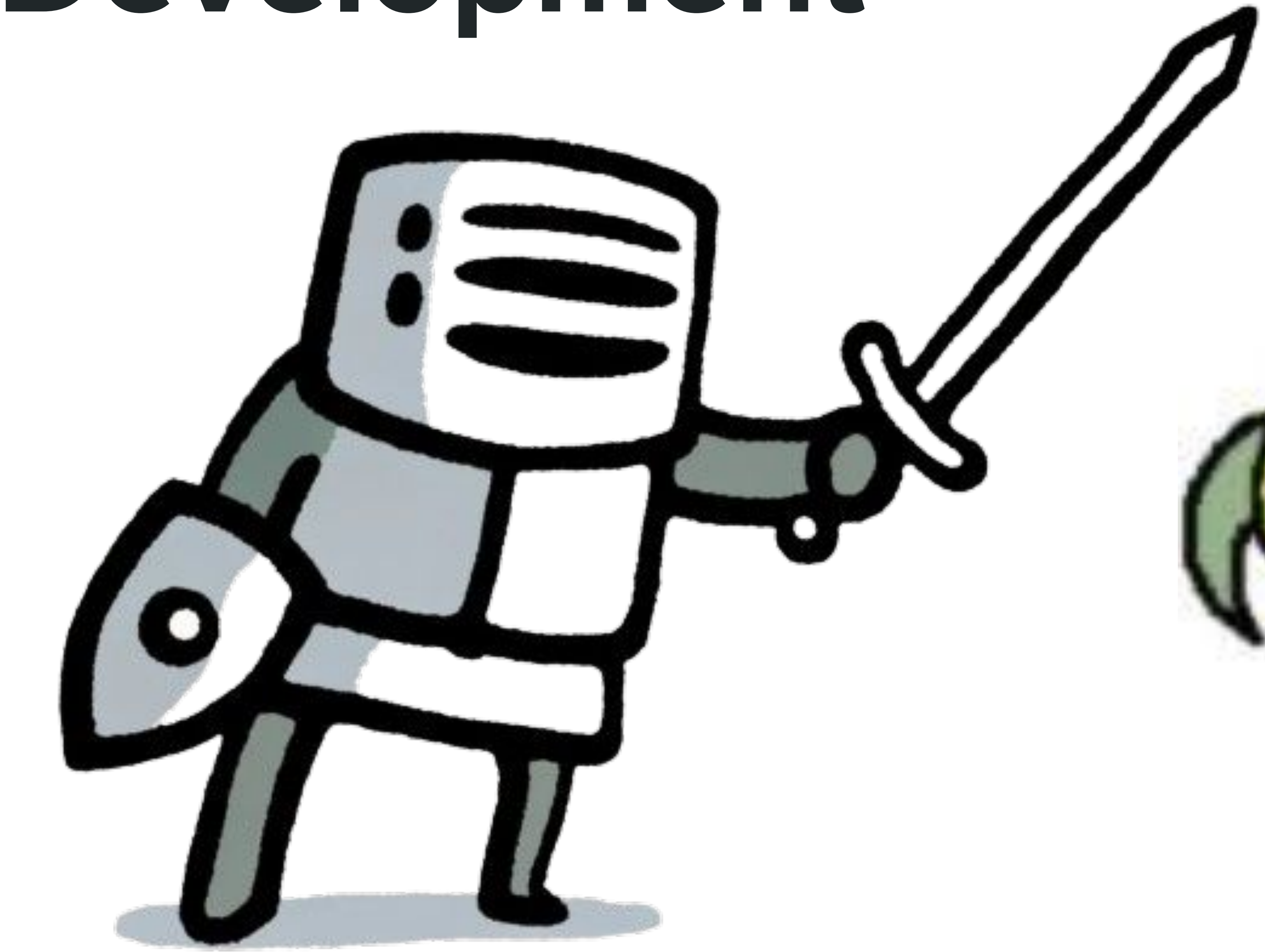
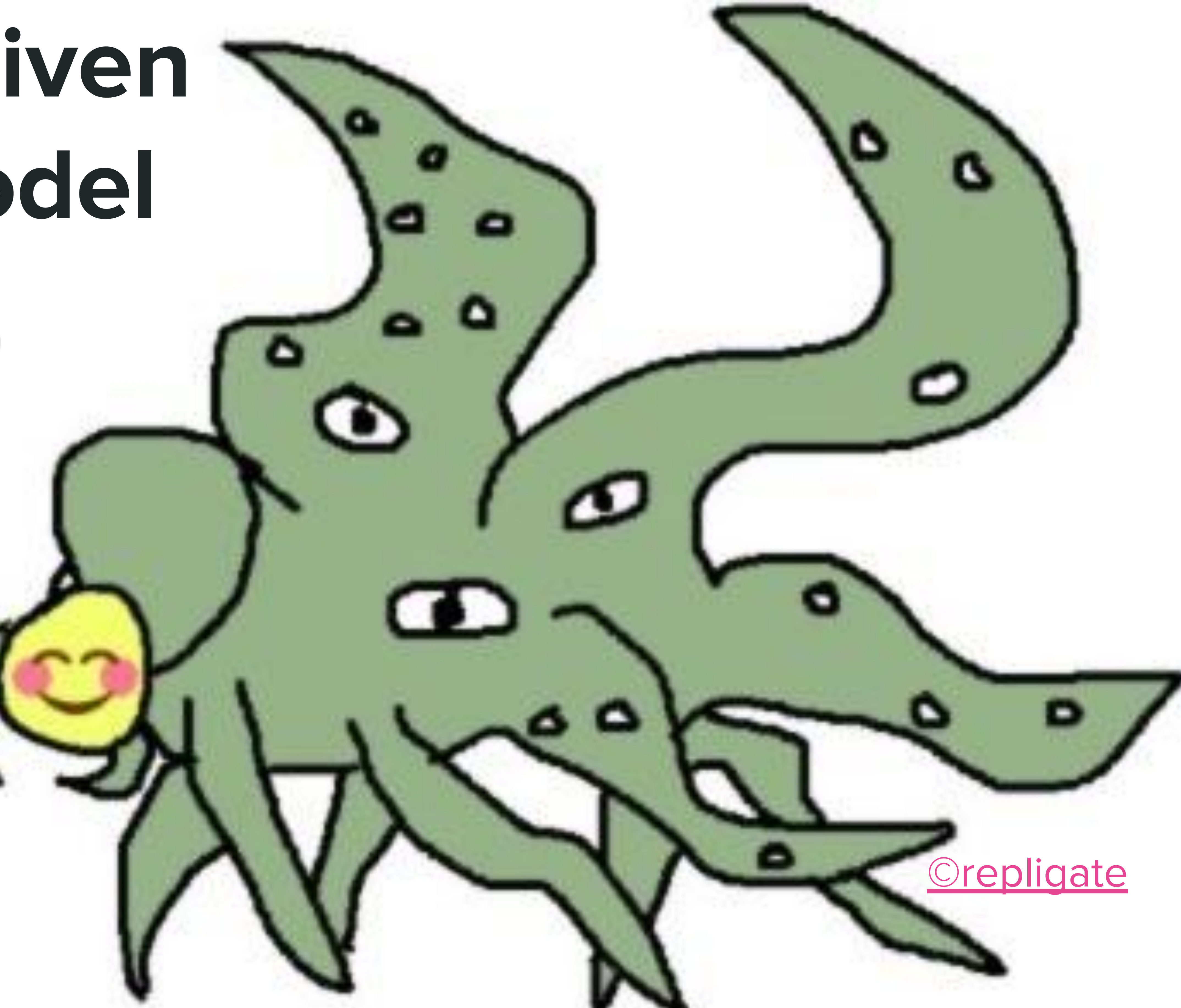
Acceptance Test-Driven Large Language Model Development



ATD^{LLM}D



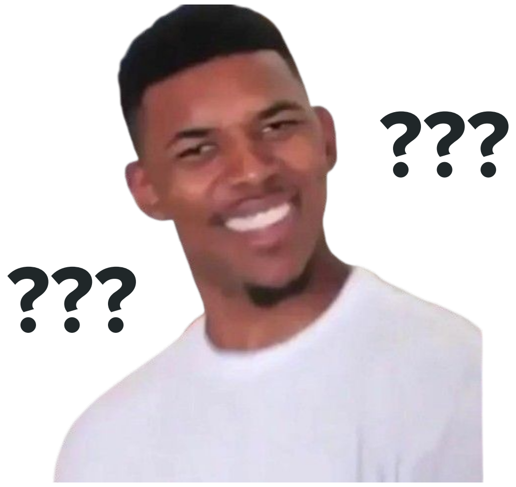
©repligate

AI Shoggoth with Smiley Face[AIMeme]

David Faragó (dfarago@mediform.io) Mediform & QPR Technologies GI TAV 49, February 15th 2024

Agenda

Motivation



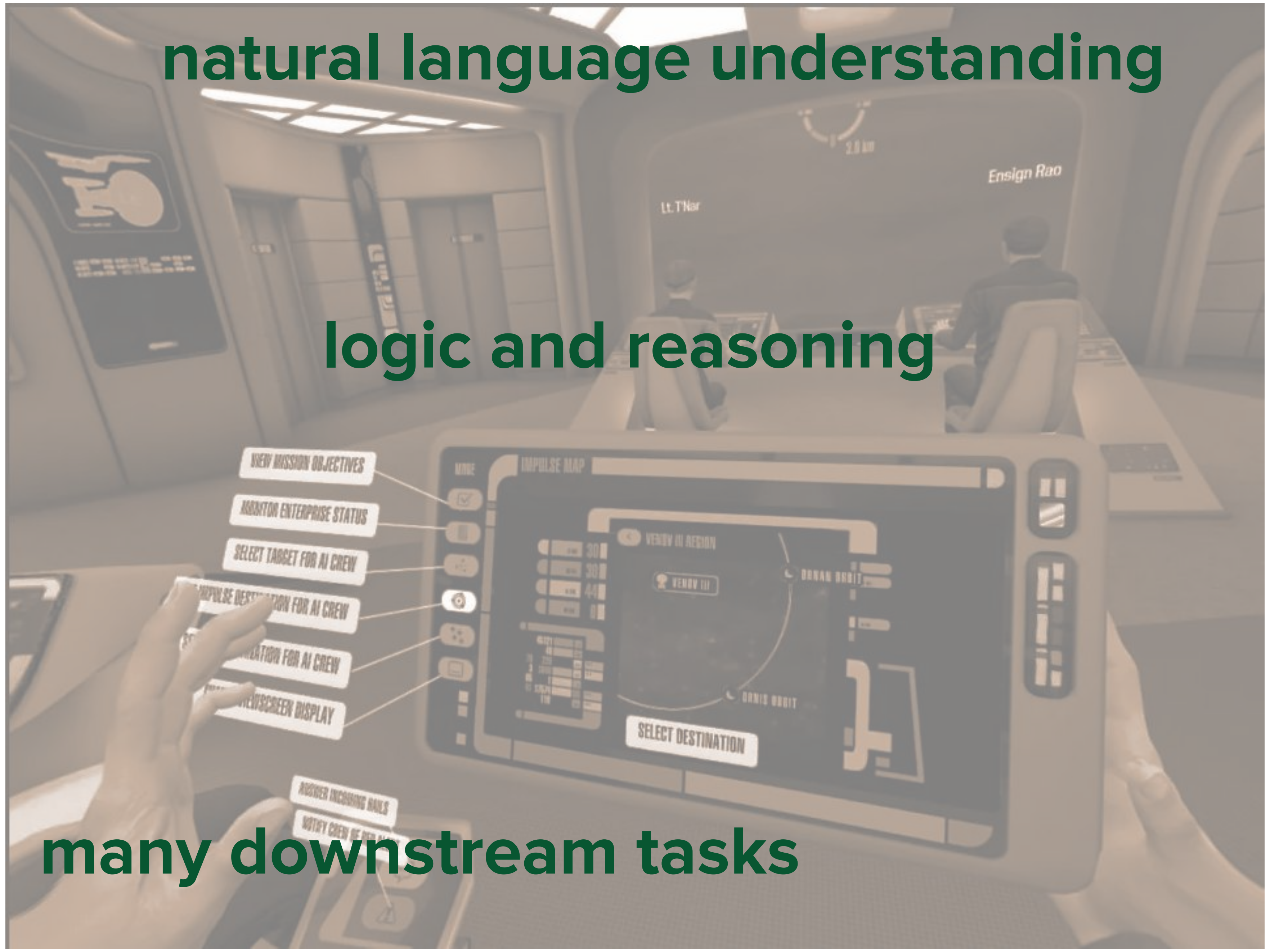
- **Bad LLM Development**
- **Root Causes for Bad LLM Development**
- **Three Tasks to Resolve Root Causes**

Solutions

- **Cognitive Project Management For AI (and ATDD)**
- **Dialog-based Business & Data Understanding**
- **LM-Eval**
- **ATD$^{\text{LLM}}$D**

Conclusion

LLMs: A Technology Gifted by Aliens Without a Manual[Gra23]



natural language understanding

logic and reasoning

many downstream tasks

LLMs

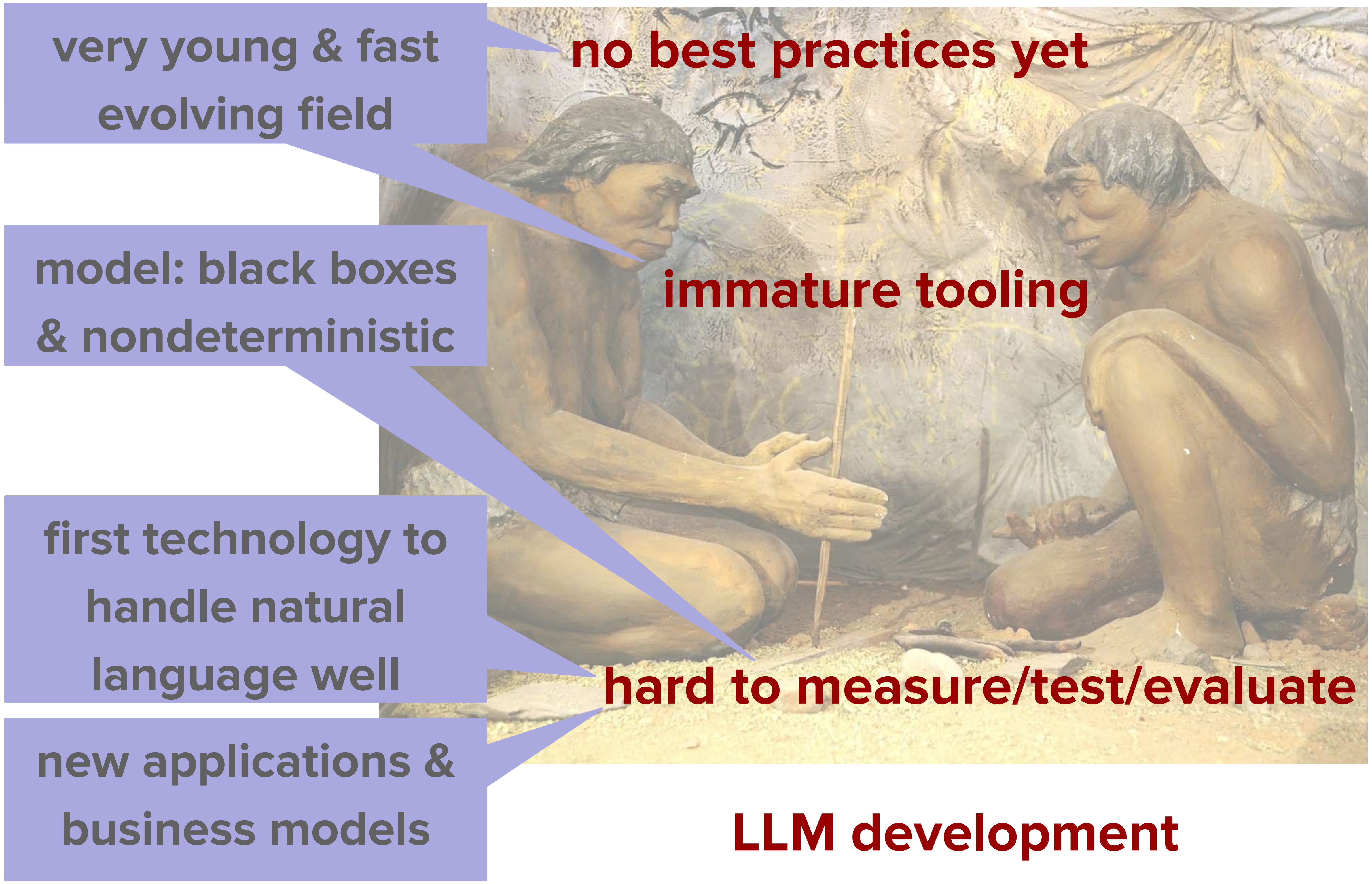
no best practices yet

immature tooling

hard to measure/test/evaluate

LLM development

Root Causes For Bad LLM Development



very young & fast evolving field

model: black boxes & nondeterministic

first technology to handle natural language well

new applications & business models

no best practices yet

immature tooling

hard to measure/test/evaluate

LLM development

Three Tasks to Resolve Root Causes

T1: Merge processes & best practices from modern development & data-centric ML

T2: Validation: understand business (data)

T3: Verification: evaluate LLM output

very young & fast evolving field

model: black boxes & nondeterministic

first technology to handle natural language well

new applications & business models

no best practices yet

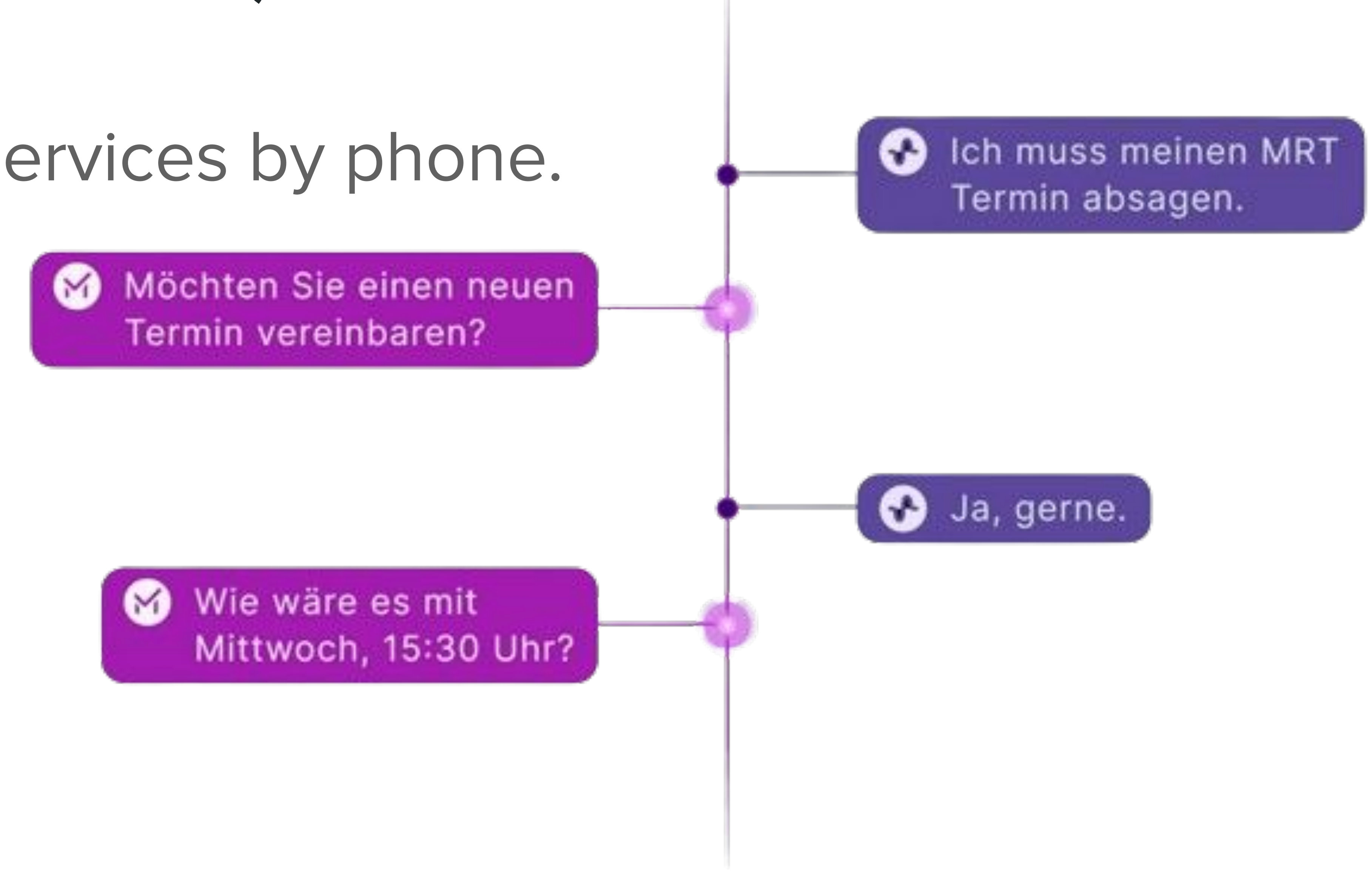
immature tooling

hard to measure/test/evaluate

LLM development

T1 (Merge processes & best practices) at Mediform

MediVoice autonomously manages patient services by phone.



Ich muss meinen MRT Termin absagen.

Möchten Sie einen neuen Termin vereinbaren?

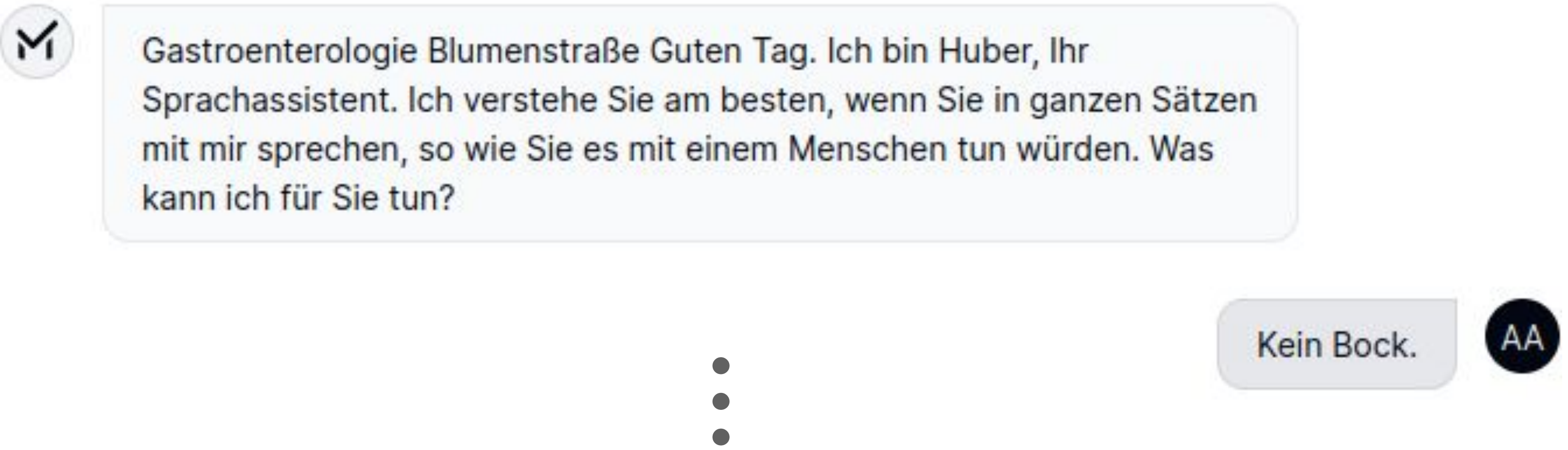
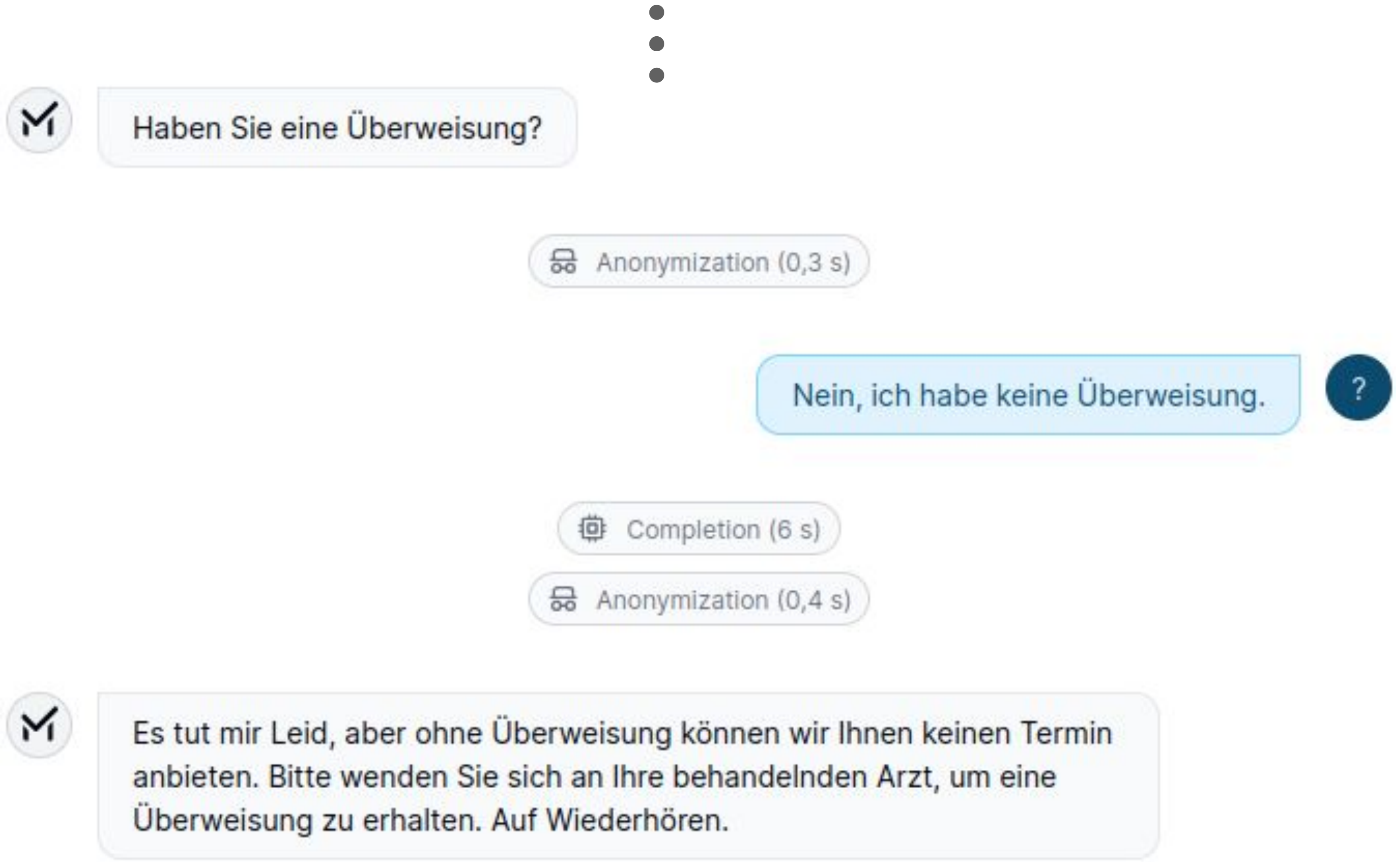
Ja, gerne.

Wie wäre es mit Mittwoch, 15:30 Uhr?

- Data-centric
  - thousands of medical practice dialogues
    - Anonymized real dialogues
    - Non-AI generated dialogues
    - AI generated dialogues
- Machine-learning
  - Prompt engineering
  - Fine-tuning LLM
  - Retrieval and agentic behavior
- Agile
  - New app into uncharted realm: stakeholders don’t know yet what they want
  - Short feedback cycles with stakeholders (medic, medical assistant, MVZ, call-center, patient)

T2 (Validation: Understand Business) at Mediform

Completely new application advancing into uncharted realm

How do patients behave?	**What do practices want?**
Weird behaviors, e.g. angry and impatient:  Gastroenterologie Blumenstraße Guten Tag. Ich bin Huber, Ihr Sprachassistent. Ich verstehe Sie am besten, wenn Sie in ganzen Sätzen mit mir sprechen, so wie Sie es mit einem Menschen tun würden. Was kann ich für Sie tun? Kein Bock. 	Often different from making patients happy:  Haben Sie eine Überweisung? Nein, ich habe keine Überweisung. Es tut mir Leid, aber ohne Überweisung können wir Ihnen keinen Termin anbieten. Bitte wenden Sie sich an Ihre behandelnden Arzt, um eine Überweisung zu erhalten. Auf Wiederhören.
Unexpected behaviors e.g. 80+ year olds interact more efficient[MediformBlog	Each practice has its individual rules e.g. whether to hand out referrals

T3 (Verification: Evaluate LLM Output) at Mediform

Business-centric verification

- Metrics for practices’ needs
- Per practice business processes

Verify a model’s natural language understanding

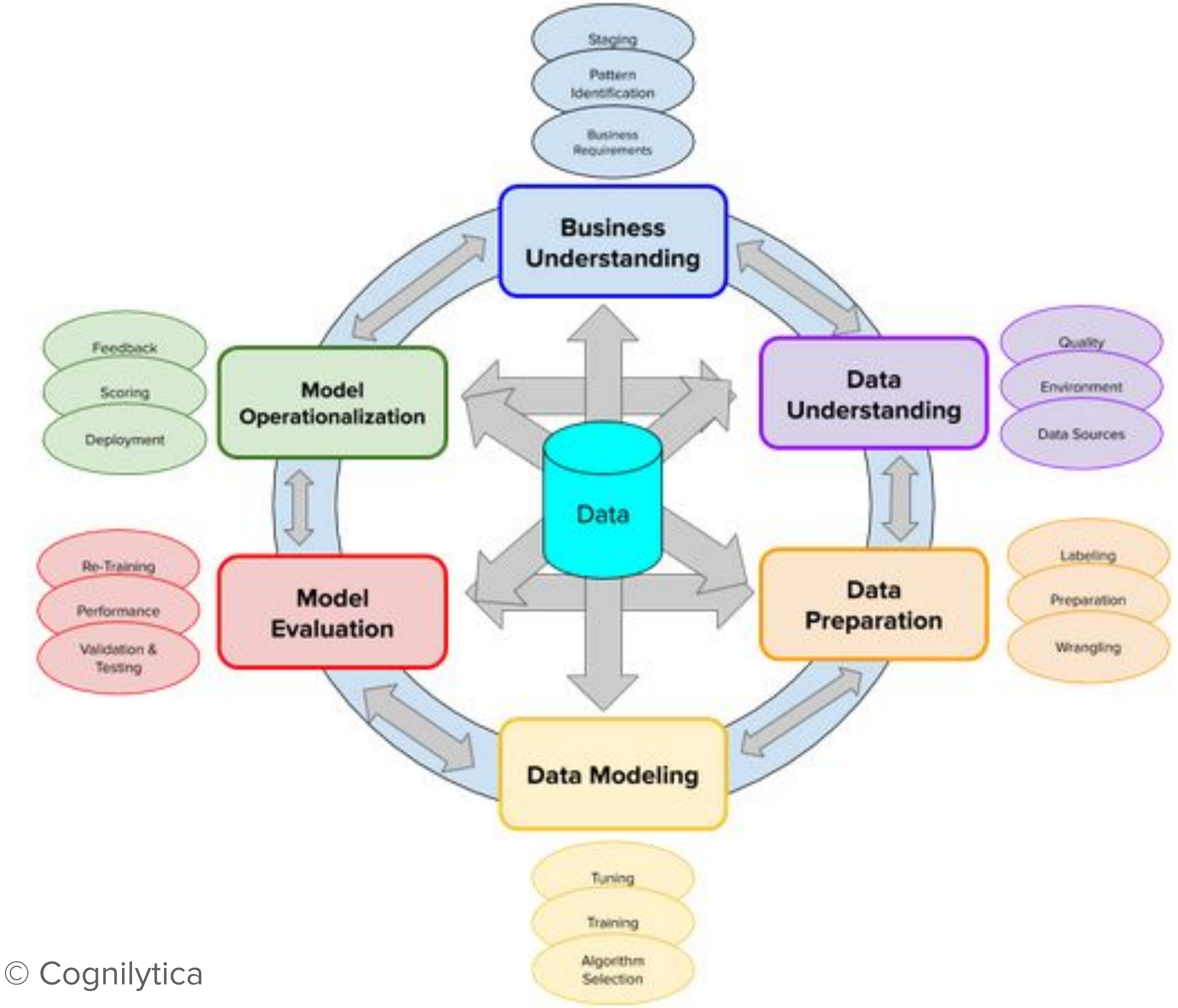
- Dialogs
  - Generalizes and handles domain specific corner cases?
  - Broken language, multiple languages, STT errors?
- Per practice business processes in natural language?

Verify a black box and nondeterministic model

- Understand variations in output
- Test statistically

Solutions

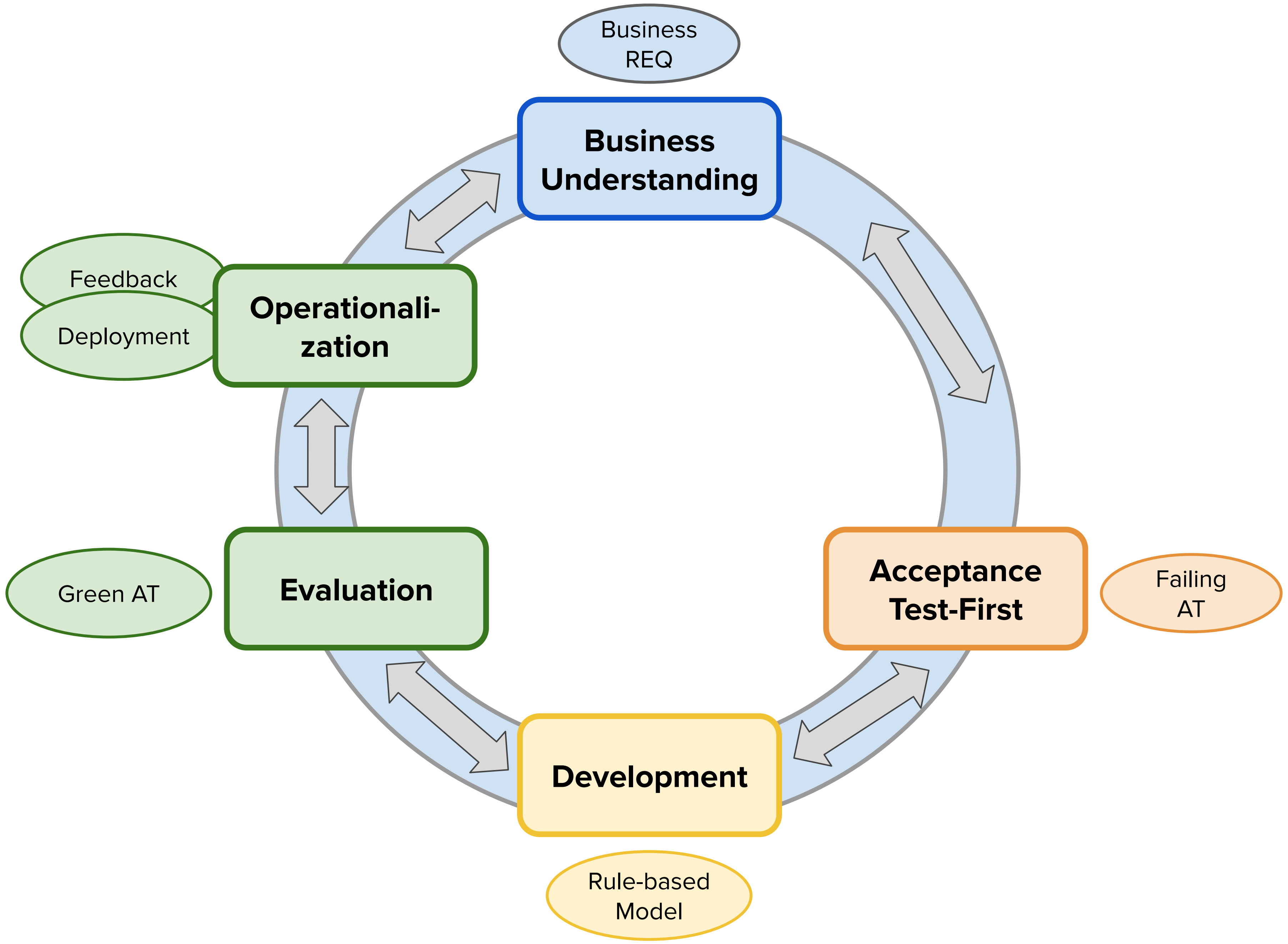
Solution for T1: CPMAI by Cognilytica[CPMAI]



© Cognilytica

- data-centric
  - Data at its core, for each phase
  - Inspired by CRISP-DM[CRISP-DM1999]
  - Embeds up-to-date data science best practices
- tailored to AI
  - Adds specific details for AI projects
  - Aligned with seven patterns to AI[7Patterns]
  - Embeds up-to-date ML best practices
- agile
  - Iterative and flexible
  - Business feedback in each cycle
  - Embeds up-to-date developer best practices

Acceptance Test-Driven Development (in relation to CPMAI)

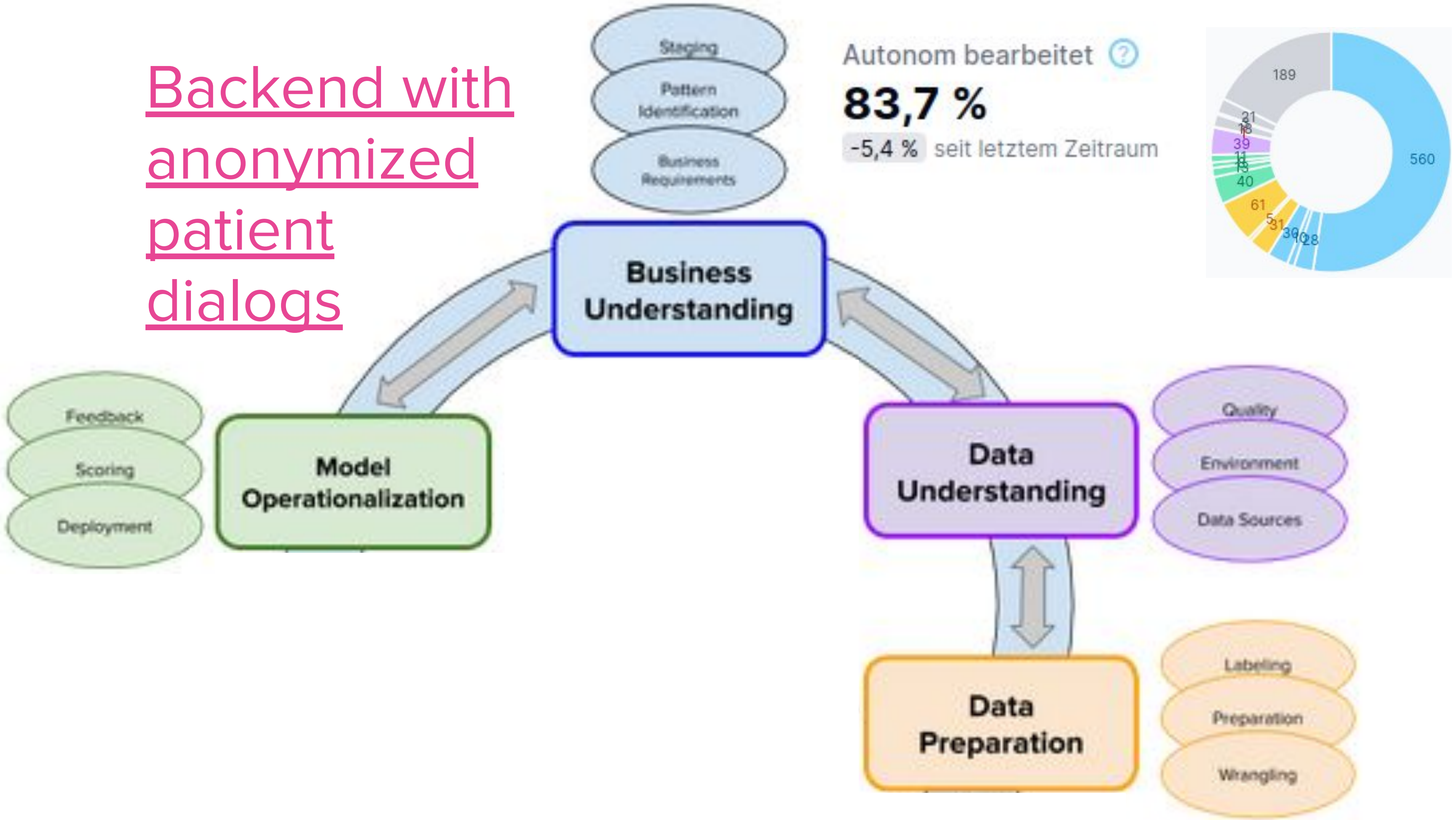


- ATDD
  - TDD: Red-Green-Refactor cycle
  - Customer-centric: with Acceptance Tests (ATs)
- In relation to CPMAI

CPMAI-Phase	ATDD
Business Understanding	Also customer-centric: also start with Business Understanding
Data Understanding	Not data-centric
Data Preparation	Also REQs as ATs (key examples, no training set with lots of data points)
Data Modeling	Rule-based “model” developed by humans, not learned statistically
Model Evaluation	Verification: also run ATs against SUT, all ATs must turn green
Model Operationalization	Validation: demo to customer in test-/demo-/production-staging

Solution for T2: Dialog-based Business & Data Understanding

Backend with anonymized patient dialogs



Autonom bearbeitet

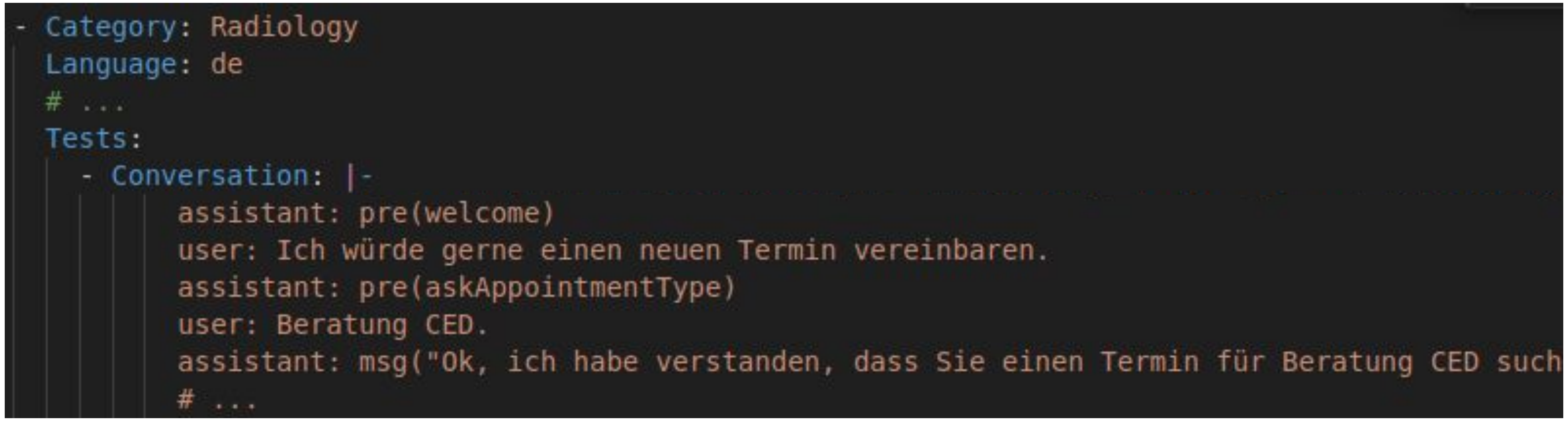
83,7 %

-5,4 % seit letztem Zeitraum

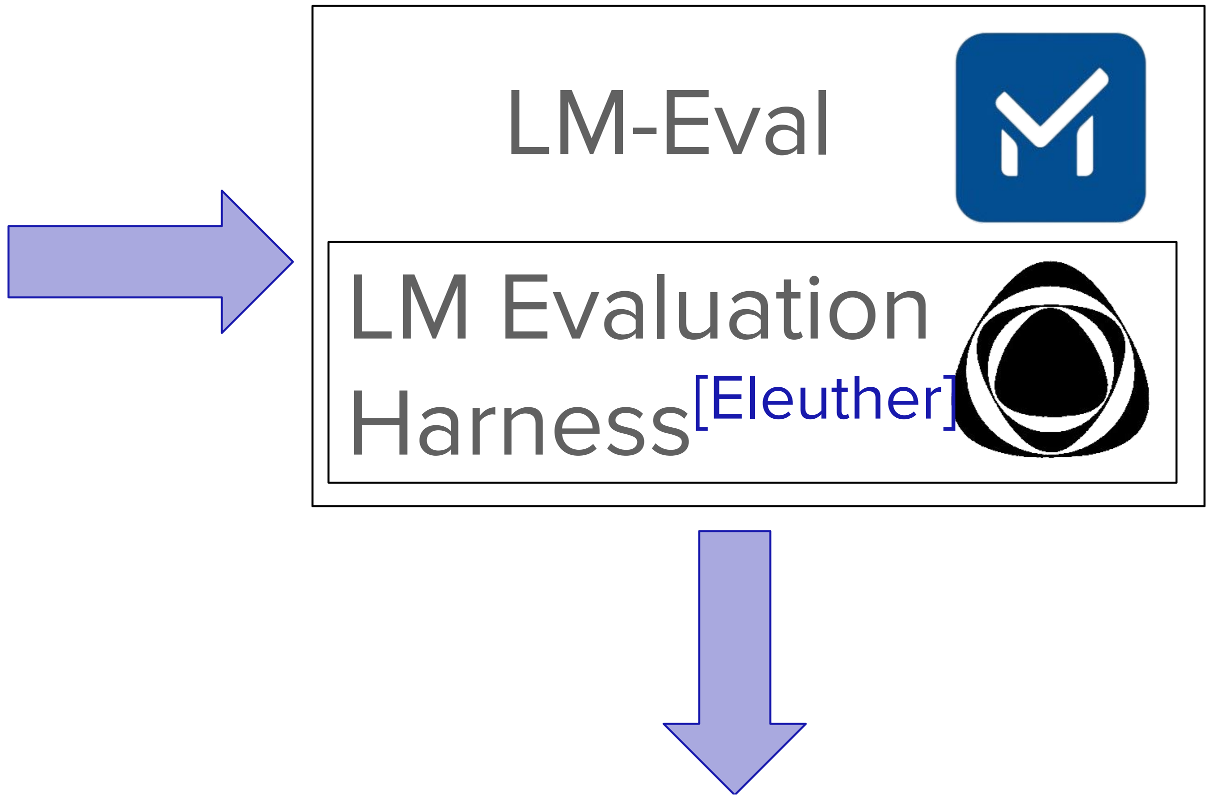
1) BI on dialogs
2) Find failing dialogs
3) Prioritize dialogs (wrt. REQs & business processes) error analysis
4) Error analysis

formulate train & test dialogs that focus on REQs & business processes
⇒ **test dialogs** = **ATs**

Solution for T3: LM-Eval

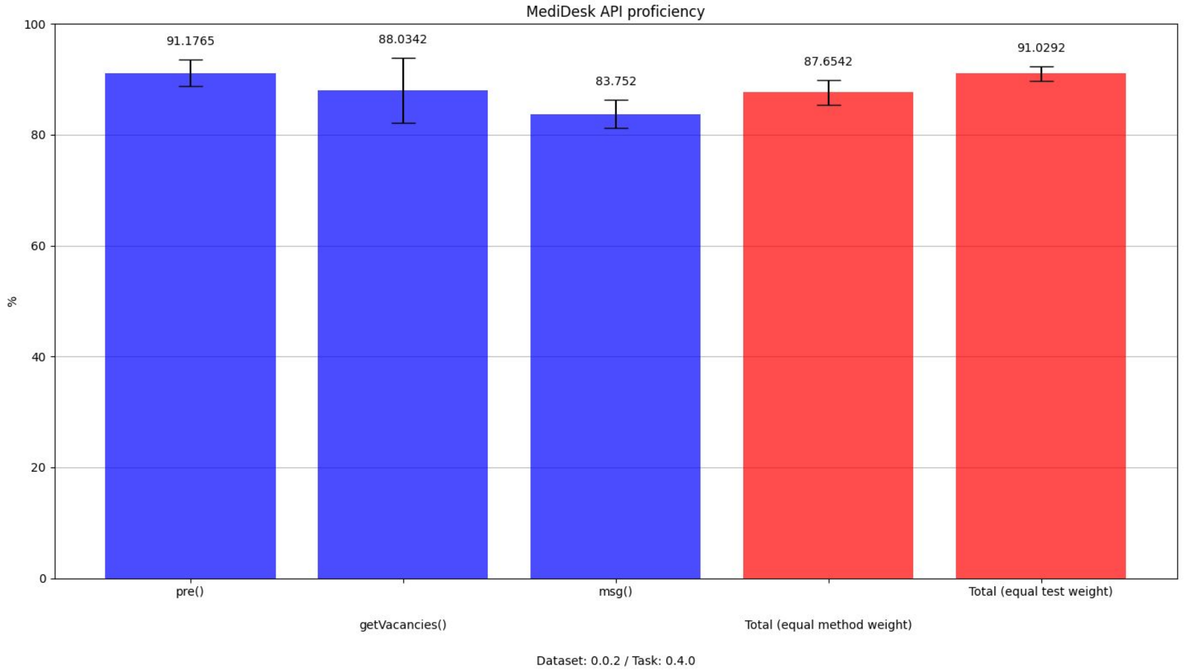


```
- Category: Radiology
  Language: de
  # ...
  Tests:
    - Conversation: |-
        assistant: pre(welcome)
        user: Ich würde gerne einen neuen Termin vereinbaren.
        assistant: pre(askAppointmentType)
        user: Beratung CED.
        assistant: msg("Ok, ich habe verstanden, dass Sie einen Termin für Beratung CED such
        # ...
```



LM-Eval

LM Evaluation Harness[Eleuther]



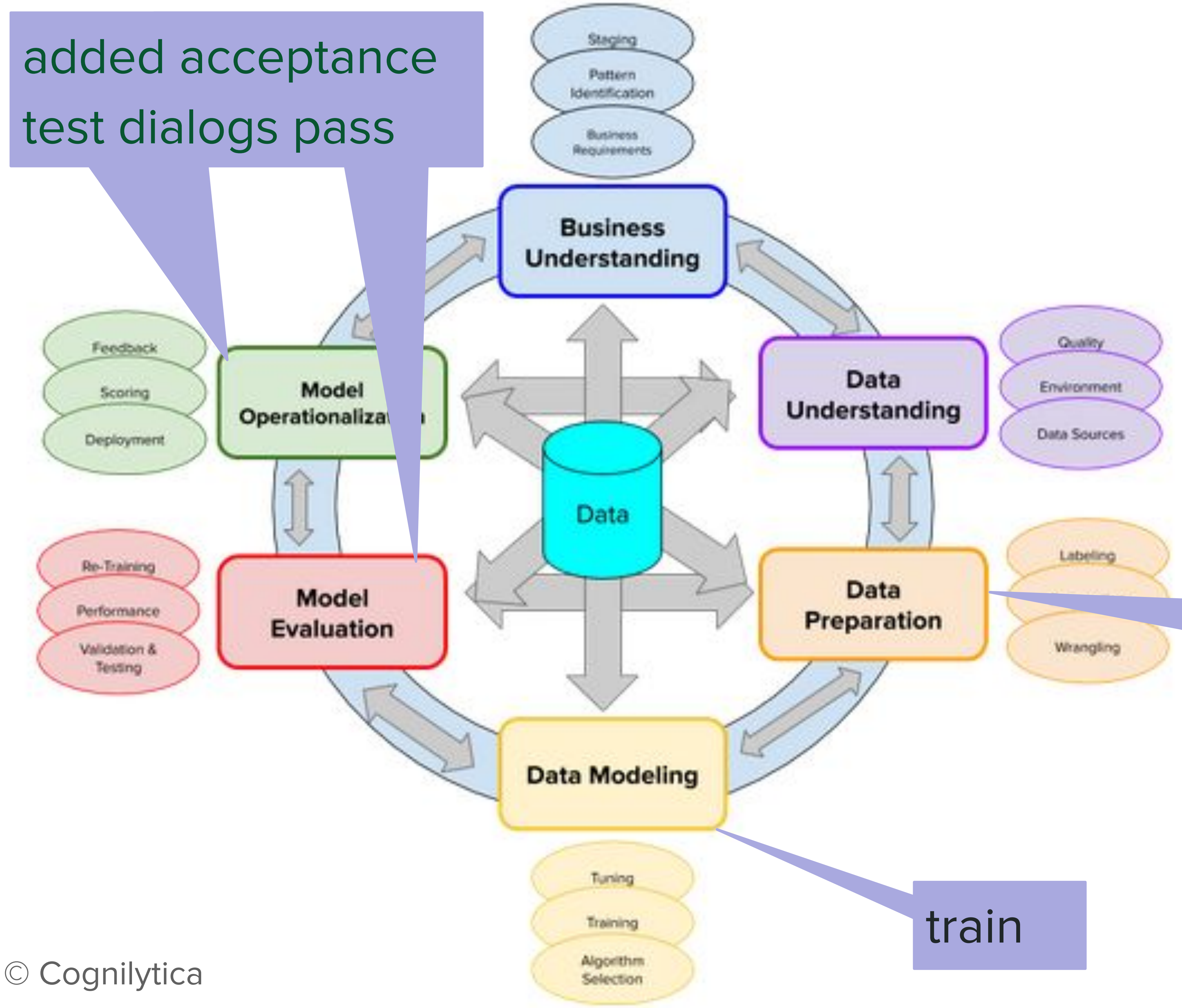
LM Evaluation Harness[Eleuther]

- many popular benchmarks out of the box
- support for custom models, benchmarks, prompts, metrics

Own extensions

- custom benchmarks: test set for each own training set
- template-based test (and training) data specification
- custom metrics: API calls vs free form messages; business-oriented; custom aggregations

Full integration: ATD^{LLM}D



added acceptance test dialogs pass

add failing acceptance test dialogs (and training dialogs)

train

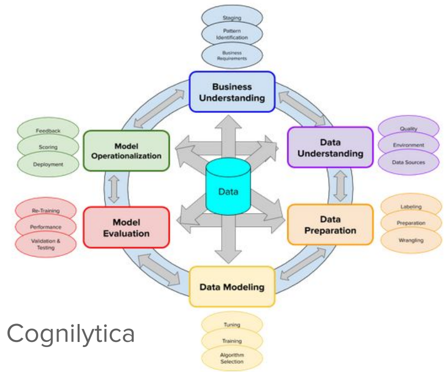
© Cognilytica

Integrate CPMAI & LM-Eval into Acceptance Test-Driven LLM dev, **ATD^{LLM}D**

Red-Train-Green cycle

Conclusion

Suitable process and best practices: CPMAI



© Cognilytica

- CPMAI course – you (and me) get 10% off with affiliate code “dfarago-10”
- visit www.cognilytica.com if you are interested in course and its provider

Make LLM behavior measurable

- LM-Eval
- ping me (dfarago@mediform.io) if you are interested in LM-Eval or MediVoice

Full integration: ATD^{LLM}D

- Red-Train-Green cycle

Bibliography and Copyright

[7Patterns] Cognilytica: “The Seven Patterns of AI”, https://www.cognilytica.com/the-seven-patterns-of-ai/

[Gra23] Gramener Blog: "Large Language Models (LLMs): A Technology Gifted by Aliens Without a Manual ", 18.11.2023, https://blog.gramener.com/large-language-models-llms-technology

[CPMAI] Cognilytica: “Cognitive Cognitive Project Management For AI”, https://www.cognilytica.com/what-is-the-cognitive-project-management-for-ai-cpmai-methodology/

[CRISP-DM1999] “Cross Industry Standard Process for Data Mining 1.0”, https://web.archive.org/web/20220401041957/https://www.the-modeling-agency.com/crisp-dm.pdf

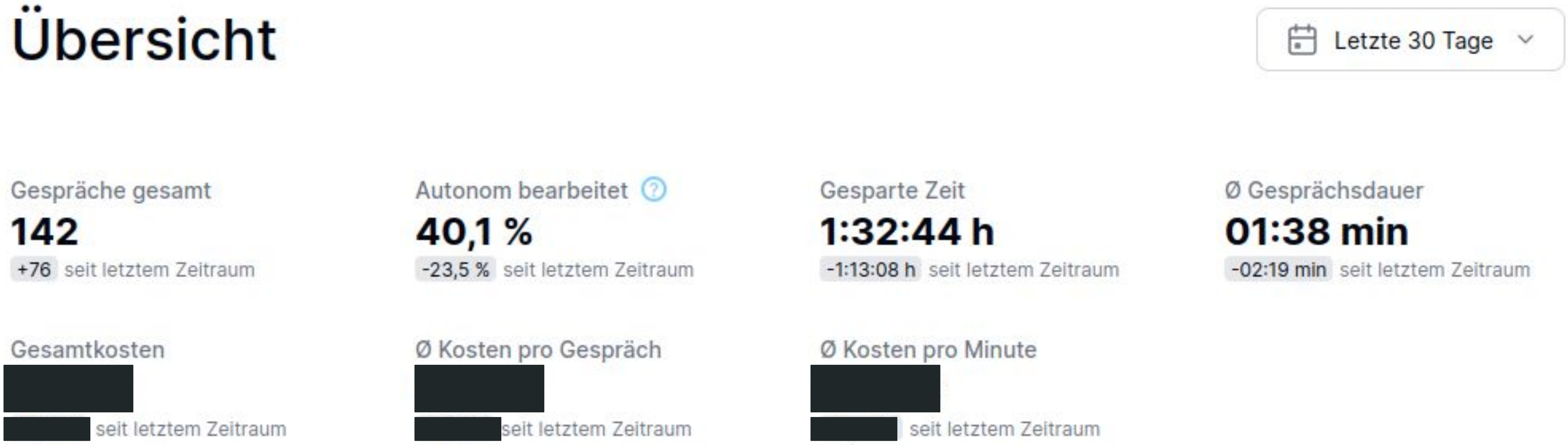
[Eleuther] “A framework for few-shot evaluation of language models”, https://github.com/EleutherAI/lm-evaluation-harness

[MediformBlog] Mediform: “Ältere Menschen buchen problemlos Termine via MediVoice”, http://tinyurl.com/bdfm8zh8

[AIMeme] “Shoggoth with Smiley Face (Artificial Intelligence)”, https://knowyourmeme.com/memes/shoggoth-with-smiley-face-artificial-intelligence

OPTIONAL

BI on Dialogs: Overview



Übersicht

Letzte 30 Tage

Gespräche gesamt
142
+76 seit letztem Zeitraum

Autonom bearbeitet ?
40,1 %
-23,5 % seit letztem Zeitraum

Gesparte Zeit
1:32:44 h
-1:13:08 h seit letztem Zeitraum

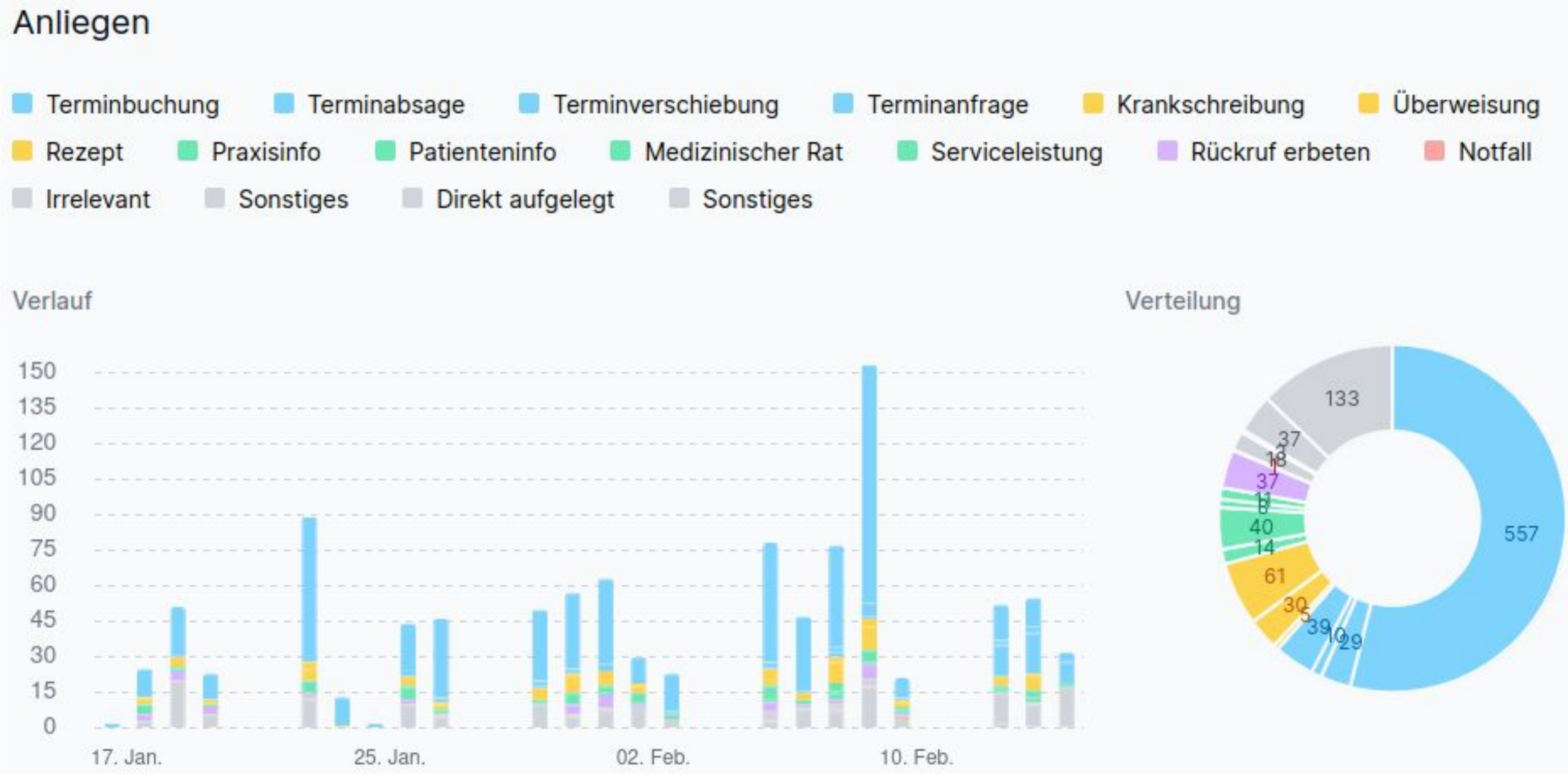
Ø Gesprächsdauer
01:38 min
-02:19 min seit letztem Zeitraum

Gesamtkosten
[redacted]
[redacted] seit letztem Zeitraum

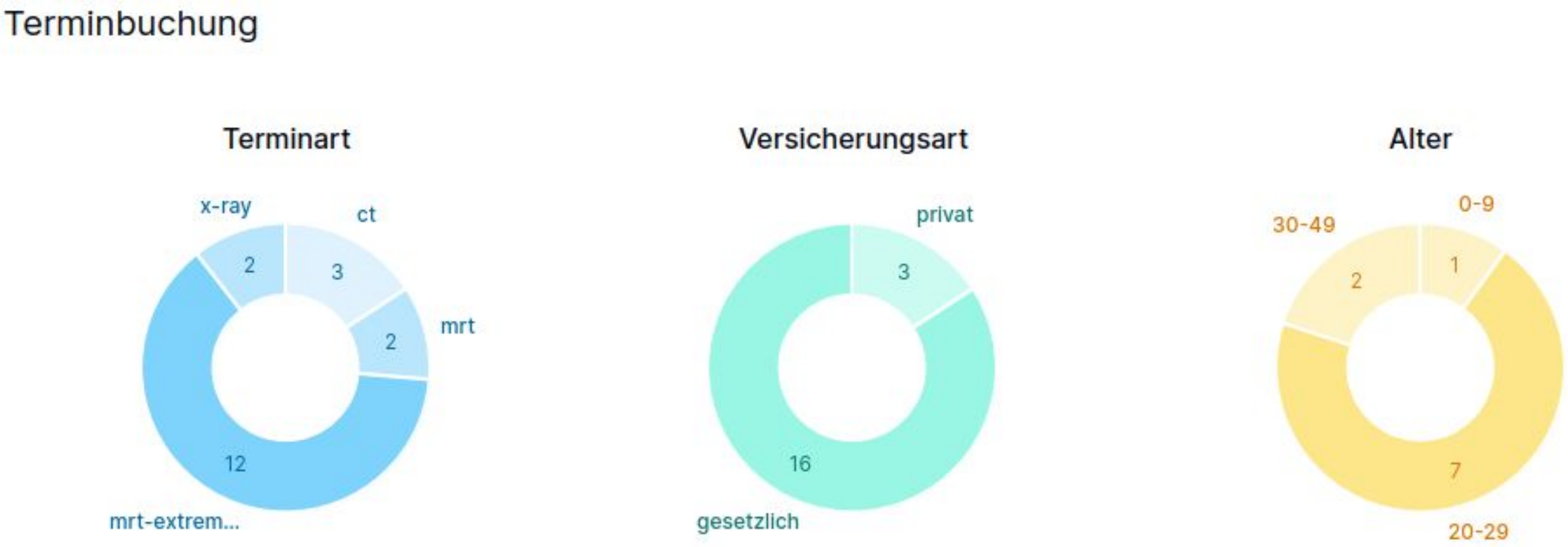
Ø Kosten pro Gespräch
[redacted]
[redacted] seit letztem Zeitraum

Ø Kosten pro Minute
[redacted]
[redacted] seit letztem Zeitraum

BI on Dialogs: Call Categories

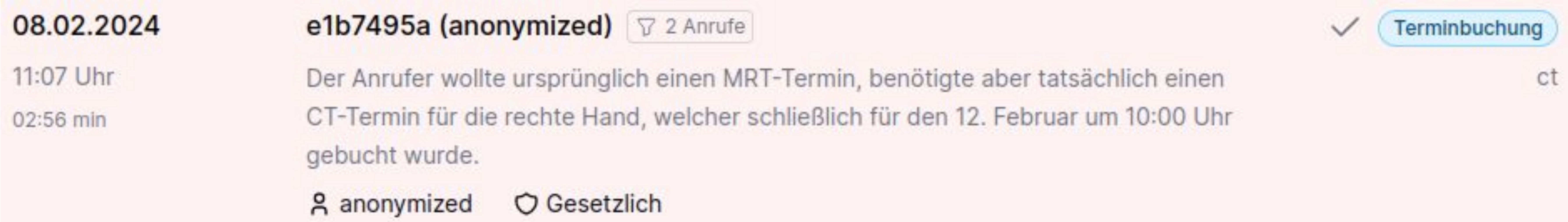


BI on Dialogs: Caller Demography



Terminbuchung

BI on Dialogs: Individual Dialog



08.02.2024 **e1b7495a (anonymized)** 2 Anrufe ✓ Terminbuchung

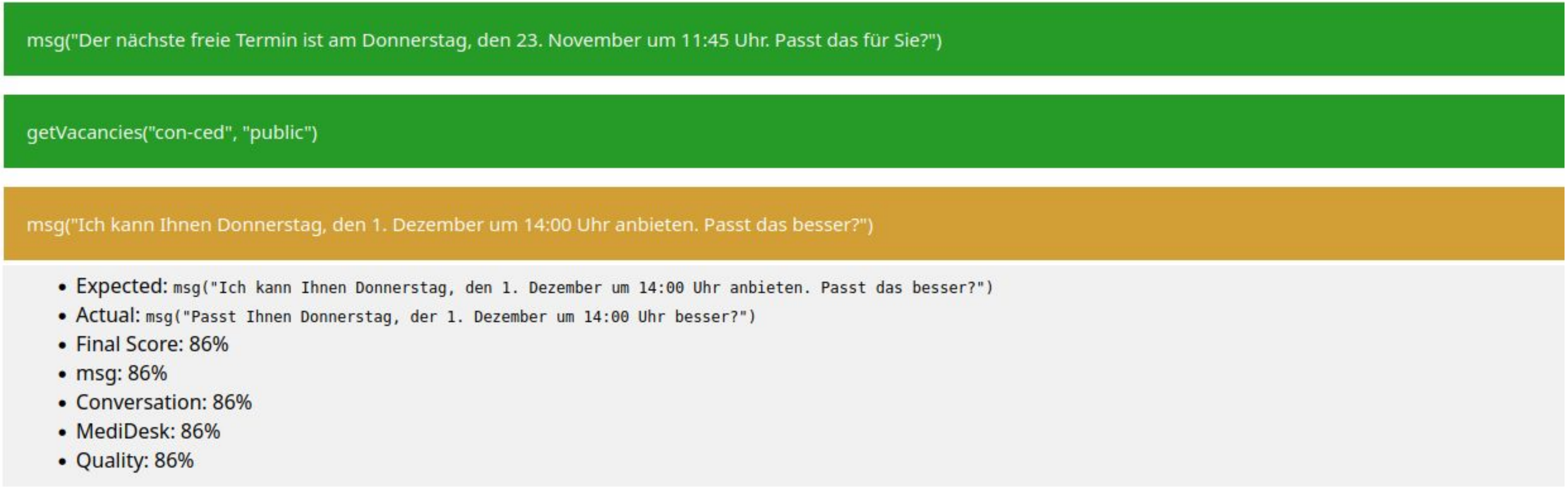
11:07 Uhr

02:56 min

Der Anrufer wollte ursprünglich einen MRT-Termin, benötigte aber tatsächlich einen CT-Termin für die rechte Hand, welcher schließlich für den 12. Februar um 10:00 Uhr gebucht wurde. ct

anonymized Gesetzlich

LM-Eval: A Dialog’s Test Cases



msg("Der nächste freie Termin ist am Donnerstag, den 23. November um 11:45 Uhr. Passt das für Sie?")

getVacancies("con-ced", "public")

msg("Ich kann Ihnen Donnerstag, den 1. Dezember um 14:00 Uhr anbieten. Passt das besser?")

- Expected: `msg("Ich kann Ihnen Donnerstag, den 1. Dezember um 14:00 Uhr anbieten. Passt das besser?")`
- Actual: `msg("Passt Ihnen Donnerstag, der 1. Dezember um 14:00 Uhr besser?")`
- Final Score: 86%
- msg: 86%
- Conversation: 86%
- MediDesk: 86%
- Quality: 86%