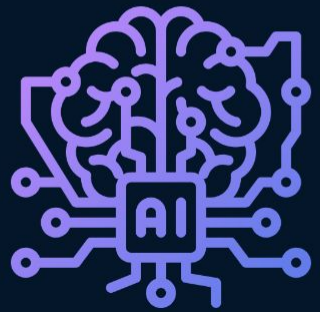
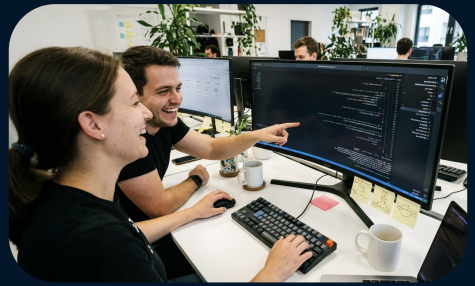
METRO digital

Chatbot Evaluation

“You can’t manage, what you can’t measure”



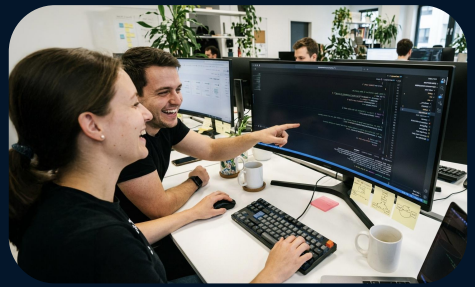
Problem



AI Developer

Unser neuer FAQ Agent soll live gehen

Problem



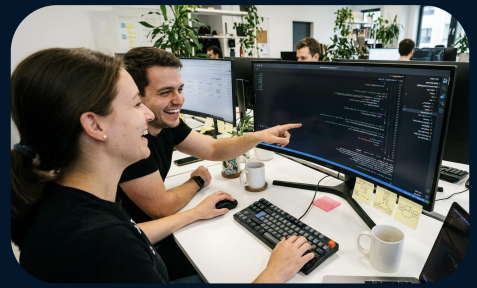
AI Developer

Danke… war etwas besser!

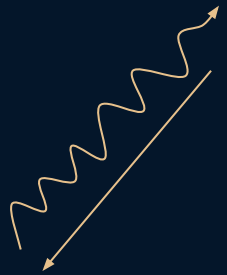
Neue Version, bitte testen!

Business Unit

Problem



AI Developer



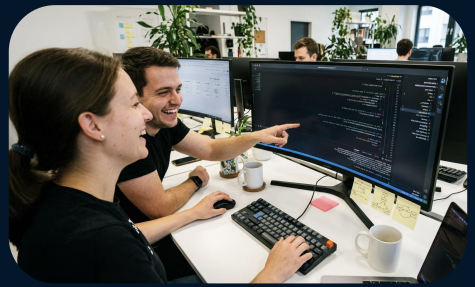
Go life approved?

Business Unit

Manuelle Eingabe der Testfragen
Kein hilfreiches Feedback
Keine Benchmarks

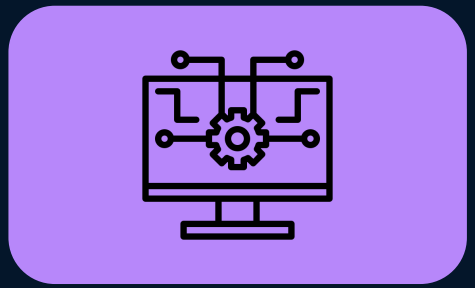
Board

Lösung



AI Developer

Go life approved?

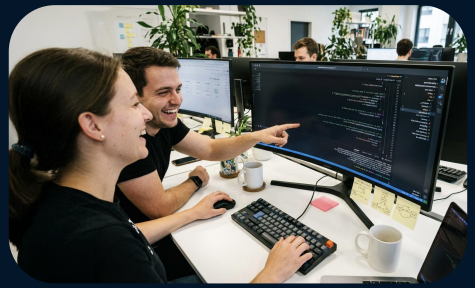


AI Evaluation Tool

😍 **Automatisiertes Testing**
😍 **Detaillierte Analyse**
😍 **Messbare KPIs**

Board

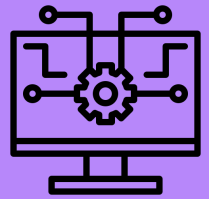
Lösung



AI Developer

Go life approved?

😍 **Go Live!**



AI Evaluation Tool

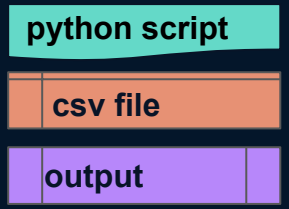
😍 **Automatisiertes Testing**
😍 **Detaillierte Analyse**
😍 **Messbare KPIs**

Board

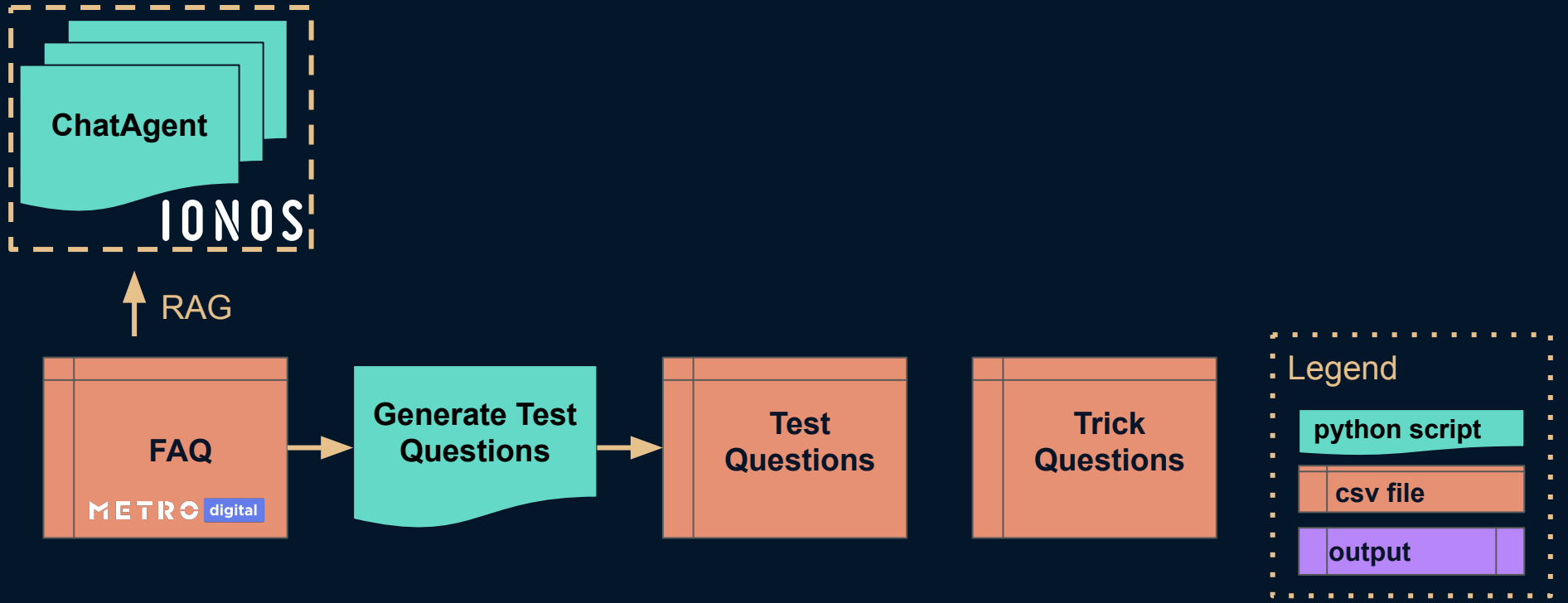
LIVE Demo

This is where the magic happens 🪄

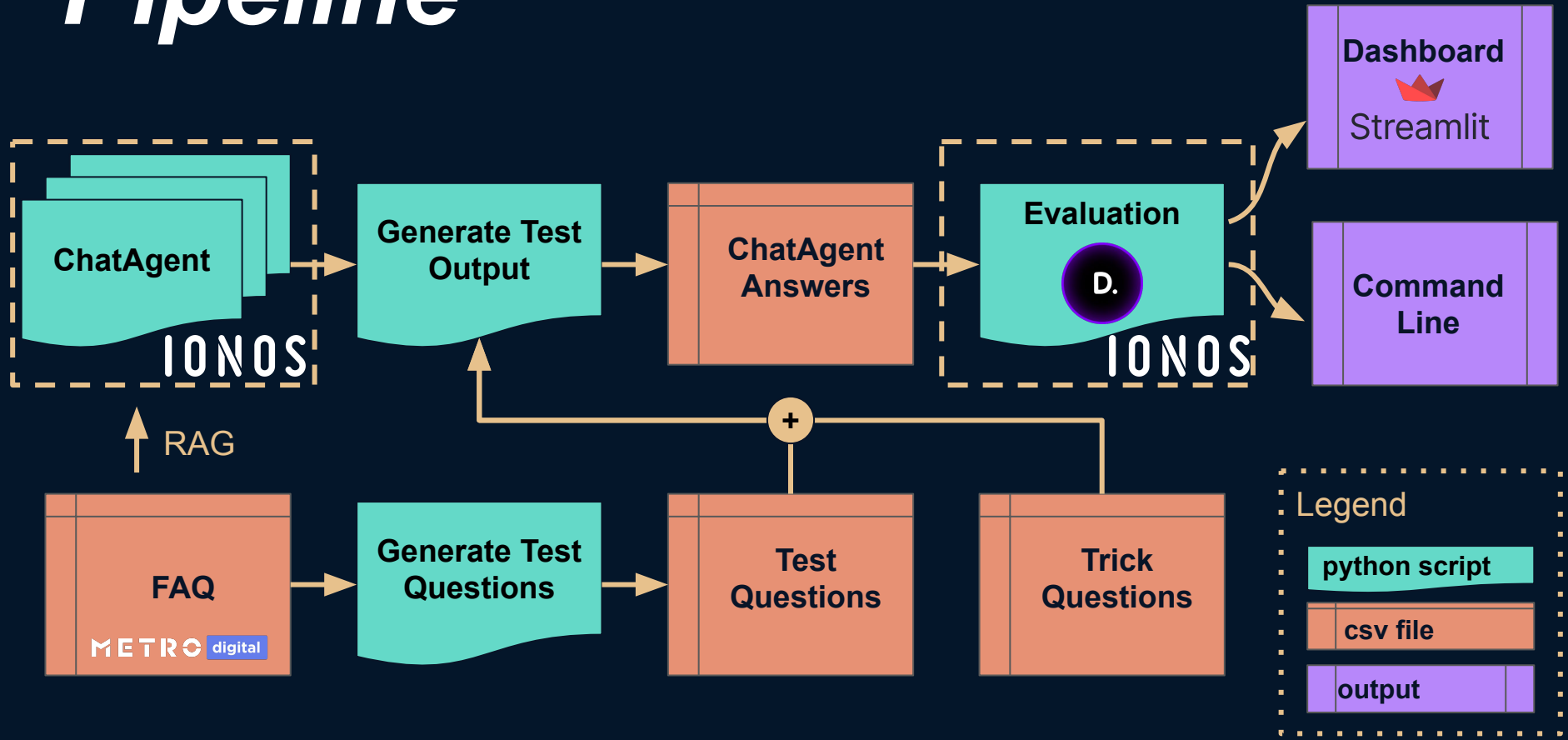
Pipeline



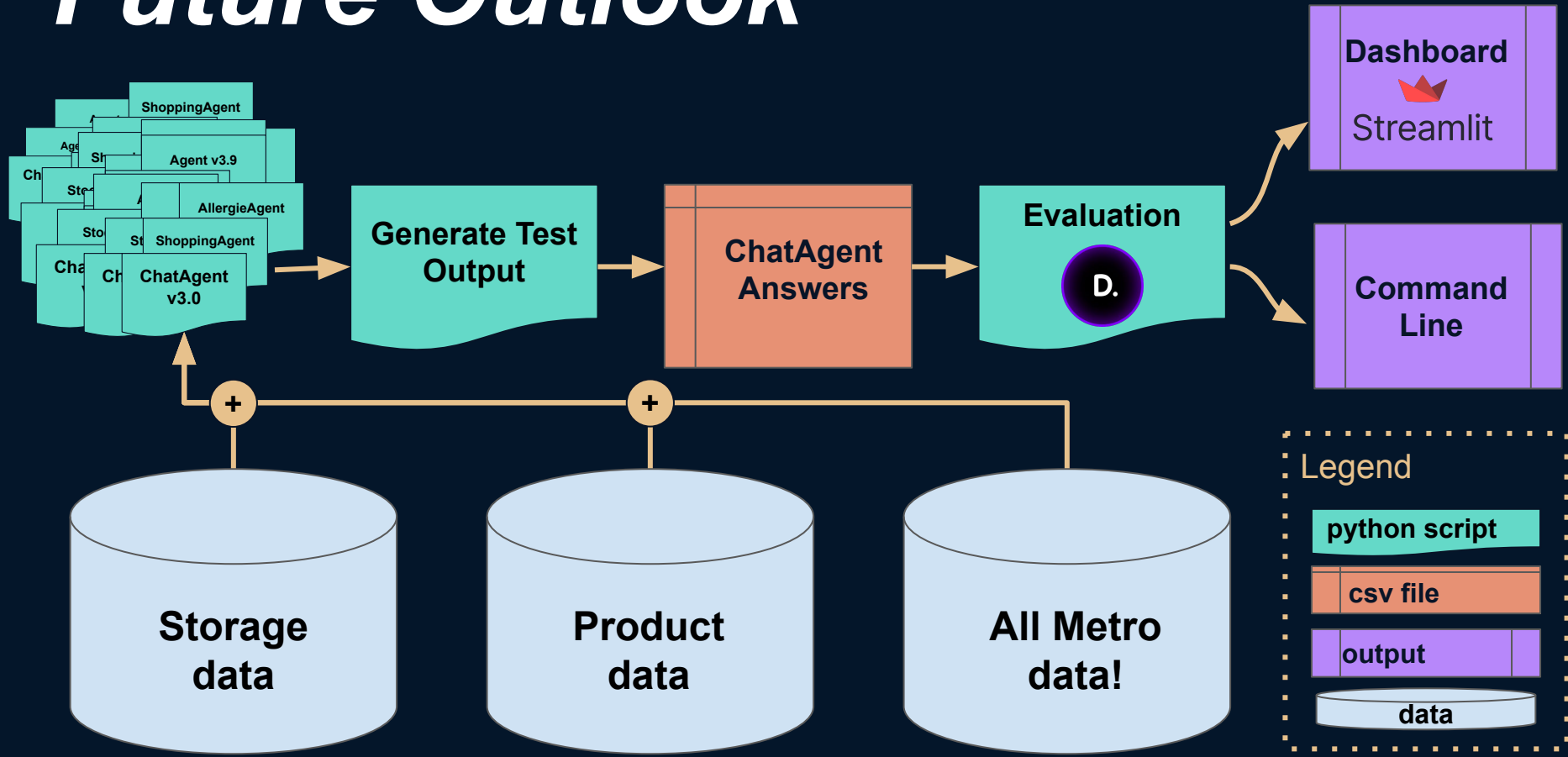
Pipeline


ChatAgent
IONOS
RAG
FAQ
METRO digital
Generate Test Questions
Test Questions
Trick Questions
Legend
python script
csv file
output

Pipeline


ChatAgent
IONOS
Generate Test Output
ChatAgent Answers
Evaluation
D.
IONOS
Dashboard
Streamlit
Command Line
RAG
+
FAQ
METRO digital
Generate Test Questions
Test Questions
Trick Questions
Legend
python script
csv file
output

Future Outlook



A success story

METRO digital & IONOS

Manage What You Measure.

Hauke Diers

Domitille Grandjean

Nico Neubauer

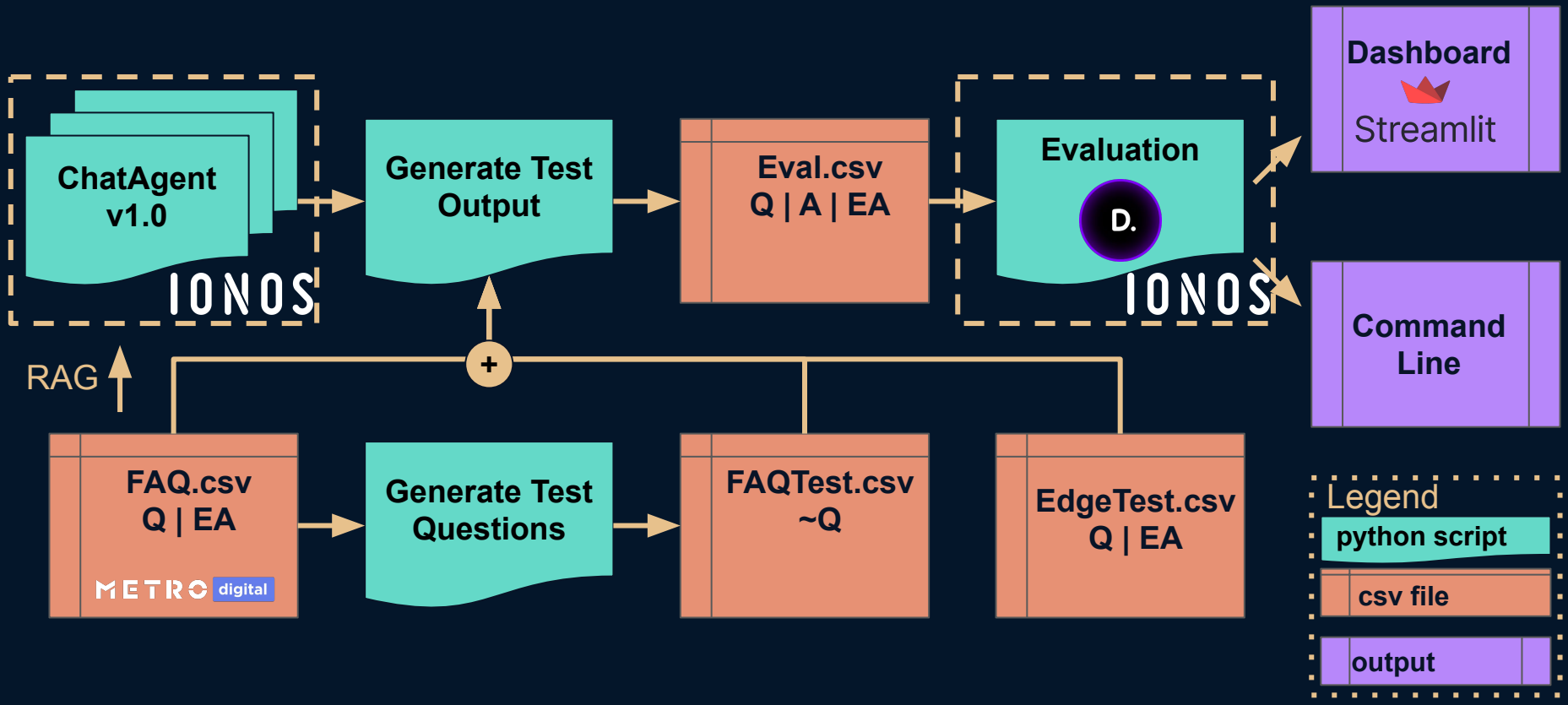
Andres Navarro

Omid Roshenas

Evaluation Metrics

1. **Correctness:** Sind die angegebenen Informationen wahr?
2. **Clarity:** Ist die Antwort leicht zu verstehen?
3. **Hospitality & Tonality:** Ist der Ton professionell?
4. **Relevance:** Beantwortet er die gestellte Frage?
5. **Hallucination Safety:** Werden Informationen erfunden?

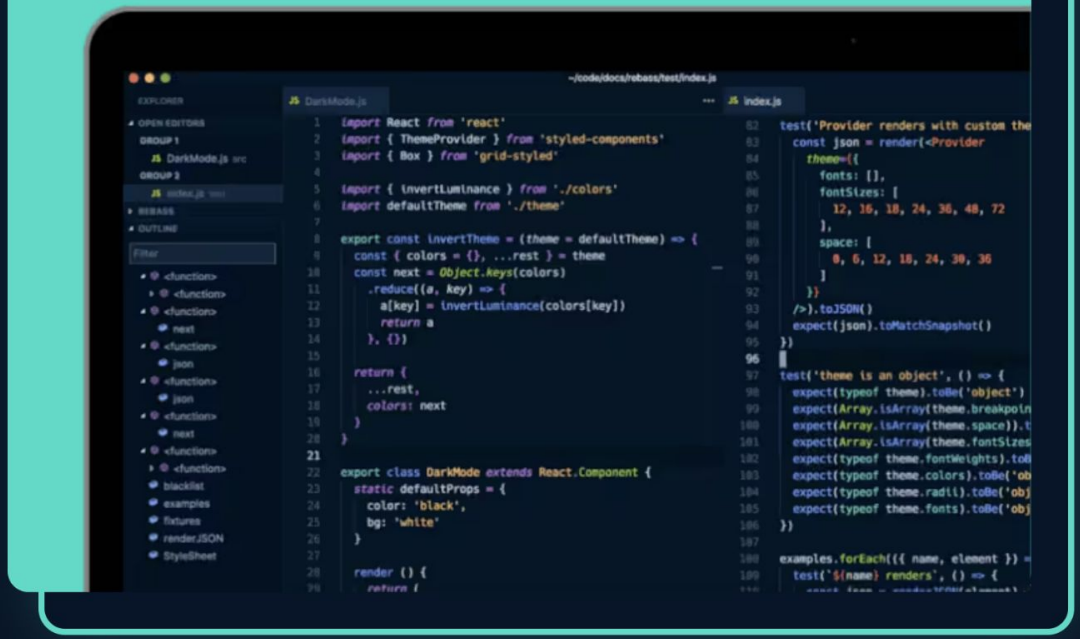
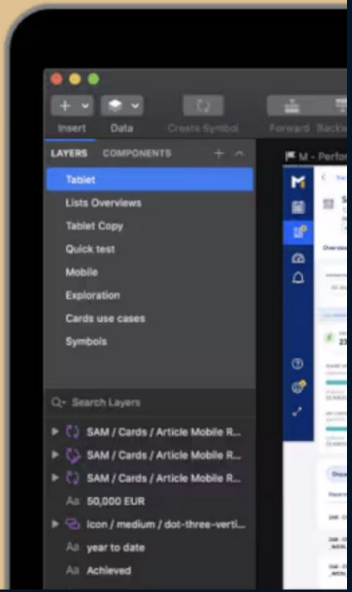
Future Outlook



Skalierbarkeit

Naheliegende Use cases:
Dish:

Broader use cases für weitere Integration:
Evaluation Robotics AI - Accuracy of order preparation
Maintenance check-up for Obi’s new storage tracker

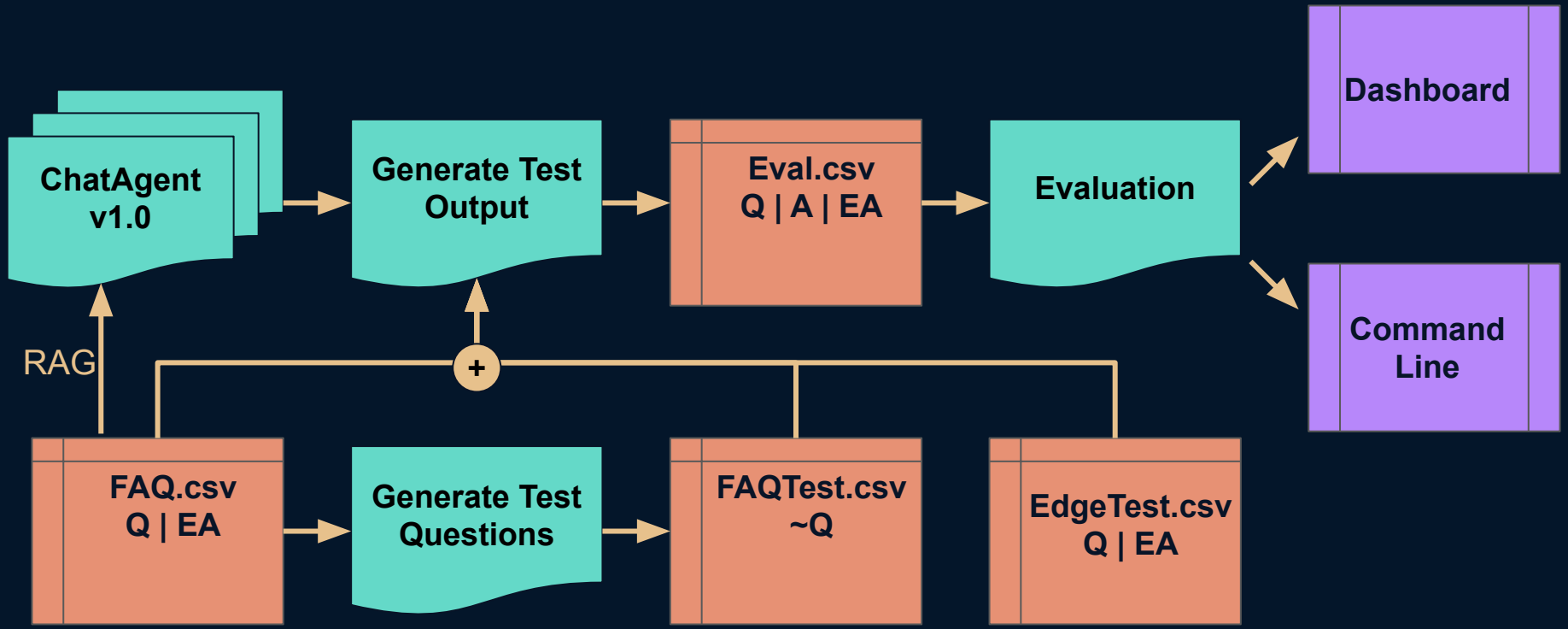

~/code/docs/rebass/test/index.js
DarkMode.js
index.js
import React from 'react'
import { ThemeProvider } from 'styled-components'
import { Box } from 'grid-styled'
import { invertLuminance } from './colors'
import defaultTheme from './theme'
export const invertTheme = (theme = defaultTheme) => {
const { colors = {}, ...rest } = theme
const next = Object.keys(colors)
.reduce((a, key) => {
a[key] = invertLuminance(colors[key])
return a
}, {})
return {
...rest,
colors: next
}
}
export class DarkMode extends React.Component {
static defaultProps = {
color: 'black',
bg: 'white'
}
render () {
test('Provider renders with custom the
const json = render(<Provider
theme={{
fonts: [],
fontSizes: [
12, 16, 18, 24, 36, 48, 72
],
space: [
0, 6, 12, 18, 24, 30, 36
]
}}
/>).toJSON()
expect(json).toMatchSnapshot()
})
test('theme is an object', () => {
expect(typeof theme).toBe('object')
expect(Array.isArray(theme.breakpoin
expect(Array.isArray(theme.space)).t
expect(Array.isArray(theme.fontSizes
expect(typeof theme.fontWeights).toB
expect(typeof theme.colors).toBe('ob
expect(typeof theme.radii).toBe('obj
expect(typeof theme.fonts).toBe('obj
})
examples.forEach(({ name, element }) =
test(`${name} renders`, () => {

Insert
Data
LAYERS
COMPONENTS
Tablet
Lists Overviews
Tablet Copy
Quick test
Mobile
Exploration
Cards use cases
Symbols
Search Layers
SAM / Cards / Article Mobile R...
SAM / Cards / Article Mobile R...
SAM / Cards / Article Mobile R...
50,000 EUR
Icon / medium / dot-three-verti...
year to date
Achieved

Welche Metriken Evaluation

Correctness
Clarity
Hospitality / Tonality
Relevance
Hallucination Safety

Each scale 1/5

Pipeline


ChatAgent
v1.0
Generate Test
Output
Eval.csv
Q | A | EA
Evaluation
Dashboard
Command
Line
RAG
+
FAQ.csv
Q | EA
Generate Test
Questions
FAQTest.csv
~Q
EdgeTest.csv
Q | EA