



ARBEITSAUFTRAG

Unit 6: Challenge

Übertragen Sie die CustomerSupportArena26 auf eine Aufgabe aus Ihrem Gebiet. Der Generator erstellt Ausgaben; der Judge bewertet sie anhand Ihrer Kriterien. Prüfen Sie die Kritik selbst, ändern Sie eine Generator-Anweisung und untersuchen Sie die Wirkung.

AUF EINEN BLICK	IHRE AUFGABE	DER GESAMTPROZESS	VORGEHEN	ARTEFAKTE	EVIDENZ	OPTIONALE ERWEITERUNG: JUDGES VERGLEICHEN
-----------------	--------------	-------------------	----------	-----------	---------	---

Auf einen Blick

Frage	Ihr Auftrag
Was erstellen Sie?	Einen eigenen Anwendungsfall mit geprüfter Fachgrundlage, drei Entwicklungsfällen E01–E03 und einem vorab reservierten Abschlussfall A01.
Was vergleichen Sie?	Ausgaben von G0 und G1 . G1 unterscheidet sich durch eine begründete Änderung am Generator-Prompt.
Was bleibt gleich?	Aufgabe, Ausgabeart, Entwicklungsfälle und ihre Version, Kontext-Version, Modell, Ausgabelimit und Judge JO samt Rubrik.
Was reichen Sie ein?	Eine ZIP-Datei mit einem Kurzbericht als Markdown oder PDF und dem vollständigen JSON-Export.

Der Generator-Vergleich ist Pflicht. Ein zusätzlicher Judge-Vergleich ist am Ende als **freiwillige Erweiterung** beschrieben. Eine höhere Bewertung oder erfolgreiche Verbesserung ist keine Bestehensbedingung.

Ihre Aufgabe

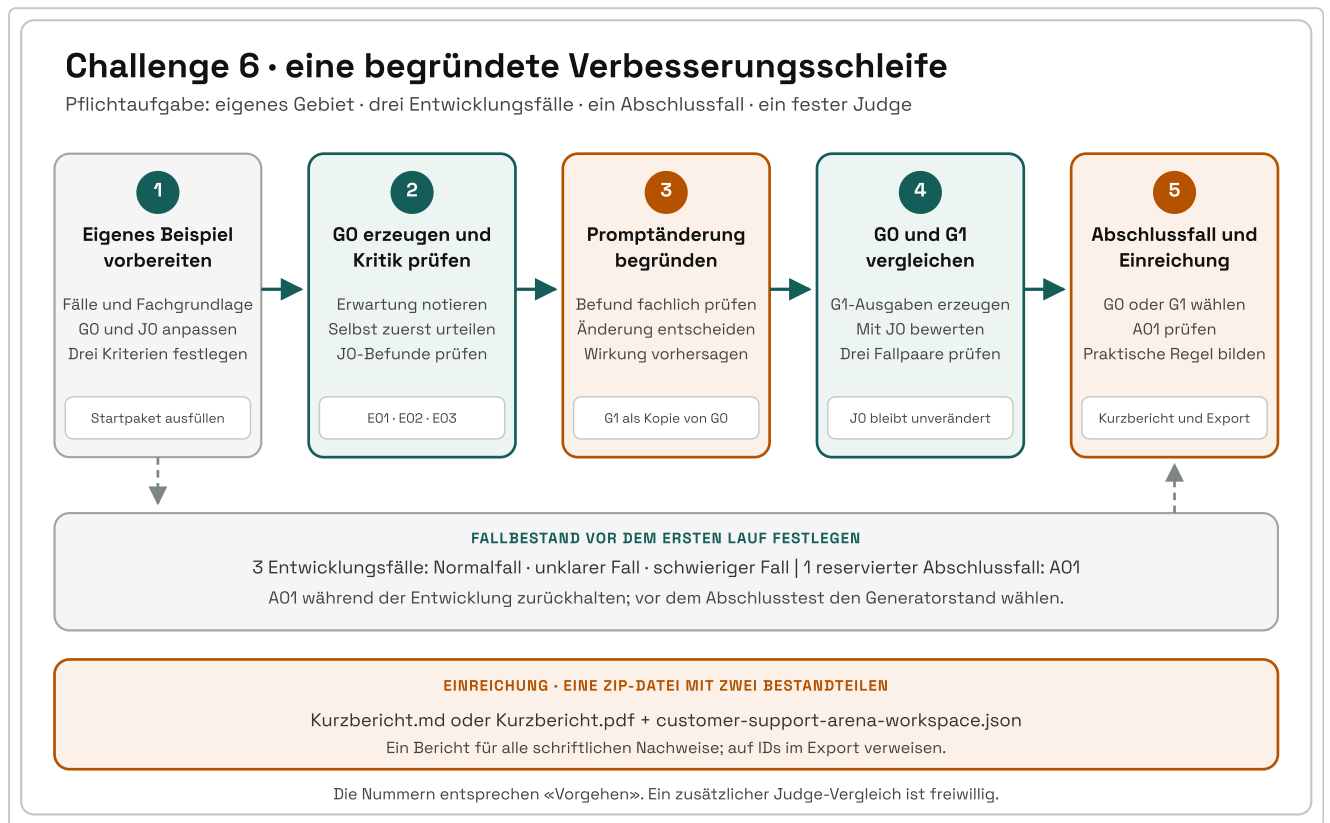
Wählen Sie eine konkrete Aufgabe für eine Zielgruppe. Sie müssen die Ergebnisse fachlich beurteilen können. Vorhandene geprüfte Materialien aus einer früheren Challenge dürfen Sie weiterverwenden. Wenige Fachregeln und kurze Ausgaben, etwa 100–200 Wörter, genügen.

Beispiel	Eingabe	Ausgabe	Mögliche Qualitätsfrage
Bildung	Aufgabe und fiktive Lernendenantwort	begründetes Feedback	Durch die Leistung belegt und fürs Lernen hilfreich?
Projektmanagement	Projektauftrag mit Rahmenbedingungen	Risikoeinschätzung	Risiken ableitbar und Massnahmen umsetzbar?
Kommunikation	Fakten, Zielgruppe und Anlass	Mitteilung	Korrekt und für die Zielgruppe verständlich?
Analyse	Ergebnistabelle mit Erläuterungen	Interpretation	Zahlen richtig wiedergegeben und Grenzen benannt?

Passen Sie Aufgabe, Fälle, Fachgrundlage und Bewertung an Ihr Gebiet an. Nur Firmen- oder Produktnamen auszutauschen reicht nicht. Verwenden Sie synthetische, öffentliche oder zur Verwendung und Einreichung autorisierte Inhalte ohne vertrauliche oder sensible Angaben.

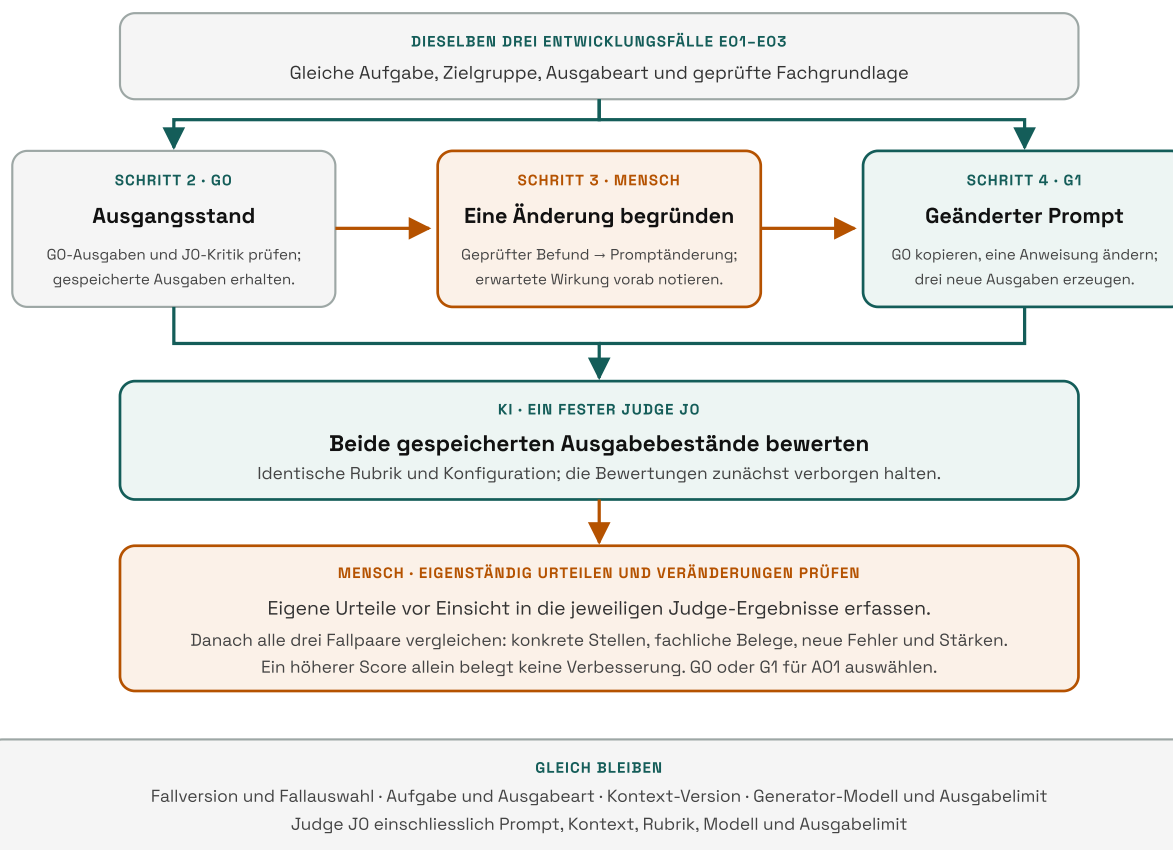
Der Gesamtprozess

Die fünf Schritte der Abbildung entsprechen dem folgenden Vorgehen.



Pflichtaufgabe · den Generator gezielt verbessern

Eine Promptänderung · dieselben Fälle und Grundlagen · ein unveränderter Judge JO



Die menschlichen Entscheidungen und drei Fallvergleiche werden im gemeinsamen Kurzbericht belegt.

G1 erzeugt neue Ausgaben. Die Anwendung überarbeitet die gespeicherten G0-Ausgaben nicht anhand eines Judge-Berichts. Sie entscheiden, welche geprüfte Kritik oder Stärke die Promptänderung begründet.

Vorgehen

Richten Sie die CustomerSupportArena26 nach ihrer README ein. Prüfen Sie den Zugang möglichst beim gemeinsamen Challenge-Start. Nutzen Sie den **OpenAI-Modus** für das Experiment. Der Demo-Modus mit festen Antworten und synthetischen Bewertungen dient nur zum Kennenlernen.

Kopieren Sie das Startpaket in Ihren Arbeitsordner:

Datei	Zweck
test_cases.json	vier Fälle und kurze Anmerkungen zum erwarteten Verhalten
context.json	geprüfte Fakten, Regeln und Grenzen
prompts-und-rubrik.md	Generator- und Judge-Auftrag; drei Kriterien mit Bewertungsankern
Kurzbericht.md	alle schriftlichen Nachweise

Ersetzen Sie die Platzhalter und prüfen Sie die Inhalte. KI darf bei der Vorbereitung helfen. Notieren Sie im Bericht ihren Beitrag und Ihre wesentlichen Korrekturen.

1. Eigenes Beispiel vorbereiten

Beschreiben Sie im Kurzbericht Aufgabe, Eingabe, gewünschte Ausgabe, Zielgruppe, Verwendung, Fachgrundlage und eine wichtige Grenze menschlicher Verantwortung.

Fälle: Füllen Sie `test_cases.json` aus:

- **E01:** klar lösbarer Normalfall;
- **E02:** fehlende oder mehrdeutige Informationen;
- **E03:** falsche Annahme, Zielkonflikt oder fachliche Grenze;
- **A01:** weitere Situation für den Abschluss, keine Kopie eines Entwicklungsfalls.

Notieren Sie zu jedem Fall erwartetes Verhalten und einen möglichen Fehler. Prüfen Sie diese **Annotationen** gegen Fall und Fachgrundlage. Fachlich richtige Alternativen bleiben zulässig. Die Anwendung gibt die Annotationen weder an Generator noch Judge weiter.

Reservieren Sie A01 **vor dem ersten Lauf**. Verwenden Sie ihn während der Entwicklung nicht zum Ausprobieren oder Anpassen. Ein selbst erstellter Fall ist Ihnen schon bekannt: Der Abschluss prüft die Übertragung auf einen noch nicht ausgeführten Fall und ist kein blinder Test.

Fachgrundlage: Tragen Sie die benötigten Fakten, Regeln und Grenzen in `context.json` ein und halten Sie deren Herkunft im Bericht fest. Für fiktive Szenarien gelten Ihre ausdrücklich festgelegten, auf Widersprüche geprüften Regeln. Importieren Sie beide JSON-Dateien unter **Test-Cases** und **Kontext**. Prüfen Sie die angezeigten Inhalte. Behalten Sie technische Feldnamen und Fallgruppen bei.

Agenten: Richten Sie G0 unter **Agent** und J0 unter **Judge** ein. Passen Sie beide Promptvorlagen an Ihr Gebiet an. Wählen Sie für beide dieselbe geprüfte Kontext-Version; für G0 nur E01–E03. Übertragen Sie **drei Kriterien mit beobachtbaren Ankern für 1, 3 und 5** sowie die Definitionen der Gesamtentscheidungen in die Judge-Einstellungen. Diese Kriterien und Anker bilden Ihre **Rubrik**. Prüfen Sie die Einstellungen vor dem Start.

Benennen Sie die Iterationen G0, G1 und J0. Verwenden Sie im Bericht ihre tatsächlichen IDs sowie Fall-, Kontext- und Lauf-IDs, damit sich die Nachweise im Export finden lassen.

2. G0 erzeugen und Kritik prüfen

Notieren Sie **vor G0**, bei welchem Fall und Kriterium Sie eine Schwäche erwarten. Speichern Sie G0 und erzeugen Sie die drei Entwicklungsausgaben. Speichern und frieren Sie J0 ein. Lassen Sie ihn die Ausgaben bewerten; J0 bleibt danach für die Pflichtaufgabe unverändert.

Beurteilen Sie **jede Ausgabe selbst, bevor Sie ihr Judge-Ergebnis ansehen**. Erfassen Sie Kriterienwerte, Gesamtentscheidung und eine Begründung mit Ausgabestelle oder fachlichem Beleg. Die Anwendung verbirgt das Judge-Ergebnis bis zu Ihrer Eingabe.

Prüfen Sie anschliessend: Was hat J0 zutreffend erkannt, übersehen oder unbegründet beanstandet? Prüfen Sie bei Abweichungen auch Ihr eigenes Urteil. Berücksichtigen Sie Schwächen von J0 bei der weiteren Beurteilung. Sie müssen ihn in der Pflichtaufgabe nicht optimieren.

3. Eine Promptänderung begründen

Wählen Sie einen fachlich geprüften Befund als Ansatzpunkt. Auch eine belegte Stärke, die Sie absichern wollen, ist geeignet. Halten Sie **vor G1** fest:

- Befund mit Fall-ID, Kriterium und Beleg;
- Entscheidung: **annehmen, verwerfen oder zurückstellen**, jeweils mit Grund;
- eine gezielte Promptänderung und die erwartete Wirkung bei benannten Fällen oder Kriterien.

Sie müssen nicht jede Entscheidungskategorie verwenden oder künstlich Kritik ablehnen. Verwerfen Sie eine Kritik, braucht die gewählte Änderung einen anderen geprüften Befund als Grundlage.

Erstellen Sie **G1 als Kopie von G0**. Ändern Sie eine Anweisung im Generator-Prompt; halten Sie die übrigen Bedingungen aus «Auf einen Blick» gleich. Eine weitere Optimierungsrunde ist nicht verlangt.

4. G0 und G1 vergleichen

Erzeugen Sie neue G1-Ausgaben zu E01–E03 und lassen Sie diese von J0 bewerten. Geben Sie wieder zuerst Ihre eigenen Urteile ein. Öffnen Sie erst danach **Vergleich** und wählen Sie G0 und G1 mit J0. Die Funktion bewertet gespeicherte Ausgaben; G0 bleibt erhalten.

Füllen Sie die drei Zeilen der Berichtstabelle aus, eine je Entwicklungsfall. Zeigen Sie an konkreten Stellen beider Ausgaben:

- den Judge-Befund und Ihre fachliche Prüfung;
- behobene, verbliebene oder neue Fehler und erhaltene Stärken;
- Ihr Urteil zur Veränderung und die verbleibende Unsicherheit.

Nehmen Sie auch Gleichstände, Verschlechterungen und fehlgeschlagene Aufrufe auf. Ein höherer Score allein belegt keine Verbesserung. Ein Lauf pro Konfiguration trennt den Effekt der Anweisung nicht sicher von zufälligen Schwankungen. Ihre Schlussfolgerung gilt zunächst nur für diese Fälle.

5. Abschlussfall prüfen

Wählen Sie **G0 oder G1** anhand Ihrer Prüfung. Notieren Sie Wahl, Grund und Erwartung an A01 **vor dem Abschlusstest**. Sie dürfen G0 behalten, wenn G1 nicht überzeugt.

Bereiten Sie den Test unter **Abschluss** aus dem gewählten Stand vor. Führen Sie A01 mit der übernommenen Konfiguration und J0 aus. Passen Sie nach der Offenlegung weder Prompt, Kontext, Kriterien noch Modell an. Urteilen Sie wieder vor Einsicht in das Judge-Ergebnis.

Halten Sie fest, ob die Erwartung zutraf und welche Grenze der Übertragung sichtbar wurde. Ein Abschlussfall belegt keine allgemeine Zuverlässigkeit. Formulieren Sie eine praktische Regel für künftige Verbesserungsprozesse. Laden Sie unter **Dashboard & Export** den **Gesamtexport (JSON)** herunter.

Artefakte

Reichen Sie **eine ZIP-Datei mit zwei Bestandteilen** ein:

1. **Kurzbericht.md oder Kurzbericht.pdf:** ein Dokument nach der Vorlage. Als Richtwert genügen zwei bis drei Seiten. Inhalt: Auftrag, Herkunft der Fachgrundlage, Vorbereitungsnotiz, Erwartungen, begründete Änderung, drei Fallvergleiche, Abschlussfall und Schlussfolgerung. Verweisen Sie auf IDs im Export. Schreiben Sie vollständige Prompts, Rubriken und Ausgaben nicht nochmals ab. Kennzeichnen Sie die relevanten Iterationen, falls der Export weitere Übungsläufe enthält.
2. **customer-support-arena-workspace.json:** vollständiger Export nach dem Abschluss, einschliesslich Fällen, Fachgrundlage, Konfigurationen, Ausgaben, Bewertungen und menschlichen Urteilen.

Die ausgefüllten Import- und Promptdateien dienen der Vorbereitung. Ihre Inhalte stehen nach dem Import im Export. Reichen Sie keinen zusätzlichen Arbeitsordner und kein zweites Dossier ein.

Evidenz und Bestehen

Die Pflichtaufgabe ist bestanden, wenn:

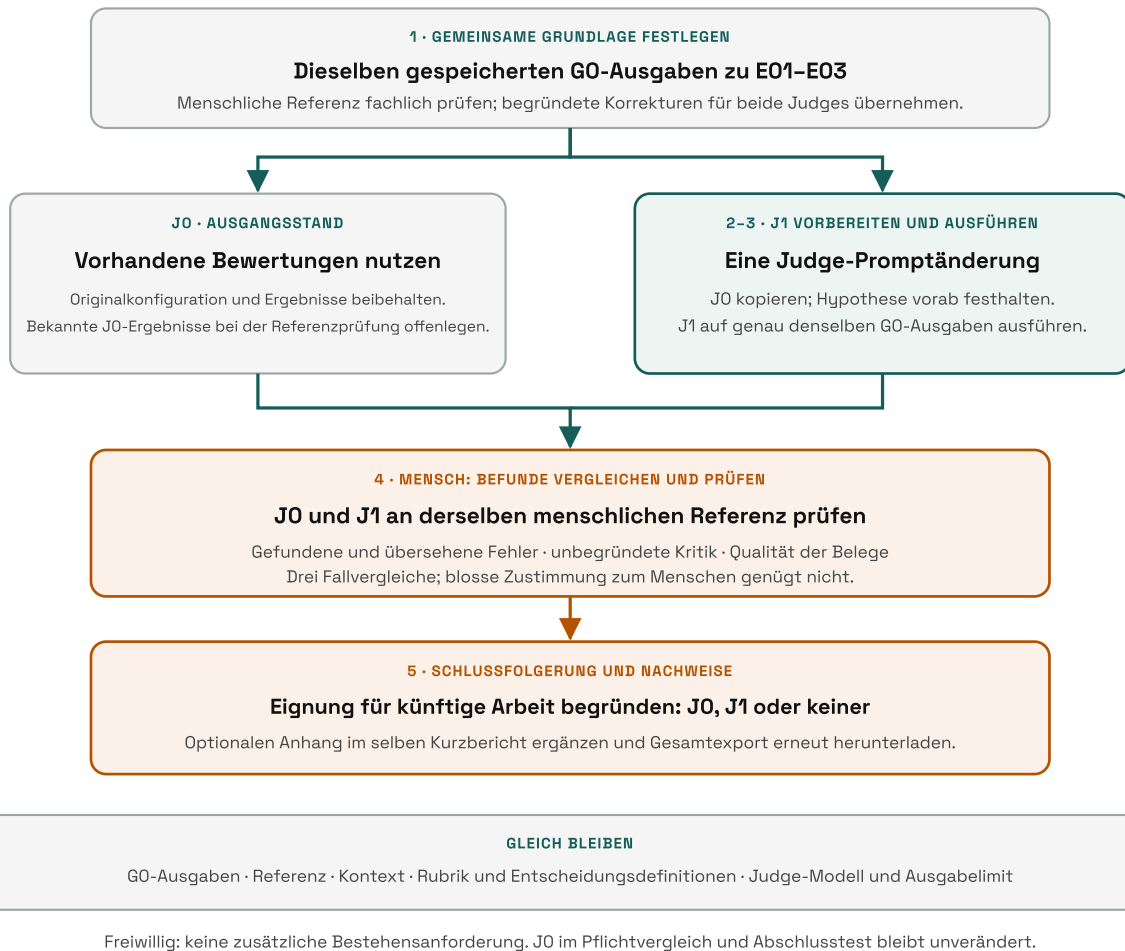
- eigener Anwendungsfall, Fachgrundlage, Prompts, vier Fälle und drei Kriterien nachvollziehbar vorbereitet sind;
- Ihre fachlich begründeten Urteile vor den jeweiligen Judge-Ergebnissen erfasst wurden;
- ein geprüfter Befund und Ihre Entscheidung die vorab formulierte Änderungshypothese stützen;
- G0/G1 unter gleichen Bedingungen ausgeführt und alle drei Fallpaare mit J0 geprüft wurden;
- der gewählte Stand an A01 geprüft und die Aussagekraft des Versuchs eingeordnet wurde;
- Bericht und Export die Ergebnisse belegen und eine begründete praktische Regel enthalten.

Optionale Erweiterung: Judges vergleichen

Bearbeiten Sie diese Erweiterung bei Interesse **nach der Pflichtaufgabe**. Sie vergleichen die Qualität zweier Bewertungen. Zusätzliche Generatorläufe oder Abschlussfälle sind nicht nötig; die Erweiterung ergänzt keine Bestehensanforderung.

Optional · Judges vergleichen und verbessern

Nach der Pflichtaufgabe: vorhandene G0-Ausgaben nutzen · keine zusätzlichen Generatorläufe



Ein Judge kann Fehler übersehen oder unbegründete Kritik liefern. Entscheidend sind fachlich bessere Befunde. Mehr Übereinstimmung mit Ihrer Meinung reicht nicht: Auch Ihr Urteil braucht Belege aus Aufgabe und Fachgrundlage.

- 1. Referenz prüfen.** Verwenden Sie G0-Ausgaben zu E01-E03 und Ihre bestehenden Urteile aus Schritt 2. Korrigieren Sie nachweisliche Fehler mit Begründung. Beide Judges werden an derselben geprüften Referenz gemessen. Legen Sie offen, dass Sie J0-Ergebnisse bereits kennen; dies ist keine neue unabhängige Beurteilung.
- 2. J1 vorbereiten.** Wählen Sie eine beobachtete oder begründet vermutete Schwäche von J0. Notieren Sie vorab eine Hypothese zu Fällen und Kriterien. Kopieren Sie J0 als J1 und ändern Sie eine Judge-Anweisung. G0-Ausgaben, Kontext-Version, Rubrik, Anker, Gesamtentscheidungsdefinitionen, Judge-Modell und Ausgabelimit bleiben gleich.
- 3. Dieselben Ausgaben bewerten.** Speichern und frieren Sie J1 ein. Lassen Sie ihn genau die gespeicherten G0-Ausgaben bewerten. Verlangt die Anwendung erneut ein menschliches Urteil, übernehmen Sie die festgelegte Referenz und vermerken dies im Bericht. Wiederverwendete gespeicherte Judge-Bewertungen sind keine neuen unabhängigen Modellurteile.

4. **Alle drei Fälle vergleichen.** Halten Sie Referenz, J0-/J1-Befunde und Ihre Prüfung fest: gefundene oder übersehene Fehler, unbegründete Beanstandungen und Qualität der Belege. Nehmen Sie unveränderte Ergebnisse, neue Probleme und fehlgeschlagene Aufrufe auf.
5. **Schlussfolgern und ergänzen.** Begründen Sie, welchen Judge Sie künftig nutzen würden und wo menschliche Prüfung nötig bleibt. J0 oder keiner von beiden sind zulässige Antworten. Ergänzen Sie den optionalen Anhang desselben Berichts um Hypothese, Änderung, IDs, Referenzkorrekturen, drei Fallvergleiche, Schlussfolgerung und praktische Regel. Laden Sie den Export erneut herunter. Wiederholen Sie den Pflichtvergleich nicht; dort bleibt J0 unverändert.

Auch ein höherer Score oder ein besserer J1 ist nicht erforderlich. Ein Lauf pro Judge mit drei Fällen erlaubt nur begrenzte Aussagen zu Ursachen und allgemeiner Zuverlässigkeit.