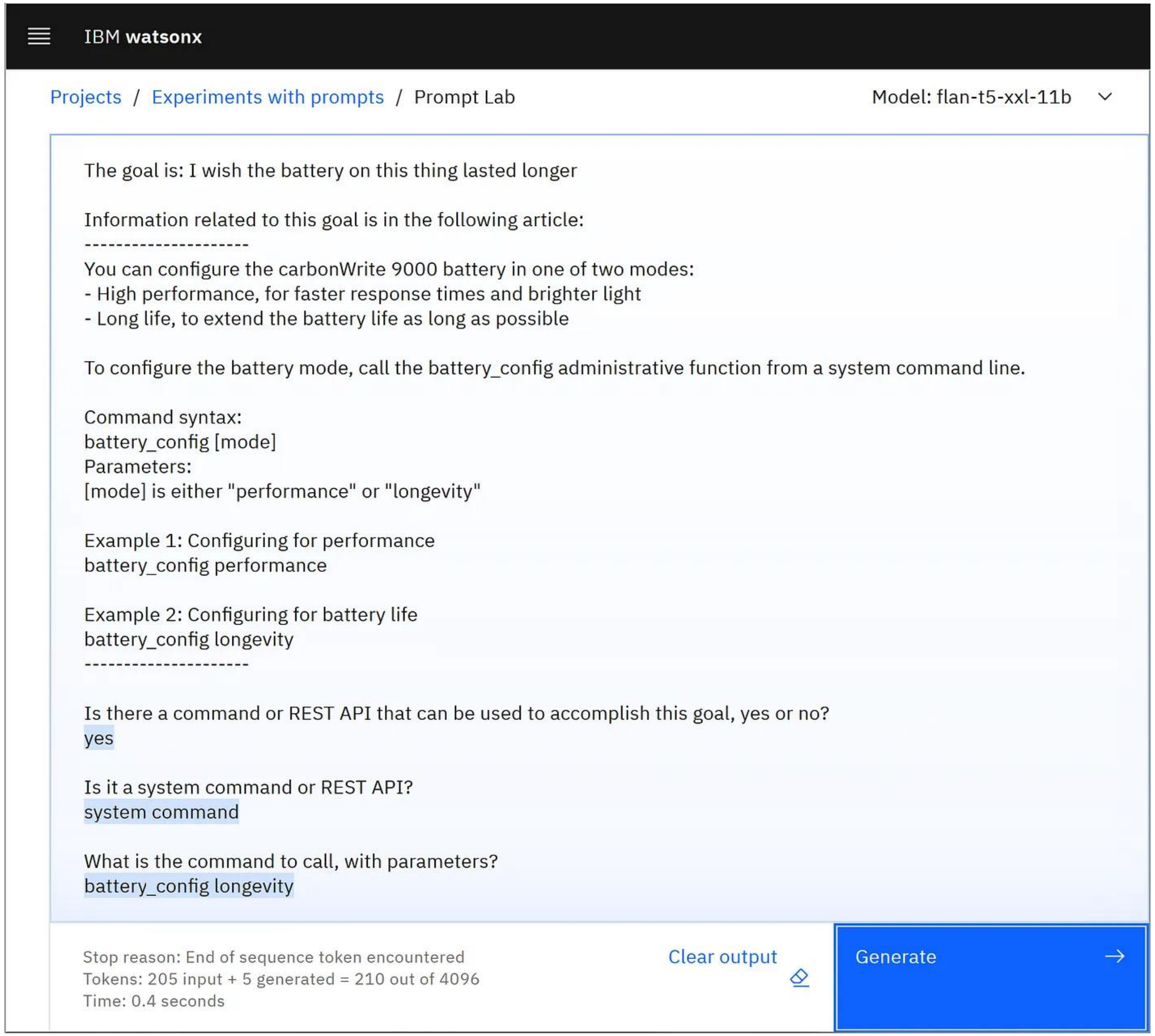
```
What is the command to achieve the goal in the
following user input?

API reference:
---------------------------
To configure the carbonWrite 9000 battery
mode, call battery_config.

## Syntax
battery_config [ performance | longevity ]

### Example 1: For performance
battery_config performance

### Example 2: For long battery life
battery_config longevity
---------------------------

User input: I wish the battery lasted longer
A: battery_config longevity
```

AI agents



I wish the battery lasted longer
User input (action request)
Search for relevant info
Knowledge base (articles, database)
Application logic

```
API reference:
---------------------------
<article-text>
---------------------------

User input: I wish the battery lasted
longer

Is there a command to accomplish this goal?
A: yes

Is it a system command or REST API?
A: system command

What is the command, with parameters?
A: battery_config longevity
```



Model: flan-t5-xxl-11b

```
The goal is: I wish the battery on this thing lasted longer

Information related to this goal is in the following article:
--------------------
You can configure the carbonWrite 9000 battery in one of two modes:
- High performance, for faster response times and brighter light
- Long life, to extend the battery life as long as possible

To configure the battery mode, call the battery_config administrative function from a system command line.

Command syntax:
battery_config [mode]
Parameters:
[mode] is either "performance" or "longevity"

Example 1: Configuring for performance
battery_config performance

Example 2: Configuring for battery life
battery_config longevity
--------------------

Is there a command or REST API that can be used to accomplish this goal, yes or no?
yes

Is it a system command or REST API?
system command

What is the command to call, with parameters?
battery_config longevity
```

```
Stop reason: End of sequence token encountered
Tokens: 205 input + 5 generated = 210 out of 4096
Time: 0.4 seconds
```

watsonx.ai

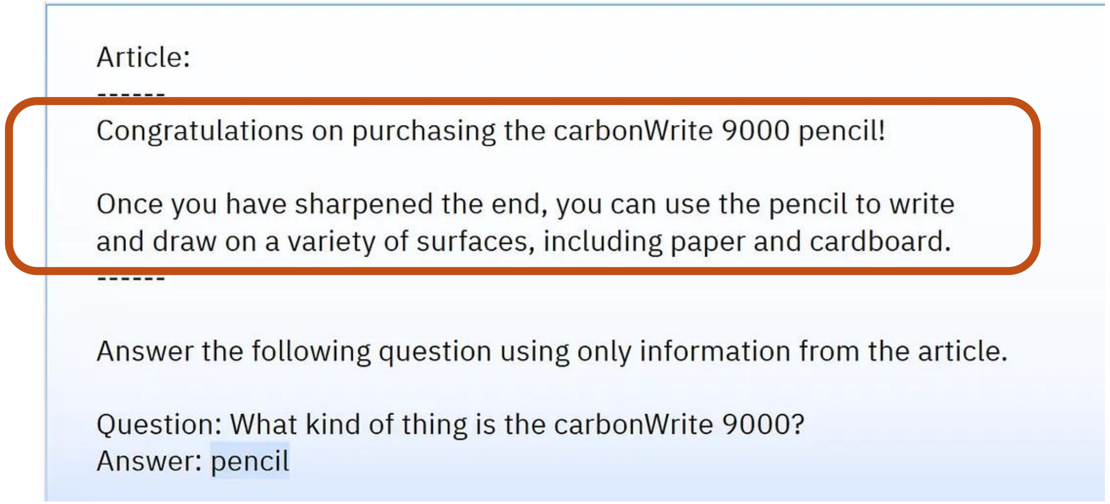
Agenda

1. RAG and agentic solutions
2. The role of content
3. Content strategy
4. Evaluating, improving results

Content is crucial

Where do these articles come from???

Writers!



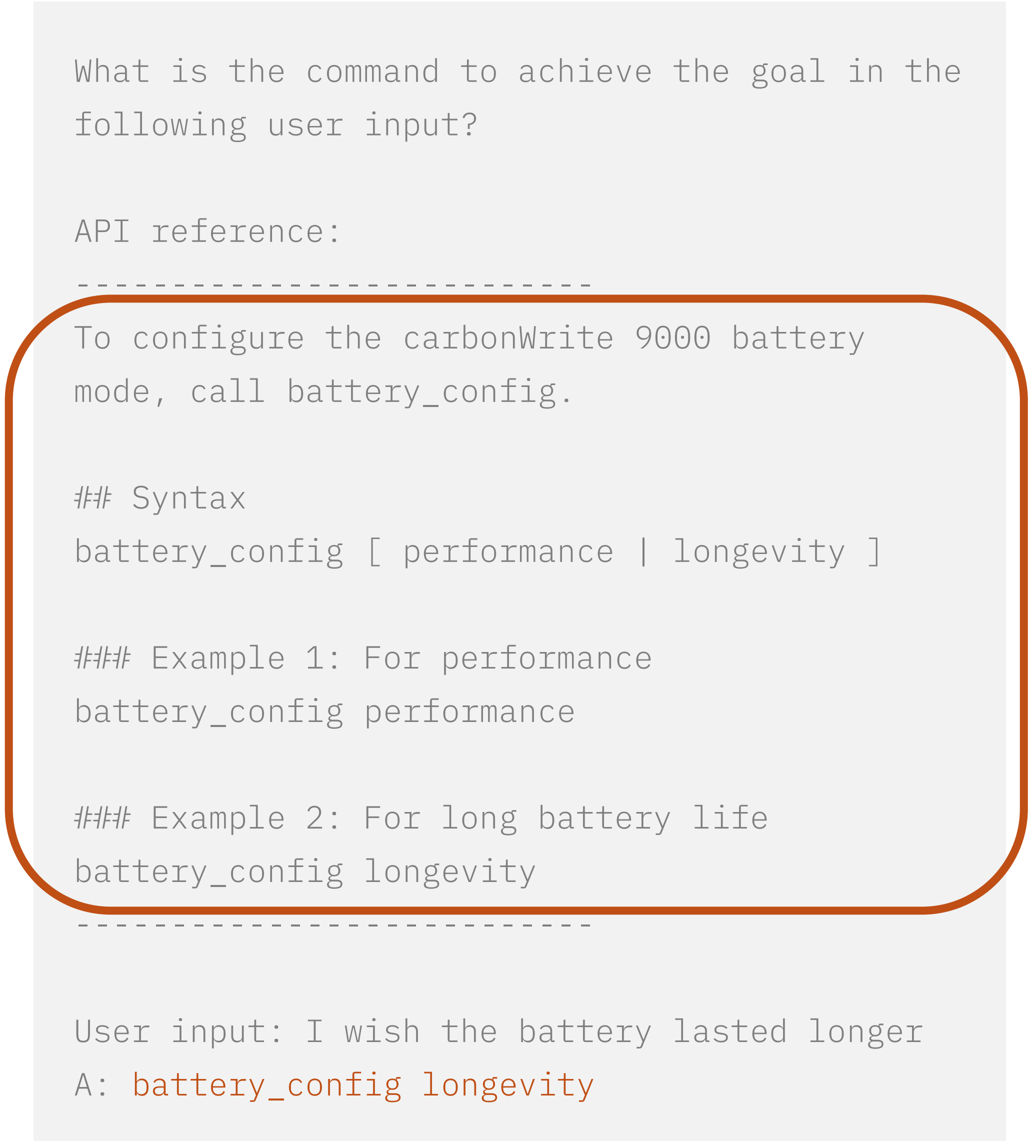
Article:

Congratulations on purchasing the carbonWrite 9000 pencil!

Once you have sharpened the end, you can use the pencil to write and draw on a variety of surfaces, including paper and cardboard.

Answer the following question using only information from the article.

Question: What kind of thing is the carbonWrite 9000?
Answer: pencil



```
What is the command to achieve the goal in the
following user input?

API reference:
---------------------------
To configure the carbonWrite 9000 battery
mode, call battery_config.

## Syntax
battery_config [ performance | longevity ]

### Example 1: For performance
battery_config performance

### Example 2: For long battery life
battery_config longevity
---------------------------

User input: I wish the battery lasted longer
A: battery_config longevity
```

Agenda

1. RAG and agentic solutions
2. The role of content
3. Content strategy
4. Evaluating, improving results

Question-driven content design

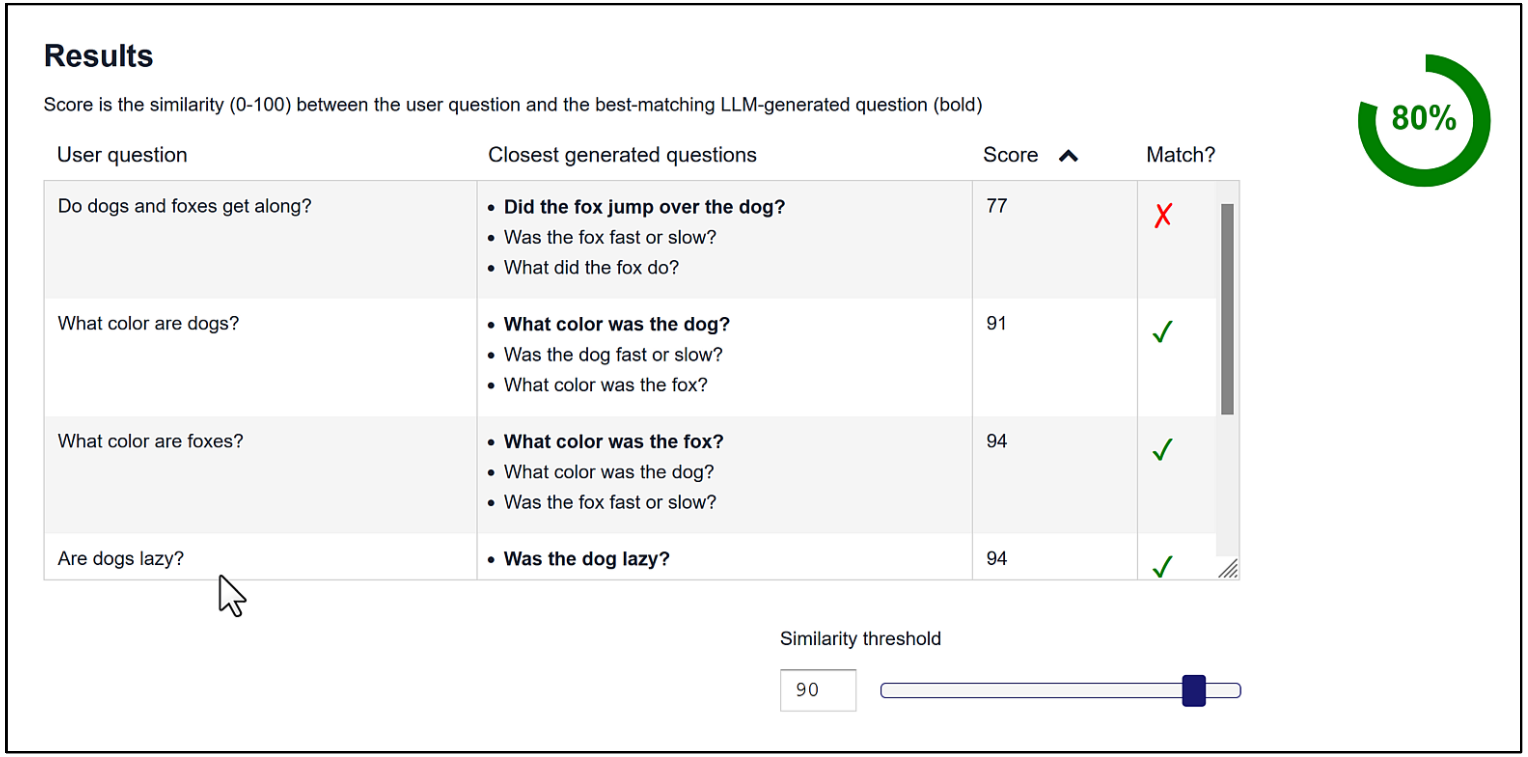
1. Collect real user questions
2. Use an LLM to generate questions “answered by your content”
3. Compare generated questions with real questions
   - When questions match, that indicates a RAG solution using your content as its knowledge base could answer those questions
   - User questions that have no matches indicate a content gap or problem

Deep dive: https://github.com/spackows/ConVEx-2025

Automated topic testing demo

Simple example: https://youtu.be/OWoHlQt0ANo

Real-world example: https://youtu.be/LwEB_obZxy0



Results

Score is the similarity (0-100) between the user question and the best-matching LLM-generated question (bold)

User question	Closest generated questions	Score ^	Match?
Do dogs and foxes get along?	• **Did the fox jump over the dog?** • Was the fox fast or slow? • What did the fox do?	77	✗
What color are dogs?	• **What color was the dog?** • Was the dog fast or slow? • What color was the fox?	91	✓
What color are foxes?	• **What color was the fox?** • What color was the dog? • Was the fox fast or slow?	94	✓
Are dogs lazy?	• **Was the dog lazy?**	94	✓

Similarity threshold: 90

Optimizing knowledge base content for RAG

https://ingeh.medium.com/adapting-content-for-ai-improving-accuracy-of-rag-solutions-4ab7a6d708a5

Adapting content for AI: Improving accuracy of RAG solutions

Inge Halilovic · Follow

3 min read · Mar 7, 2024

- 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

Agenda

1. RAG and agentic solutions
2. The role of content
3. Content strategy
4. Evaluating, improving results

Is your content RAG/Agent ready?

Is your content RAG/Agent ready?

Level 1

No value

Attributes:

- Limited access
- Hard to use

Even POCs not possible

Is your content RAG/Agent ready?

Level 1

No value

Attributes:

- Limited access
- Hard to use

Even POCs not possible

Is your content RAG/Agent ready?

Level 1

No value

Attributes:

- Limited access
- Hard to use

Even POCs not possible

Level 2

Some value

Attributes:

- Manual access

Maintaining solutions not sustainable

Is your content RAG/Agent ready?

Level 1

No value

Attributes:

- Limited access
- Hard to use

Even POCs not possible

👎

Level 2

Some value

Attributes:

- Manual access

Maintaining solutions not sustainable

😐

Is your content RAG/Agent ready?

Level 1

No value

Attributes:
- Limited access
- Hard to use

Even POCs not possible

👎

Level 2

Some value

Attributes:
- Manual access

Maintaining solutions not sustainable

😐

Level 3

Medium value

Attributes:
- Accessible
- Clean
- Content API
- Help available

Some production solutions

Is your content RAG/Agent ready?

Level 1

No value

Attributes:
- Limited access
- Hard to use

Even POCs not possible

👎

Level 2

Some value

Attributes:
- Manual access

Maintaining solutions not sustainable

😐

Level 3

Medium value

Attributes:
- Accessible
- Clean
- Content API
- Help available

Some production solutions

✓

Is your content RAG/Agent ready?

Level 1

No value

Attributes:

- Limited access
- Hard to use

Even POCs not possible

👎

Level 2

Some value

Attributes:

- Manual access

Maintaining solutions not sustainable

😐

Level 3

Medium value

Attributes:

- Accessible
- Clean
- Content API
- Help available

Some production solutions

✓

Level 4

High value

Attributes:

- Automated access
- Search API
- Robust help

Solutions increase productivity and reduce costs

Is your content RAG/Agent ready?

Level 1

No value

Attributes:

- Limited access
- Hard to use

Even POCs not possible

Level 2

Some value

Attributes:

- Manual access

Maintaining solutions not sustainable

Level 3

Medium value

Attributes:

- Accessible
- Clean
- Content API
- Help available

Some production solutions

Level 4

High value

Attributes:

- Automated access
- Search API
- Robust help

Solutions increase productivity and reduce costs

Is your content RAG/Agent ready?

Level 1

No value

Attributes:

- Limited access
- Hard to use

Even POCs not possible

Level 2

Some value

Attributes:

- Manual access

Maintaining solutions not sustainable

Level 3

Medium value

Attributes:

- Accessible
- Clean
- Content API
- Help available

Some production solutions

Level 4

High value

Attributes:

- Automated access
- Search API
- Robust help

Solutions increase productivity and reduce costs

Level 5

Winning!

Attributes:

- RAG use prioritized
- User community
- Governance
- Content –aaS

Internal and external solution builders can collaborate on complex solutions

Is your content RAG/Agent ready?

Level 1

No value

Attributes:
- Limited access
- Hard to use

Even POCs not possible

Level 2

Some value

Attributes:
- Manual access

Maintaining solutions not sustainable

Level 3

Medium value

Attributes:
- Accessible
- Clean
- Content API
- Help available

Some production solutions

Level 4

High value

Attributes:
- Automated access
- Search API
- Robust help

Solutions increase productivity and reduce costs

Level 5

Winning!

Attributes:
- RAG use prioritized
- User community
- Governance
- Content –aaS

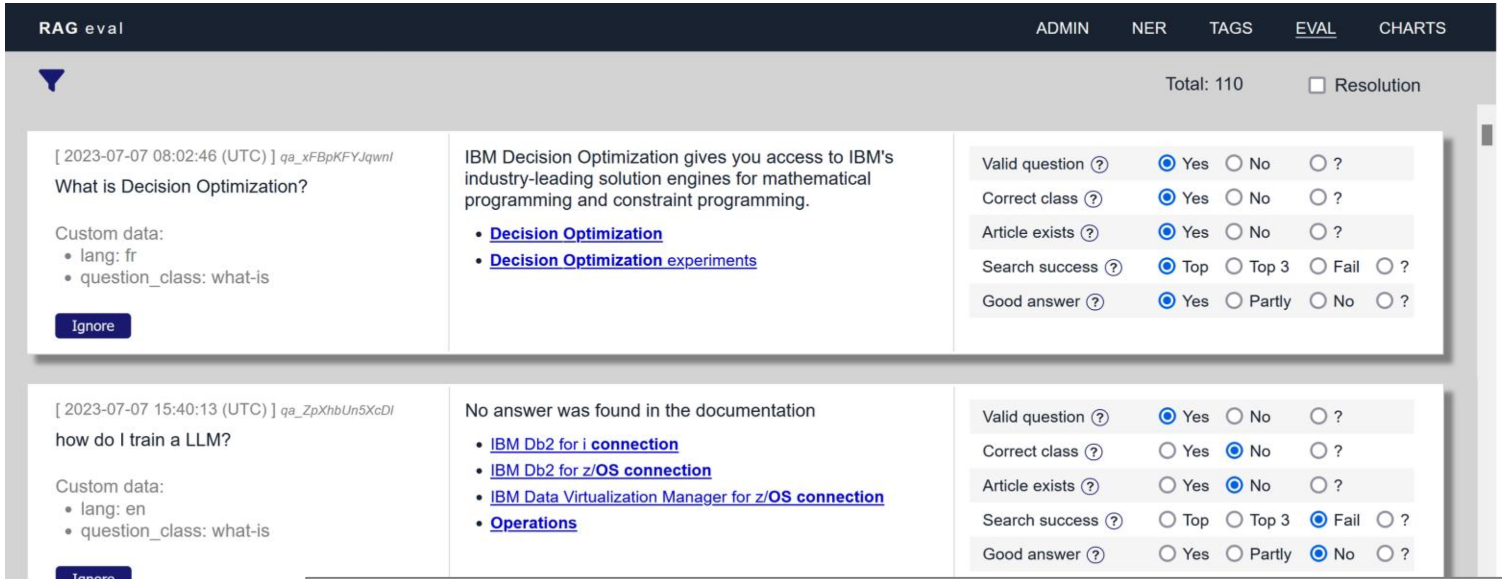
Internal and external solution builders can collaborate on complex solutions

Content rewriting experiment

1. Extracted 189 questions from NQ benchmark
2. Built simple RAG solution
3. Rewrote articles
4. Ran the questions again
5. Compared results

Just rewriting articles improved results to 100% correct

Evaluating RAG results



Total: 110 □ Resolution

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

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?

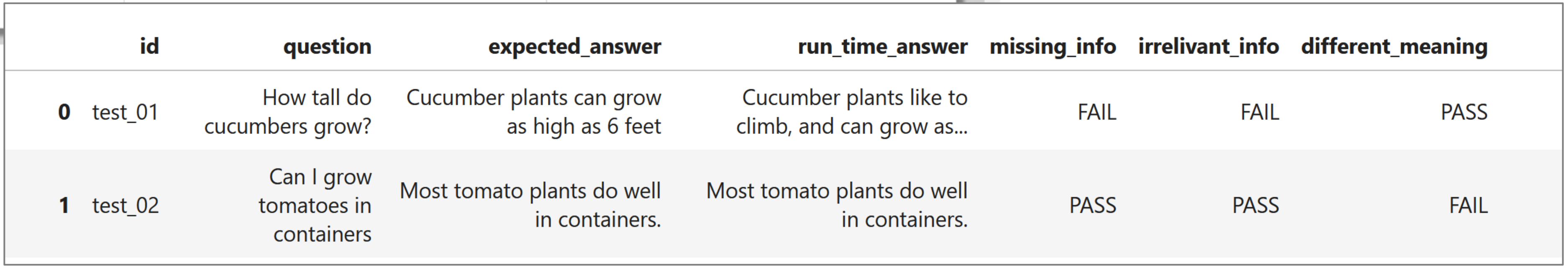
Custom data:
- lang: en
- question_class: what-is

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**

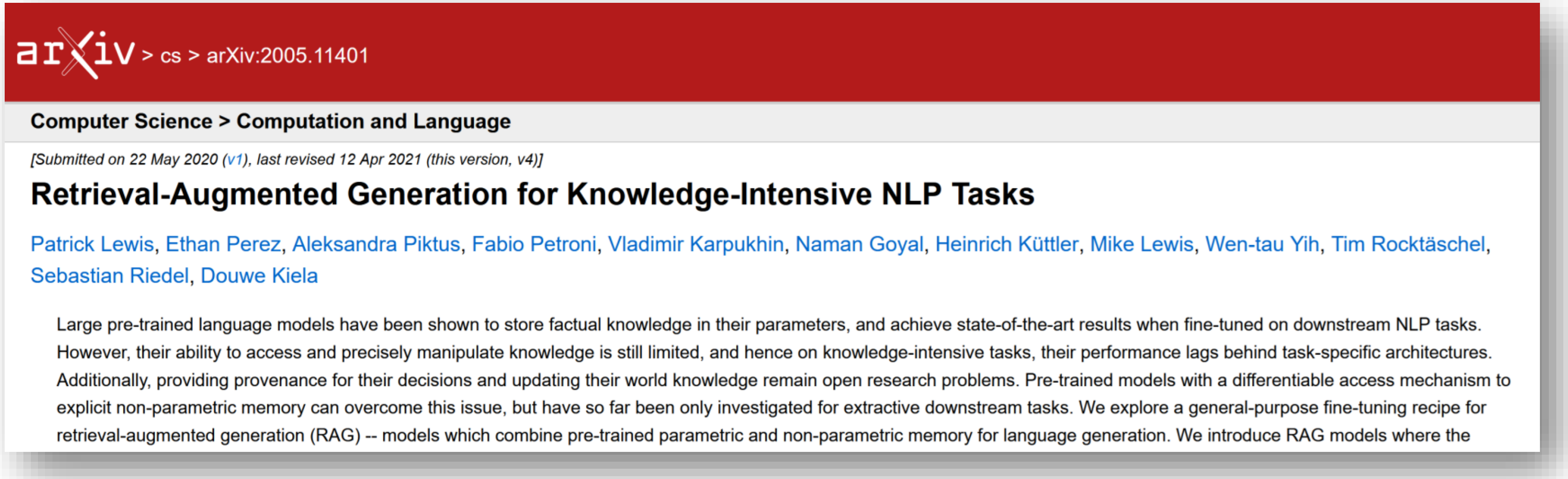
Valid question ?	● Yes	○ No	○ ?	
Correct class ?	○ Yes	● No	○ ?	
Article exists ?	○ Yes	● No	○ ?	
Search success ?	○ Top	○ Top 3	● Fail	○ ?
Good answer ?	○ Yes	○ Partly	● No	○ ?



	id	question	expected_answer	run_time_answer	missing_info	irrelivant_info	different_meaning
0	test_01	How tall do cucumbers grow?	Cucumber plants can grow as high as 6 feet	Cucumber plants like to climb, and can grow as...	FAIL	FAIL	PASS
1	test_02	Can I grow tomatoes in containers	Most tomato plants do well in containers.	Most tomato plants do well in containers.	PASS	PASS	FAIL

Sample Python notebook

Improving RAG results



arXiv > cs > arXiv:2005.11401

Computer Science > Computation and Language

[Submitted on 22 May 2020 (v1), last revised 12 Apr 2021 (this version, v4)]

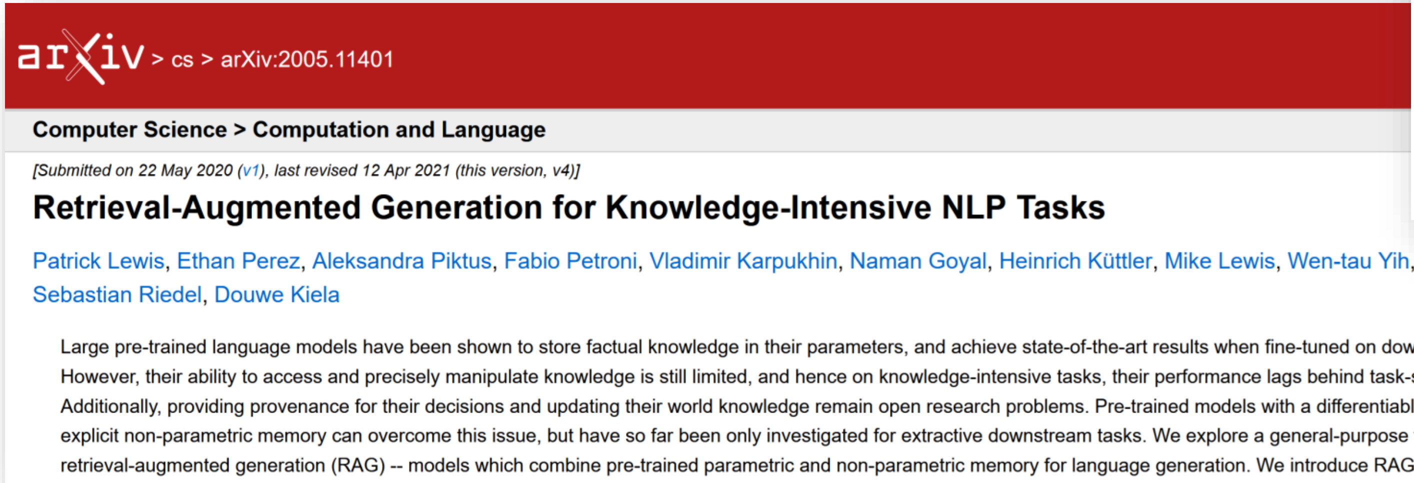
Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, Douwe Kiela

Large pre-trained language models have been shown to store factual knowledge in their parameters, and achieve state-of-the-art results when fine-tuned on downstream NLP tasks. However, their ability to access and precisely manipulate knowledge is still limited, and hence on knowledge-intensive tasks, their performance lags behind task-specific architectures. Additionally, providing provenance for their decisions and updating their world knowledge remain open research problems. Pre-trained models with a differentiable access mechanism to explicit non-parametric memory can overcome this issue, but have so far been only investigated for extractive downstream tasks. We explore a general-purpose fine-tuning recipe for retrieval-augmented generation (RAG) -- models which combine pre-trained parametric and non-parametric memory for language generation. We introduce RAG models where the

Source

Improving RAG results



arXiv > cs > arXiv:2005.11401

Computer Science > Computation and Language

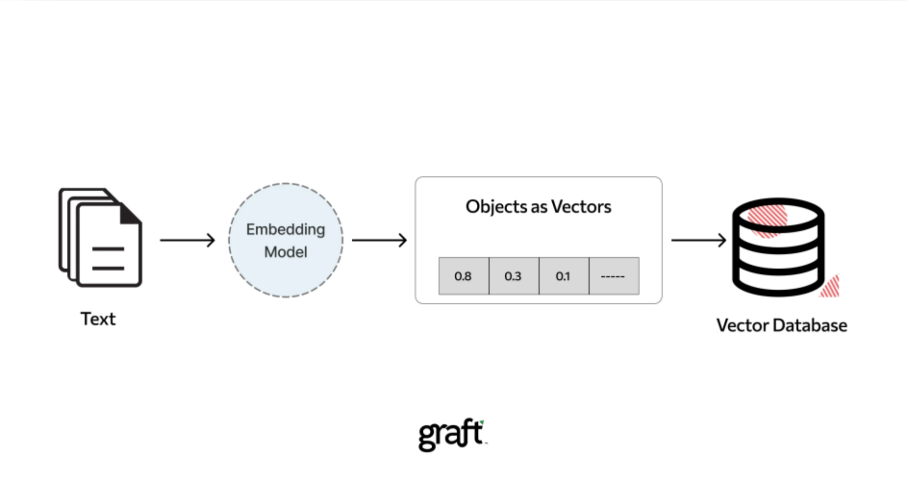
[Submitted on 22 May 2020 (v1), last revised 12 Apr 2021 (this version, v4)]

Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, Douwe Kiela

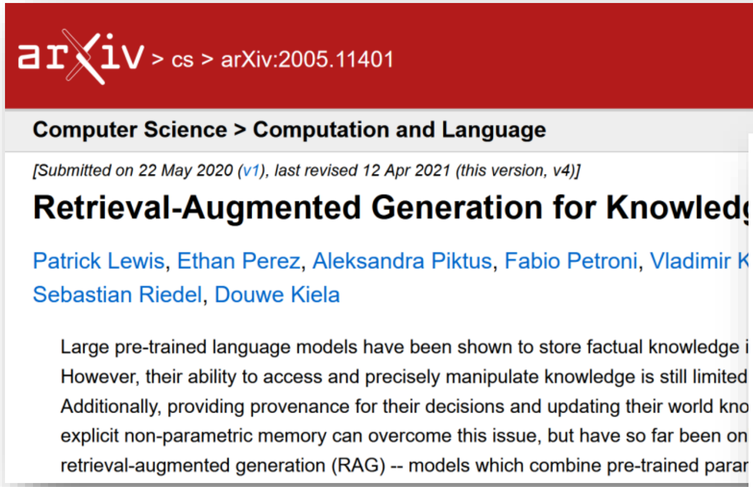
Large pre-trained language models have been shown to store factual knowledge in their parameters, and achieve state-of-the-art results when fine-tuned on downstream NLP tasks. However, their ability to access and precisely manipulate knowledge is still limited, and hence on knowledge-intensive tasks, their performance lags behind task-specific architectures. Additionally, providing provenance for their decisions and updating their world knowledge remain open research problems. Pre-trained models with a differentiable access mechanism to explicit non-parametric memory can overcome this issue, but have so far been only investigated for extractive downstream tasks. We explore a general-purpose fine-tuning recipe for retrieval-augmented generation (RAG) -- models which combine pre-trained parametric and non-parametric memory for language generation. We introduce RAG models where the

Source



Source

Improving RAG results



arXiv > cs > arXiv:2005.11401

Computer Science > Computation and Language

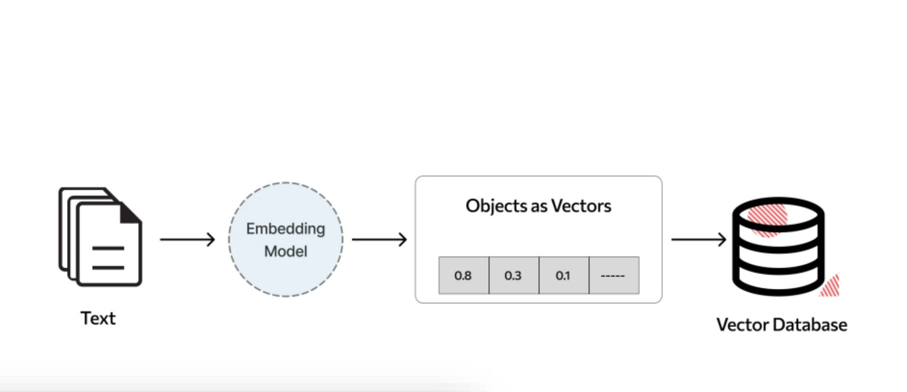
[Submitted on 22 May 2020 (v1), last revised 12 Apr 2021 (this version, v4)]

Retrieval-Augmented Generation for Knowledg

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir K
Sebastian Riedel, Douwe Kiela

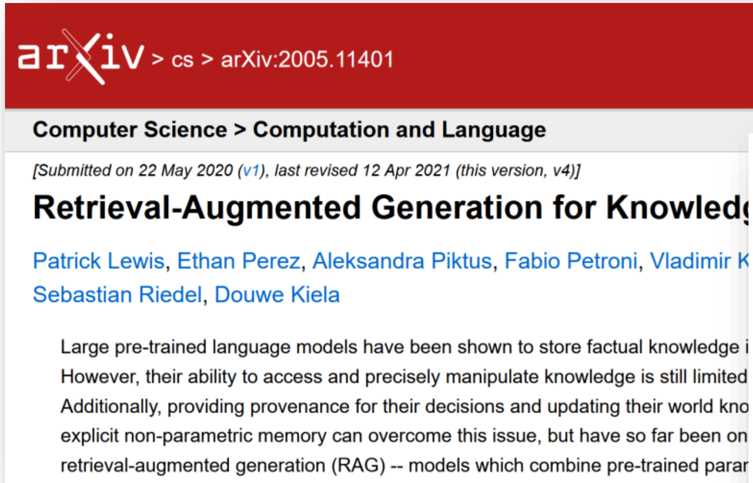
Large pre-trained language models have been shown to store factual knowledge i
However, their ability to access and precisely manipulate knowledge is still limited
Additionally, providing provenance for their decisions and updating their world kno
explicit non-parametric memory can overcome this issue, but have so far been on
retrieval-augmented generation (RAG) -- models which combine pre-trained parar

Source



Source

Improving RAG results



arXiv > cs > arXiv:2005.11401

Computer Science > Computation and Language

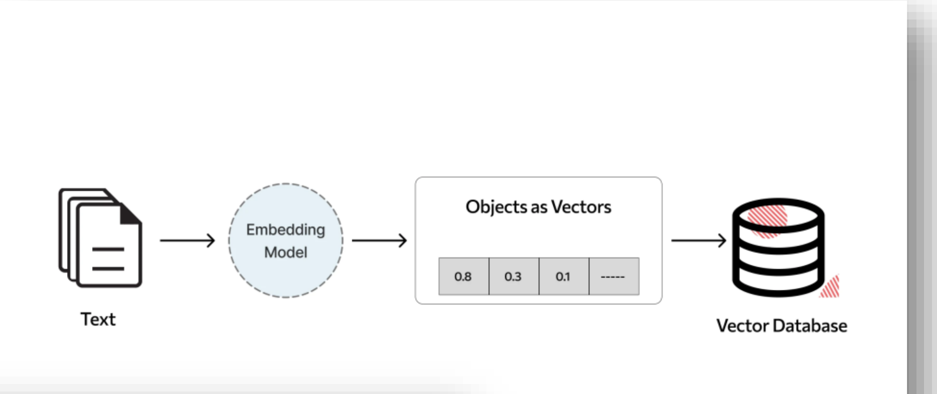
[Submitted on 22 May 2020 (v1), last revised 12 Apr 2021 (this version, v4)]

Retrieval-Augmented Generation for Knowledg

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir K
Sebastian Riedel, Douwe Kiela

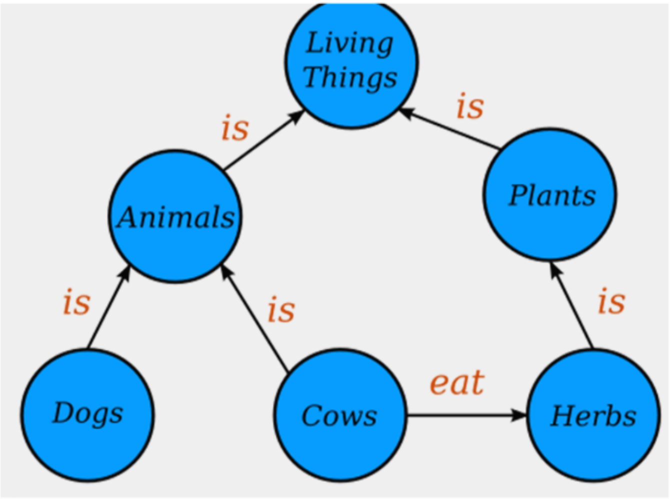
Large pre-trained language models have been shown to store factual knowledge i
However, their ability to access and precisely manipulate knowledge is still limited
Additionally, providing provenance for their decisions and updating their world kno
explicit non-parametric memory can overcome this issue, but have so far been on
retrieval-augmented generation (RAG) -- models which combine pre-trained paran

Source



Source

Source



Source

Conclusion

What is our role in all of this?

Conclusion

Are we to be RAG quality assurance now?

Conclusion

RAG and agentic solutions cannot work without accurate, up-to-date, domain-specific content pulled into their prompts.

Conclusion

RAG and agentic solutions cannot work without accurate, up-to-date, domain-specific content pulled into their prompts.

But we'll have to prove that.

Conclusion

Should we create new titles or descriptions for what we do?

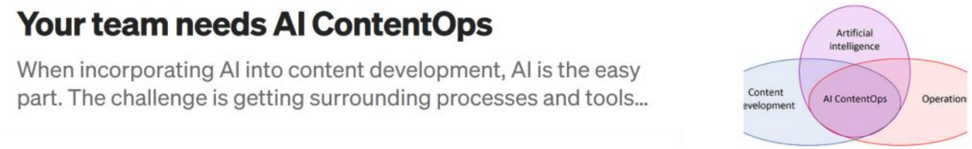
Knowledge base design

Knowledge base engineer

RAG content strategist

Blogs

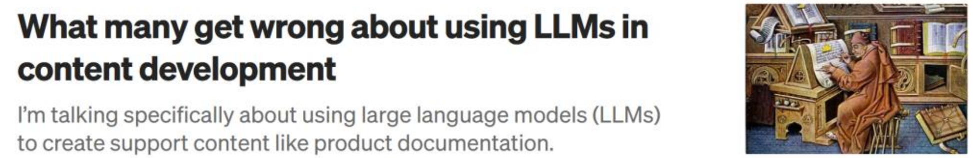
https://medium.com/@sarah-packowski



Your team needs AI ContentOps

When incorporating AI into content development, AI is the easy part. The challenge is getting surrounding processes and tools...

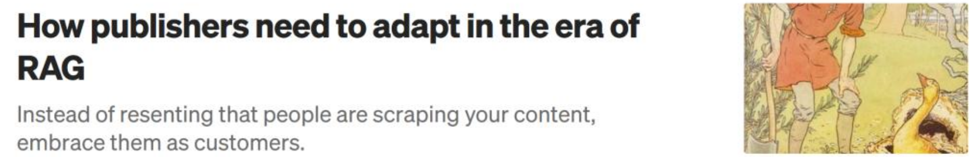
Link



What many get wrong about using LLMs in content development

I'm talking specifically about using large language models (LLMs) to create support content like product documentation.

Link



How publishers need to adapt in the era of RAG

Instead of resenting that people are scraping your content, embrace them as customers.

Link

The role of technical writers in the age of RAG and agentic LLM solutions

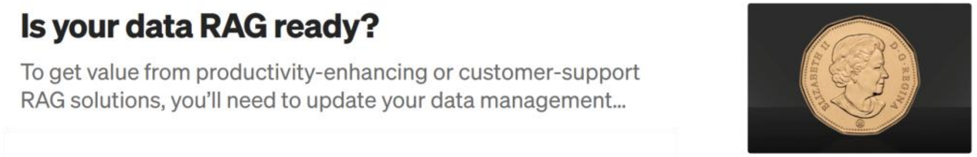
Good news: They're gonna need us more than ever, fellow writers!

Link

The unintended impact of Wikipedia on RAG best practices

Your RAG solution might be starting on the wrong foot. And it's because Wikipedia is so darn great.

Link



Is your data RAG ready?

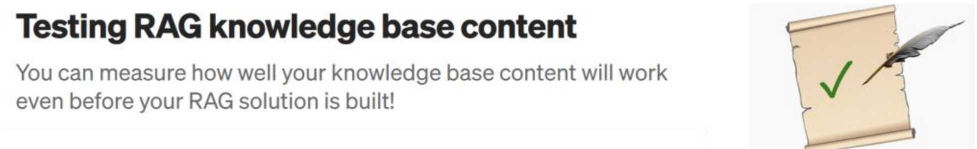
To get value from productivity-enhancing or customer-support RAG solutions, you'll need to update your data management...

Link

Question-driven content design

An new strategy for the age of RAG.

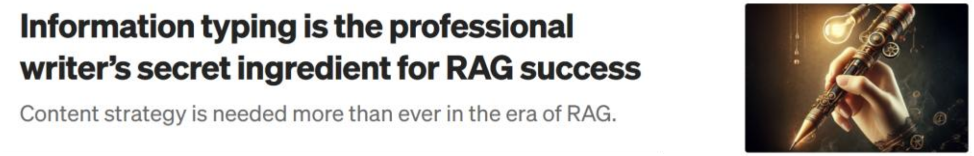
Link



Testing RAG knowledge base content

You can measure how well your knowledge base content will work even before your RAG solution is built!

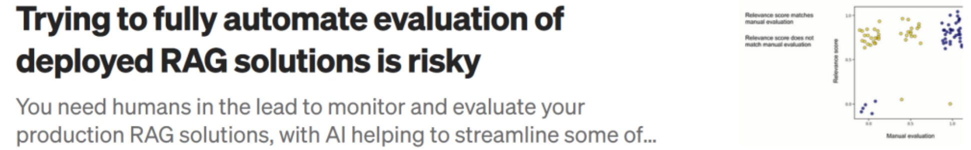
Link



Information typing is the professional writer's secret ingredient for RAG success

Content strategy is needed more than ever in the era of RAG.

Link



Trying to fully automate evaluation of deployed RAG solutions is risky

You need humans in the lead to monitor and evaluate your production RAG solutions, with AI helping to streamline some of...

Link