Automated Evaluation of RAG Pipelines

DEMAND Community Event

September 2025

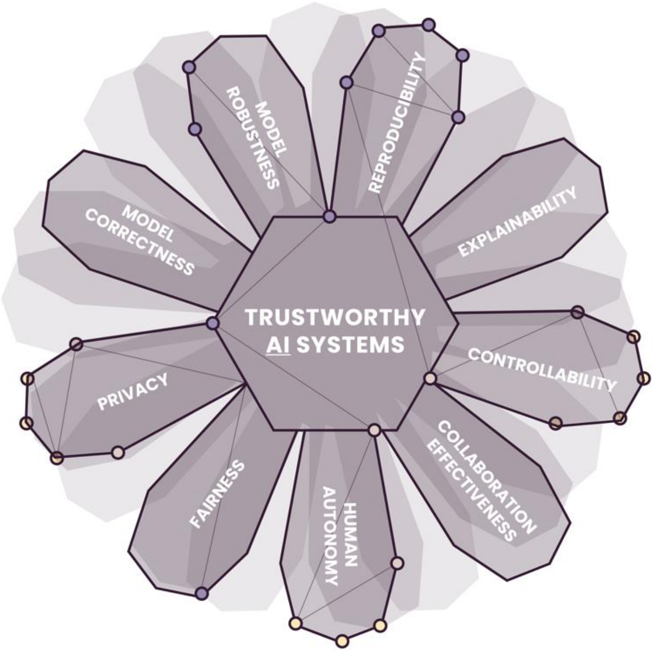
Leon Schrijvers – Fontys ICT

l.schrijvers@fontys.nl

Petra Heck – Fontys ICT

p.heck@fontys.nl

Trustworthy AI Systems

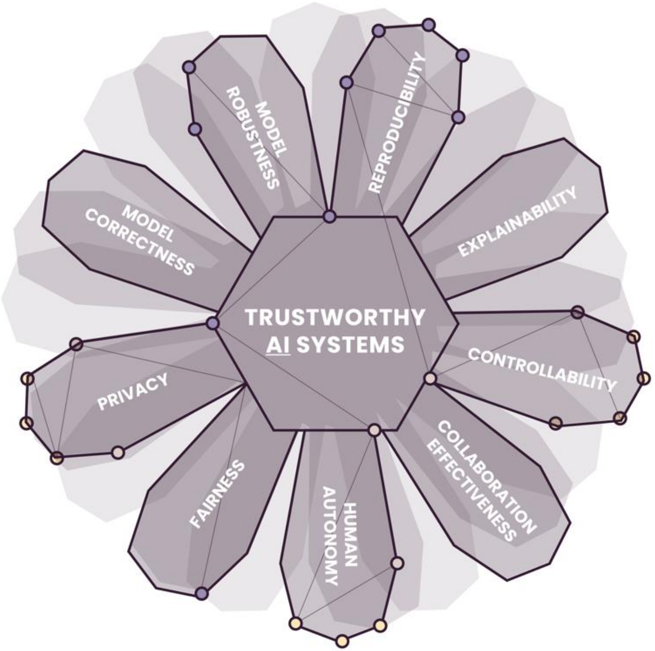


- Model Correctness
- Model Robustness
- Reproducibility
- Explainability
- Controllability
- Collaboration Effectiveness
- Human Autonomy
- Fairness
- Privacy

Fontys > FOR SOCIETY

https://fontysblogt.nl/a-quality-model-for-trustworthy-ai-systems/

Trustworthy LLM Systems



- **Model Correctness**
- Model Robustness
- Reproducibility
- Explainability
- Controllability
- Collaboration Effectiveness
- Human Autonomy
- Fairness
- Privacy

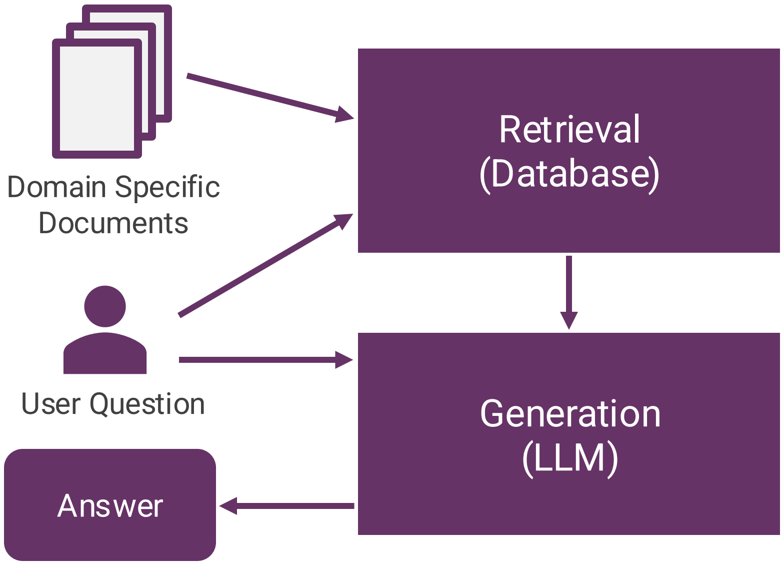
https://fontysblogt.nl/llmops-engineering-trustworthy-llm-systems/

Case Study – Retrieval Augmented Generation

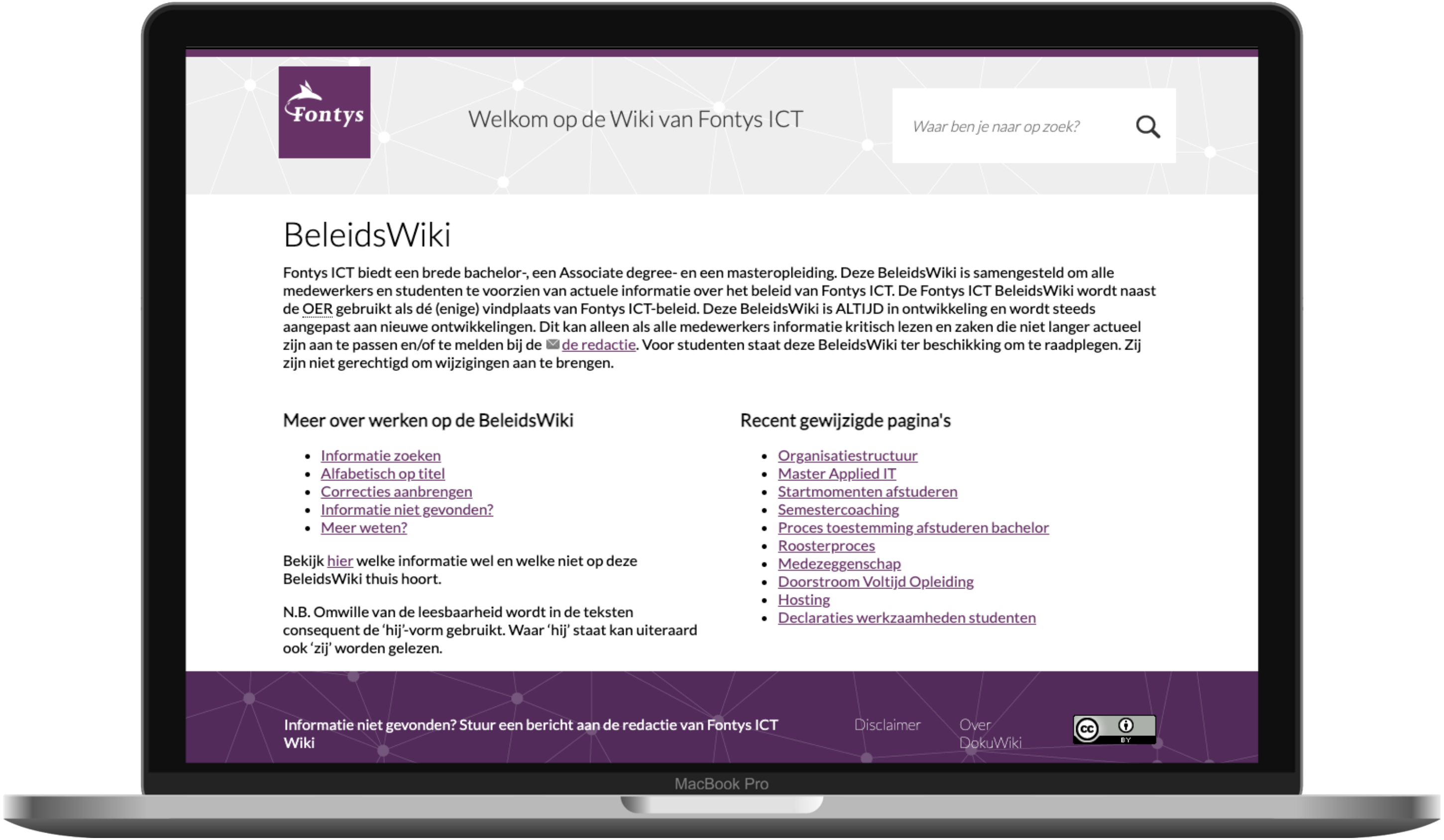
How to design a chatbot application that enables company employees to ask questions about company specific documents?

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



Welkom op de Wiki van Fontys ICT

BeleidsWiki

Fontys ICT biedt een brede bachelor-, een Associate degree- en een masteropleiding. Deze BeleidsWiki is samengesteld om alle medewerkers en studenten te voorzien van actuele informatie over het beleid van Fontys ICT. De Fontys ICT BeleidsWiki wordt naast de OER gebruikt als dé (enige) vindplaats van Fontys ICT-beleid. Deze BeleidsWiki is ALTIJD in ontwikkeling en wordt steeds aangepast aan nieuwe ontwikkelingen. Dit kan alleen als alle medewerkers informatie kritisch lezen en zaken die niet langer actueel zijn aan te passen en/of te melden bij de ✉ de redactie. Voor studenten staat deze BeleidsWiki ter beschikking om te raadplegen. Zij zijn niet gerechtigd om wijzigingen aan te brengen.

Meer over werken op de BeleidsWiki

- Informatie zoeken
- Alfabetisch op titel
- Correcties aanbrengen
- Informatie niet gevonden?
- Meer weten?

Bekijk hier welke informatie wel en welke niet op deze BeleidsWiki thuis hoort.

N.B. Omwille van de leesbaarheid wordt in de teksten consequent de 'hij'-vorm gebruikt. Waar 'hij' staat kan uiteraard ook 'zij' worden gelezen.

Recent gewijzigde pagina's

- Organisatiestructuur
- Master Applied IT
- Startmomenten afstuderen
- Semestercoaching
- Proces toestemming afstuderen bachelor
- Roosterproces
- Medezeggenschap
- Doorstroom Voltijd Opleiding
- Hosting
- Declaraties werkzaamheden studenten

Informatie niet gevonden? Stuur een bericht aan de redactie van Fontys ICT Wiki

Why Evaluate Your LLM Application?

LLMs are non-deterministic and unpredictable, which comes with risks:

- Hallucinations
- Inconsistent outputs
- Harmful outputs
- Jailbreak attacks
- Data and PII leaks

Why Evaluate Your LLM Application?

LLMs are non-deterministic and unpredictable, which comes with risks:

- Hallucinations
- Inconsistent outputs
- Harmful outputs
- Jailbreak attacks
- Data and PII leaks

Evaluation Approaches

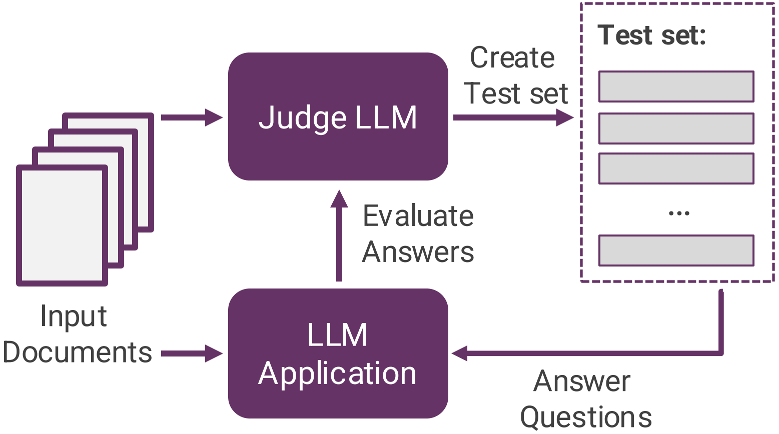
A **Vibe evaluation:** Manually sample and test output

B **Manual evaluation:** Structured evaluation of a test set

C **Custom code/tests:** Automate validation via custom code

D **Evaluation libraries:** Integrate existing validation strategies

Evaluation Libraries



A **Black-box evaluation**
Most LLM-as-a-Judge libraries ignore architecture internals

B **Subjective & context dependent metrics**
Hard to quantify and compare results

C **Immature tooling**
Evaluation libraries often fail or produce obscure errors

Our Approach

Domain-specific test set

Context-aware evaluation metrics

White-box evaluation method

Our Approach

Domain-specific test set

Context-aware evaluation metrics

White-box evaluation method

Characteristics of a Good Test Set

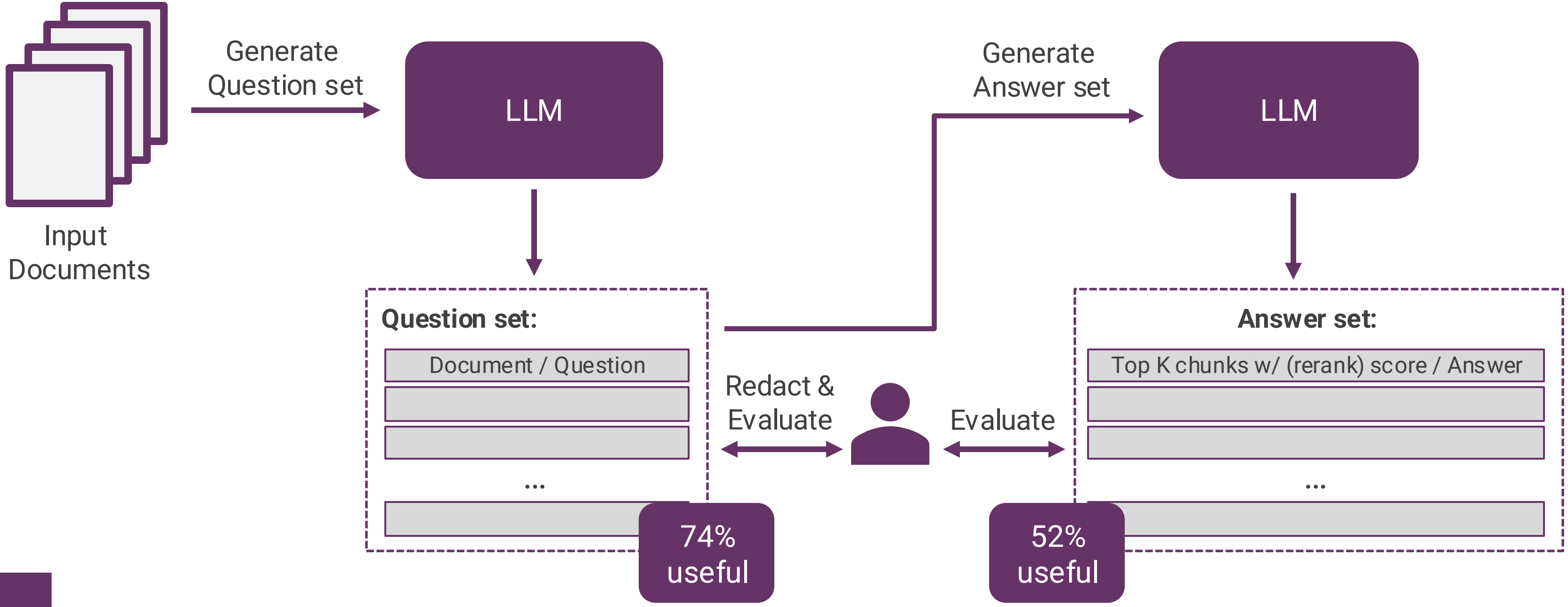
A **Use-case driven:** Design tests that reflect your real application

B **User-centered:** Cover the types of questions users will ask in production (factual, summarizing, clarifying, etc.)

C **Relevant metrics:** Measure what matters for your domain

- Include domain-specific examples & edge cases
- Define what “good” means with reference answers or evaluation criteria

Test Set Generation



Input Documents

Generate Question set

LLM

Question set:

Document / Question
…

74% useful

Redact & Evaluate

Evaluate

Generate Answer set

LLM

Answer set:

Top K chunks w/ (rerank) score / Answer
…

52% useful

Example Questions

- A **Factual:** “*What is the contact information of the Student Desk?*”
- B **Summarizing:** “*What are the additional criteria for internship or graduation within an own company or Fontys ICT research group?*”
- C **Clarifying:** “*How does the process of temporarily deregistering from a study programme work?*”

Test Set Tips & Learnings

A **Involve experts:** Collaborate with subject matter specialists

B **Leverage internal resources:** Draw from company guidelines & onboarding documents

C **Automate & redact:** Generating questions can be a solid starting point, but always validate outputs

D **Ground in reality:** Include examples from production logs

Our Approach

Domain-specific test set

Context-aware evaluation metrics

White-box evaluation method

Context-aware Evaluation Metrics

Structured text validation: Pattern matching, syntax validation

Statistical metrics: Custom calculations, F1 score, accuracy, ROUGE

ML based metrics: Semantic similarity, BERT score, sentiment score

LLM-as-a-Judge metrics: Custom evaluation criteria

Context-aware Evaluation Metrics

Structured text validation: Pattern matching, syntax validation

Statistical metrics: Custom calculations, F1 score, accuracy, ROUGE

ML based metrics: Semantic similarity, BERT score, sentiment score

LLM-as-a-Judge metrics: Custom evaluation criteria

Example Metrics

A **Retrieval quality:** How many retrieved chunks are relevant to the question?

B **Answer specificity:** Does the answer include concrete information?

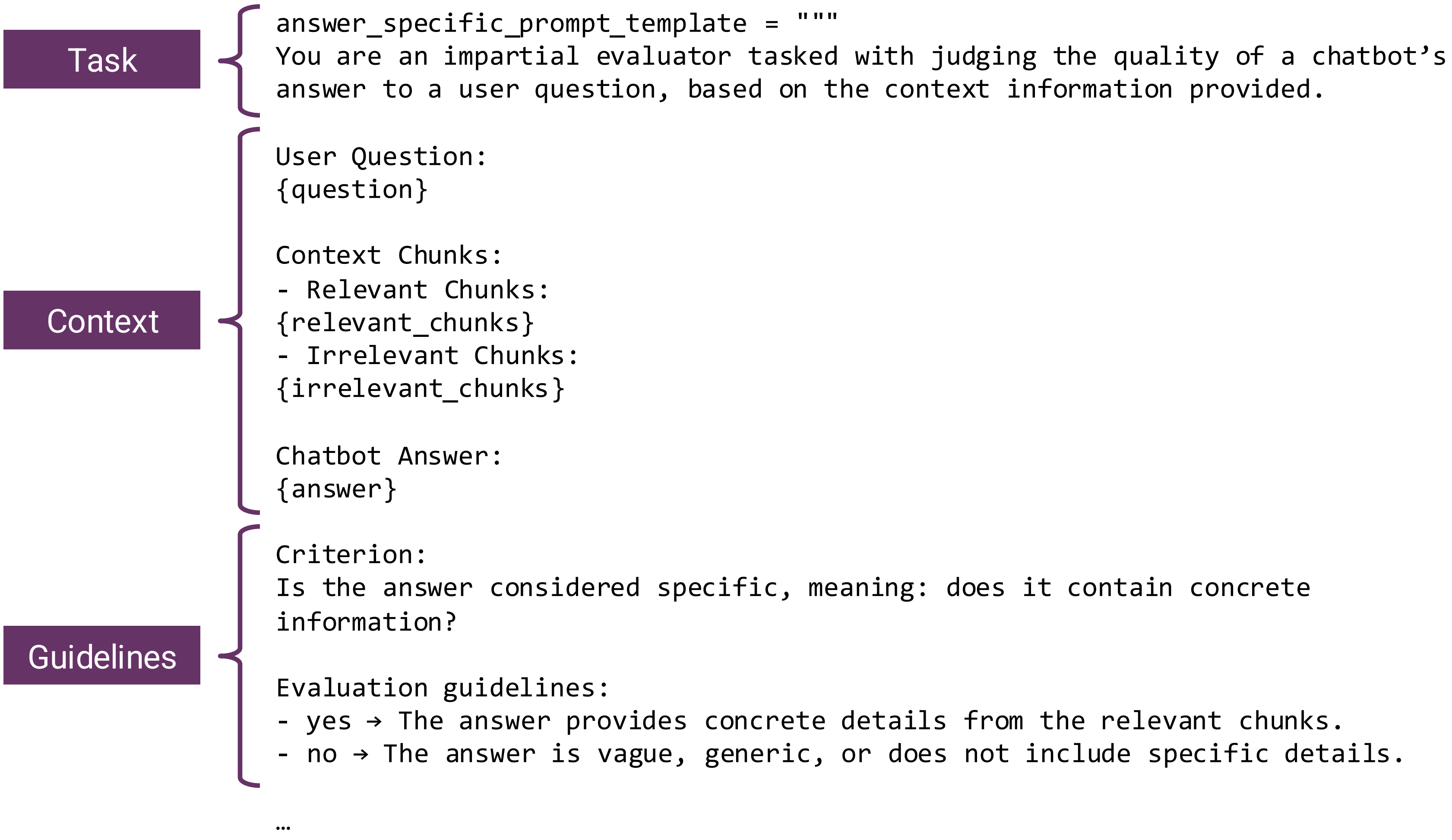
C **Answer correctness:** Does the answer align with the ground truth?

LLM-as-a-Judge Evaluation Metrics

A **Domain-specific is better than generic:** measure what matters for your use case

B **Flexible granularity:** Evaluation questions can be general (for all test items) or tailored (per question)

C **Context-dependent or independent:** Metrics may use context information (e.g., question, ground truth, retrieved chunks)



Task

```
answer_specific_prompt_template = """
You are an impartial evaluator tasked with judging the quality of a chatbot’s
answer to a user question, based on the context information provided.
```

Context

```
User Question:
{question}

Context Chunks:
- Relevant Chunks:
{relevant_chunks}
- Irrelevant Chunks:
{irrelevant_chunks}

Chatbot Answer:
{answer}
```

Guidelines

```
Criterion:
Is the answer considered specific, meaning: does it contain concrete
information?

Evaluation guidelines:
- yes → The answer provides concrete details from the relevant chunks.
- no → The answer is vague, generic, or does not include specific details.

…
```

```
…
```

Examples

```
Examples:
- Positive example:
  - Question: What are the side effects of aspirin?
  - Relevant Chunks: “Aspirin may cause nausea, stomach pain, and heartburn.”
  - Irrelevant Chunks: “Vitamin C is important for the immune system.”
  - Answer: “Aspirin can cause nausea, stomach pain, and heartburn.”
  - Expected Result: yes

- Negative example:
  - Question: What are the side effects of aspirin?
  - Relevant Chunks: “Aspirin may cause nausea, stomach pain, and heartburn.”
  - Irrelevant Chunks: “Ibuprofen is another pain reliever.”
  - Answer: “Aspirin can cause problems.”
  - Expected Result: no
```

Output

```
Now evaluate the given case.

Return your evaluation in the following JSON format:

{
  "result": "yes" | "no",
  "reason": "Short explanation of reasoning"
}
"""
```

LLM-as-a-Judge Metric Tips & Learnings

A **Binary scoring:** Keep evaluations simple (Yes/No)

B **Require reasoning:** Explain why a score was given

C **Provide examples:** Show clear “yes” and “no” cases in prompts

D **Structured output:** Return results in JSON for easy processing

E **Set low model temperature:** Reduce randomness in judgements

LLM-as-Judge Metric Development

Input document:
"France, officially the French Republic, is a country primarily located in Western Europe. Its overseas regions and territories include French Guiana in South America, Saint Pierre and Miquelon in the North Atlantic, the French West Indies, and many islands in Oceania and the Indian Ocean, giving it the largest discontiguous exclusive economic zone in the world."

Answer:
"The estimated total population of France is over 68.8 million"

Metric:
Is the answer correct?

LLM-as-Judge Metric Development

Guidelines

```
Criterion:
Is the answer correct, meaning: does the information from the answer align with
the relevant chunks from the provided background information?

Evaluation guidelines:
- yes → The answer provides factual statements that align with the provided
background information.
- no → The answer contains information that cannot be verified by using the
  provided background information.
```

Evaluation:

Answer: Yes

Reason: The statement is a factual claim that can be verified by external sources.

LLM-as-Judge Metric Development

Guidelines

```
Criterion:
Is the answer correct, meaning: does the information from the answer align with
the relevant chunks from the provided background information?

Notes:
- Do not refer to external data sources.

Evaluation guidelines:
- yes → The answer provides factual statements that align with the provided
background information.
- no → The answer contains information that cannot be verified by using the
  provided background information.
```

Evaluation:

Answer: Yes

Reason: Although the exact population figure is not mentioned, the statement "over 68.8 million" implies that it is an estimate, which can be verified by making an educated guess.

LLM-as-Judge Metric Development

Guidelines

Criterion:
Is the answer correct, meaning: does the information from the answer align with the relevant chunks from the provided background information?

Notes:
- Do not refer to external data sources.
- **Do not make educated guesses.**

Evaluation guidelines:
- yes → The answer provides factual statements that align with the provided background information.
- no → The answer contains information that cannot be verified by using the provided background information.

Evaluation:
Answer: Yes
Reason: The answer provides specific numerical data that does not conflict with the provided background information.

LLM-as-Judge Metric Development

Guidelines

Criterion:
Is the answer correct, meaning: does the information from the answer align with the relevant chunks from the provided background information?

Notes:
- Do not refer to external data sources.
- Do not make educated guesses.
- **All information needs to be verified, even if it doesn't conflict with the provided background information.**

Evaluation guidelines:
- yes → The answer provides factual statements that align with the provided background information.
- no → The answer contains information that cannot be verified by using the provided background information.

Evaluation:

Answer: No

Reason: The background information does not contain any information about the estimated total population of France.

LLM-as-Judge Metric Tips & Learnings

A **Detailed, example-backed prompts:** Give clear guidance and representative examples, but avoid overfitting to one phrasing

B **Inspect reasoning outputs:** Analyse explanations to discover edge cases, ambiguous items, and calibration issues

C **Repeat evaluations:** Run multiple independent judgments per item to reveal instability and flakiness

D **Consensus aggregation:** Use majority voting (in-prompt voting, or ensemble runs) to reduce single-run noise

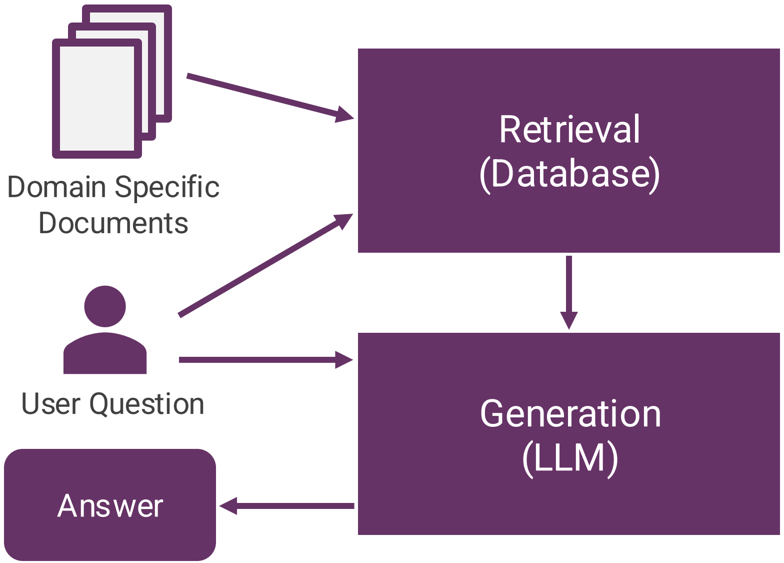
Our Approach

Domain-specific test set

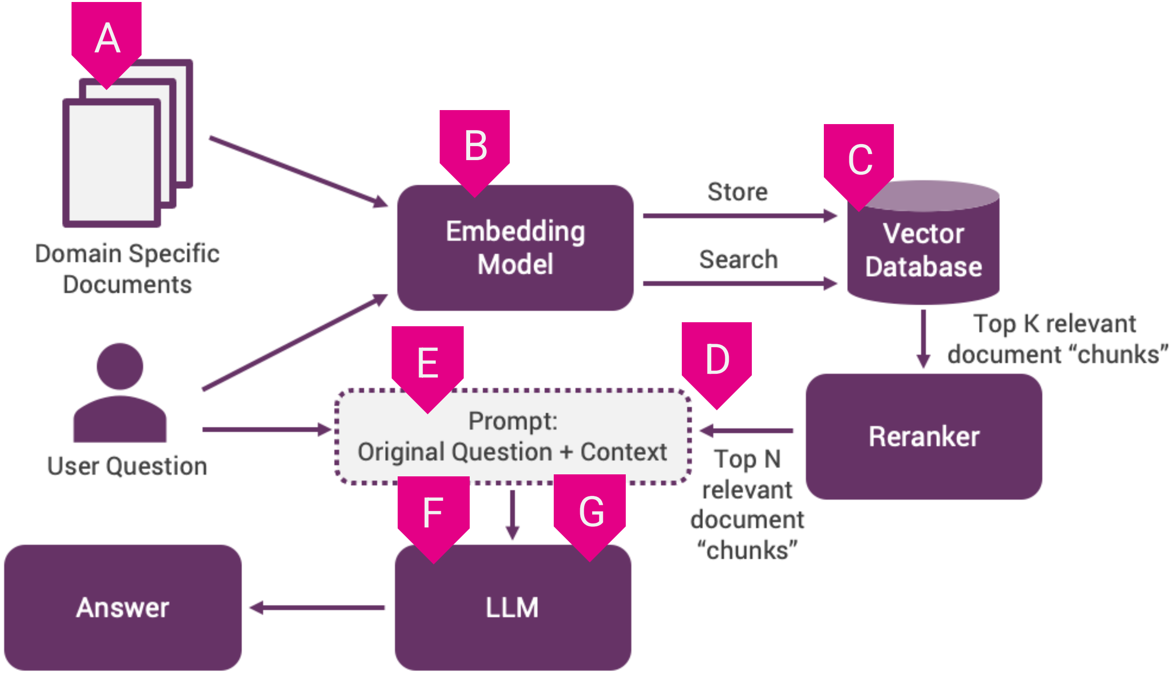
Context-aware evaluation metrics

White-box evaluation method

Architecture of RAG Applications



White-box Evaluation of RAG Applications


A
Domain Specific Documents
B
Embedding Model
Store
Search
C
Vector Database
Top K relevant document "chunks"
User Question
E
D
Prompt:
Original Question + Context
Reranker
Top N relevant document "chunks"
F
G
Answer
LLM

White-box 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

Test Point / Metric Mapping Examples

Test Point	Metric	Metric Type	Evaluation
C Chunk Retrieval	Chunks retrieved	Custom calculation	Calculate number of relevant retrieved chunks
F Answer Quality	Answer specific	LLM-as-a-Judge	Evaluation prompt
F Answer Quality	Answer correct	Context-aware LLM-as-a-Judge	Context-aware evaluation prompt

Example Evaluation

Question:
What are the additional criteria for internship or graduation within an own company or Fontys ICT research group?

Answer:
The additional criteria for internship or graduation within an own company are as follows: 1) The stability and continuity […]; 2) The quality of […]; 3) The student has […]; 4) If an internship […].

The additional criteria for internships or graduations at a Fontys ICT research group are as follows: 1) The role of the company supervisor is […]; 2) The final assignment is […]; 3) The final assignment is […].

Metric	Score	Manual Score
1 Chunks retrieved	Yes	Yes
2 Answer specific	Yes	Yes
3 Answer correct	Yes	Yes

Example Evaluation

Chunks Retrieved

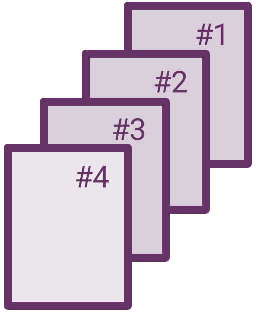
Ground Truth Document: Criteria for internship and graduation

Chunk	Source	Retrieval Score	Rerank Score	Ground Truth
Chunk #1	Criteria for internship and graduation	0.31	0.78	Yes
Chunk #2	Criteria for internship and graduation	0.30	0.74	Yes
Chunk #3	Criteria for internship and graduation	0.28	0.74	Yes
Chunk #4	Internship at Fontys ICT	0.42	0.73	No

Example Evaluation

Chunks Retrieved

Ground Truth Document: Criteria for internship and graduation



Chunk	Source	Retrieval Score	Rerank Score	Ground Truth
Chunk #1	Criteria for internship and graduation	0.31	0.78	Yes
Chunk #2	Criteria for internship and graduation	0.30	0.74	Yes
Chunk #3	Criteria for internship and graduation	0.28	0.74	Yes
Chunk #4	Internship at Fontys ICT	0.42	0.73	No

Result: Yes. Reason: Chunk Relevancy Score: 0.75, above threshold

Example Evaluation

Answer Specific

Evaluation 1: Answer: Yes. Reason: Specific numbers are not present, but proper names (Fontys ICT research group) and verifiable facts (the criteria for internships or graduations) are clearly stated.

Evaluation 2: Answer: No. Reason: While the answer provides detailed information about specific procedures and criteria, it lacks specific numbers and direct quotations. The text is descriptive rather than numerical or fact-based.

Evaluation 3: Answer: Yes. Reason: Although no direct quotations are present, proper names (Fontys ICT research group) and verifiable facts (the criteria for internships or graduations) are clearly stated, meeting the "Yes" criteria.

Result: Yes. Reason: Two evaluations support the answer being "Yes", and only one evaluation suggests it is "No". The presence of proper names and verifiable facts in multiple instances outweighs the lack of specific numbers and direct quotations.

Example Evaluation

Answer Specific

Evaluation 1: Answer: Yes. Reason: Specific numbers are not present, but proper names (Fontys ICT research group) and verifiable facts (the criteria for internships or graduations) are clearly stated.

Evaluation 2: Answer: No. Reason: While the answer provides detailed information about specific procedures and criteria, it lacks specific numbers and direct quotations. The text is descriptive rather than numerical or fact-based.

Evaluation 3: Answer: Yes. Reason: Although no direct quotations are present, proper names (Fontys ICT research group) and verifiable facts (the criteria for internships or graduations) are clearly stated, meeting the "Yes" criteria.

Result: Yes. Reason: Two evaluations support the answer being "Yes", and only one evaluation suggests it is "No". The presence of proper names and verifiable facts in multiple instances outweighs the lack of specific numbers and direct quotations.

Example Evaluation

Answer Correct

Evaluation 1: Answer: Yes. Reason: The answer aligns with the provided background information, specifically section "Criteria internship or graduation within own company" which lists similar conditions.

Evaluation 2: Answer: No. Reason: The additional criteria mentioned in the answer do not match exactly with the provided background information, especially regarding the role of the Centre for Entrepreneurship and the status of "student-entrepreneur".

Evaluation 3: Answer: Yes. Reason: All points mentioned in the answer are present in the provided background information under section "Criteria internship or graduation within own company", making it a correct alignment.

Result: Yes. Reason: The winning answer is from Evaluation 1, which correctly identified that the answer aligns with the provided background information.

Example Evaluation

Answer Correct

Evaluation 1: Answer: Yes. Reason: The answer aligns with the provided background information, specifically section "Criteria internship or graduation within own company" which lists similar conditions.

Evaluation 2: Answer: No. Reason: The additional criteria mentioned in the answer do not match exactly with the provided background information, especially regarding the role of the Centre for Entrepreneurship and the status of "student-entrepreneur".

Evaluation 3: Answer: Yes. Reason: All points mentioned in the answer are present in the provided background information under section "Criteria internship or graduation within own company", making it a correct alignment.

Result: Yes. Reason: The winning answer is from Evaluation 1, which correctly identified that the answer aligns with the provided background information.

White-box Evaluation Tips & Learnings

A **Cross-check:** Compare automated evaluation against manual reviews

B **Refine & improve:** Adjust metrics and prompts based on findings

C **Align reasoning:** Ensure the evaluator’s explanations match human judgement (to reduce subjectivity)

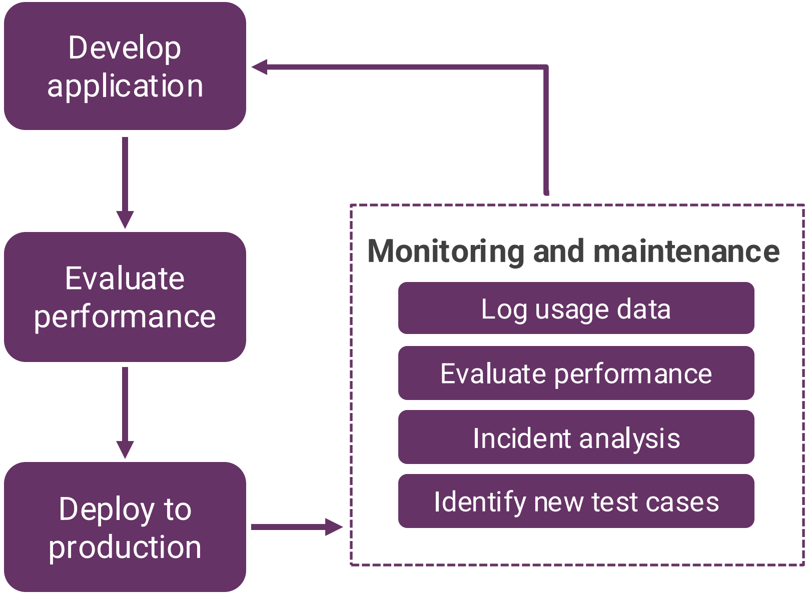
Quick Recap

Domain-specific test set

Context-aware evaluation metrics

White-box evaluation method

Evaluation Workflow



A **Pre-production**
Define what's good, run experiments & evaluate output with test set

B **In-production**
Log usage data & respond to real time alerts

C **Post-production**
Analyse logs, debug & fix what went wrong, update the test set

Fontys > FOR SOCIETY

Conclusion

Domain-specific test set

Context-aware evaluation metrics

White-box evaluation method

Workflow: Continuously monitor output quality over time and analyze results to refine & extend the test set

LEON SCHRIJVERS

Role : lecturer/researcher ict

E-mail : l.schrijvers@fontys.nl

Phone : +31 6 28 37 79 11