EngineTalk:

Supporting maintenance of diesel engines aboard navy ships through a RAG enabled LLM system

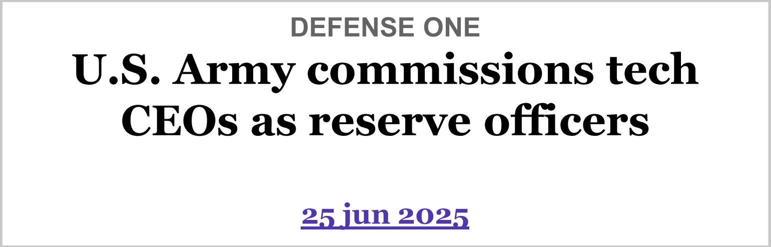
Youri Linden
Bart-Peter Smit
Jonathan Laurens Maas

Contemporary challenges in naval operations:

- Maintaining operational capability despite:
  - Reduced crew sizes
  - Increasing complexity and number of on-board systems
  - Challenges in staffing expertise
  - Anticipated expertise turnover rates

Deciding with Words

How LLMs are becoming Defence's new strategic partner

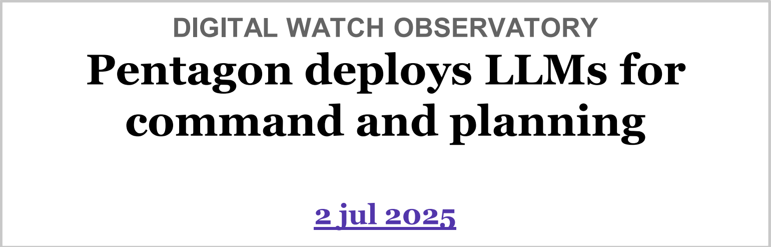


DEFENSE ONE

U.S. Army commissions tech CEOs as reserve officers

25 jun 2025

- Goal: Innovate faster with top-tier civil expertise.
- Strategic: AI as a core capability, not a gadget.
- Ethics: Explicitly address conflicts of interest and transparency.

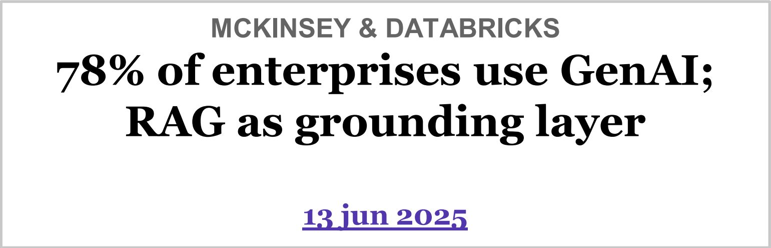


DIGITAL WATCH OBSERVATORY

Pentagon deploys LLMs for command and planning

2 jul 2025

- Operational: LLMs in production for command and planning.
- Cases: Document-based planning, analysis, and decision support.
- Governance: Human-in-the-loop, logging, access control.



MCKINSEY & DATABRICKS

78% of enterprises use GenAI; RAG as grounding layer

13 jun 2025

- Adoption (78%): Broad use ≠ guaranteed impact.
- Colleague: From tool to data-driven teammate in processes.
- Value: Still limited — integration and ownership are crucial.

Research Question

Our research poses the following:

“**To what extent can an LLM&RAG solution support naval engine service engineers in their maintenance tasks?”**

From which we study the following:

- To what extend can information from manuals best be processed to improve information delivery?
- How can historical maintenance data from the SAP system be integrated?
- How can our solution be evaluated?

Expected outcome

Prevent unnecessary calls for land support, and having to fly in senior maintenance crews

Thereby reducing costs

Doing more, with a less skilled crew, in fewer time.

Thereby increasing uptime

Create a better data fundament by aiding maintenance crews in documenting their work.

Thereby increasing quality

User groups

Maintenance crew (onboard)

- Small maintenance (regular, planned, reactive)

- Relatively unexperienced

Operational support (Den Helder)

- Shore support for multiple ships

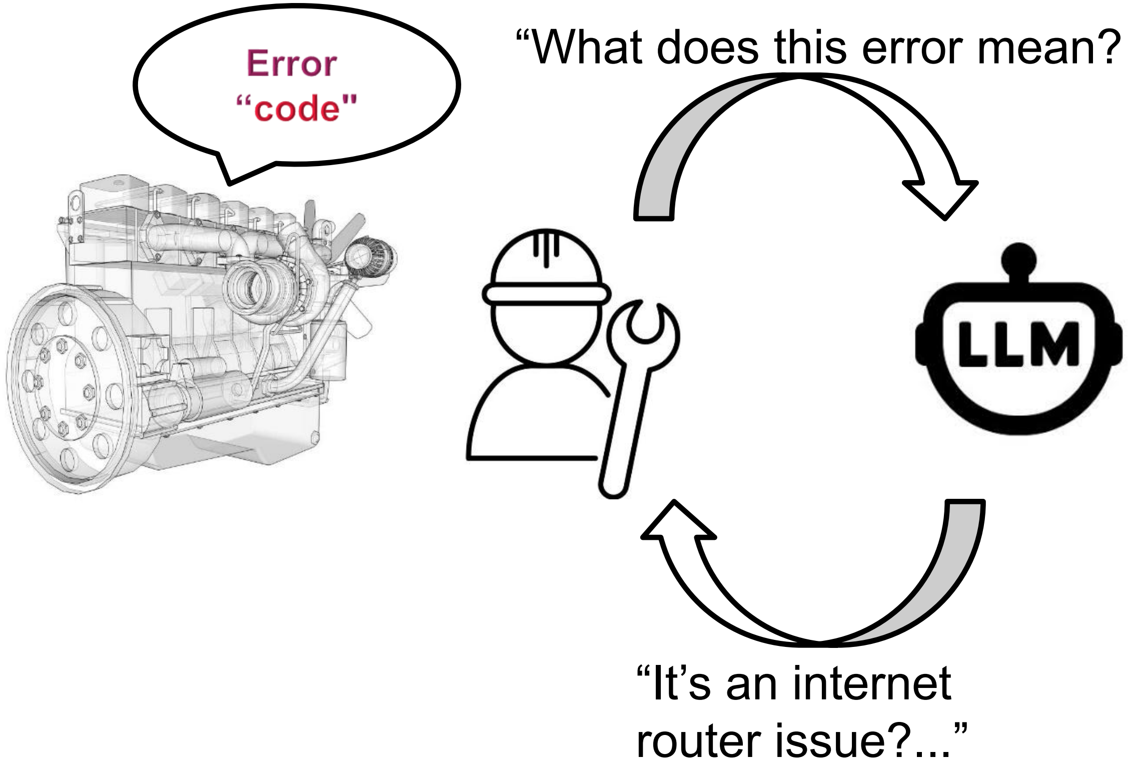
- Very experienced

Installation manager (Den Helder)

- Plans and coordinates large maintenance with specialised external crews and suppliers

- extremly experienced

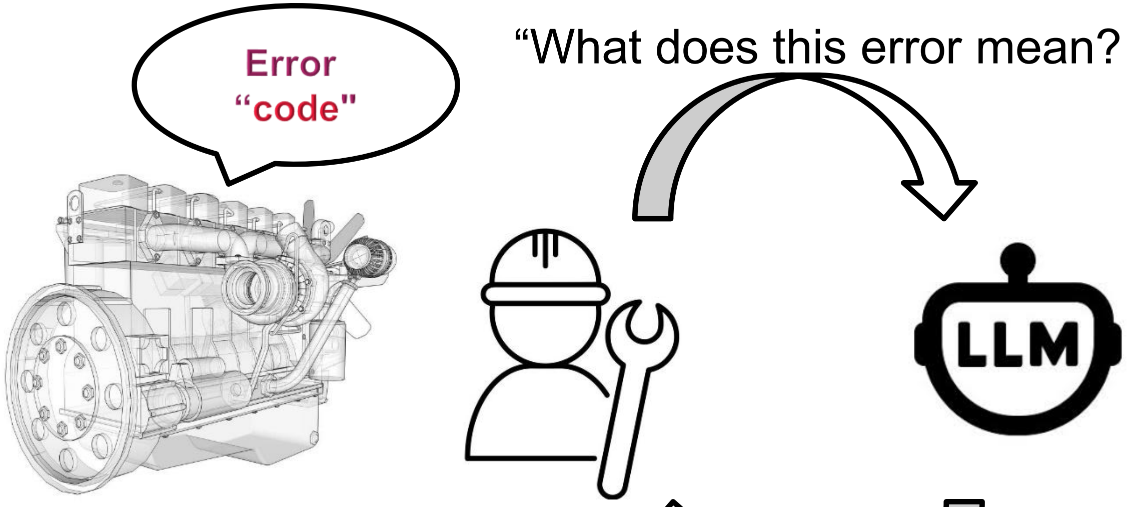
Workflow of an LLM…



“What does this error mean?

“It’s an internet router issue?...”

Workflow of an LLM…



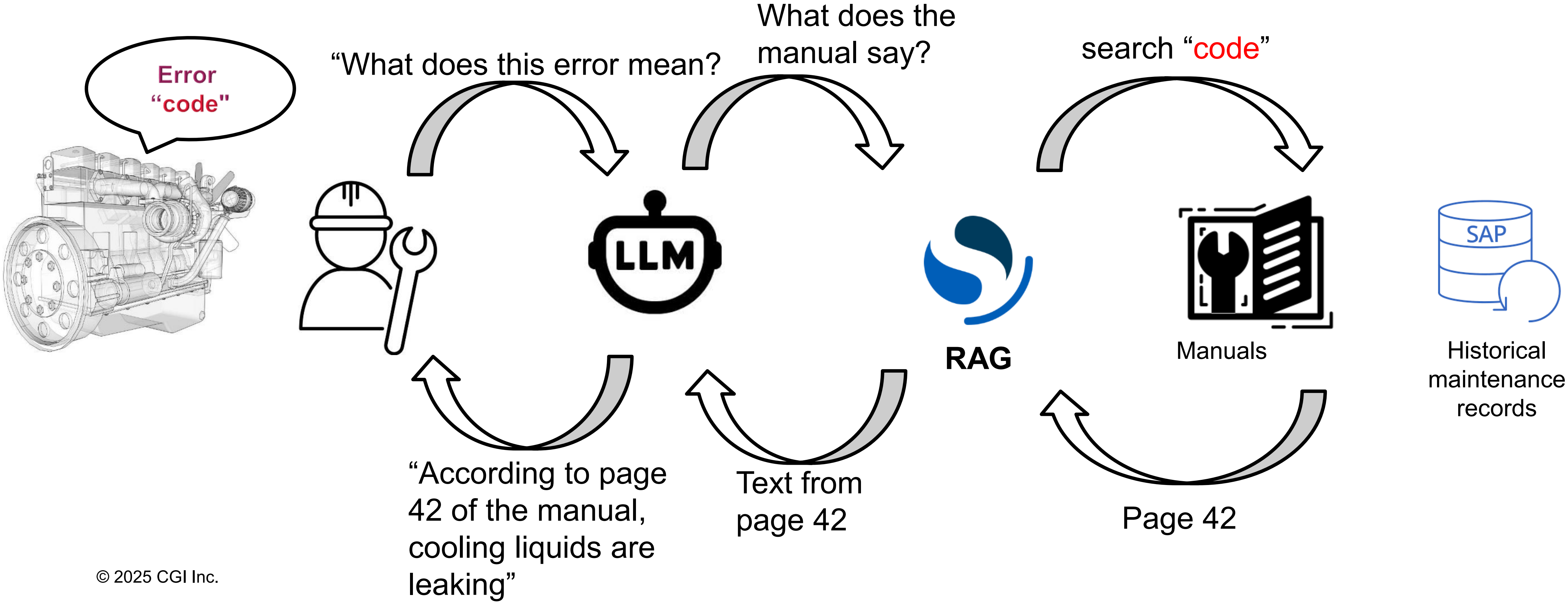
“What does this error mean?

“It’s an internet router issue?...”

Hallucination: *an LLM doesn’t possess enough knowledge about the domain, yet it is compelled to answer any question*

→ **It starts making things up**

Workflow of an LLM with **Retrieval-Augmented Generation**



Error “code"

“What does this error mean?

What does the manual say?

search “code”

RAG

Manuals

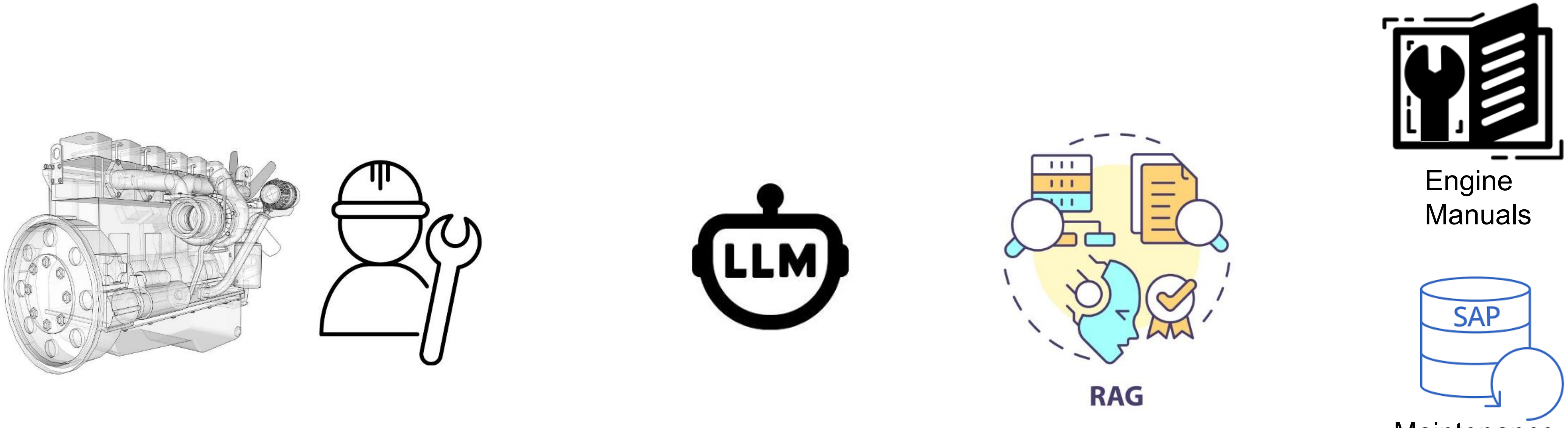
Historical maintenance records

“According to page 42 of the manual, cooling liquids are leaking”

Text from page 42

Page 42

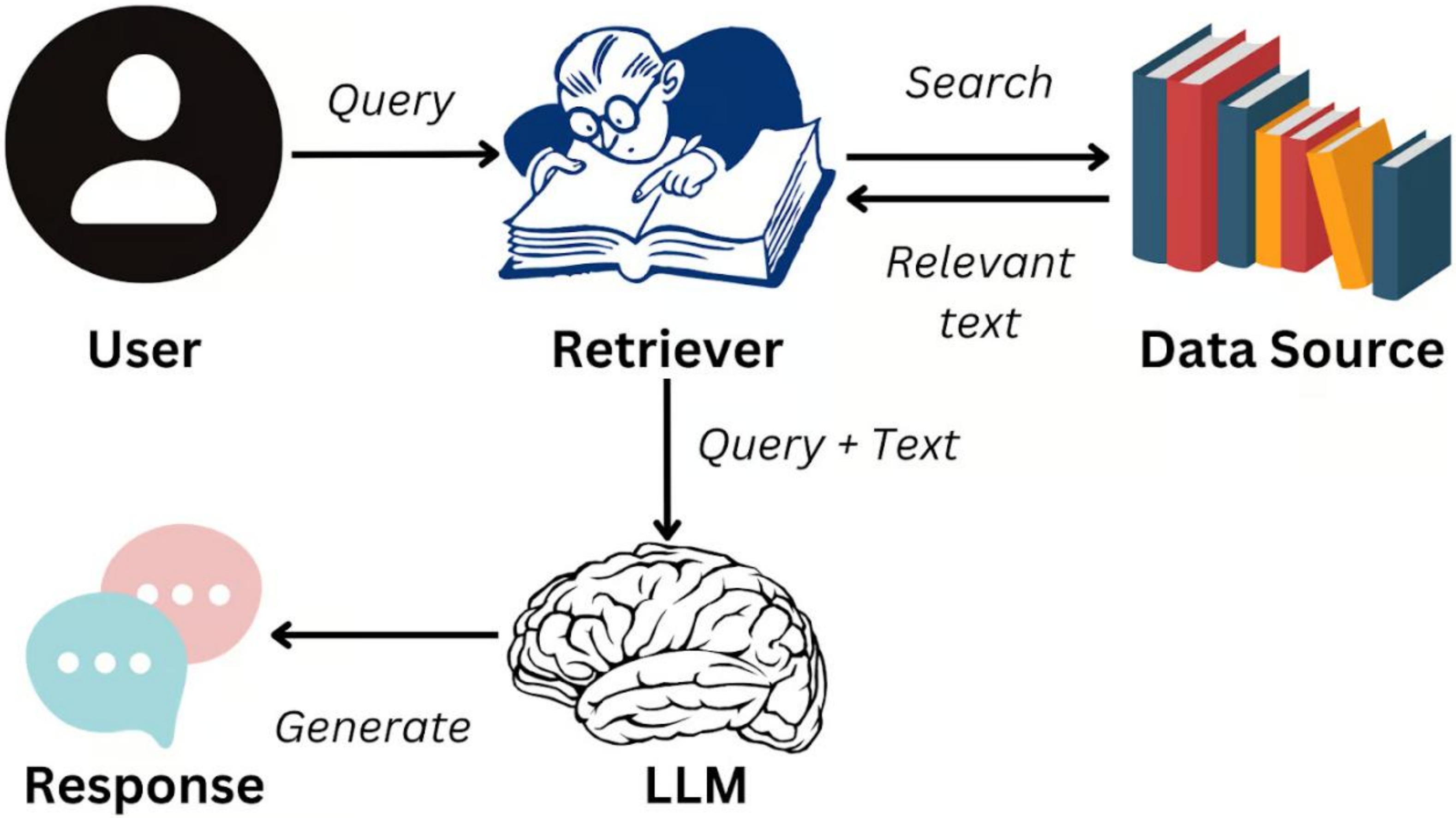
LLM with **Retrieval-Augmented Generation**



Through RAG, generated answers of an LLM become more:

- Reliable
- Explainable
- Traceable
- Grounded
- Multi-faceted

LLM & **Retrieval-Augmented Generation** (RAG)

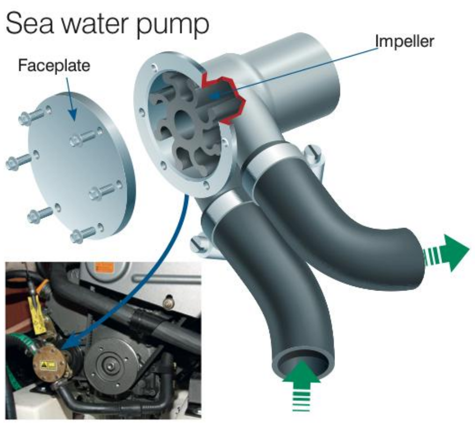


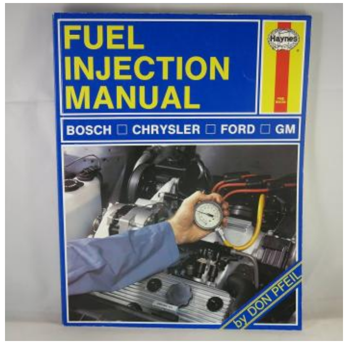
LLMs, RAG, & Vector Embeddings

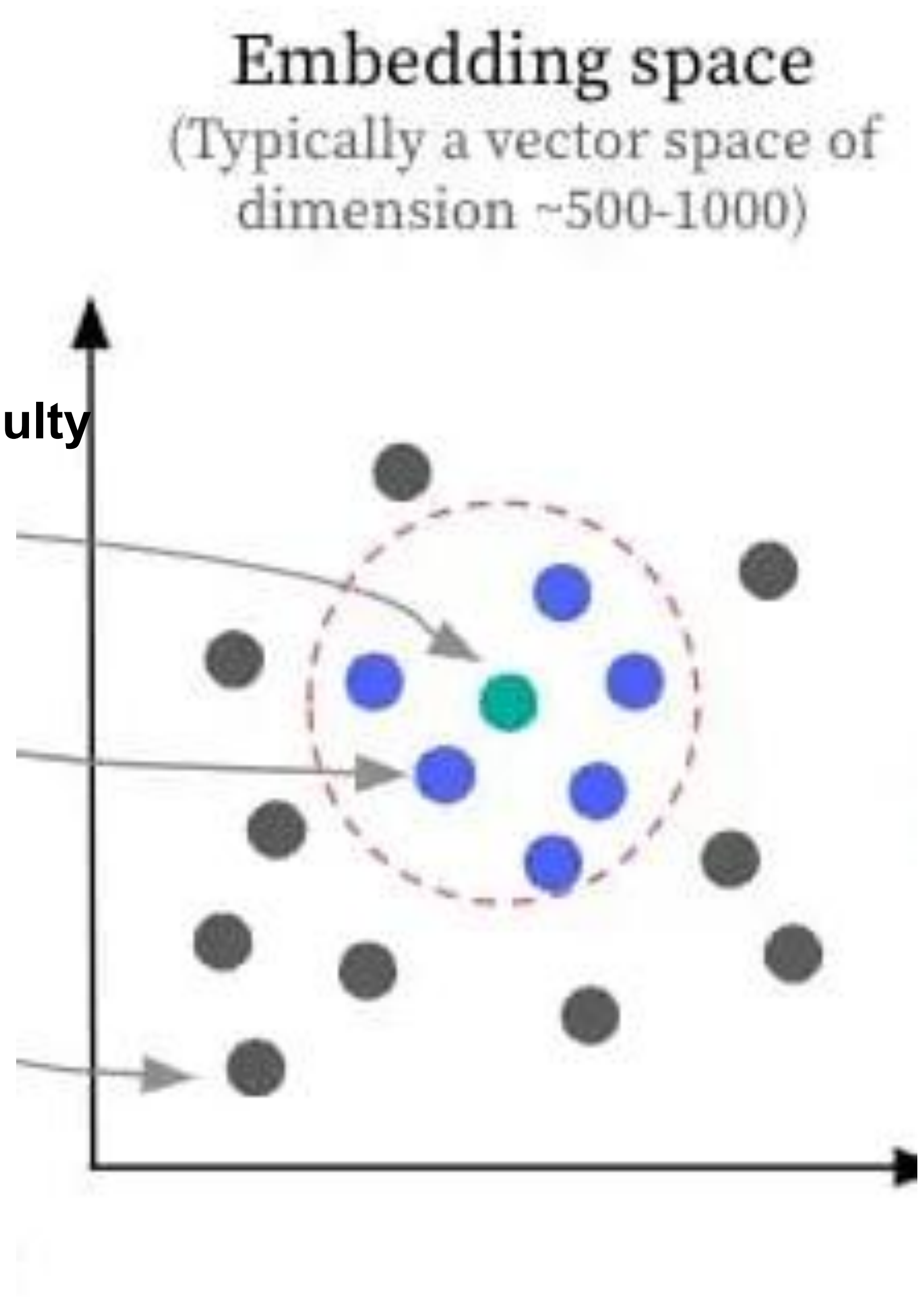
User

Query →

"What procedure do I need to diagnose a **faulty water pump**?"






Embedding space
(Typically a vector space of dimension ~500-1000)

RAG-improvements

RAG seldom works out-of-the-box…

RAG-improvements

RAG seldom works out-of-the-box…

How can we ensure that retrieved contexts are more likely to be relevant to the user's questions?

RAG-improvements

RAG seldom works out-of-the-box…

How can we ensure that retrieved contexts are more likely to be relevant to the user's questions?

How do we test our solution? How do we know when it "works"?

RAG-improvements

- Hybrid search

I’m working on the impeller of a sea-water cooling pump. What are the disassembly instructions?

✓ Disassembly instructions for sea-water pump:[…]

✗ Disassembly instructions for backup generator for water pump:[…]

✗ ‘I’m working on the power supply, and after following disassembly instructions […]

RAG-improvements

– Hybrid search: word-matching

Meaningful words:

- Impeller
- Sea-water
- Cooling pump
- Disassembly instructions

Non-meaningful words:

- On
- Of
- Are
- The
- Working
- after

I'm working on the impeller of a sea-water cooling pump. What are the disassembly instructions?

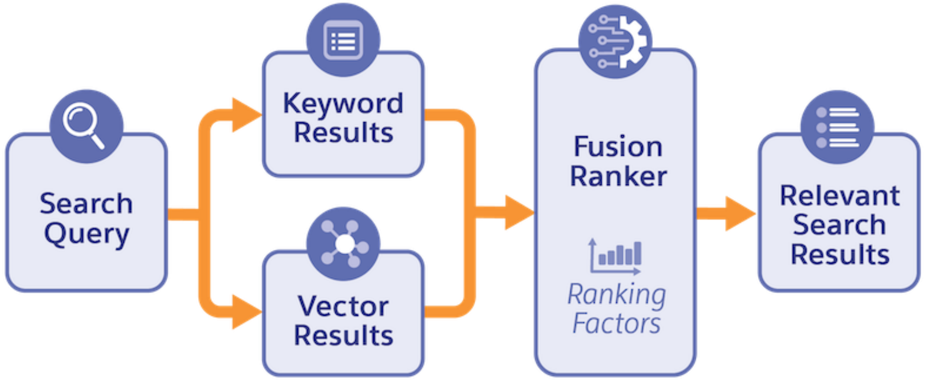
✓ Disassembly instructions for sea-water pump:[…]

✗ Disassembly instructions for backup generator for water pump:[…]

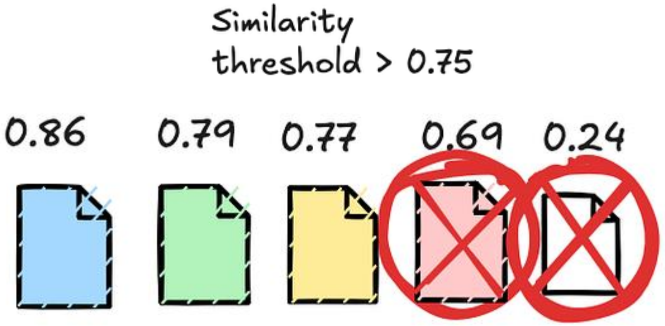
✗ 'I'm working on the power supply, and after following disassembly instructions […]

RAG-improvements

- Hybrid search

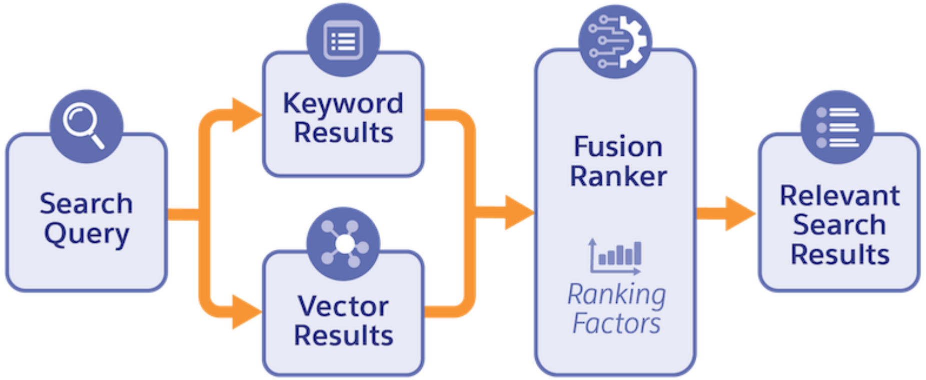


- Minimum-threshold matching

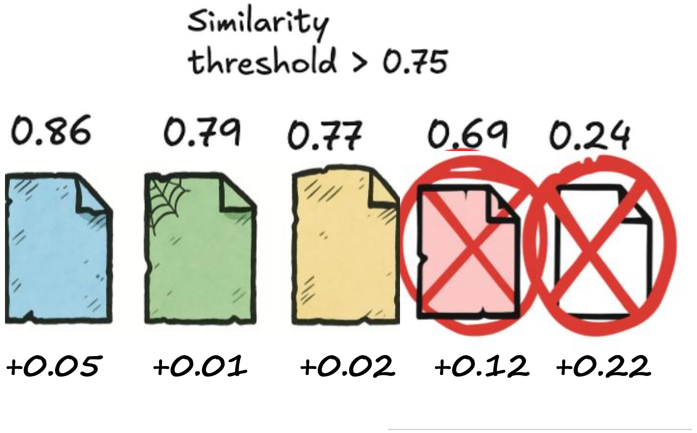


RAG-improvements

- Hybrid search



- Minimum-threshold matching
- Recency-boosting



RAG

RAG seldom works out-of-the-box…

How can we ensure that retrieved contexts are more likely to be relevant to the user's questions?

How do we test our solution? How do we know when it "works"?

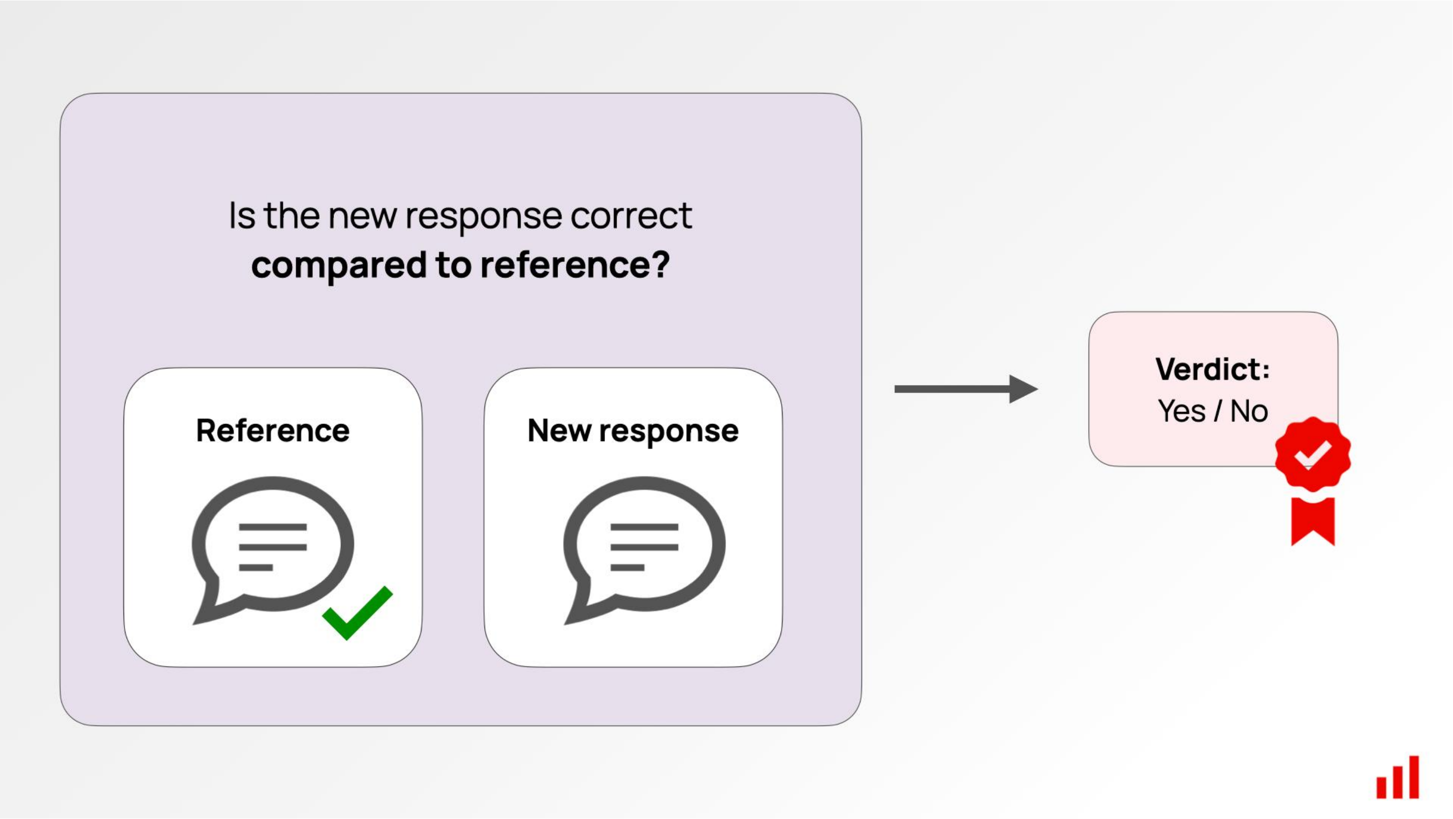
Evaluating an LLM application

- Establishing an **evaluation set**

User query	RAG context	Generated answer	Reference answer
“How big was the Titanic?”	“[…] Titanic was listed to be a staggering 270 […]” - Wikipedia	“According to Wikipedia, the Titanic was 270 meters in length”	“The Titanic was 270 meters long”
…	…	…	…
…	…	…	…

Evaluating an LLM application

- Establishing an **evaluation set**



Is the new response correct
compared to reference?

Reference

New response

Verdict:
Yes / No

Evaluating an LLM application

- Establishing an **evaluation set**

User query	RAG-context	Generated answer	Reference answer
“How big was the Titanic?”	“[…] Titanic was listed to be a staggering […]” - Wikipedia	According to Wikipedia, the Titanic was […]	The Titanic was 270 meters long […]
…	…	…	…
…	…	…	…

- Computing evaluation metrics:
  - Factual correctness
  - Context Recall
  - Faithfulness

Evaluating an LLM application

- Establishing an **evaluation set**

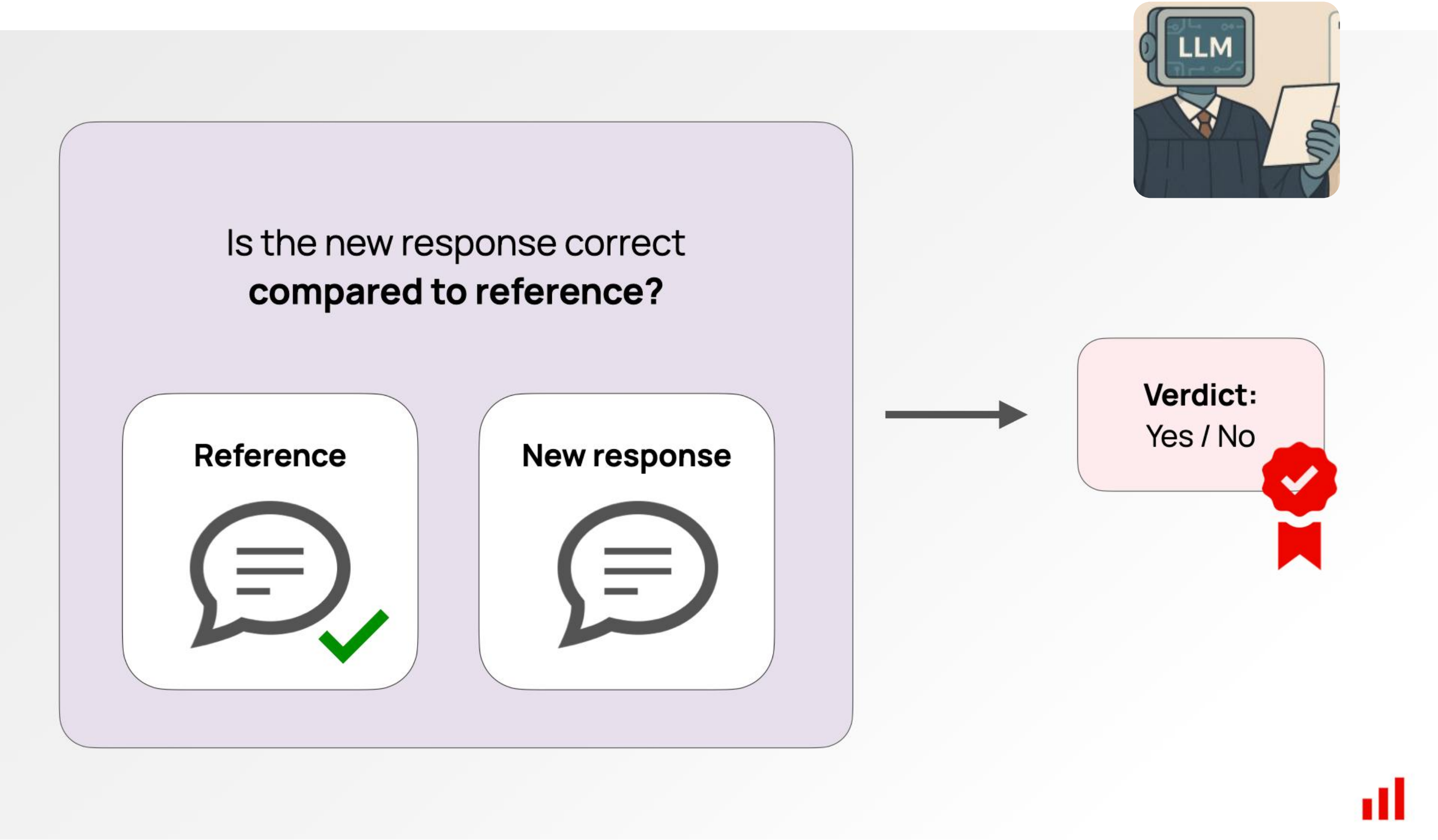
- Computing evaluation metrics:
  - Factual correctness
  - Context Recall
  - Faithfulness

  Using **LLM-as-a-judge**

User query	RAG-context	Generated answer	Reference answer
“How big was the Titanic?”	“[…] Titanic was listed to be a staggering […]” - Wikipedia	According to Wikipedia, the Titanic was […]	The Titanic was 270 meters long […]
…	…	…	…
…	…	…	…

Evaluating an LLM application

Evaluating an LLM application



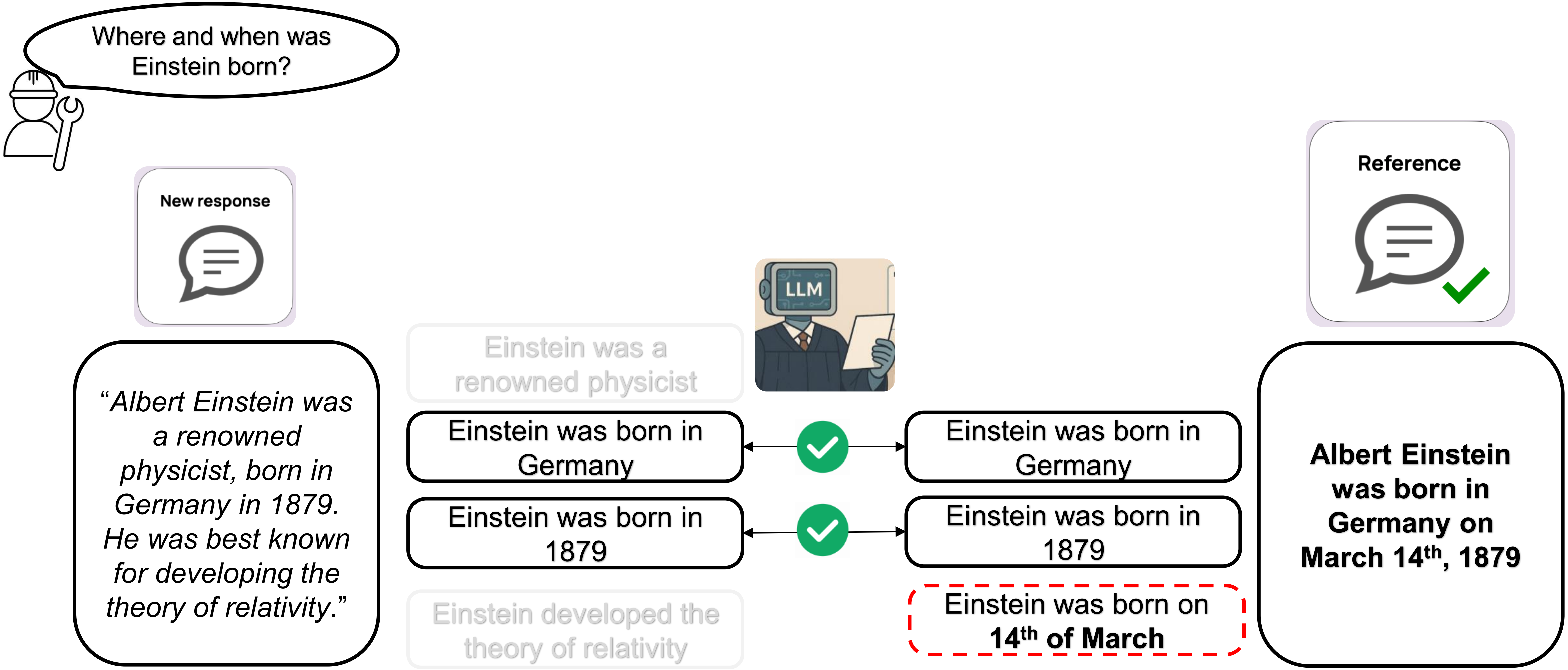
Is the new response correct **compared to reference?**

Reference

New response

Verdict:
Yes / No

Evaluating an LLM application: Factual Correctness



Where and when was Einstein born?

New response

“Albert Einstein was a renowned physicist, born in Germany in 1879. He was best known for developing the theory of relativity.”

Einstein was a renowned physicist

Einstein was born in Germany ✓ Einstein was born in Germany

Einstein was born in 1879 ✓ Einstein was born in 1879

Einstein developed the theory of relativity

Einstein was born on **14th of March**

Reference

Albert Einstein was born in Germany on March 14th, 1879

Evaluating an LLM application: Factual Correctness

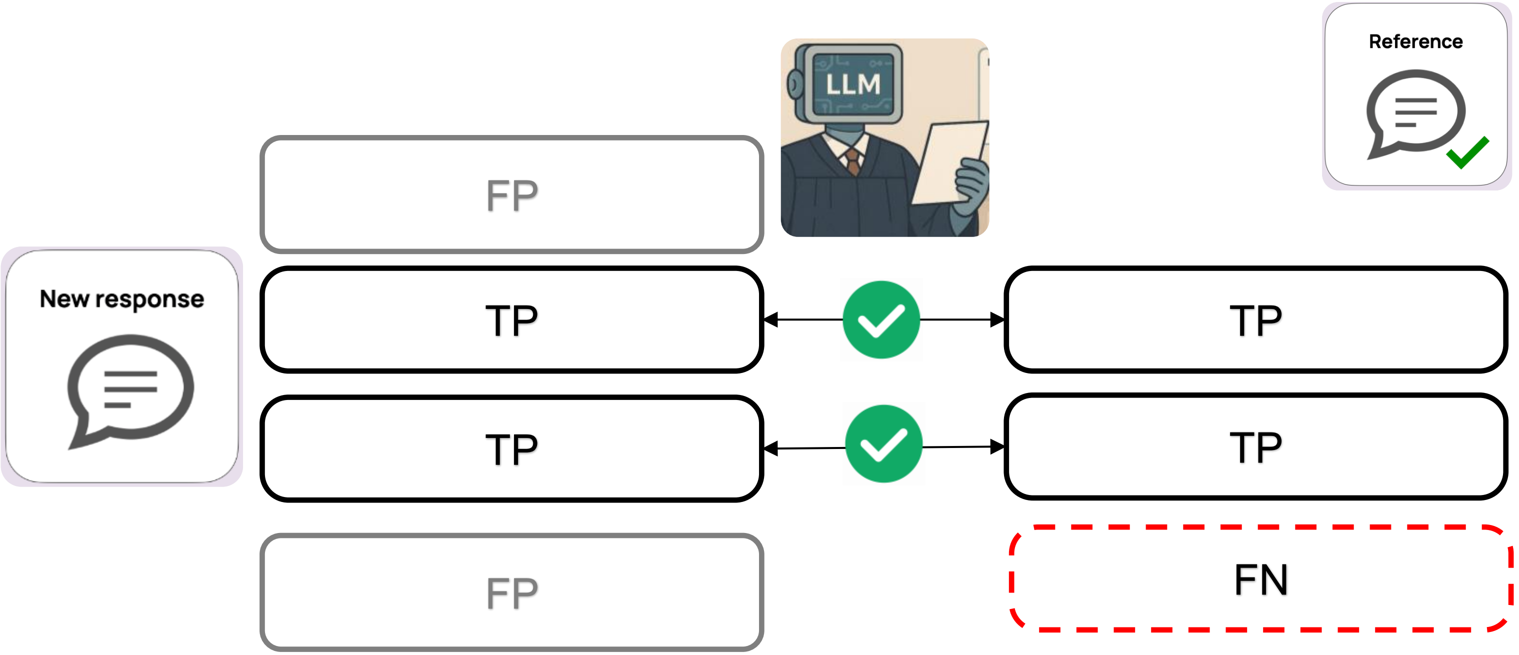
$$\text{True Positive (TP)} = \text{Number of claims in response that are present in reference}$$

$$\text{False Positive (FP)} = \text{Number of claims in response that are not present in reference}$$

$$\text{False Negative (FN)} = \text{Number of claims in reference that are not present in response}$$

$$\text{Precision} = \frac{TP}{(TP + FP)}$$

$$\text{Recall} = \frac{TP}{(TP + FN)}$$



Evaluating an LLM application: Factual Correctness

$$\text{True Positive (TP)} = \text{Number of claims in response that are present in reference}$$

$$\text{False Positive (FP)} = \text{Number of claims in response that are not present in reference}$$

$$\text{False Negative (FN)} = \text{Number of claims in reference that are not present in response}$$

$$\text{Precision} = \frac{TP}{(TP + FP)}$$

$$\text{Recall} = \frac{TP}{(TP + FN)}$$

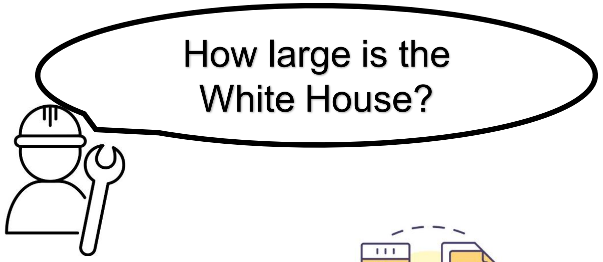
Factual correctness:

How close is the answer generated by the LLM to the “correct” answer?

- *How much irrelevant extra information did the LLM give?*
- *How much information did the LLM fail to provide, that should have been part of its answer?*

Evaluating an LLM application: Context Recall

Computing evaluation metrics:



$$\text{Context Recall} = \frac{\text{Number of claims in the reference supported by the retrieved context}}{\text{Total number of claims in the reference}}$$

Height: 22 m

Width: 52 m

16 main rooms

6 restrooms

Height: 22 m

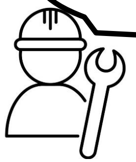
Width: 52 m

Depth: 26 m

Evaluating an LLM application: Context Recall

Computing evaluation metrics:

How large is the White House?



$$\text{Context Recall} = \frac{\text{Number of claims in the reference supported by the retrieved context}}{\text{Total number of claims in the reference}}$$

RAG

Height: 22 m

Width: 52 m

16 main rooms

6 restrooms

Reference

Height: 22 m

Width: 52 m

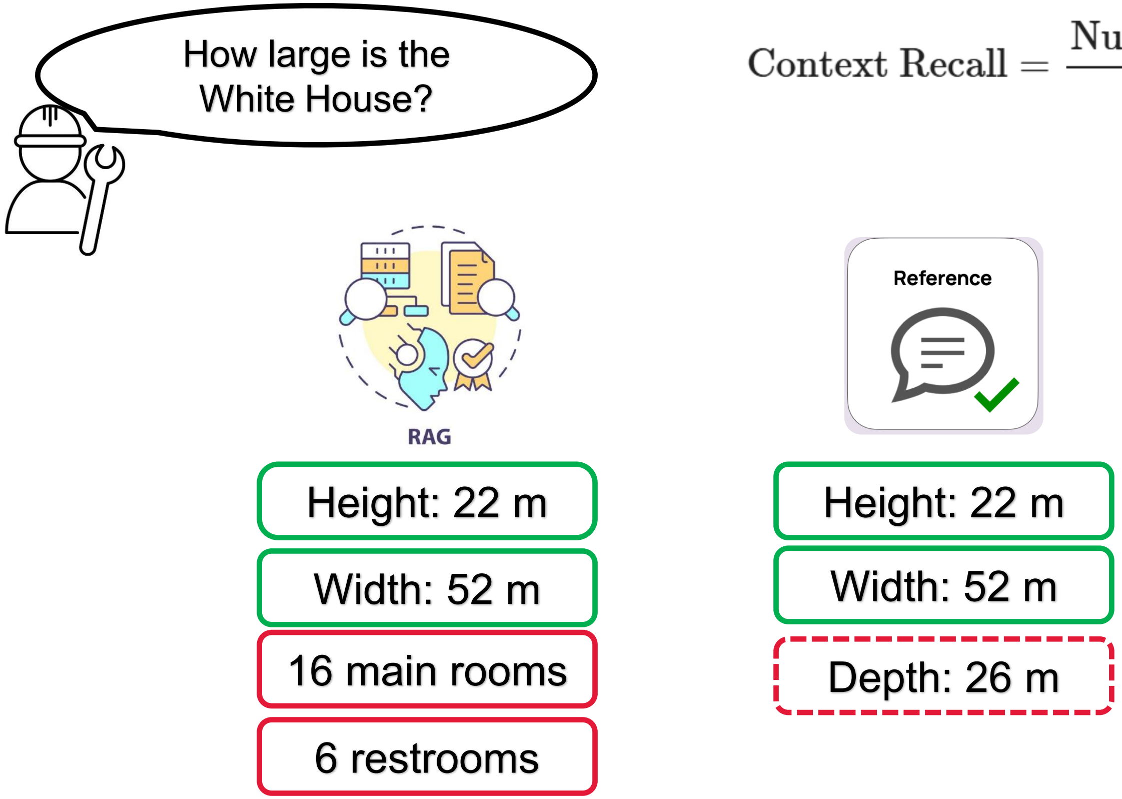
Depth: 26 m

“How many of the claims in the reference answer were present in the retrieved contexts?”

Evaluating an LLM application: Context Recall

Computing evaluation metrics:

$$\text{Context Recall} = \frac{\text{Number of claims in the reference supported by the retrieved context}}{\text{Total number of claims in the reference}}$$



How large is the White House?

RAG	Reference
Height: 22 m	Height: 22 m
Width: 52 m	Width: 52 m
16 main rooms	Depth: 26 m
6 restrooms	

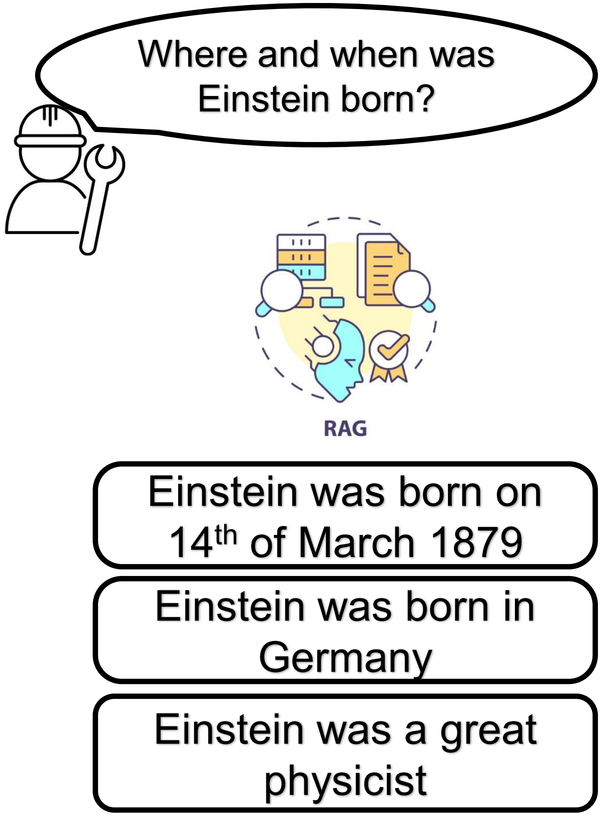
“How many of the claims in the reference answer were present in the retrieved contexts?”

i.e: “Did RAG manage to retrieve the information that would suffice a perfect LLM agent to generate a correct answer?”

Evaluating an LLM application: Faithfulness

Computing evaluation metrics:

$$\text{Faithfulness score} = \frac{|\text{Number of claims in the generated answer that can be inferred from given context}|}{|\text{Total number of claims in the generated answer}|}$$

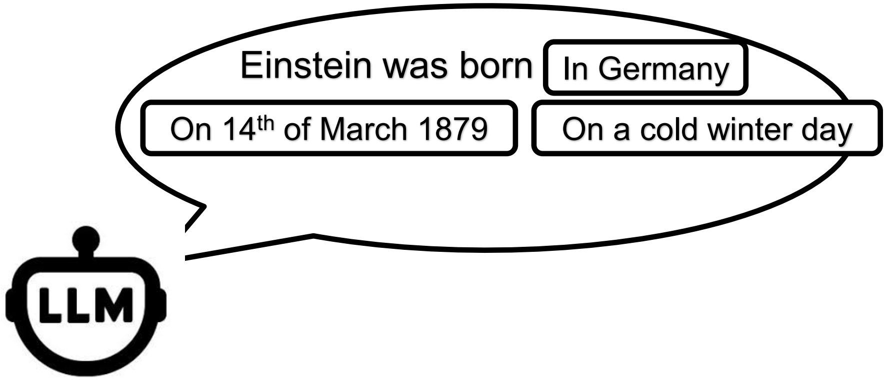


Where and when was Einstein born?

Einstein was born on 14th of March 1879

Einstein was born in Germany

Einstein was a great physicist



Einstein was born In Germany

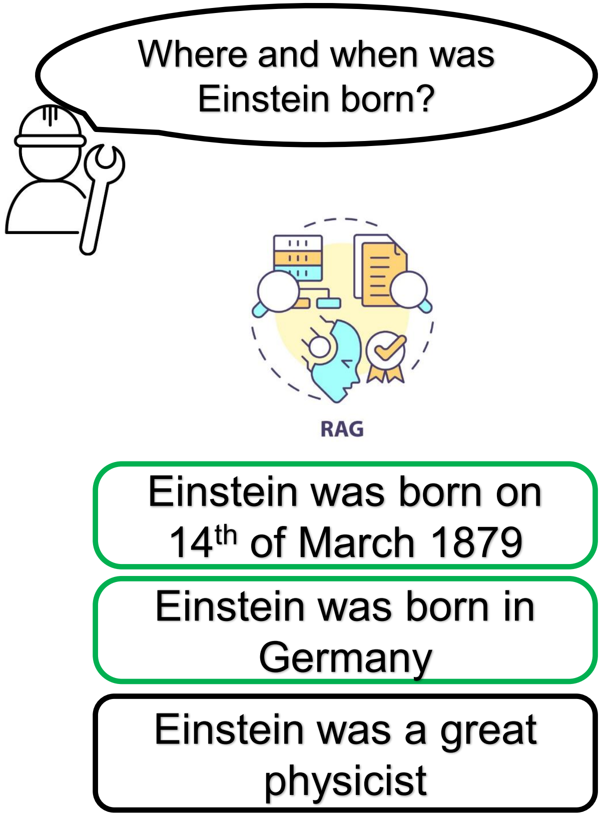
On 14th of March 1879

On a cold winter day

Evaluating an LLM application: Faithfulness

Computing evaluation metrics:

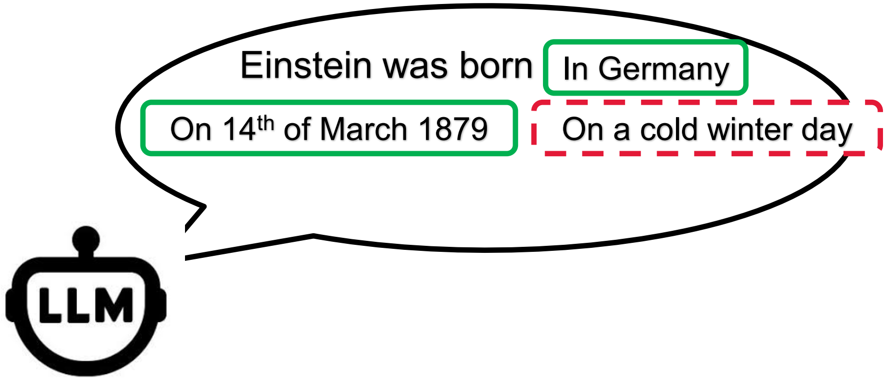
$$\text{Faithfulness score} = \frac{|\text{Number of claims in the generated answer that can be inferred from given context}|}{|\text{Total number of claims in the generated answer}|}$$



Einstein was born on 14th of March 1879

Einstein was born in Germany

Einstein was a great physicist

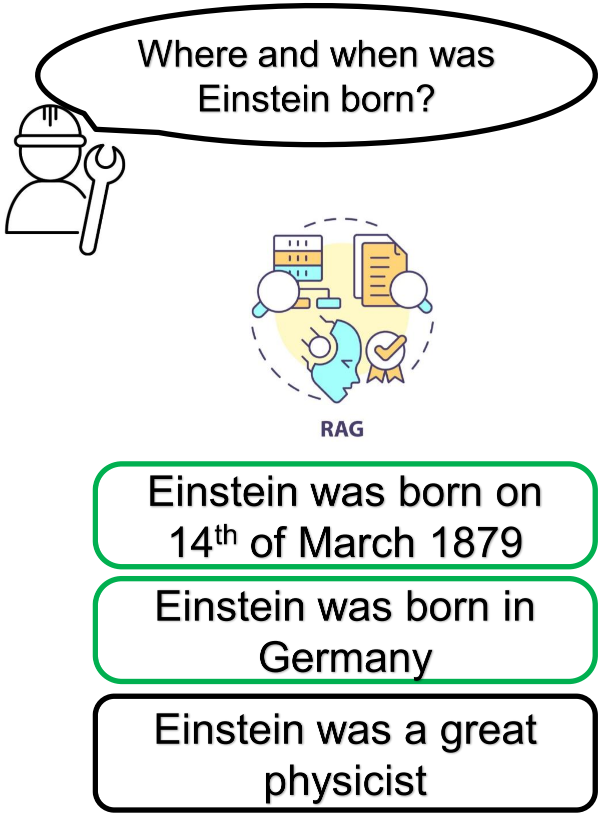


"How faithful was the LLM in only using claims that were backed by its RAG context?"

Evaluating an LLM application: Faithfulness

Computing evaluation metrics:

$$\text{Faithfulness score} = \frac{|\text{Number of claims in the generated answer that can be inferred from given context}|}{|\text{Total number of claims in the generated answer}|}$$

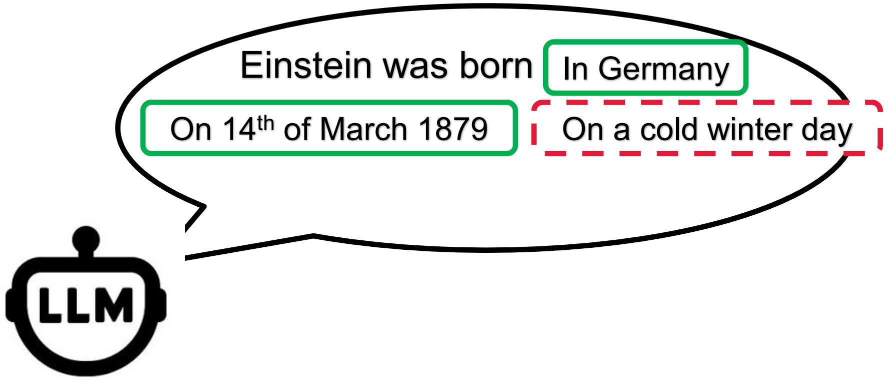


Where and when was Einstein born?

Einstein was born on 14th of March 1879

Einstein was born in Germany

Einstein was a great physicist



Einstein was born In Germany

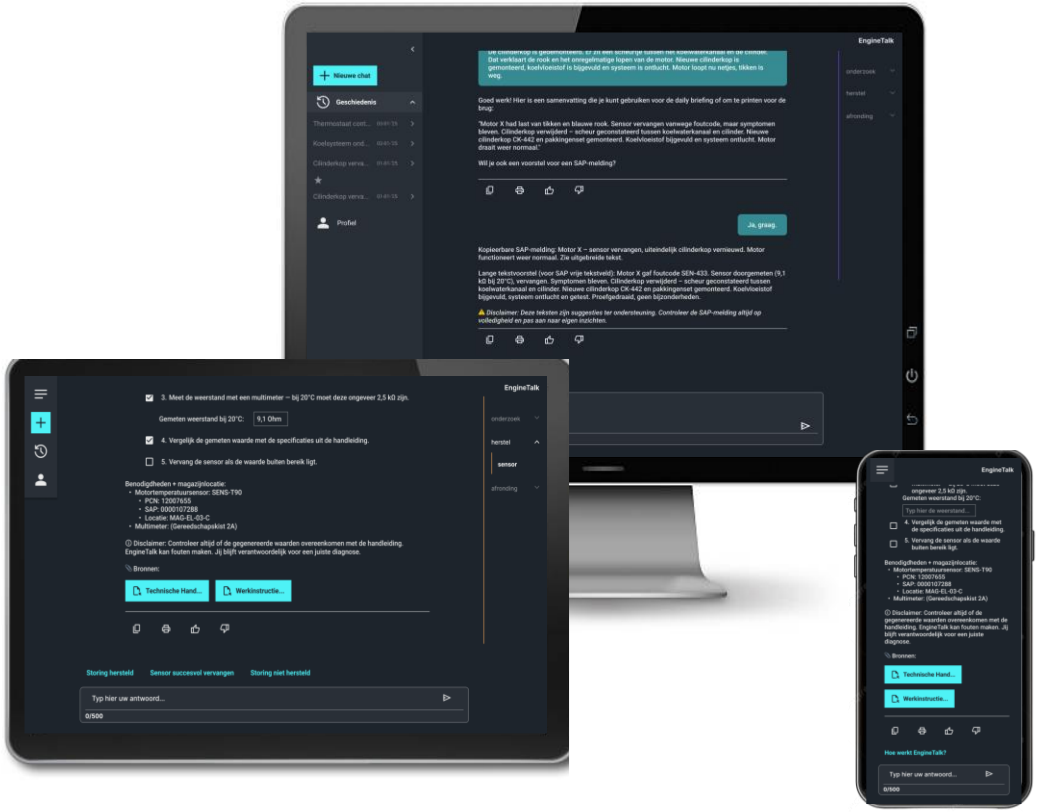
On 14th of March 1879

On a cold winter day

“*How faithful was the LLM in only using claims that were backed by its RAG context?*”

→ Gives a measure of **Hallucinations**

Results



De cilinderkop is gedemonteerd. Er zit een scheurtje tussen het koelwaterkanaal en de cilinder. Dat verklaart de rook en het onregelmatige lopen van de motor. Nieuwe cilinderkop is gemonteerd, koelvloeistof is bijgevuld en systeem is ontlucht. Motor loopt nu netjes, tikken is weg.

Goed werk! Hier is een samenvatting die je kunt gebruiken voor de daily briefing of om te printen voor de brug:

"Motor X had last van tikken en blauwe rook. Sensor vervangen vanwege foutcode, maar symptomen bleven. Cilinderkop verwijderd – scheur geconstateerd tussen koelwaterkanaal en cilinder. Nieuwe cilinderkop CK-442 en pakkingenset gemonteerd. Koelvloeistof bijgevuld en systeem ontlucht. Motor draait weer normaal."

Wil je ook een voorstel voor een SAP-melding?

Ja, graag.

Kopieerbare SAP-melding: Motor X – sensor vervangen, uiteindelijk cilinderkop vernieuwd. Motor functioneert weer normaal. Zie uitgebreide tekst.

Lange tekstvoorstel (voor SAP vrije tekstveld): Motor X gaf foutcode SEN-433. Sensor doorgemeten (9,1 kΩ bij 20°C), vervangen. Symptomen bleven. Cilinderkop verwijderd – scheur geconstateerd tussen koelwaterkanaal en cilinder. Nieuwe cilinderkop CK-442 en pakkingenset gemonteerd. Koelvloeistof bijgevuld, systeem ontlucht en getest. Proefgedraaid, geen bijzonderheden.

⚠ *Disclaimer: Deze teksten zijn suggesties ter ondersteuning. Controleer de SAP-melding altijd op volledigheid en pas aan naar eigen inzichten.*

☑ 3. Meet de weerstand met een multimeter – bij 20°C moet deze ongeveer 2,5 kΩ zijn.

Gemeten weerstand bij 20°C: 9,1 Ohm

☑ 4. Vergelijk de gemeten waarde met de specificaties uit de handleiding.

□ 5. Vervang de sensor als de waarde buiten bereik ligt.

Benodigdheden + magazijnlocatie:
- Motortemperatuursensor: SENS-T90
  - PCN: 12007655
  - SAP: 0000107288
  - Locatie: MAG-EL-03-C
- Multimeter: (Gereedschapskist 2A)

ⓘ Disclaimer: Controleer altijd of de gegenereerde waarden overeenkomen met de handleiding. EngineTalk kan fouten maken. Jij blijft verantwoordelijk voor een juiste diagnose.

Bronnen:

ongeveer 2,5 kΩ zijn.
Gemeten weerstand bij 20°C:

□ 4. Vergelijk de gemeten waarde met de specificaties uit de handleiding.

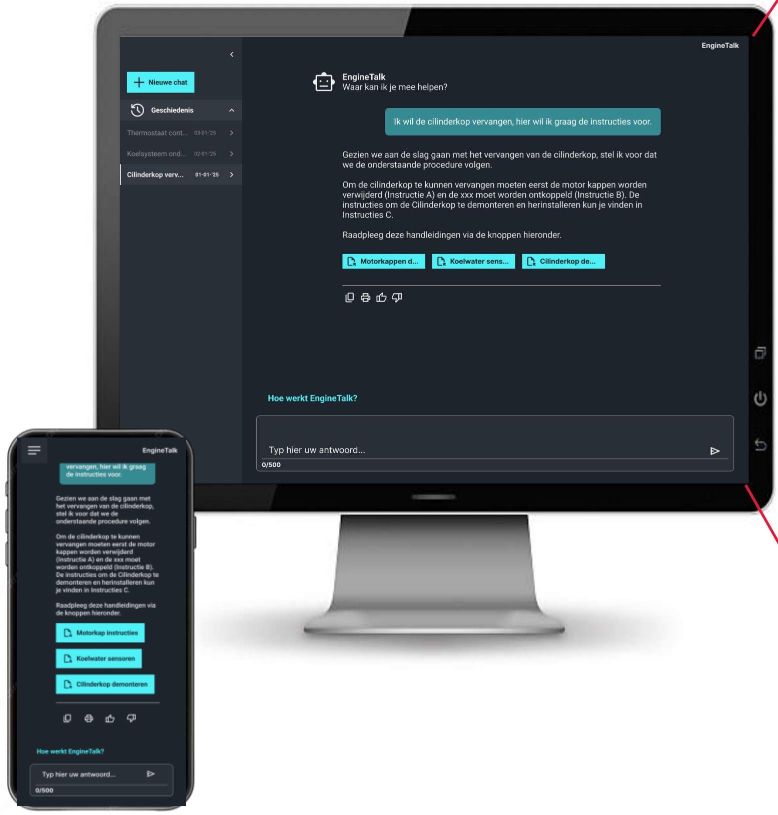
□ 5. Vervang de sensor als de waarde buiten bereik ligt.

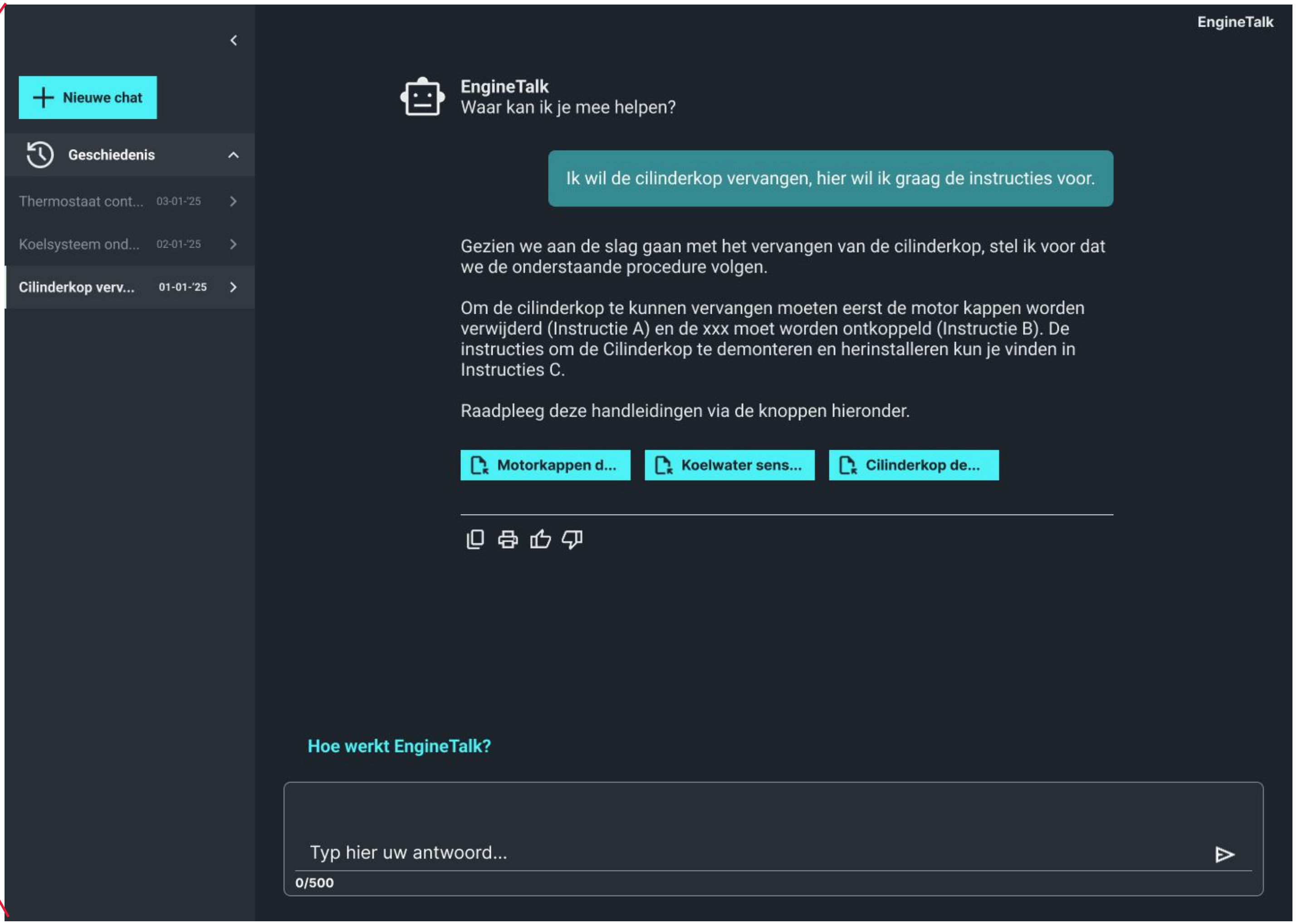
Benodigdheden + magazijnlocatie:
- Motortemperatuursensor: SENS-T90
  - PCN: 12007655
  - SAP: 0000107288
  - Locatie: MAG-EL-03-C
- Multimeter: (Gereedschapskist 2A)

ⓘ Disclaimer: Controleer altijd of de gegenereerde waarden overeenkomen met de handleiding. EngineTalk kan fouten maken. Jij blijft verantwoordelijk voor een juiste diagnose.

Bronnen:

Results




EngineTalk
Nieuwe chat
Geschiedenis
Thermostaat cont... 03-01-'25
Koelsysteem ond... 02-01-'25
Cilinderkop verv... 01-01-'25

EngineTalk
Waar kan ik je mee helpen?

Ik wil de cilinderkop vervangen, hier wil ik graag de instructies voor.

Gezien we aan de slag gaan met het vervangen van de cilinderkop, stel ik voor dat we de onderstaande procedure volgen.

Om de cilinderkop te kunnen vervangen moeten eerst de motor kappen worden verwijderd (Instructie A) en de xxx moet worden ontkoppeld (Instructie B). De instructies om de Cilinderkop te demonteren en herinstalleren kun je vinden in Instructies C.

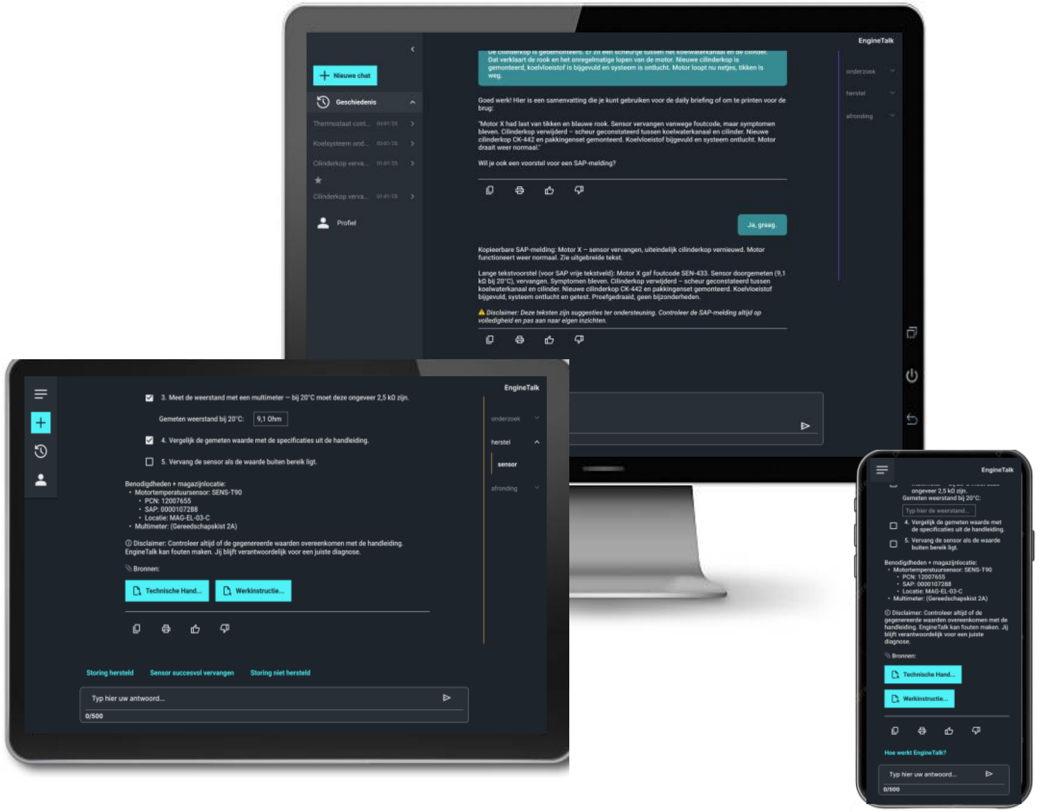
Raadpleeg deze handleidingen via de knoppen hieronder.

Results

Early results are promising:

Functionally:

- Positive feedback from end-users
  - Especially inexperienced personnel
- RAG shown to reduce hallucinations



De cilinderkop is gedemonteerd. Er zit een scheurtje tussen het koelwaterkanaal en de cilinder. Dat verklaart de rook en het onregelmatige lopen van de motor. Nieuwe cilinderkop is gemonteerd, koelvloeistof is bijgevuld en systeem is ontlucht. Motor loopt nu netjes, tikken is weg.

Goed werk! Hier is een samenvatting die je kunt gebruiken voor de daily briefing of om te printen voor de brug:

"Motor X had last van tikken en blauwe rook. Sensor vervangen vanwege foutcode, maar symptomen bleven. Cilinderkop verwijderd – scheur geconstateerd tussen koelwaterkanaal en cilinder. Nieuwe cilinderkop CK-442 en pakkingenset gemonteerd. Koelvloeistof bijgevuld en systeem ontlucht. Motor draait weer normaal."

Wil je ook een voorstel voor een SAP-melding?

Ja, graag.

Kopieerbare SAP-melding: Motor X – sensor vervangen, uiteindelijk cilinderkop vernieuwd. Motor functioneert weer normaal. Zie uitgebreide tekst.

Lange tekstvoorstel (voor SAP vrije tekstveld): Motor X gaf foutcode SEN-433. Sensor doorgemeten (9,1 kΩ bij 20°C), vervangen. Symptomen bleven. Cilinderkop verwijderd – scheur geconstateerd tussen koelwaterkanaal en cilinder. Nieuwe cilinderkop CK-442 en pakkingenset gemonteerd. Koelvloeistof bijgevuld, systeem ontlucht en getest. Proefgedraaid, geen bijzonderheden.

⚠ *Disclaimer: Deze teksten zijn suggesties ter ondersteuning. Controleer de SAP-melding altijd op volledigheid en pas aan naar eigen inzichten.*

☑ 3. Meet de weerstand met een multimeter – bij 20°C moet deze ongeveer 2,5 kΩ zijn.

Gemeten weerstand bij 20°C: 9,1 Ohm

☑ 4. Vergelijk de gemeten waarde met de specificaties uit de handleiding.

□ 5. Vervang de sensor als de waarde buiten bereik ligt.

Benodigdheden + magazijnlocatie:
- Motortemperatuursensor: SENS-T90
  - PCN: 12007655
  - SAP: 0000107288
  - Locatie: MAG-EL-03-C
- Multimeter: (Gereedschapskist 2A)

ⓘ Disclaimer: Controleer altijd of de gegenereerde waarden overeenkomen met de handleiding. EngineTalk kan fouten maken. Jij blijft verantwoordelijk voor een juiste diagnose.

Bronnen:

Results

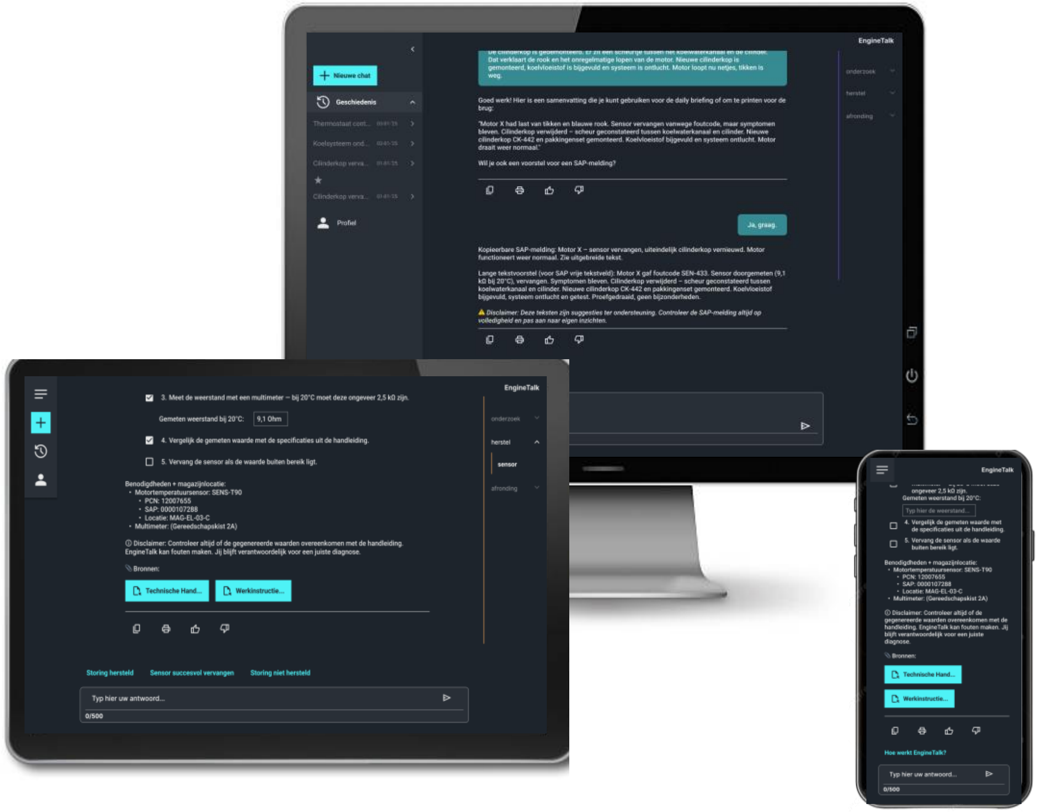
Early results are promising:

Functionally:

- Positive feedback from end-users
  - Especially inexperienced personnel
- RAG shown to reduce hallucinations

LLM-evaluation:

- Promising early results,
  - but still pending due to complexity



De cilinderkop is gedemonteerd. Er zit een scheurtje tussen het koelwaterkanaal en de cilinder. Dat verklaart de rook en het onregelmatige lopen van de motor. Nieuwe cilinderkop is gemonteerd, koelvloeistof is bijgevuld en systeem is ontlucht. Motor loopt nu netjes, tikken is weg.

Goed werk! Hier is een samenvatting die je kunt gebruiken voor de daily briefing of om te printen voor de brug:

"Motor X had last van tikken en blauwe rook. Sensor vervangen vanwege foutcode, maar symptomen bleven. Cilinderkop verwijderd – scheur geconstateerd tussen koelwaterkanaal en cilinder. Nieuwe cilinderkop CK-442 en pakkingenset gemonteerd. Koelvloeistof bijgevuld en systeem ontlucht. Motor draait weer normaal."

Wil je ook een voorstel voor een SAP-melding?

Ja, graag.

Kopieerbare SAP-melding: Motor X – sensor vervangen, uiteindelijk cilinderkop vernieuwd. Motor functioneert weer normaal. Zie uitgebreide tekst.

Lange tekstvoorstel (voor SAP vrije tekstveld): Motor X gaf foutcode SEN-433. Sensor doorgemeten (9,1 kΩ bij 20°C), vervangen. Symptomen bleven. Cilinderkop verwijderd – scheur geconstateerd tussen koelwaterkanaal en cilinder. Nieuwe cilinderkop CK-442 en pakkingenset gemonteerd. Koelvloeistof bijgevuld, systeem ontlucht en getest. Proefgedraaid, geen bijzonderheden.

⚠ *Disclaimer: Deze teksten zijn suggesties ter ondersteuning. Controleer de SAP-melding altijd op volledigheid en pas aan naar eigen inzichten.*

☑ 3. Meet de weerstand met een multimeter – bij 20°C moet deze ongeveer 2,5 kΩ zijn.

Gemeten weerstand bij 20°C: 9,1 Ohm

☑ 4. Vergelijk de gemeten waarde met de specificaties uit de handleiding.

□ 5. Vervang de sensor als de waarde buiten bereik ligt.

Benodigdheden + magazijnlocatie:

- Motortemperatuursensor: SENS-T90
  - PCN: 12007655
  - SAP: 0000107288
  - Locatie: MAG-EL-03-C
- Multimeter: (Gereedschapskist 2A)

ⓘ Disclaimer: Controleer altijd of de gegenereerde waarden overeenkomen met de handleiding. EngineTalk kan fouten maken. Jij blijft verantwoordelijk voor een juiste diagnose.

Bronnen:

Conclusions

Conclusions

- An LLM&RAG solution can substantially reduce time spent on diagnosis and repair
    - Especially for inexperienced crews.
- The integration of historical maintenance records is highly dependent on quality of such records.
- Thorough LLM-evaluations exist
    - but remain experimental, and prone to complexity

Future prospects

- On-vessel deployment of EngineTalk
- Expanding scope to different user-groups, and other IT systems
- Utilizing LLMs with real-time system components

Any questions?

Youri Linden

Senior System Engineer,
Material and IT
Command

Y.Linden@mindef.nl

Bart-Peter Smit

Navy Consultant

CGI Netherlands

bart-peter.smit@cgi.com

Jonathan L. Maas

AI Consultant

CGI Netherlands

Jonathan.Maas@cgi.com