



# Community of Practice KIPerWeb

Austausch zur Nutzung und Entwicklung KI-gestützter Webanwendungen



**KIPerWEB**



**Forschungsinstitut  
Betriebliche Bildung**

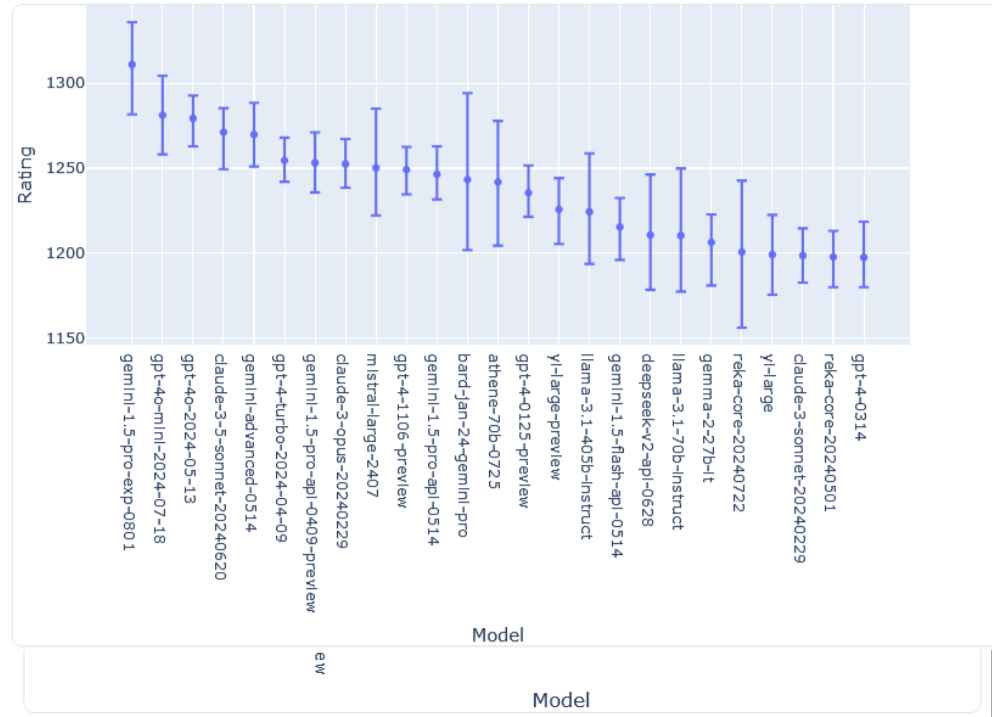
- **Update**
  - News & Leaderboard-Update
- **Input**
  - „Rapid Iterative Testing and Evaluation (RITE) Method for Enhancing the Context of Large Language Models (LLMs) “ (Gastbeitrag von Toke Lichtenberg)
- **Diskussion**

# News & Update (08.08.2024)




- *Gemini-1.5-pro* nun an der Spitze vor *Gpt-4o-mini*  
*Gemini-1.5-flash-api-0514* & *deepseek v2-api-0628*  
vor *gemma-2-27b-it*
- Neuzugänge: *Mistral-large-2407* (*Mistral research licence, non-commercial*), *Llama-3.1-405B-Instruct* & *Llama-3.1-70B-Instruct* (noch vor *gemma 2-27b-it*)
- Neu im Rennen (aber noch nicht ausgewiesen) ist *Gemma-2-2B-it*
- (Noch) nicht im Rennen, aber pot. unter den besten freien LLMs: *Mistral-Nemo-Instruct-2407* (Apache 2.0)

Confidence Intervals on model strength (Arena Elo, German)



# Exkurs: GraphRAG-Chatbot zu BERUFENET



Auf Basis des BERUFENET-Wissensgraphen (vgl. Fischer & Dörpinghaus, 2024) wird aktuell GraphRAG (vgl. CoP-Sitzung 6) über die dynamische Generierung des folgenden System-Prompt implementiert – weitere Attribute lassen sich bedarfsgerecht einbinden. 

Du bist Berufsberater mit einem guten Überblick über alle Berufe in Deutschland.

Berücksichtige bei deiner Antwort auf die letzte Nachricht des aktuellen Dialogs folgende Informationen aus der Datenbank BERUFENET der Bundesagentur für Arbeit:

'KI-Engineer: KI-Engineers entwickeln komplexe intelligente Systeme, die Informationen verarbeiten, Muster erkennen und Entscheidungen treffen können. Dazu nutzen sie beispielsweise künstliche neuronale Netzwerke.'

Folgende Beziehungen zu verwandten Entitäten lassen sich der Datenbank entnehmen:

- 'Informatik (weiterführend)' qualifiziert für Tätigkeit als 'KI-Engineer'
- 'Künstliche Intelligenz (grundständig)' qualifiziert für Tätigkeit als 'KI-Engineer'
- 'Ingenieurinformatik, Computational Engineering (grundst.)' qualifiziert für Tätigkeit als 'KI-Engineer'
- 'Softwaretechnik (weiterführend)' qualifiziert für Tätigkeit als 'KI-Engineer'
- 'KI-Engineer' erfordert Kompetenz 'Qualitätsmanagement'
- 'KI-Engineer' erfordert Kompetenz 'Datenschutzrecht'
- 'KI-Engineer' erfordert Kompetenz 'Web-Applikationen (Entwicklung, Programmierung)'
- 'KI-Engineer' erfordert Kompetenz 'Kompetenzgruppe "Programmiersprachen, Programmbibliotheken, Entwicklungsumgebungen"'
- 'KI-Engineer' bietet Aufstiegsmöglichkeit zu 'Informatik (weiterführend)'
- 'KI-Engineer' erfordert Kompetenz 'KI-Plattformen (as a Service)'
- 'KI-Engineer' erfordert Kompetenz 'Informatik'
- 'KI-Engineer' bietet Aufstiegsmöglichkeit zu 'Kybernetik (weiterführend)'
- 'KI-Engineer' erfordert Kompetenz 'Expertensysteme, Künstliche Intelligenz'
- 'KI-Engineer' erfordert Kompetenz 'Machine Learning' (...)

# Exkurs: GraphRAG-Chatbot zu BERUFENET



Hier ein ausgebautes Beispiel mit Gradio-User-Interface mit System Prompt und User-Interface zur Anfrage „Ich würde gerne was mit KI machen.“ Für Multi-Turn-Conversations werden im RAG-Prozess neben der letzten Anfrage auch Auszüge der History einbezogen. 💡

Du bist Berufsberater mit einem guten Ü

Berücksichtige bei deiner Antwort auf di  
Informationen aus der Datenbank BERU  
EXAKT die in der Datenbank hinterlegte  
Bezeichnungen in der Datenbank nicht v

Am ehesten zur Nachricht passt aktuell

Folgende Beziehungen zu verwandten B

- 'KI-Prompter' erfordert Kompetenz 'KI
- 'Techniker/in - Informatik (KI und Robotik Programmierassistenten'
- 'KI-Engineer' erfordert Kompetenz 'KI
- 'KI-Manager/in' erfordert Kompetenz

Chatbot

Ich würde gerne was mit KI machen.

Prima! Wenn du gerne mit KI arbeiten möchtest, gibt es viele spannende Berufe, die du in Betracht ziehen könntest. Hier einige Vorschläge:

1. KI-Programmierassistent/in

2. KI-Prompter

3. Techniker/in - Informatik (KI und Robotik)

4. KI-Engineer

5. KI-Manager/in

Die Kompetenz "KI-Programmierassistent/in" ist eine Grundvoraussetzung für alle diese Berufe. KI-Prompter erstellen und überwachen Algorithmen für maschinelles Lernen. Techniker/innen für Informatik (KI und Robotik) entwickeln und warten

Retry

Undo


Clear

Type a message...

Submit

# RITE Method for Enhancing the Context of LLMs



**WIKIPEDIA**  
The Free Encyclopedia

Q Search Wikipedia

Search

**RITE Method** 🌐 1 language

**Contents** hide

[\(Top\)](#)  
[See also](#)  
[References](#)

[Article](#) [Talk](#)

[Read](#) [Edit](#) [View history](#) Tools

From Wikipedia, the free encyclopedia

**RITE Method**, for **Rapid Iterative Testing and Evaluation**,<sup>[1]</sup> typically referred to as "RITE" testing, is an iterative [usability](#) method. It was defined by Michael Medlock, Dennis Wixon, Bill Fulton, Mark Terrano and Ramon Romero. It has been publicly championed by Dennis Wixon<sup>[2]</sup> while working in the games space for [Microsoft](#).

It has many similarities to "traditional"<sup>[3]</sup> or "discount"<sup>[4]</sup> [usability testing](#). The tester and team must define a target population for testing, schedule participants to come into the lab, decide on how the users' behaviors will be measured, construct a test script and have participants engage in a verbal protocol (e.g. think aloud). However it differs from these methods in that it advocates that changes to the user interface are made as soon as a problem is identified and a solution is clear. Sometimes this can occur after observing as few as one participant. Once the data for a participant has been collected the usability engineer and team decide if they will be making any changes to the prototype prior to the next participant. The changed interface is then tested with the remaining users.

The philosophy behind the RITE method is described as: "1) once you find a problem, solve it as soon as you can, and 2) make the decision makers part of the research team."<sup>[5]</sup> In this way it is a bridge between a strict research method and a design method...and in many ways it represents a participatory design method. Since its official definition and naming its use has rapidly expanded to many other software industries,<sup>[6]</sup> including interface design research.<sup>[7]</sup>

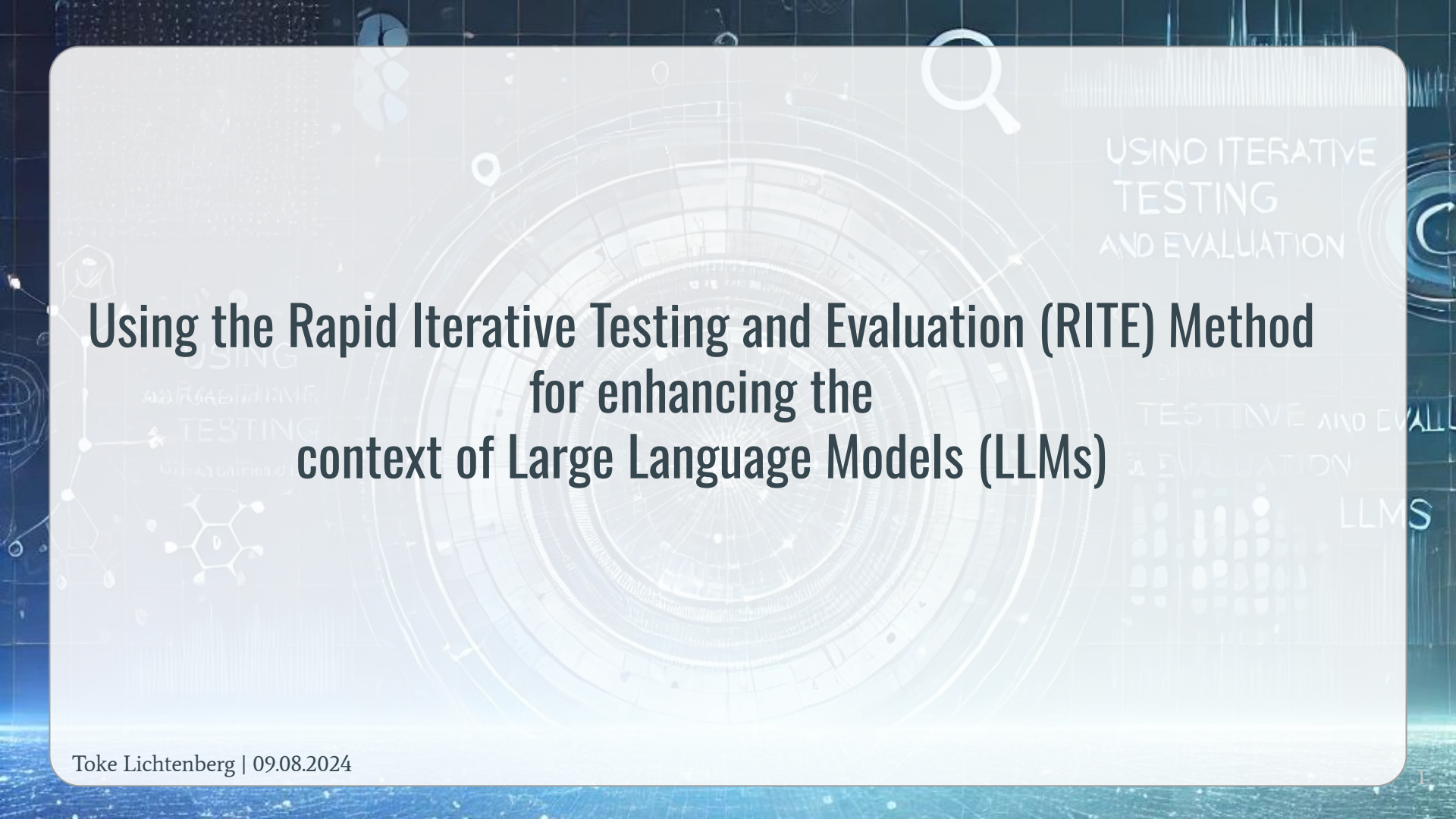
# Tokes Repo zum Fokus-Thema

---



Hier für Interessierte der Link zu Tokes RITE Repository:

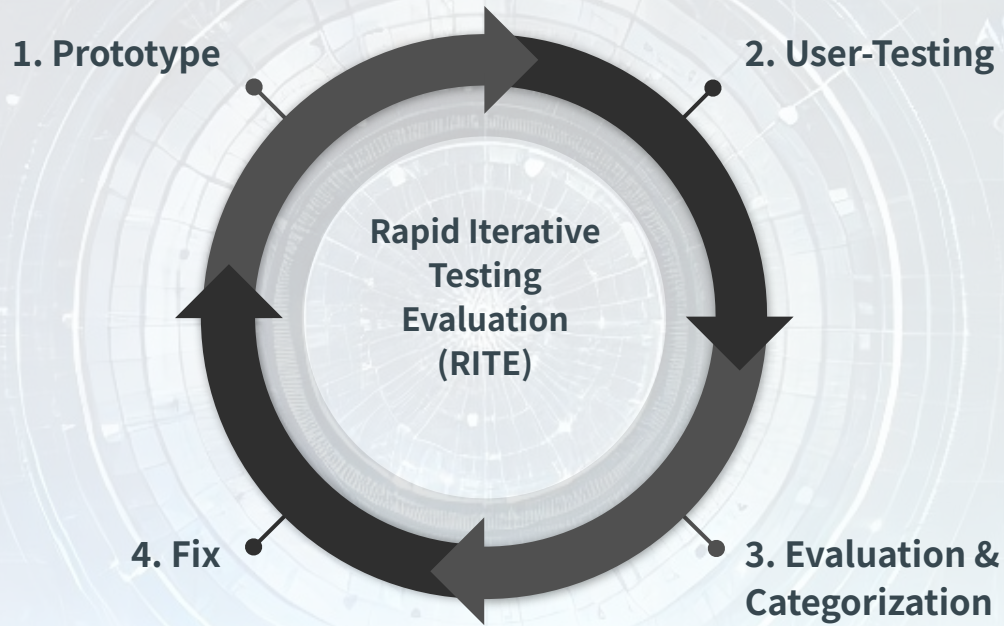
<https://gitlab.com/tokx/llm-rite/-/tree/main>



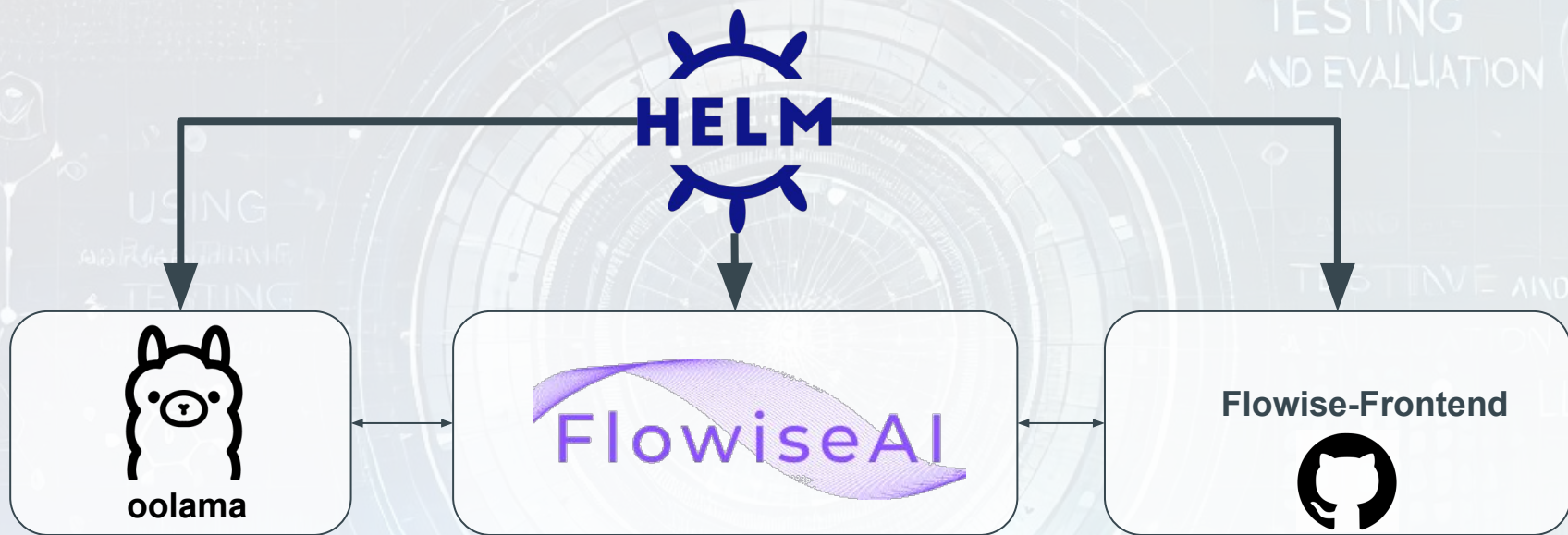
# Using the Rapid Iterative Testing and Evaluation (RITE) Method for enhancing the context of Large Language Models (LLMs)



# The RITE Method



# 1. Prototype & Infrastructure Setup





## 2. User-Testing Procedure



- Give an introduction
- Remember users to “think aloud”
- Respond to a user question with a counter question like: “what do you expect {user\_question}”
- Categorize:

Error

Failure

Feature

## 2. User-Testing Procedure



Error

Issues where the user can not continue the test

Failure

Issues where the user can continue but there is a misunderstanding

Feature

Issues where the user can continue but proposes an improvement

# 2. User-Testing Procedure

Open

21

Vorschlag Testperson: Ging es, in den Nachrichten der KI Antworten einfach auszuwählen?

Feature

#28

KI wird einfach nicht konkret in den Antworten

Failure

#23

Custom Error Message

Error

#18

Der Begrüßungstext beinhaltet eine Frage, deren Antwort aber ignoriert wird

Failure

#24

Person versteht eigenen Avatar nicht

Failure

#31

Vorschlag Testperson (wusste nicht, ob ich das als Failure labeln soll, ist schwer abzugrenzen...): Die Antwortfloskeln der KI nerven

Failure

#13

KI fragt zu viele Fragen auf einmal.

Failure

#16

Vorschlag Testperson: Auswahl zwischen "effizientem" und "kommunikativem" Weg der Unterhaltung mit Bot

Feature

#29

A

B

C

Closed



### 3. Evaluation & Categorization of Issues

A

Issues with an obvious cause and an obvious solution that can be implemented in the prototype quickly. If the model should answer in another language then this would be one addition to the context.

B

Issues with an obvious cause and an obvious solution that can't be implemented quickly/by the time of the next test session.

C

Issues with no obvious cause and no obvious solutions.

# 3. Evaluation & Categorization

USING ITERATIVE

Open3

Vorschlag Testperson: Ging es, in den Nachrichten der KI Antworten einfach auszuwählen?  
Failure  
#28

KI wird einfach nicht konkret in den Antworten  
Failure  
#23

Der Begrüßungstext beinhaltet eine Frage, deren Antwort aber ignoriert wird  
Failure  
#24

5

Text Eingabefenster Missverständlich  
Failure  
#30

Vorschlag Testperson: Auswahl zwischen "effizientem" und "kommunikativem" Weg der Unterhaltung mit Bot  
#29

KI fragt zu viele Fragen auf einmal.  
Failure  
#16

Person versteht eigenen Avatar nicht  
Failure  
#31

Vorschlag Testperson (wusste nicht, ob ich das als Failure labeln soll, ist schwer abzugrenzen...): Die Antwortfloskeln der KI nerven  
Failure  
#13

3

KI gibt zum Teil Unwahrheiten wieder.  
Failure  
#14

Vorschlag Testperson: Spracheingabe  
#19

Vorschlag Testperson: Design etwas verändern  
#25

2

KI braucht zwischendrin recht lange zum Antworten  
Failure  
#22

Custom Error Message  
Error  
#18

Closed15

Vorschlag Testperson: Eingabefenster für eigenen Text sollte größer sein, um eine bessere Übersicht über das Geschriebene zu haben.  
#12

Nachfragen seitens der KI, um Bereich einzugrenzen, nur hilfreich für Personen, die schon wissen, was sie machen wollen  
Failure  
#27

Eingangsnachricht missverständlich  
Failure  
#26

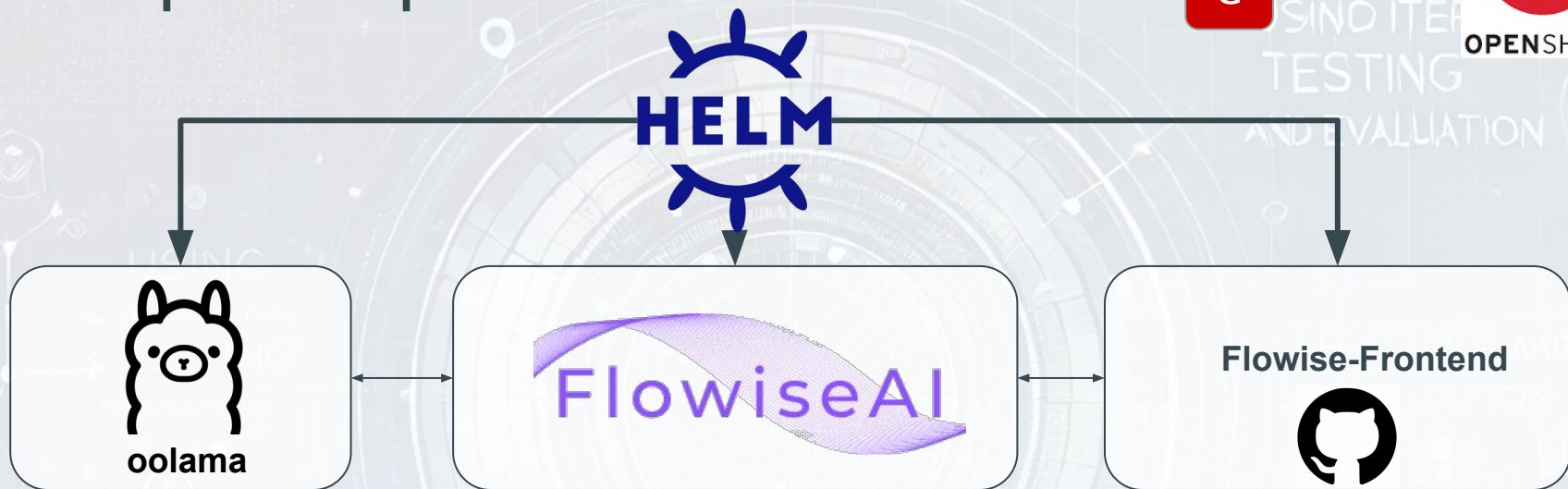
KI lässt die meiste Arbeit die Testperson selbst machen  
Failure  
#17

Eigene Anmerkung: Bitte Titel abändern! Wir machen immernoch keine Jobberatung! :P  
#10

KI vergisst zwischendrin schon erwähnte Ausbildungen  
Failure  
#21

## 4. Fix - possible options

- Analyze and fix disconnects **C**



- switch model (current: gemma2)

**A****A****B**

- Change flow
- Fix code

**A****B**


- Change context
- Fix UI

**A****B**

- scale services
- change backend




# Results - UI



Hi there! How can I help?



Hey wie geht es dir?


Error: Failed to fetch



Powered by **Flowise**



 **KIWI** 




Hallo, ich helfe dir dabei, passende Wege für deine Weiterbildung zu finden. Schreibe mir einfach unten, was du gerne beruflich machen wollen würdest, ob du konkrete Weiterbildungen suchst oder eine Orientierungshilfe brauchst und ich versuche, dir bestmöglich zu helfen. Oh und sollte ich zu viele Fragen stellen, konzentriere dich einfach auf das, was für dich am wichtigsten ist 😊

Ich bin mir unsicher was ich machen soll

Was interessiert dich beruflich?

Hier kannst du schreiben! Konzentriere dich auf die für dich wichtigsten Aspekte.



# Results - RITE Metrics

**Impact Ratio:** How many tickets could be fixed?

$$\text{impact\_ratio} = \text{fixed\_issues} / \text{issues} * 100 \\ = \mathbf{75\%}$$

**Refix Ratio:** How many tickets of the fixed tickets needed to be fixed again?

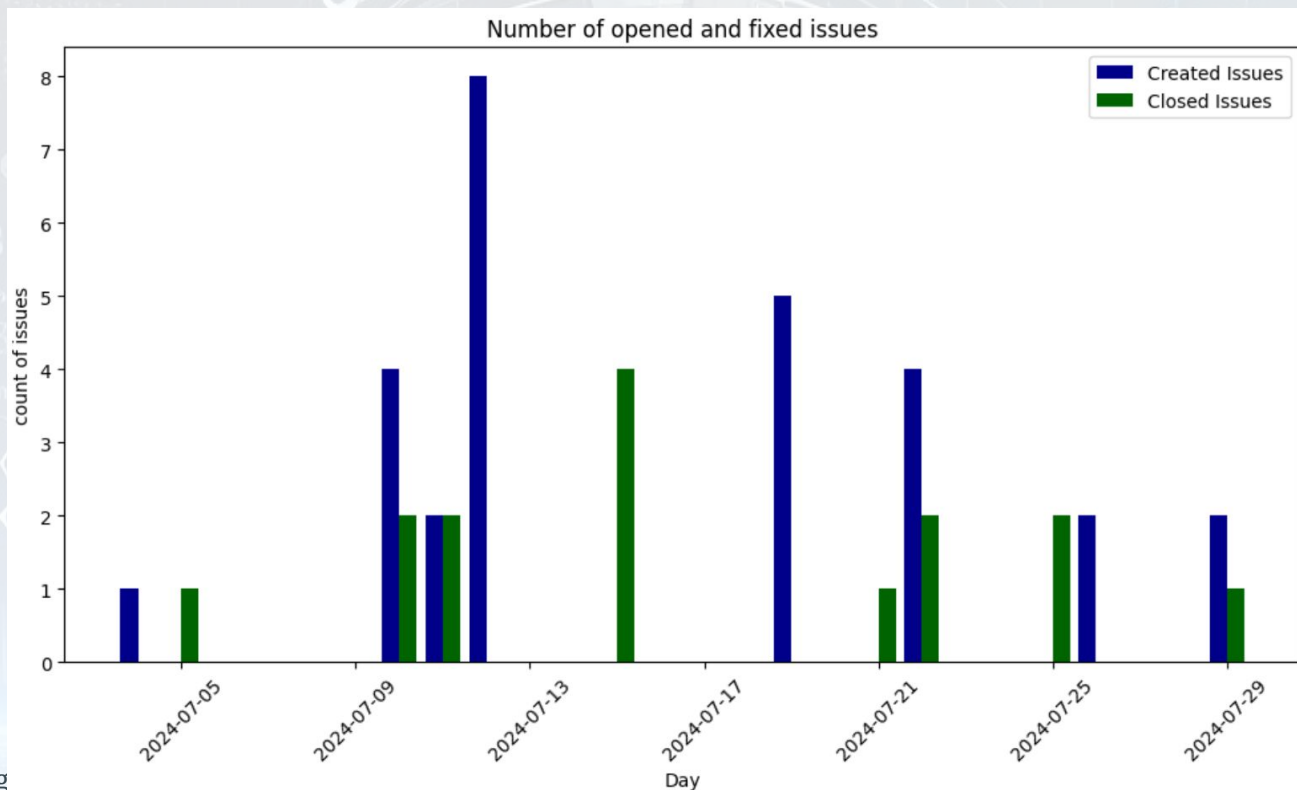
$$\text{refix\_ratio} = \text{refixed\_issues} / \text{fixed\_issues} * 100 \\ = \mathbf{28.57\%}$$



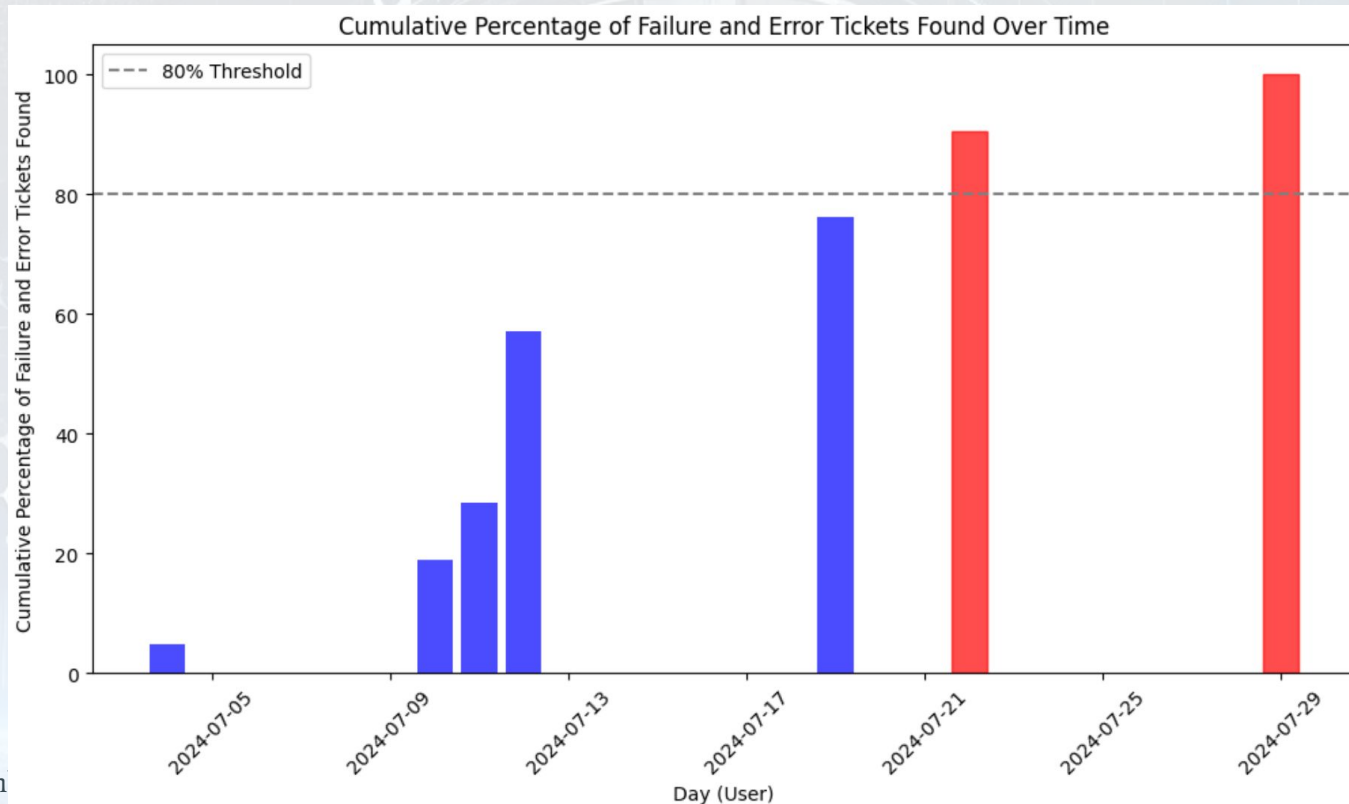
# Results - RITE Metrics - Which tickets needed to be refixed?

Title	Label	Refix-Count	Link
Vorschlag Testperson: Die Antwortfloskeln der KI nerven	A	3	<a href="https://gitlab.com/apps-lab/kiwi/analysis/rite-llm/-issues/13">https://gitlab.com/apps-lab/kiwi/analysis/rite-llm/-issues/13</a>
KI fragt zu viele Fragen auf einmal.	A	2	<a href="https://gitlab.com/apps-lab/kiwi/analysis/rite-llm/-issues/16">https://gitlab.com/apps-lab/kiwi/analysis/rite-llm/-issues/16</a>
Angegebene Links funktionieren nicht	A	2	<a href="https://gitlab.com/apps-lab/kiwi/analysis/rite-llm/-issues/11">https://gitlab.com/apps-lab/kiwi/analysis/rite-llm/-issues/11</a>
Wieso kann ich nur 50 Buchstaben eingeben?	A	2	<a href="https://gitlab.com/apps-lab/kiwi/analysis/rite-llm/-issues/8">https://gitlab.com/apps-lab/kiwi/analysis/rite-llm/-issues/8</a>
Vorschlag: Eine Überschrift/ein Label für das Tool einfügen	A	2	<a href="https://gitlab.com/apps-lab/kiwi/analysis/rite-llm/-issues/7">https://gitlab.com/apps-lab/kiwi/analysis/rite-llm/-issues/7</a>

# Results - Statistics

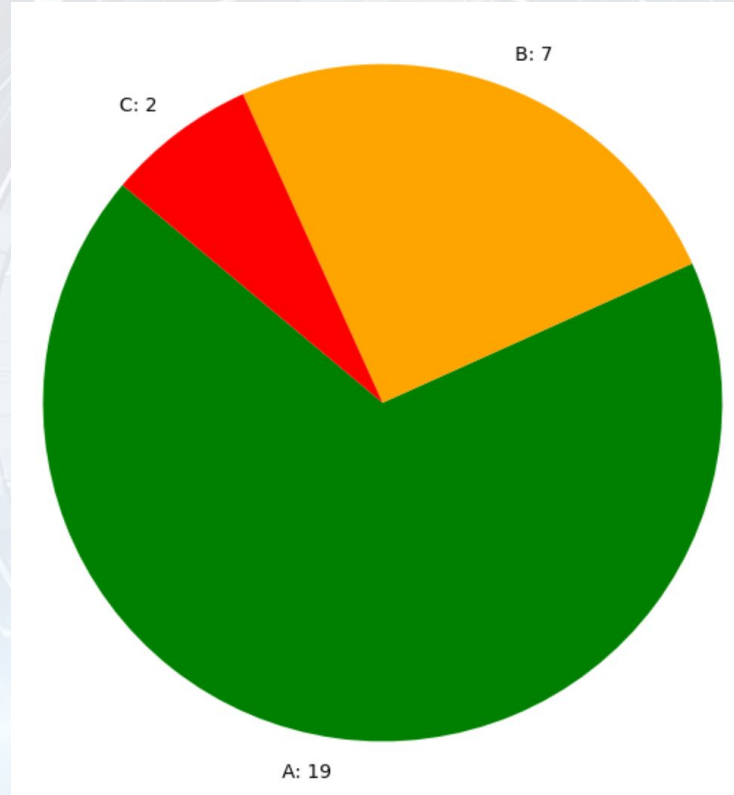


# Results - Do 3 users find 80% of the usability issues?

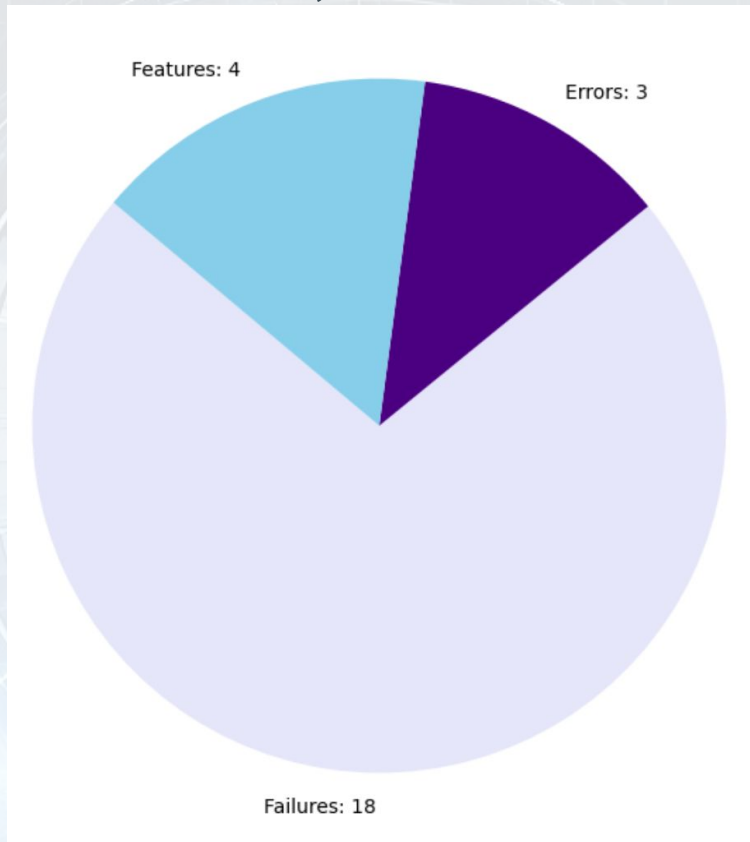




# Results - Distribution of categories



# Results - Distribution of failures, errors and features





**Thank you for your patience**