

LLM Engineering: Building Production-Ready LLM-enabled systems

Leon Schrijvers – l.schrijvers@fontys.nl / Petra Heck – p.heck@fontys.nl
Fontys Knowledge Center Applied AI For Society / Fontys ICT
DEMAND Research Project – <https://demand.nl>



FONTYS KNOWLEDGE CENTER

APPLIED AI FOR SOCIETY

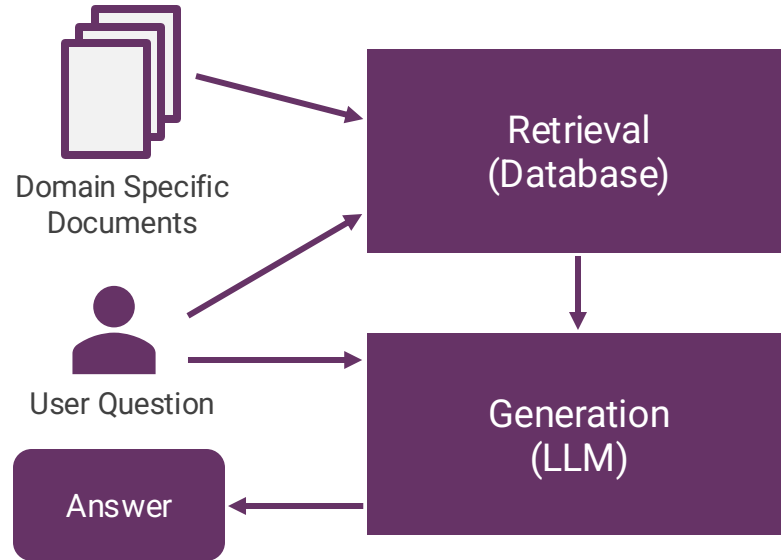
EXPLORE THE FUTURE OF
TRUSTWORTHY AI

Case Study – Retrieval Augmented Generation

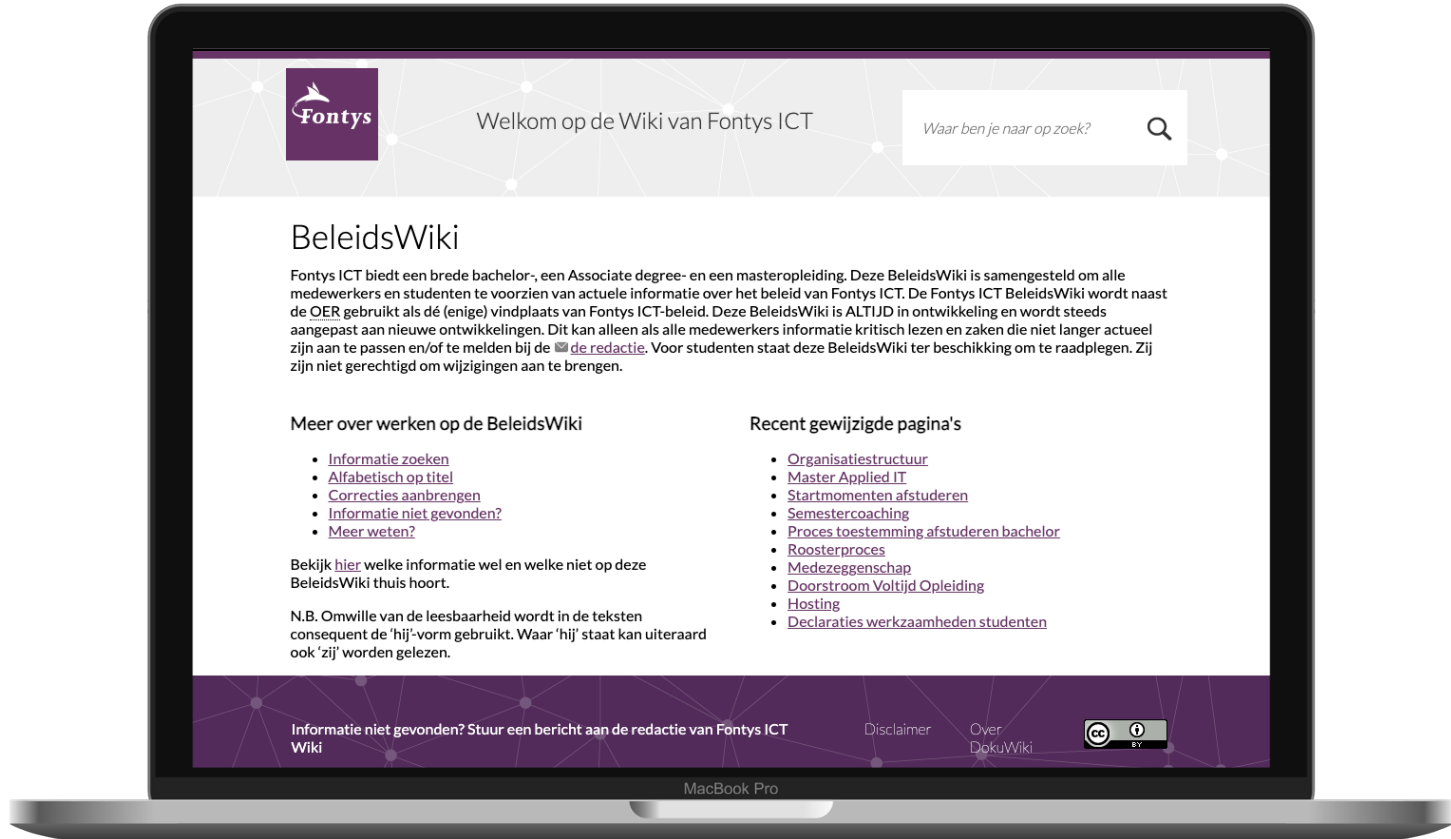
How to design a chatbot application that enables company employees to ask questions about company specific documents?

Solution Characteristics:

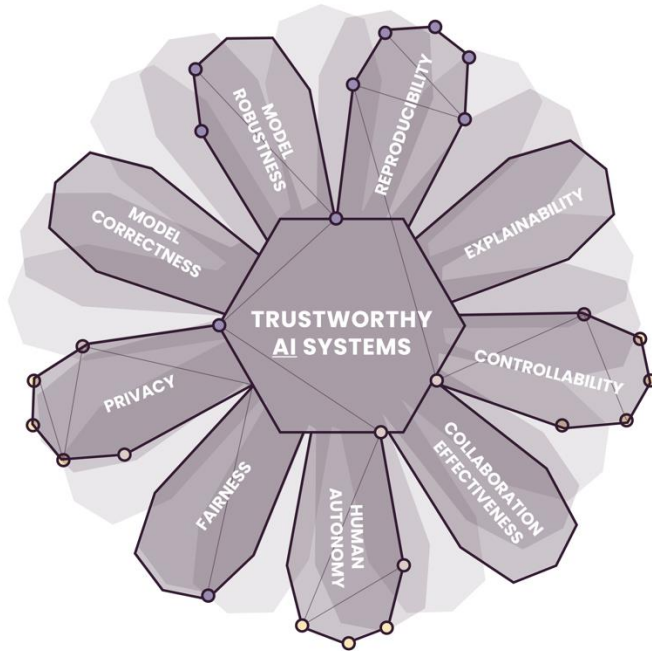
- A Reliable results
- B Everything runs local
- C Easy maintainable
- D Easy to use for non-technical users



Case Study – Retrieval Augmented Generation



Trustworthy AI Systems



Model Correctness

Model Robustness

Reproducibility

Explainability

Controllability

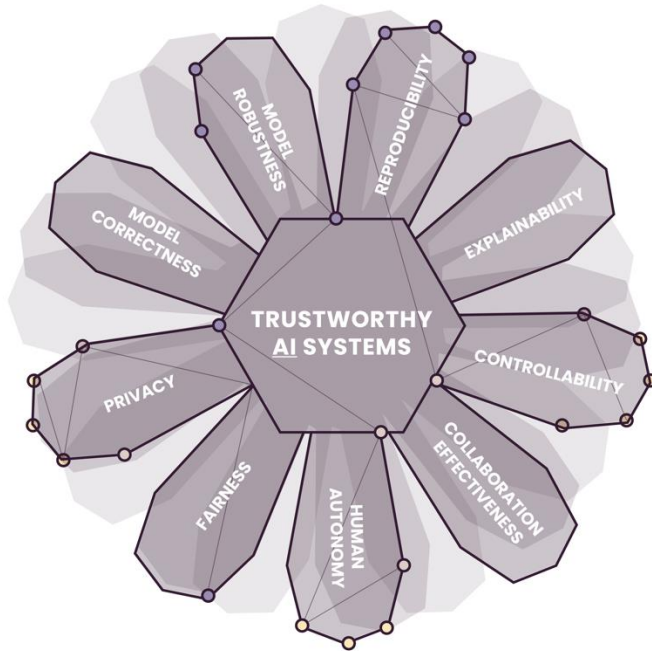
Collaboration Effectiveness

Human Autonomy

Fairness

Privacy

Trustworthy AI Systems



Model Correctness

Model Robustness

Reproducibility

Explainability

Controllability

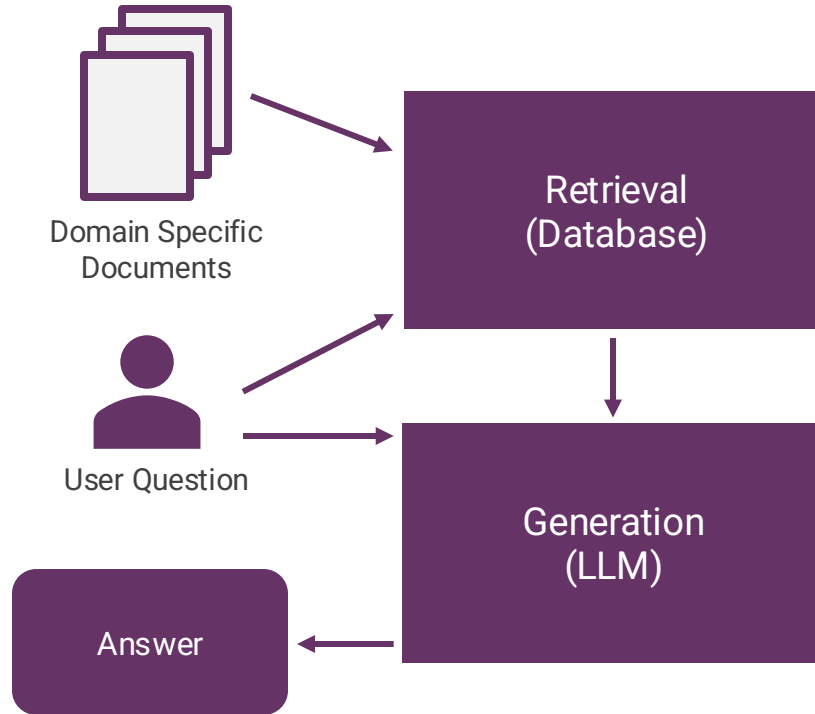
Collaboration Effectiveness

Human Autonomy

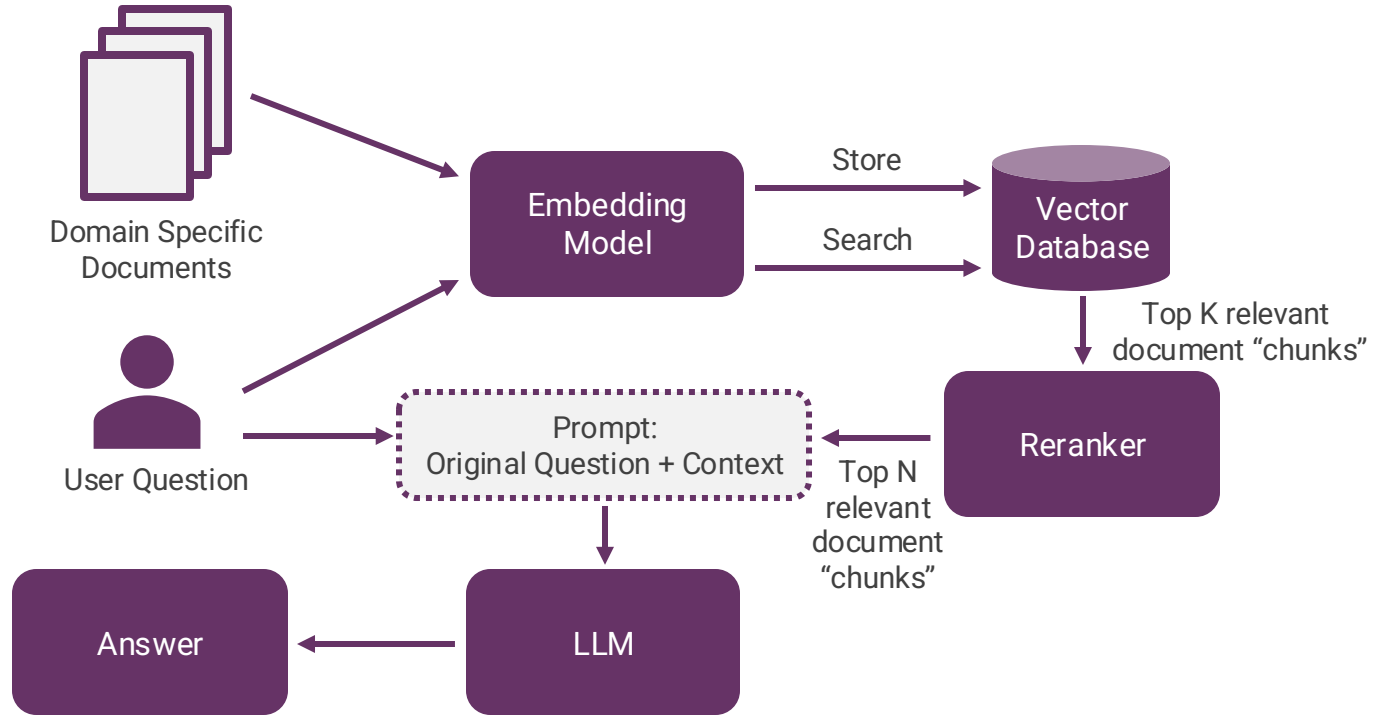
Fairness

Privacy

Architecture of RAG Applications



Architecture of RAG Applications

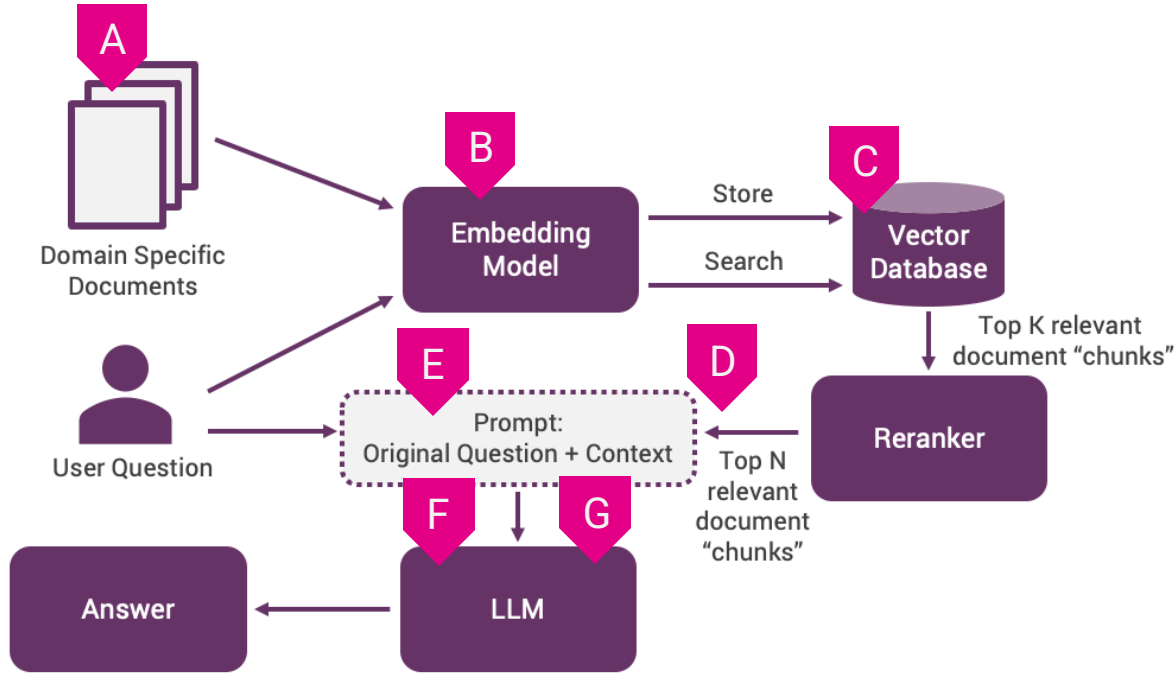


It's not hard to build something that works from a technical point of view.

It's not hard to build something that works from a technical point of view.

The real challenge lies in assessing how well it performs and what 'well' even means in your problem domain.

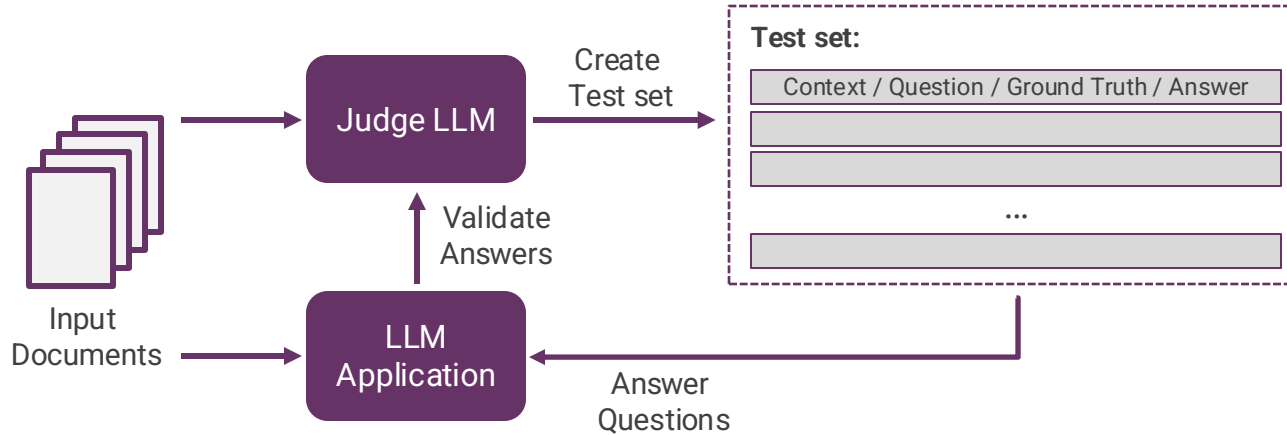
Validation of RAG Applications



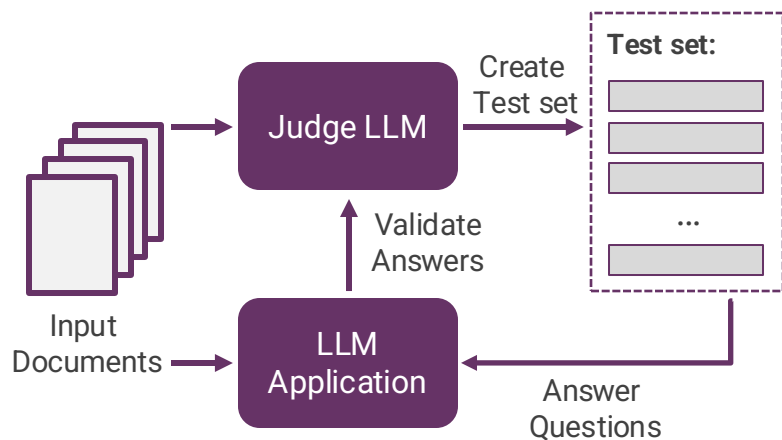
Test Points

- A** Input document quality
- B** Document chunking & embedding
- C** Chunk retrieval
- D** Context construction
- E** Prompt engineering
- F** Answer quality
- G** LLM updates

Automated Validation of RAG Pipelines



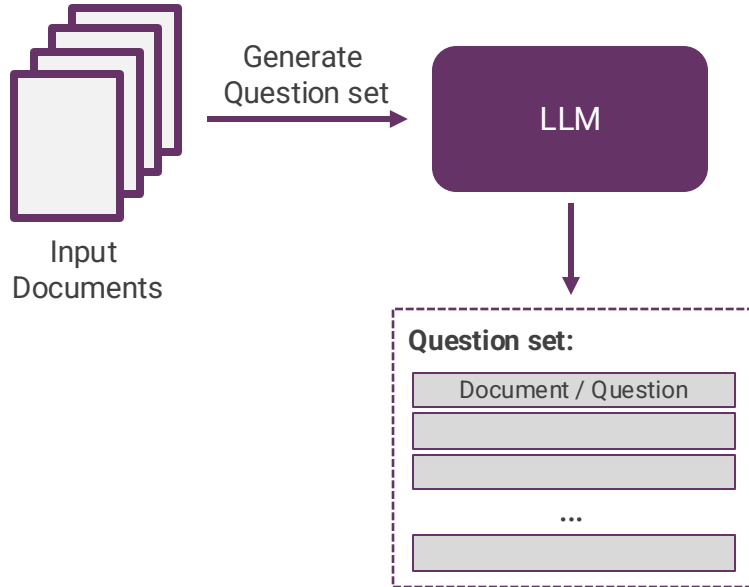
Automated Validation of RAG Pipelines



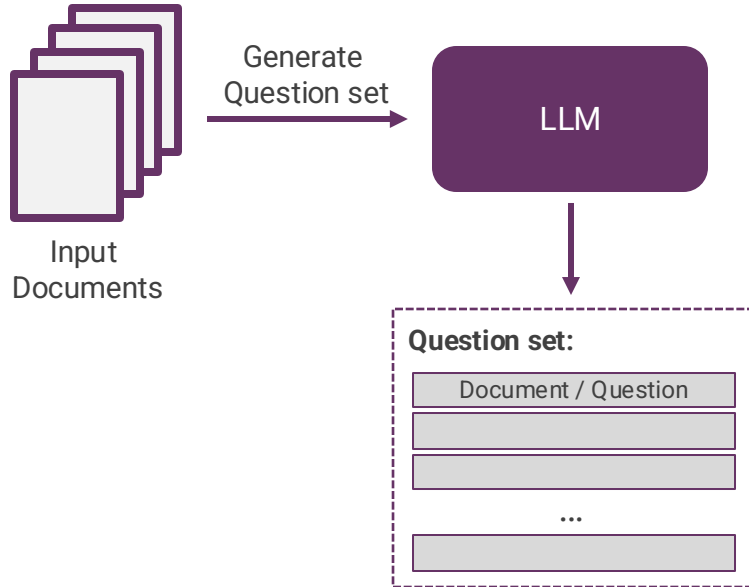
- 1 Many frameworks follow the LLM-as-a Judge approach, trying to limit manual work
- 2 Many frameworks cannot be used with local LLMs, but rely on OpenAI inference quality
- 3 LLM metrics are often subjective, making it hard to quantify and compare results
- 4 Frameworks aren't that mature yet, causing many obscure errors and high failure rates

**Automated validation of RAG pipelines
with local LLMs is just not there yet.**

Question Generation

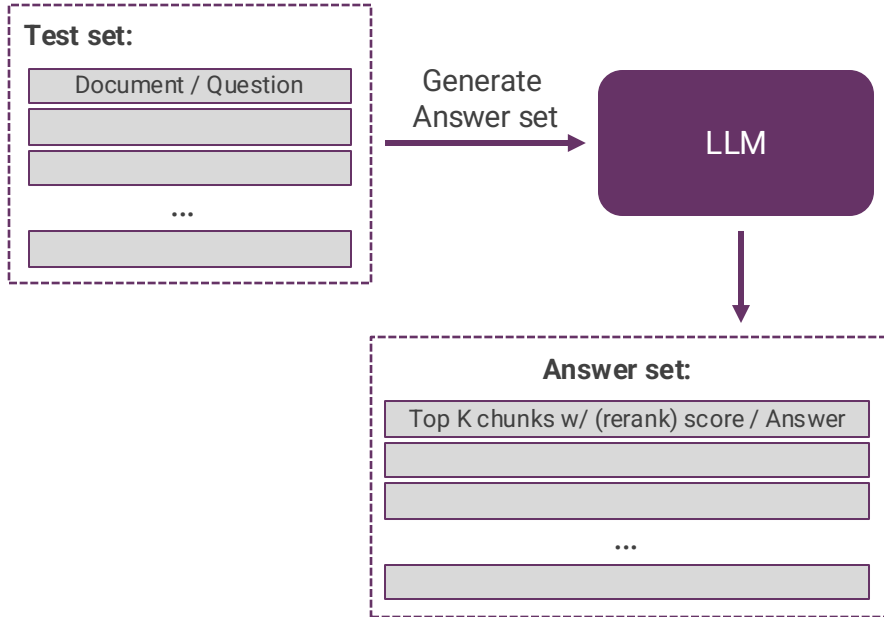


Question Validation

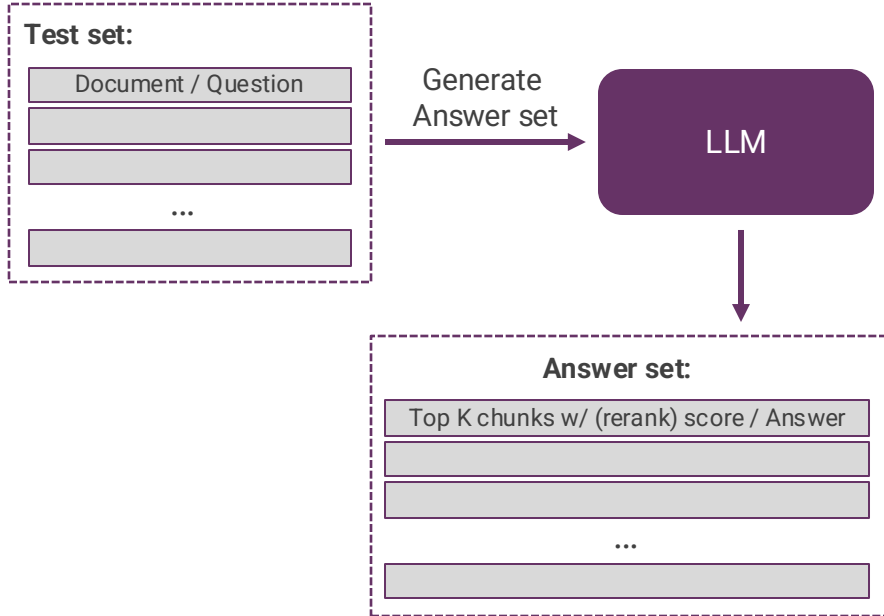


Validation Criteria	Score (n=100)
1 Is the question specific?	83%
2 Is the question relevant?	84%
3 Can the question be answered with the input document?	77%
Useful questions:	74%

Answer Generation



Answer Validation



Validation Criteria	Score (n=100)
1 Are relevant chunks retrieved?	72%
2 Does the answer contain information from chunks?	78%
3 Is the answer correct?	61%
4 Is the answer specific?	81%
5 Is the answer complete?	57%
6 Does the answer only contain relevant information?	61%
Useful answers:	
52%	

**‘Numbers going up’ is great, but it is
better to understand what is going on
under the hood**

So, what is going on then?

Answer is misleading, containing totally incorrect (hallucinated) information

Answer is partly correct, but is missing essential information

Answer is partly correct, but is also cluttered with additional, irrelevant information



Validation Criteria	Score (n=100)
1 Are relevant chunks retrieved?	72%
2 Does the answer contain information from chunks?	78%
3 Is the answer correct?	61%
4 Is the answer specific?	81%
5 Is the answer complete?	57%
6 Does the answer only contain relevant information?	61%
Useful answers:	52%

Test Point Example 1

Question

What is the role of the graduation coordinators in the graduation process at Fontys ICT?

Answer

The role of the graduation coordinators in the graduation process at Fontys ICT involves ensuring that students have sufficient opportunities to transfer to Master's study programmes that match the substantive choices they have made within the study programme. They maintain contacts with partner universities, inform (prospective) students about the possibilities and conditions for transfer, provide input on academic preparation, pre-master as minor programmes and subsequent masters for 'study navigator', advise on substantive connection with adjoining masters of basic profiles and specialisations, and can be combined by 1 person with the semester coordinator-ship Academic Preparation.

Validation Criteria

1	Are relevant chunks retrieved?	X
2	Does the answer contain information from chunks?	X
3	Is the answer correct?	X
4	Is the answer specific?	✓
5	Is the answer complete?	X
6	Does the answer only contain relevant information?	X

Test Point Example 1

Question

What is the role of the graduation coordinators in the graduation process at Fontys ICT?

Analysis

- No relevant chunks are retrieved
- Answer is mostly based on one irrelevant retrieved chunk

Test points

- B: Document chunking & embedding
- C: Chunk retrieval

Possible solution

- Change chunking strategy
- Extend search strategy

Test Point Example 2

Question

What can I do in an emergency situation within Fontys?

Answer

In an emergency situation within Fontys, you can call 119 (08850-77119) immediately. This will connect you to the emergency room, which can alert the appropriate BHV officers at the scene and support you with the national emergency services. You should also watch the film on how you can help keep our locations safe. For general questions or additional information, please contact the BHV team leader at your location.

Validation Criteria

1	Are relevant chunks retrieved?	✓
2	Does the answer contain information from chunks?	✓
3	Is the answer correct?	✓
4	Is the answer specific?	✓
5	Is the answer complete?	✗
6	Does the answer only contain relevant information?	✓

Test Point Example 2

Question

What can I do in an emergency situation within Fontys?

Analysis

- The most important piece of information is missing in the answer ('Call 112 immediately')
- Input document doesn't contain a clear, structured emergency protocol

Test points

- A: Input document quality
- F: Answer quality

Possible solution

- Rewrite input document so that it consists of a better structure

Test Point Example 3

Question

What is the purpose of adding notes in StudyCoach according to the guidelines provided?

Answer

The purpose of adding notes in StudyCoach is to provide an overview of relevant learning analytics that give lecturers insight into a student's learning process. By making these notes, the tool visually displays information from Canvas across subjects, which helps coaches and students identify 'At Risk' students and make early adjustments. Additionally, semester coaches can take notes on a student's behavior or things that may influence their learning process, such as personal circumstances. These notes are important for guidance in the study and can be accessed by both the coach and the student during the entire program.

Validation Criteria

1	Are relevant chunks retrieved?	✓
2	Does the answer contain information from chunks?	✓
3	Is the answer correct?	✗
4	Is the answer specific?	✓
5	Is the answer complete?	✓
6	Does the answer only contain relevant information?	✗

Test Point Example 3

Question

What is the purpose of adding notes in StudyCoach according to the guidelines provided?

Analysis

- The subject is covered in multiple documents
- Multiple similar terms with slightly different meaning are being used

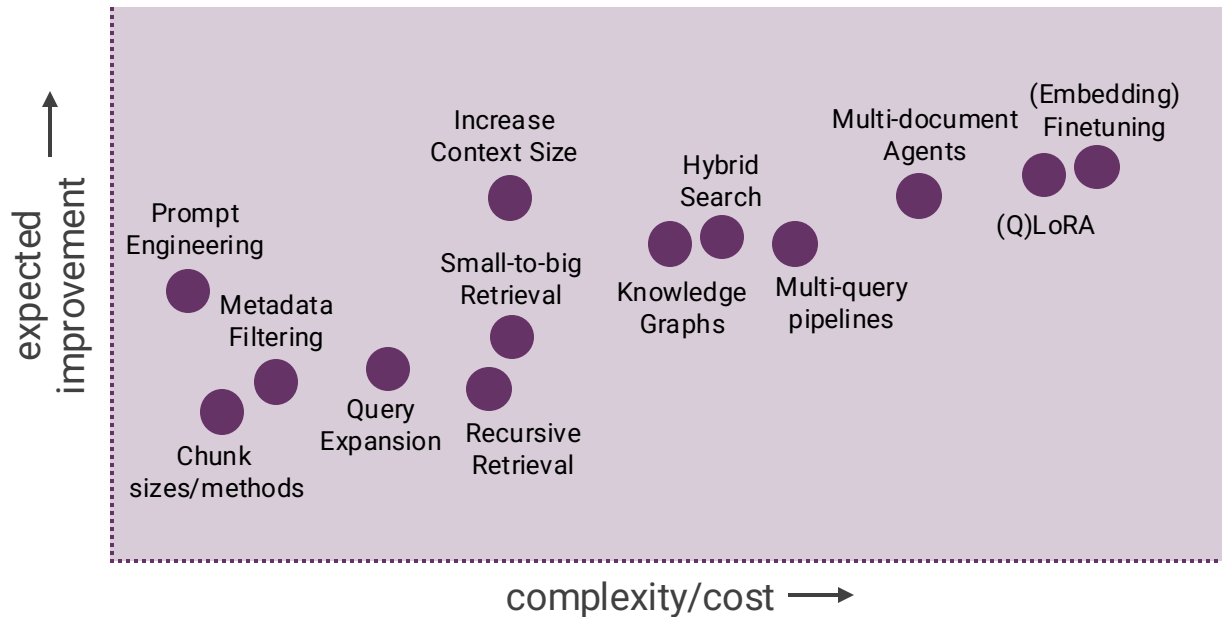
Test points

- A: Input document quality
- C: Chunk retrieval

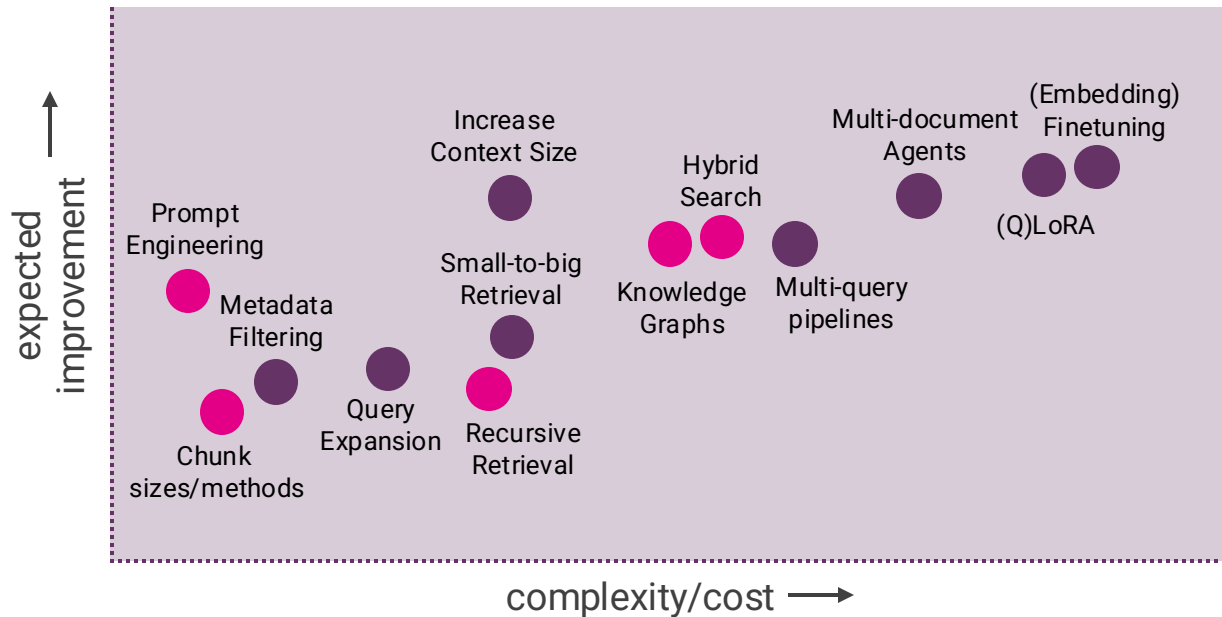
Possible solution

- Rewrite input documents to be more coherent
- Recursive retrieval
- Use semantic structures (e.g. knowledge graphs)

RAG Improvements



RAG Improvements

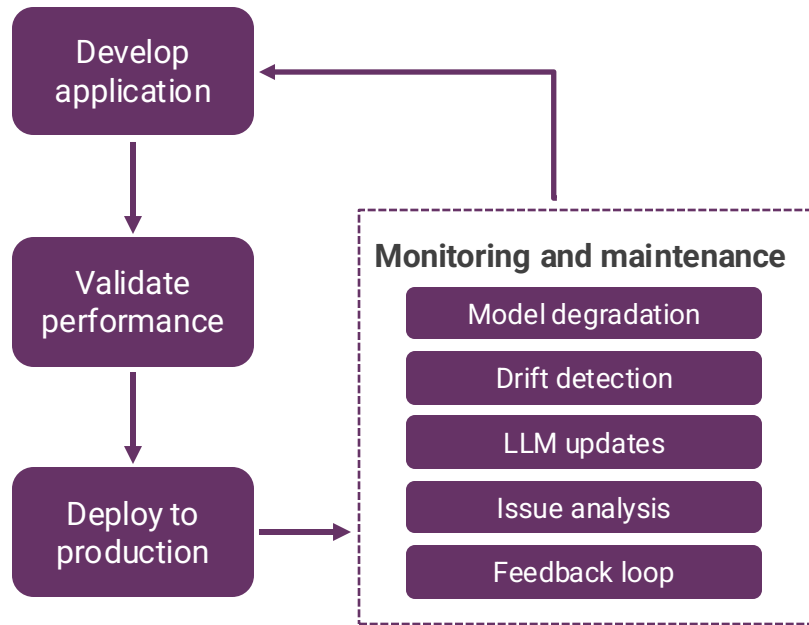


**So, you have achieved good
performance in production? Awesome!**

**So, you have achieved good
performance in production? Awesome!**

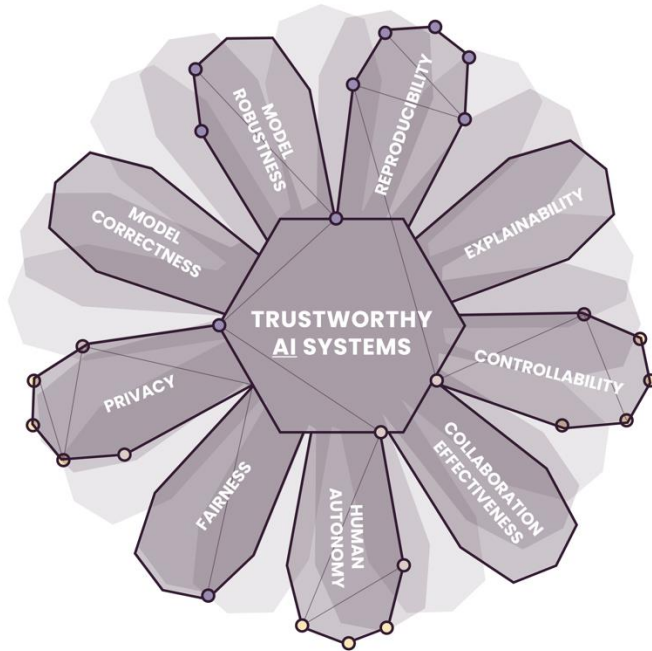
**Don't forget to keep monitoring performance
over time and to keep the human in the loop.**

LLMOps in Practice



- 1 Model Drift**
Gradual decline in model performance due to changes in underlying data patterns over time
- 2 Monitoring & Tracing**
Detect performance degradation and identify root causes of issues in real-time
- 3 Maintenance**
In case of LLM base model updates, detect regressions on prompts via a regression test set
- 4 Feedback Loop**
Enables continuous learning and improvement by using real-world data to refine the model

Trustworthy AI Systems



Model Correctness

Model Robustness

Reproducibility

Explainability

Controllability

Collaboration Effectiveness

Human Autonomy

Fairness

Privacy

Thank you! Questions?

Leon Schrijvers – l.schrijvers@fontys.nl / Petra Heck – p.heck@fontys.nl
Fontys Knowledge Center Applied AI For Society / Fontys ICT
DEMAND Research Project – <https://demand.nl>



FONTYS KNOWLEDGE CENTER

APPLIED AI FOR SOCIETY

EXPLORE THE FUTURE OF
TRUSTWORTHY AI