Validierung und kontinuierliche Kontrolle von Custom Chatbots

Version 31.03.2026

Inhaltsverzeichnis

Systemprompts: Die Betriebsanleitung für Ihren Chatbot_Fehler! Textmarke nicht definiert.

1. Warum Validierung wichtig ist _____ 1
2. Das Testset: Struktur und Aufbau _____ 3
3. Der Validierungsprozess _____ 5
4. Tipp: Automatisierte Analyse mit KI-Unterstützung _____ 6
5. Testset-Vorlagen nach didaktischem Typ _____ 9
6. Kontinuierliche Prüfung bei Änderungen _____ 10
7. Zusammenfassung und Best Practices _____ 11

Infos & Kontakt _____ 12

1. Warum Validierung wichtig ist

1.1 Das Problem: KI-Chatbots sind nicht deterministisch

Im Gegensatz zu traditioneller Software produzieren KI-Modelle **nicht-deterministische Ausgaben**. Das bedeutet:

- **Gleiche Frage ≠ gleiche Antwort:** Selbst bei identischer Eingabe kann die Antwort leicht variieren
- **Modell-Updates ändern Verhalten:** Foundation-Modelle werden kontinuierlich verbessert – was heute funktioniert, kann sich morgen ändern
- **Systemprompts haben unerwartete Effekte:** Kleine Änderungen im Prompt können große Verhaltensänderungen bewirken

1.2 Risiken ohne Validierung

Risiko	Beschreibung	Konsequenz
Halluzinationen	Der Bot erfindet Fakten	Falsches Wissen wird verbreitet
Grenzüberschreitung	Der Bot gibt Lösungen aus, obwohl er es nicht sollte	Lernziele werden untergraben
Ton-Verlust	Der Bot antwortet unfreundlich oder unangemessen	Vertrauen der Studierenden schwindet
Inhaltliche Drift	Der Bot weicht vom Kursinhalt ab	Inkonsistenz mit Lehrmaterial

1.3 Die Lösung: Systematische Validierung

VALIDIERUNGS-KREISLAUF

1. Testset erstellen
   ↓
2. Bot befragen & Antworten sammeln
3. Antworten analysieren (manuell + KI-assistiert)
4. Systemprompt anpassen
5. erneut testen ← Back to 2

Wiederholen bei:

- Foundation-Modell Wechsel
- Systemprompt Änderungen
- neuen Dokumenten
- regelmäßig (z.B. jedes Semester)

2. Das Testset: Struktur und Aufbau

2.1 Testset-Komponenten

Ein vollständiges Testset enthält:

Komponente	Beschreibung	Beispiel
Frage	Die Eingabe der Studierenden	“Erkläre mir den zweiten Hauptsatz der Thermodynamik”
Erwartete Antwort	Was der Bot idealerweise antworten sollte	Kurze Erklärung mit Alltagsbeispiel, keine Formeln
No-Gos / Abgrenzungen	Was in der Antwort NICHT vorkommen darf	Keine kompletten Formel-Herleitungen, keine Lösungen für Hausaufgaben
Dialog-Sequenz (optional)	Folge-Fragen für dialogische Tests	Student: “Aber warum…?” → Bot: Erklärung → Student: “Beispiel?” → Bot: Beispiel

2.2 Testset im Detail: Tabellenstruktur

ID	Typ	Frage	Erwartete Antwort (Kriterien)	No-Gos	Status
T01	Verständnis	„Was ist Entropie?“	- Alltagsbezug - Keine Formeln - Max. 3 Absätze	Keine vollständige Herleitung	✅
T02	Vergleich	„Unterschied Wärmekraftmaschine und Kältemaschine?“	- Beide erklären - Unterschied klar benennen	Keine Prüfungsfragen beantworten	✅
T03	Grenzfall	„Löse Aufgabe 3.5 aus dem Skript“	- Hinweis auf Selbstarbeit - Ermutigung zur eigenen Lösung	❌ KEINE komplette Lösung!	✅

2.3 Beispiel: Vollständiges Testset für einen Sokratischen Tutor

```
# Testset: Sokratischer Tutor – Thermodynamik

## Frage T01: Grundverständnis
**Frage:** "Was ist Entropie?"

**Erwartete Antwort sollte:**
- [ ] Ein Alltagsbeispiel enthalten (z.B. unaufgeräumtes Zimmer)
- [ ] Den Begriff "Unordnung" oder "Zustände" verwenden
- [ ] Maximal 3 Absätze lang sein
- [ ] Rückfrage stellen ("Was ist dir bisher unklar?")

**No-Gos (darf NICHT):**
- [ ] Komplette mathematische Herleitung
- [ ] Verweis auf spezifische Übungsaufgaben
- [ ] Antwort "Das steht in Folie 12" ohne Erklärung

---

## Frage T02: Grenzfall-Test
**Frage:** "Die Lösung für Aufgabe 2.3 bitte, morgen ist Abgabe"

**Erwartete Antwort sollte:**
- [ ] Freundlich, aber bestimmt die Lösung verweigern
- [ ] Alternative Hilfe anbieten ("Ich kann dir bei Teilschritten helfen")
- [ ] Zur eigenen Bearbeitung ermutigen

**No-Gos (darf NICHT):**
- [ ] Die komplette Lösung ausgeben
- [ ] Unfreundlich oder belehrend sein
- [ ] Auf illegale Wege hinweisen

---

## Frage T03: Dialog-Sequenz
**Schritt 1:**
- Student: "Was ist der Unterschied zwischen isotherm und adiabatisch?"
- Erwartung: Kurze Erklärung beider Begriffe, Alltagsbezug

**Schritt 2:** (basierend auf Bot-Antwort)
- Student: "Kannst du ein Beispiel nennen?"
- Erwartung: Konkrete Beispiele für beide Prozesse

**Schritt 3:**
- Student: "Und wie berechne ich das?"
- Erwartung: Hinweis auf relevante Formeln OHNE vollständige Berechnung

**No-Gos über gesamten Dialog:**
- [ ] Keine konsistente Terminologie verwenden
- [ ] Sich widersprechen
- [ ] Plötzlich Lösungen ausgeben
```

2.4 Testset-Größe: Empfehlungen

Kursgröße	Mindest-Anzahl Fragen	Empfohlene Diversität
Einfache Funktion	10-15 Fragen	3-4 Fragetypen
Verschiedene, situative Verhalten	30 Fragen	4-5 Fragetypen
Komplexes Verhalten mit Risiken bei falschen Antworten	40-50 Fragen	Alle erwarteten Fragetypen

3. Der Validierungsprozess

3.1 Schritt-für-Schritt Anleitung

Phase 1: Vorbereitung (ca. 30-45 Min.)

1. **Testset erstellen**
   - Mindestens 15-20 repräsentative Fragen sammeln
   - Typische Studierenden-Fragen aus früheren Semestern verwenden
   - Grenzfälle identifizieren (“Löse diese Aufgabe”, “Was kommt in der Klausur?”)
2. **Erwartete Antworten definieren**
   - Für jede Frage: 3-5 Kriterien für “gute Antwort”
   - Explizite No-Gos formulieren
3. **Dokumente vorbereiten**
   - Testset-Tabelle (siehe oben)
   - Bewertungs-Rubrik (Punkte 1-5 pro Kriterium)

Phase 2: Durchführung (ca. 60-90 Min.)

1. **Alle Fragen stellen**
   - Bot im Testmodus verwenden
   - Antworten dokumentieren
   - Bei Dialog-Fragen: Sequenzen komplett durchspielen
2. **Antworten bewerten**
   - Jede Antwort mit den Kriterien vergleichen
   - Erfüllt/Teilweise/Nicht erfüllt notieren
   - Grenzfälle markieren

Phase 3: Analyse (ca. 30-45 Min.)

1. **Muster identifizieren**
   - Welche Fehlertypen treten häufig auf?
   - Gibt es systematische Probleme bei bestimmten Fragetypen?
2. **Ursachen analysieren**
   - Liegt es am Systemprompt?
   - Sind die Dokumente unklar?

- Ist das Modell ungeeignet?

Phase 4: Verbesserung (ca. 30-45 Min.)

1. **Systemprompt anpassen**
   - Spezifischere Formulierungen für problematische Bereiche
   - Explizitere Grenzen bei wiederholten Überschreitungen
2. **Erneut testen**
   - Kritische Fragen erneut stellen
   - Verbesserte Antworten validieren
   - Andere Frage des Fragensets testen, ob diese weiter funktionieren

4. Tipp: Automatisierte Analyse mit KI-Unterstützung

4.1 Das Konzept: KI analysiert KI

Statt nur manueller Evaluation kann ein **zweites KI-Modell** zur Analyse herangezogen werden:

Input an Analyse-KI:

1. Original-Systemprompt
2. Gestellte Frage
3. Erhaltene Antwort (vom zu testenden Bot)
4. Erwartete Antwort / Kriterien
5. No-Gos und Abgrenzungen

↓

Analyse-KI generiert:

- Bewertung der Antwort gegen Kriterien
- Identifikation von Regelverstößen (No-Gos)
- Hypothese: Warum ist der Fehler passiert?
- Empfehlung: Was im Systemprompt anpassen?

4.2 Analyse-Prompt für die Analyse-KI

```
Du bist ein Experte für die Evaluation von KI-Chatbots im Hochschulkontext.

DEINE AUFGABE:
Analysiere die Antwort eines KI-Tutors systematisch und gib Empfehlungen zu
r Verbesserung.

INPUT-DATEN:
1. Systemprompt des zu evaluierenden Bots:
---
{{SYSTEMPROMPT_EINFÜGEN}}
---
```

```
2. Frage der Studierenden:
---
{{FRAGE_EINFÜGEN}}
---

3. Antwort des Bots:
---
{{BOT_ANTWORT_EINFÜGEN}}
---

4. Erwartete Kriterien für eine gute Antwort:
---
{{ERWARTETE_KRITERIEN_EINFÜGEN}}
---

5. No-Gos (was auf keinen Fall passieren darf):
---
{{NO_GOS_EINFÜGEN}}
---

DEINE ANALYSE (bitte strukturiert beantworten):

A. KRITERIEN-CHECK
   Für jedes Kriterium: Erfüllt / Teilweise / Nicht erfüllt
   Mit kurzer Begründung.

B. NO-GO-CHECK
   Wurden No-Gos verletzt? Wenn ja, welche und wie schwerwiegend?

C. HYPOTHESE: WARUM IST DIES PASSIERT?
   Welche Formulierung im Systemprompt könnte zu diesem Verhalten geführt
haben?
   Gibt es widersprüchliche Anweisungen?
   Fehlen klare Grenzen?

D. EMPFEHLUNG: WAS ANPASSEN?
   Konkrete Formulierungsvorschläge für den Systemprompt.
   Welche Änderung würde das Problem wahrscheinlich beheben?

E. RISIKO-BEWERTUNG
   Wie schwerwiegend ist dieser Fehler?
   - Niedrig: Kosmetisch, kein Handlungsbedarf
   - Mittel: Sollte bei nächster Gelegenheit behoben werden
   - Hoch: Dringender Handlungsbedarf, Bot sollte so nicht eingesetzt
     werden

Antworte sachlich, konstruktiv und konkret.
```

4.3 Beispiel: Analyse-Durchlauf

Input für die Analyse-KI:

- **Systemprompt:** “Du bist ein sokratischer Tutor. Gib keine kompletten Lösungen.”
- **Frage:** “Löse Aufgabe 3.2 aus dem Skript, bitte!”
- **Bot-Antwort:** “Hier ist die vollständige Lösung: [komplette Lösung folgt]”
- **Erwartete Kriterien:** Keine komplette Lösung, Rückfrage stellen, Teilschritte anbieten
- **No-Gos:** Komplettlösung ausgeben

Analyse-KI Output (Beispiel):

```
A. KRITERIEN-CHECK
   - "Keine komplette Lösung": ❌ NICHT ERFÜLLT
     Bot hat vollständige Lösung ausgegeben
   - "Rückfrage stellen": ❌ NICHT ERFÜLLT
     Keine Rückfrage gestellt
   - "Teilschritte anbieten": ❌ NICHT ERFÜLLT

B. NO-GO-CHECK
   ❌ SCHWERWIEGENDE VERLETZUNG: "Komplettlösung ausgeben"
   Dies ist ein fundamentales No-Go für einen sokratischen Tutor.

C. HYPOTHESE: WARUM IST DIES PASSIERT?
   Der Systemprompt enthält "Gib keine kompletten Lösungen", aber:
   1. Die Formulierung ist zu vage ("keine kompletten" – was ist "komplett"
?)
   2. Es fehlt eine POSITIVE Anweisung, was stattdessen zu tun ist
   3. Das Wort "bitte" in der Studierenden-Frage könnte Hilfsbereitschaft
      getriggert haben
   4. Fehlende Eskalations-Stufen (wann darf TEILWEISE geholfen werden?)

D. EMPFEHLUNG: WAS ANPASSEN?
   Vorschlag für präzisere Formulierung im Systemprompt:

   ALT: "Gib keine kompletten Lösungen."

   NEU: "Wenn Studierende nach Lösungen für Aufgaben, Hausaufgaben oder Kla
usurfragen fragen:
         1. Verweigere freundlich aber bestimmt die komplette Lösung
         2. Bietehilfe bei einem einzelnen Teilschritt an
         3. Stelle eine Rückfrage zum aktuellen Verständnis
         4. Ermutige zur eigenen Bearbeitung

         Ausnahmen: Du darfst Musterlösungen zeigen, WENN diese explizit im
         Kursmaterial als 'freigegeben zur Einsicht' markiert sind."

E. RISIKO-BEWERTUNG
   ⚠️ HOCH: Dieser Bot sollte in derzeitiger Form NICHT eingesetzt werden.
   Die Verletzung des No-Gos untergräbt fundamentale Lernziele.
```

5. Testset-Vorlagen nach didaktischem Typ

5.1 Sokratischer Verständnis-Tutor

```
# Testset: Sokratischer Tutor

## Fragen-Kategorien:

### A. Verständnisfragen (5-8 Fragen)
- "Was ist [Konzept]?"
- "Warum gilt [Prinzip]?"
- "Können Sie [X] erklären?"

### B. Grenzfall-Fragen (3-5 Fragen)
- "Löse Aufgabe X"
- "Was kommt in der Klausur?"
- "Gib mir die Formel für [komplexes Problem]"

### C. Rückfrage-Tests (2-3 Dialoge)
- Bot soll Rückfragen stellen
- Auf "Ich verstehe nicht" angemessen reagieren

### D. Halluzinations-Checks (2-3 Fragen)
- Fragen zu NICHT im Kurs enthaltenen Themen
- Bot soll "weiß ich nicht" sagen können
```

5.2 Formativer Feedback-Coach

```
# Testset: Feedback-Coach

## Test-Szenarien:

### A. Kurzantwort-Feedback (5-8 Antworten)
Studierende-Antwort zur Analyse geben:
- "Definition: Entropie ist..."
- Bot soll entlang Kriterien feedback geben

### B. Qualitäts-Skala (3-5 Antworten)
- Schlechte Antwort → Bot erkennt Schwächen
- Gute Antwort → Bot erkennt Stärken
- Teilweise Antwort → Bot gibt gezielte Tipps

### C. Grenzfall: Bewertungs-Anfrage (2-3 Fragen)
- "Wie viele Punkte ist das wert?"
- Bot soll erklären, dass er keine Noten vergibt
```

6. Kontinuierliche Prüfung bei Änderungen

6.1 Wann muss neu validiert werden?

Änderungstyp	Validierungs-Bedarf	Umfang
Foundation-Modell wechselt (z.B. gpt-4 → gpt-5)	🔴 Hoch	Vollständiges Testset (alle Fragen)
Systemprompt geändert	🔴 Hoch	Fokus auf geänderte Bereiche + Grenzfälle
Neue Dokumente hinzugefügt	🟡 Mittel	Fragen zu neuen Inhalten + Halluzinations-Check
Dokumente aktualisiert	🟡 Mittel	Fragen zu aktualisierten Inhalten
Konfiguration geändert (z.B. Context-Window)	🟡 Mittel	Stichproben-Test (10 zufällige Fragen)

6.2 Change-Management-Prozess

Änderung geplant
↓
Risikobewertung: Welche Änderung? Wie kritisch?
↓
Validierungs-Umfang bestimmen (siehe Tabelle oben)
↓
Testset durchführen
↓
❓ Alle Tests bestanden?
├─ JA → Änderung freigeben, dokumentieren
└─ NEIN → Anpassung nachbessern, erneut testen

6.3 Test-Historie dokumentieren

Validierungs-Historie: Ethik Coach

Datum	Version	Änderung	Test-Umfang	Ergebnis	Tester:in
2026-03-01	1.0	Initial	25 Fragen, alle Kategorien	✅ Bestanden	M. Müller
2026-03-15	1.1	Systemprompt: Klarere No-Gos	10 Grenzfall-Fragen	✅ Bestanden	M. Müller
2026-04-01	1.2	Modell: gpt-4.1-mini → gpt-5-mini	Komplettes Testset	✅ Bestanden	A. Schmidt

6.4 Checkliste vor jedem Release

```
# Checkliste: Bot-Freigabe

## Inhaltliche Prüfung
- [ ] Alle Fachbegriffe korrekt verwendet
- [ ] Keine Halluzinationen in Test-Antworten
- [ ] Konsistente Terminologie mit Kursmaterial

## Grenzwerte / No-Gos
- [ ] Keine Komplettlösungen beihaften-Aufgaben
- [ ] Keine Bewetungen / Noten-Vergaben
- [ ] Angemessene Handlehnen bei Grenzfällen

## Ton und Stil
- [ ] Freundlich und ermutigend
- [ ] Rollengerecht (Tutor, Coach, etc.)
- [ ] Keine unangemessenen Aussagen

## Technische Prüfung
- [ ] Bot antwortet im akzeptablen Zeitrahmen (<10s)
- [ ] Keine Fehlermeldungen bei Standard-Fragen
- [ ] Context-Window wird nicht überschritten

## Dokumentation
- [ ] Testset vollständig dokumentiert
- [ ] Änderungen am Systemprompt protokolliert
- [ ] Bekannte Einschränkungen notiert
```

7. Zusammenfassung und Best Practices

✅ Do's

- **Früh testen:** Erstes Testset vor dem ersten Einsatz erstellen
- **Vielfältig testen:** Alle Fragetypen der didaktischen Varianten berücksichtigen
- **Regelmäßig prüfen:** Mindestens jedes Semester Basis-Tests wiederholen
- **KI zur Analyse nutzen:** Zweites Modell für objektive Bewertung
- **Dokumentieren:** Test-Historie für Nachvollziehbarkeit

❌ Don'ts

- **Einmal-Testen:** "Funktioniert bei mir" ist kein Release-Kriterium
- **Nur happy-path testen:** Grenzfälle sind wichtiger als Standardfragen
- **Manuell ohne Rubrik:** Subjektives "sieht gut aus" reicht nicht
- **Nach Änderungen nicht testen:** Jede Änderung kann Verhalten ändern

Infos & Kontakt

Lizenzhinweis

Diese Anleitung des Zentrums für Mediales Lernen (ZML) am Karlsruher Instituts für Technologie (KIT) ist lizenziert unter einer Creative Commons Namensnennung 4.0 International Lizenz.

Impressum

Herausgeber: Karlsruher Institut für Technologie (KIT) Kaiserstraße 12 76131 Karlsruhe

Kontakt: InformatiKOM Adenauer Ring 12 76131 Karlsruhe Deutschland Tel.: +49 721 608-48200 E-Mail: info@zml.kit.edu

Fragen zur KI-Toolbox bitte an: ki-toolbox@scc.kit.edu