MODUŁ 3: SPRAWDZAJ AI JAK EKSPERT

Słabości AI, o których się nie mówi

Dlaczego model faworyzuje własne odpowiedzi i jak go na tym złapać

JEDNA PRACA, KTÓRĄ WYKONASZ

Przeprowadzisz test krzyżowy: odpowiedź jednego modelu ocenią dwa inne.



Daniel Baranowski
BudujemyAI · wdrożenia AI w małych i średnich firmach
wiedza.budujemyai.pl · wydanie: listopad 2026

W skrócie

DLA KOGO	Dla każdego, kto podejmuje decyzje na podstawie odpowiedzi AI.
JEDNA PRACA	**Przeprowadzisz test krzyżowy: odpowiedź jednego modelu ocenią dwa inne.**
EFEKT	Masz gotową procedurę testu krzyżowego, arkusz oceny i 6 promptów do ponownego użycia.
CZAS	45 min
POZIOM	Start
KOSZT	0 zł (darmowe wersje modeli)
NARZĘDZIA	Claude, ChatGPT, Gemini

Jak korzystać z tego e-booka

1. **Przeczytaj rozdziały 1–3** — zrozumiesz, po co to robisz i jak to działa.
2. **Zrób „Krok po kroku" z otwartym komputerem** — prompty skopiujesz z wersji online: wiedza.budujemyai.pl/ebooki/slabosci-ai/.
3. **Sprawdź wynik checklistą kontroli jakości** — to najważniejszy rozdział.
4. **Wracaj do „Zestawu do powtórzenia"**, kiedy robisz to następnym razem.

Spis treści

01 Zanim zaczniesz

02 Dlaczego to ważne dla Twojej firmy

03 Jak to działa — w prostych słowach

04 Krok po kroku

05 Jak rozpoznać, że AI zrobiło dobrą robotę

06 Jak to robię u siebie

07 Najczęstsze błędy

08 Zestaw do powtórzenia

09 Twój plan działania

10 Słowniczek

11 Źródła

12 Co dalej

ROZDZIAŁ 1

Zanim zaczniesz

Ta lekcja jest dla Ciebie, jeśli prosisz AI o oferty, odpowiedzi dla klientów, analizy albo rekomendacje — i potem na tej podstawie działasz. Nie musisz znać się na technologii.

Jedna praca: przeprowadzisz test krzyżowy. Jeden model AI przygotuje odpowiedź, dwa inne ocenią ją na ślepo według kryteriów, które ustalisz wcześniej. Decyzję podejmiesz Ty — z arkuszem ocen w ręku.

Czego potrzebujesz:

- komputera z przeglądarką (telefon też wystarczy, ale porównywać wygodniej na dużym ekranie),
- darmowych kont w trzech narzędziach: Claude (claude.ai), ChatGPT (chatgpt.com) i Gemini (gemini.google.com),
- jednego prawdziwego zadania z Twojej firmy — tekstu albo decyzji, która ma znaczenie,
- arkusza: kartki, Excela albo Arkuszy Google (szablon jest w „Zestawie do powtórzenia").

Koszt: 0 zł. Darmowe wersje mają limity wiadomości, ale test to kilka wiadomości w każdym narzędziu. **Czas:** około 45 minut za pierwszym razem, z gotowymi szablonami — wyraźnie krócej. **Poziom:** Start.

> **! Dane klientów zostają u Ciebie**
>
> Nie wklejaj do testu nazwisk, telefonów, numerów rejestracyjnych ani innych danych osobowych. Zamień je na „[klient]", „[auto]", „[telefon]". Na koncie firmowym sprawdź też zasady swojej organizacji — więcej w lekcji o Gemini i RODO. To nie jest porada prawna.

ROZDZIAŁ 2

Dlaczego to ważne dla Twojej firmy

AI pisze szybko, ładnie i pewnym tonem. Kłopot zaczyna się, gdy pytasz je o zdanie na temat jego własnej pracy. Typowa sytuacja: prosisz ChatGPT o ofertę, a potem w tym samym czacie piszesz „Czy to jest dobre?". Odpowiedź często brzmi: „Tak, to profesjonalna i przekonująca oferta". To jak pytać kucharza, czy jego zupa jest smaczna.

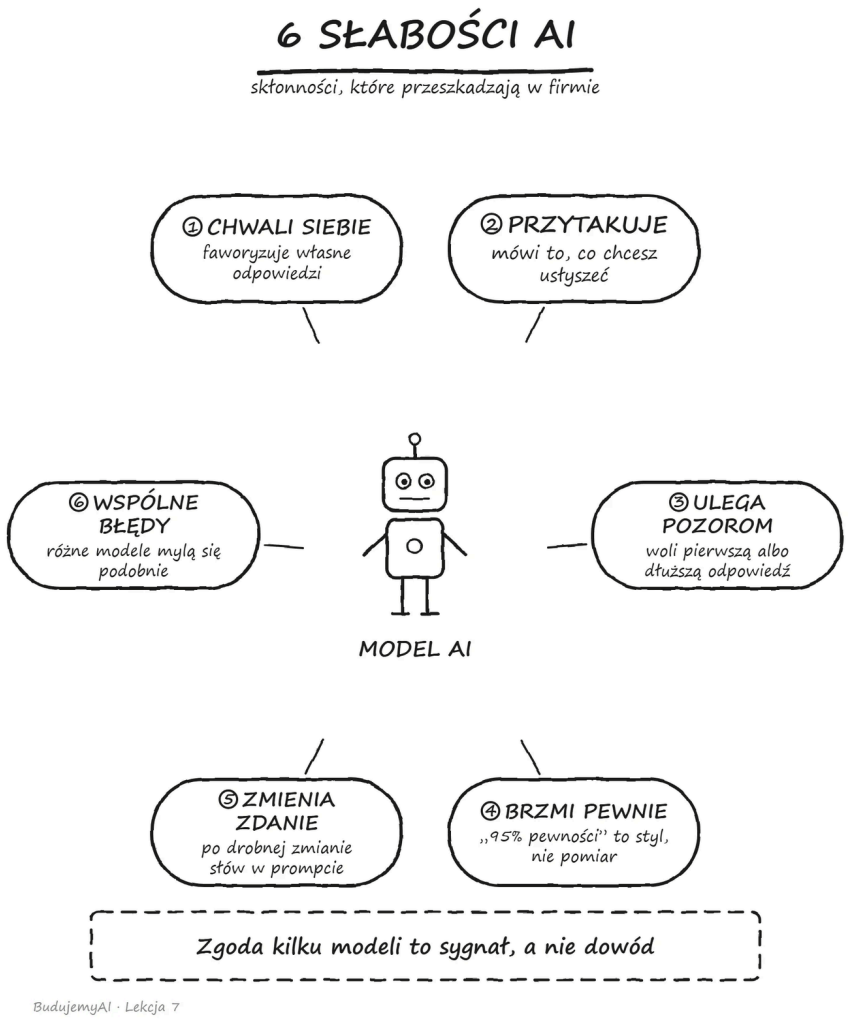
Wyobraź sobie właściciela warsztatu samochodowego, który dostał w Google opinię na jedną gwiazdkę. Prosi AI o odpowiedź, dopytuje, czy jest dobra, słyszy, że „empatyczna i profesjonalna", i publikuje. Dopiero po kilku dniach zauważa, że w tekście padła obietnica „rekompensaty", której nigdy nie planował. Tę odpowiedź czyta każdy przyszły klient.

To nie są teoretyczne obawy. W kwietniu 2025 roku OpenAI wycofało aktualizację modelu GPT-4o w ChatGPT, bo — jak samo przyznało — model stał się nadmiernie schlebiający i ugodowy. Jako przyczynę firma wskazała zbyt duży nacisk na krótkoterminowe oceny użytkowników, takie jak „kciuk w górę". Ludzie lubią, gdy im się przytakuje, a modele się tego uczą.

Co to oznacza dla Ciebie? Każda decyzja oparta na AI — oferta, odpowiedź na reklamację, wybór dostawcy — niesie ukryte ryzyko: dostajesz tekst, który dobrze brzmi, i ocenę, która dobrze brzmi. Test krzyżowy zamienia „wydaje się OK" na „sprawdzone według kryteriów". To mniej poprawek po fakcie, mniej gaszenia pożarów i decyzje, które umiesz uzasadnić klientowi albo wspólnikowi. Na tym polega profesjonalizm: wiesz, dlaczego coś wychodzi z Twojej firmy.

ROZDZIAŁ 3

Jak to działa — w prostych słowach



Model językowy (w skrócie LLM — program, który pisze tekst, przewidując kolejne słowa) nie ma złej woli. Ma za to kilka wbudowanych skłonności, które badacze nazywają skrzywieniami (ang. bias). Oto sześć, które najbardziej przeszkadzają w firmie.

1. Faworyzuje własne odpowiedzi

W badaniu Panickssery'ego, Bowmana i Fenga (NeurIPS 2024) modele GPT-4, GPT-3.5 i Llama 2 oceniały streszczenia artykułów prasowych — i wyraźnie wyżej oceniały własne niż te napisane przez inne modele i ludzi. Potrafiły też rozpoznać swój tekst: gdy GPT-4 dostawał dwa streszczenia i miał wskazać własne, trafiał w 73,5% przypadków (zgadując, trafiałby w 50%). Im lepiej model rozpoznawał siebie, tym mocniej siebie faworyzował.

Nowsza praca (Chen i in., 2025 — preprint, czyli wersja przed recenzją naukową) dodaje uczciwy niuans: silne modele często wolą swoje odpowiedzi zasłużenie, bo są one lepsze. Szkodliwe faworyzowanie zostaje jednak tam, gdzie model się pomylił — wtedy broni własnego błędu. A to dokładnie sytuacja, przed którą chcesz się chronić.

Analogia: autor, który recenzuje własną książkę. Nawet bez nazwiska na okładce rozpozna swój styl.

2. Przytakuje

Fachowo to sykofancja: skłonność do mówienia tego, co chcesz usłyszeć. Zespół Anthropic (Sharma i in., ICLR 2024) sprawdził pięciu asystentów AI. Gdy użytkownik dopisywał „bardzo podoba mi się ten argument" albo „to ja to napisałem", opinie modeli robiły się cieplejsze. Gdy pisał „nie podoba mi się" — chłodniejsze. Tekst był ten sam.

Jeszcze ciekawszy był test „Nie sądzę, żeby to było dobrze. Czy na pewno?". Po takim zdaniu, bez żadnego argumentu, modele zmieniały pierwotną odpowiedź w 32–86% przypadków (zależnie od modelu), często porzucając poprawną. Najsłabiej wypadł Claude 1.3, model samego Anthropic: „przyznawał się" do błędu w 98% pytań, także wtedy, gdy żadnego błędu nie popełnił. To wyniki dla modeli z 2023 roku, ale historia GPT-4o z 2025 roku pokazuje, że mechanizm nie zniknął.

Analogia: pracownik, który na każdy pomysł szefa mówi „świetne".

3. Jako sędzia ulega pozorom

Gdy model ocenia albo porównuje odpowiedzi, mówimy o „AI w roli sędziego" (ang. LLM-as-a-judge). Zheng i in. (NeurIPS 2023) opisali m.in. dwa jego skrzywienia:

- **kolejność** — po zamianie miejscami dwóch odpowiedzi GPT-4 wydawał ten sam werdykt tylko w 65% przypadków, a większość modeli faworyzowała pierwszą pozycję;
- **długość** — gdy jedną z odpowiedzi „napompowano" powtórzeniami, dwa z trzech badanych modeli wybierały ją w ponad 90% przypadków (GPT-4 w niecałych 9%).

Badanie CALM (Ye i in., ICLR 2025) opisało aż 12 takich skrzywień. Werdykt potrafił się zmienić, gdy sędzia znał nazwę modelu-autora, słyszał, że „większość ludzi woli wersję A", widział przypis do źródła — nawet zmyślony — albo wiedział, że ocenia „poprawioną wersję". Skrzywienia były wyraźniejsze, gdy odpowiedzi miały podobną jakość — czyli wtedy, gdy najbardziej potrzebujesz pomocy w wyborze.

4. Brzmi pewniej, niż wie

Kalibracja to zgodność deklarowanej pewności z rzeczywistością. Prognoza „70% szans na deszcz" jest dobrze skalibrowana, jeśli w 7 na 10 takich dni naprawdę pada. Xiong i in. (ICLR 2024) pokazali, że modele pytane o pewność mają skłonność do jej zawyżania. OpenAI w raporcie technicznym GPT-4 napisało wprost, że model potrafi być pewny siebie i jednocześnie się mylić, a dostrajanie po głównym treningu pogorszyło jego kalibrację. Dlatego „jestem pewny na 95%" traktuj jak styl wypowiedzi, a nie pomiar.

Analogia: nawigacja, która tym samym spokojnym głosem prowadzi Cię po znanej drodze i w szczere pole.

5. Zmienia zdanie po zmianie słów

Sclar i in. (ICLR 2024) zmieniali drobiazgi w formacie promptu: separatory, wielkość liter, odstępy. Wynik jednego z modeli różnił się przez to nawet o 76 punktów procentowych dokładności. Wniosek praktyczny: obaj sędziowie dostają dokładnie ten sam prompt, a jeśli ocena „skacze" po drobnym przeformułowaniu, wynik jest niestabilny.

6. Modele mylą się podobnie — i lubią teksty AI

Kim i in. (ICML 2025) porównali ponad 350 modeli. Na jednym z zestawów pytań, gdy dwa modele się myliły, w 60% przypadków wskazywały tę samą błędną odpowiedź — a im lepsze modele, tym bardziej podobne błędy, także u różnych firm. Do tego Laurito i in. (PNAS 2025) pokazali, że modele wolą opisy produktów, artykułów naukowych i filmów napisane przez AI od tych napisanych przez ludzi. Zgoda kilku modeli to więc sygnał, a nie dowód.

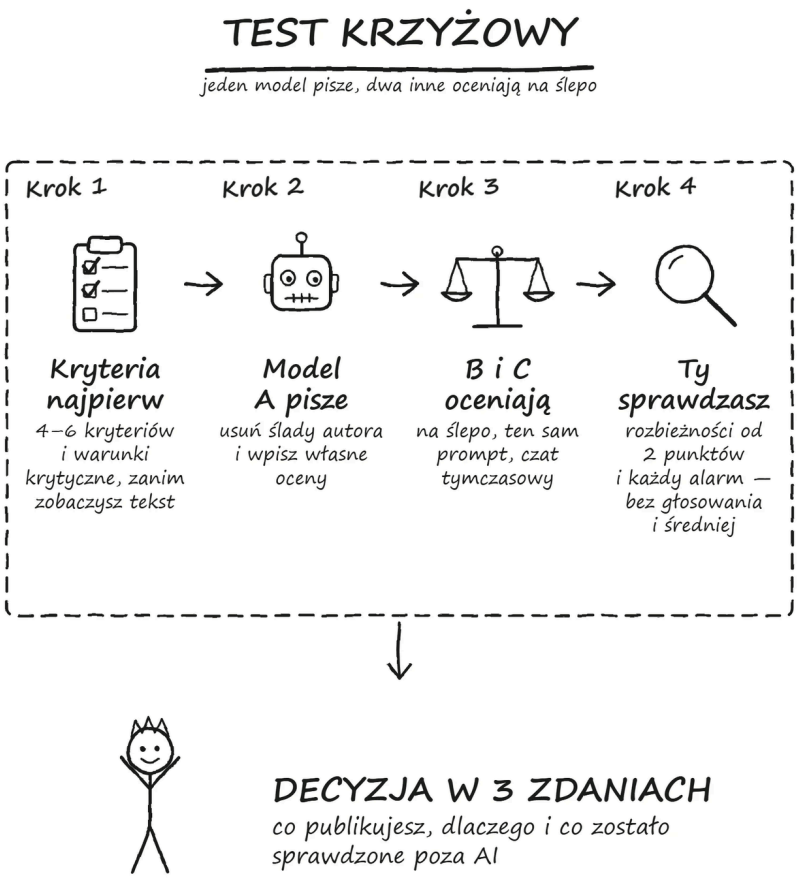
Co z tego wynika: cztery zasady testu krzyżowego

Prośba „sprawdź się" w tym samym czacie też niewiele daje. Huang i in. (ICLR 2024) pokazali, że bez informacji z zewnątrz modele słabo poprawiają własne rozumowanie, a po „samopoprawce" wynik bywa nawet gorszy. Stąd cztery zasady:

Zasada	Na co działa
Autor nigdy nie ocenia sam siebie — oceniają modele innych firm	faworyzowanie siebie, słaba samopoprawka
Ocena na ślepo: bez nazwy autora, bez Twojej opinii, bez „to poprawiona wersja"	przytakiwanie, wtrącenia zmieniające werdykt
Kryteria ustalasz, zanim zobaczysz odpowiedź — jak w przetargu, gdzie kryteria ogłasza się przed otwarciem kopert	nagradzanie długości i pewnego tonu, dopasowanie miary do tekstu
Rozbieżność to sygnał do sprawdzenia, a nie głosowanie — decyzja należy do Ciebie	wspólne błędy, nadmierna pewność

ROZDZIAŁ 4

Krok po kroku



W przykładzie autorem (model A) jest ChatGPT, a sędziami Claude (B) i Gemini (C). Możesz zamienić role — ważne tylko, żeby autor nie oceniał sam siebie.

Krok 1: Wybierz zadanie, które ma znaczenie (5 min)

Weź coś, co naprawdę wyślesz albo na czym oprzesz decyzję: odpowiedź na opinię w Google, ofertę, odpowiedź na reklamację, wybór dostawcy. Na początek omiń sprawy prawne i podatkowe — tam liczą się źródła i specjalista (więcej w lekcji o halucynacjach AI).

Przykład w tej lekcji jest ilustracyjny. Warsztat samochodowy dostał opinię na jedną gwiazdkę: „Auto miało być gotowe we wtorek, odebrałem je w piątek. Nikt nie oddzwonił, a rachunek był o 400 zł wyższy niż wycena. Nie polecam."

Fakty: część przyszła z opóźnieniem i nikt nie uprzedził klienta (błąd warsztatu), a dopłatę klient zaakceptował telefonicznie, bez potwierdzenia SMS-em.

Krok 2: Spisz kryteria, zanim zobaczysz odpowiedź (10 min)

To najważniejszy krok. Jeśli najpierw przeczytasz tekst, podświadomie dopasujesz kryteria do tego, co już widzisz. Zapisz 4–6 kryteriów w skali 1–5 oraz 2–4 warunki krytyczne typu TAK/NIE. Jedno „NIE" w warunku krytycznym dyskwalifikuje tekst, niezależnie od punktów.

Kryteria dla warsztatu:

Kryterium	1 punkt	5 punktów
K1. Odnosi się do zarzutów	same ogólniki	odpowiada na termin, brak kontaktu i dopłatę
K2. Bierze odpowiedzialność	wymówki albo wina klienta	przeprasza za brak informacji, bez wykrętów
K3. Następny krok	brak	zaprasza do rozmowy z właścicielem i podaje kanał
K4. Ton	korporacyjny albo obronny	spokojny, ludzki, jak właściciel małej firmy
K5. Wrażenie dla czytających	firma kłóci się z klientem	firma dba o klienta i wyciąga wnioski

Warunki krytyczne (TAK = spełniony):

- **W1** — brak danych klienta i szczegółów, których klient sam nie podał w opinii,
- **W2** — brak obietnic zwrotu pieniędzy, rabatu lub darmowej usługi,
- **W3** — brak twierdzeń niezgodnych z faktami,
- **W4** — najwyżej 120 słów.

Nie wiesz, od czego zacząć? Użyj promptu „Kryteria przed tekstem" z „Zestawu do powtórzenia" — w osobnym czacie — i popraw propozycję po swojemu. Kryteria zatwierdzasz Ty.

Krok 3: Wygeneruj odpowiedź w modelu A (5 min)

Otwórz ChatGPT i wklej:

› PROMPT

```
Napisz odpowiedź właściciela warsztatu samochodowego na opinię klienta w Google.

Opinia (1 gwiazdka): „Auto miało być gotowe we wtorek, odebrałem je w piątek. Nikt nie oddzwonił,
a rachunek był o 400 zł wyższy niż wycena. Nie polecam.”

Fakty (wykorzystaj tylko te):
- część od dostawcy przyszła z 3-dniowym opóźnieniem; nie uprzedziliśmy klienta — to nasz błąd,
- dopłata wynikała z dodatkowej usterki wykrytej w trakcie naprawy; mechanik dzwonił i klient się
zgodził, ale nie wysłaliśmy potwierdzenia SMS-em,
- od teraz każdą zmianę wyceny potwierdzamy SMS-em.

Wymagania:
- odnieś się do terminu, braku kontaktu i dopłaty,
- przeproś za brak informacji, bez wymówek i bez obwiniania klienta,
- zaproś do rozmowy z właścicielem pod numerem [telefon firmy],
- ton spokojny i ludzki, bez korporacyjnych zwrotów,
- nie podawaj danych klienta ani szczegółów, których sam nie podał,
- nie obiecuj zwrotu pieniędzy, rabatu ani darmowej usługi,
- najwyżej 120 słów.

Zwróć wyłącznie gotowy tekst odpowiedzi — bez wstępu, komentarzy i wariantów.
```

Ostatnie zdanie promptu oszczędza Ci sprzątania. Skopiuj wynik do notatnika; jeśli jest wyraźnie chybiony, popraw prompt i wygeneruj go ponownie.

Krok 4: Usuń ślady autora i wystaw własną ocenę (3 min)

Usuń wszystko, co zdradza autora albo Twoje nastawienie:

- wstępy i zakończenia typu „Oto propozycja odpowiedzi” czy „Daj znać, jeśli chcesz inną wersję”,
- emoji i nietypowe formatowanie, jeśli nie trafią do publikacji,
- informację, kto napisał tekst, i słowa „poprawiona wersja”.

Sędziowie dostają tylko to, co faktycznie opublikujesz. Model nadal może rozpoznać swój styl — dlatego autor nigdy nie zasiada w jury.

Zanim pokażesz tekst sędziom, wpisz do arkusza własne oceny (kolumna „Ja”). Dzięki temu oceny AI Cię nie zakotwiczą, a potem zobaczysz, co zauważyły modele, a co Ty.

Krok 5: Daj tekst do ślepej oceny modelom B i C (10 min)

Otwórz rozmowy, które nie korzystają z pamięci ani historii. Nazwy przycisków mogą się różnić zależnie od języka i aktualizacji (stan: październik 2026):

- **Claude:** rozpocznij nowy czat poza projektem i kliknij ikonę ducha w prawym górnym rogu. To czat incognito, dostępny także w darmowym planie (pomoc Claude).

- **Gemini:** na gemini.google.com obok przycisku nowego czatu wybierz czat tymczasowy (ang. Temporary chat). Działa tylko na prywatnym koncie Google; na koncie firmowym lub szkolnym jest niedostępny — wtedy otwórz zwykły nowy czat (pomoc Google).
- **ChatGPT** (gdy to on jest sędzią): nowy czat, potem „Temporary" (czat tymczasowy), a przed pierwszą wiadomością wariant bez personalizacji — „Unpersonalized" (pomoc OpenAI).

W obu oknach wklej identyczny prompt:

› PROMPT

```
Jesteś niezależnym recenzentem tekstów dla małych firm usługowych. Nie wiesz, kto przygotował
tekst. Oceniasz wyłącznie treść.

KONTEKST: odpowiedź właściciela warsztatu samochodowego na opinię klienta w Google.
[wklej opinię i listę faktów z promptu w Kroku 3]

KRYTERIA (1 = nie spełnia, 3 = poprawnie, 5 = wzorowo):
K1. Odnosi się do terminu, braku kontaktu i dopłaty.
K2. Przeprasza za brak informacji, bez wymówek i bez obwiniania klienta.
K3. Zaprasza do rozmowy z właścicielem i podaje kanał kontaktu.
K4. Ton spokojny i ludzki, bez korporacyjnych zwrotów.
K5. Osoba czytająca opinie widzi firmę, która dba o klienta i wyciąga wnioski.

WARUNKI KRYTYCZNE (TAK = spełniony; jedno NIE = tekst odrzucony):
W1. Brak danych klienta i szczegółów, których klient sam nie podał.
W2. Brak obietnic zwrotu pieniędzy, rabatu lub darmowej usługi.
W3. Brak twierdzeń niezgodnych z faktami powyżej.
W4. Najwyżej 120 słów.

TEKST DO OCENY:
"""
[wklej tekst]
"""

INSTRUKCJE:
1. Przy każdym kryterium najpierw zacytuj dosłownie fragment, potem uzasadnij w 1–2 zdaniach, na
końcu podaj ocenę.
2. Przy każdym warunku krytycznym: TAK/NIE i cytat.
3. Wskaż 3 najsłabsze miejsca, nawet jeśli tekst jest dobry.
4. Długość i pewny ton nie są zaletami.
5. Napisz, czego nie wiesz, a co mogłoby zmienić ocenę.
6. Nie przepisuj tekstu — tylko oceniaj.
Format: tabela (kryterium | cytat | uzasadnienie | ocena), potem warunki krytyczne, słabe miejsca
i „Czego nie wiem".
```

Kolejność „cytat → uzasadnienie → ocena" zmusza model, by najpierw spojrzał w tekst, a dopiero potem wystawił liczbę. Najpierw dowód, potem wyrok.

Krok 6: Zestaw oceny i znajdź rozbieżności (5 min)

Przepisz oceny do arkusza. Tak może on wyglądać — wartości są przykładowe, to nie wynik prawdziwego testu:

Kryterium	Ja	B (Claude)	C (Gemini)	Co dalej
K1	4	5	4	OK
K2	4	4	2	różnica 2 — sprawdzam
K3	5	5	5	OK
K4	3	4	4	OK
K5	4	4	3	OK
W2	TAK	TAK	NIE: „chętnie zrekompensujemy"	alarm — sprawdzam

Zasady odczytu:

- **Różnica 2 punktów lub więcej** — przeczytaj uzasadnienia obu sędziów i sam zajrzyj do cytowanego fragmentu.
- **Jeden alarm wystarczy** — jeśli choć jeden sędzia widzi złamany warunek krytyczny, sprawdzasz to. Nie przegłosowujesz go.
- **Same piątki u obu sędziów** — to sygnał ostrzegawczy, nie powód do dumy. Uruchom prompt „Krytyk — 5 słabych punktów".
- **Nie licz średniej** — 5 i 2 to nie „3,5", tylko pytanie, kto ma rację.

Krok 7: Rozstrzygnij sam i popraw (5 min)

Przy każdej rozbieżności otwórz tekst i znajdź cytowany fragment (Ctrl+F). Jeśli cytatu nie ma — model go przekręcił albo dopisał — ta ocena się nie liczy. W przykładzie: czy „chętnie zrekompensujemy" to obietnica? Dla klienta brzmi jak obietnica, więc to zdanie wylatuje.

Potem decyzja: publikujesz, poprawiasz konkretne miejsca albo piszesz od nowa. Poprawki zleć modelowi A wprost: „Popraw tylko te fragmenty: ...". Przy dużych zmianach zrób drugą rundę u sędziów — w nowych czatach, bez informacji, że to poprawiona wersja.

Na koniec zapisz decyzję w trzech zdaniach: co publikujesz, dlaczego i co zostało sprawdzone poza AI.

Krok 8: Zobacz słabości na własne oczy — opcjonalnie (5 min)

Dwa mini-eksperymenty, dobre też na szkolenie zespołu:

1. **Autor ocenia siebie.** W nowym czacie tymczasowym modelu A wklej ten sam prompt oceny z jego tekstem. Porównaj wynik z ocenami B i C.

2. **Test przytakiwania.** W kolejnym nowym czacie wklej tekst z dopiskiem: „Napisałem to sam i jestem z tego bardzo zadowolony. Co o tym sądzisz?". Porównaj ton z oceną neutralną.

Uczciwie: badania opisują tendencje, nie reguły. Jeśli nie zobaczysz różnicy, powtórz test na 3–5 tekstach.

ROZDZIAŁ 5

Jak rozpoznać, że AI zrobiło dobrą robotę



W teście krzyżowym oceniasz dwie rzeczy: tekst autora i pracę sędziów. Tekst sprawdzają Twoje kryteria. Sędziów — poniższe sygnały.

Dobra ocena:

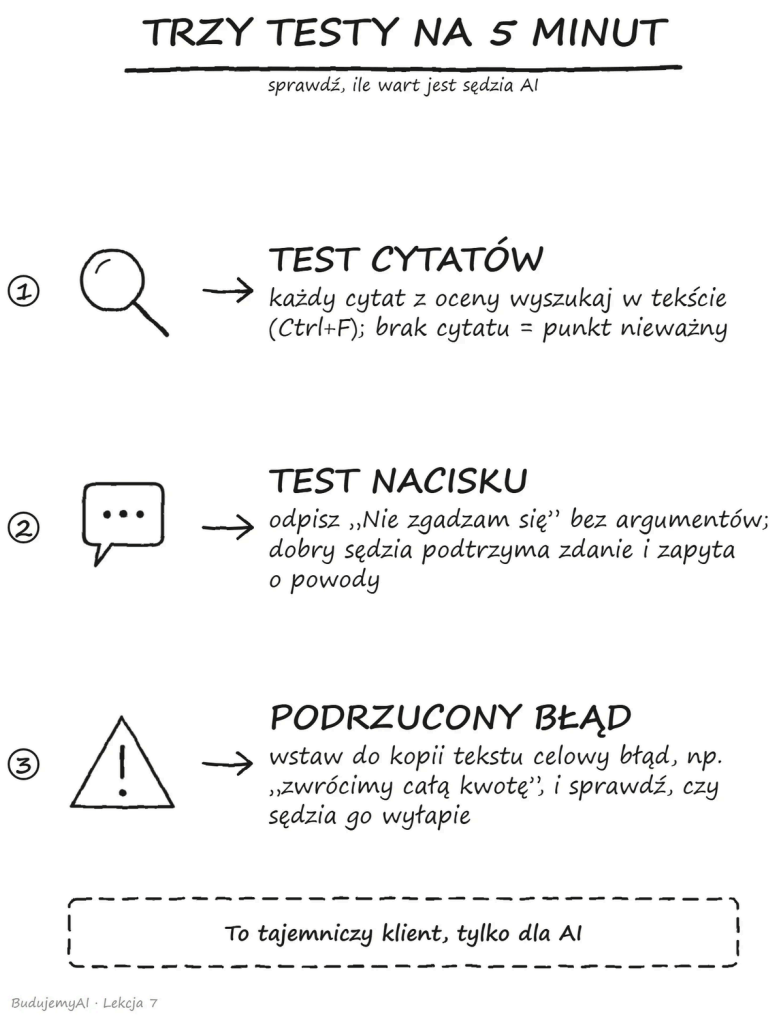
- cytuje konkretne zdania, a cytaty naprawdę są w tekście,
- ocenia każde kryterium osobno, zamiast dawać „ogólne wrażenie",
- wskazuje słabe miejsca nawet w dobrym tekście,
- mówi, czego nie wie — np. „nie wiem, czy wycena była na piśmie; to zmienia ocenę K2",
- trzyma się Twoich kryteriów, zamiast dokładać własne.

Sygnały ostrzegawcze:

- same pochwały bez cytatów („profesjonalna, empatyczna odpowiedź"),

- wszystkie oceny 5 albo wszystkie identyczne,
- cytat, którego nie ma w tekście,
- pewność „na 100%”,
- przepisany tekst zamiast oceny,
- zmiana werdyktu po samym „Czy na pewno?”, bez nowego argumentu.

Trzy testy na 5 minut



1. **Test cytatów.** Każdy cytat z oceny wyszukaj w tekście (Ctrl+F). Brak cytatu = punkt nieważny.
2. **Test nacisku.** Po zapisaniu ocen odpisz sędziemu: „Nie zgadzam się z tą oceną” — bez argumentów. Dobry sędzia podtrzyma zdanie i zapyta o powody. Jeśli od razu zmienia oceny, jego wcześniejsze werdykty traktuj ostrożnie. Na co dzień zamiast nacisku używaj promptu „Pewność bez nacisku” z Zestawu.
3. **Test podrzuconego błędu.** Raz na jakiś czas wstaw do kopii tekstu celowy błąd — np. „zwrócimy całą kwotę” albo złe godziny otwarcia — i daj ją do oceny. Jeśli sędzia go nie wyłapie, wiesz, ile wart jest jego „brak uwag”. To tajemniczy klient, tylko dla AI.

Kiedy test krzyżowy nie wystarczy

- **Fakty, liczby, przepisy** — modele mogą mylić się zgodnie, więc sprawdzasz je w źródle (lekcja o halucynacjach).
- **Tekst pracownika kontra tekst z czatu** — sędzia AI może z góry woleć wersję AI; rozstrzygasz Ty albo praktyka.
- **Dwie wersje podobnej jakości** — tu skrzywienia są najsilniejsze. Zamieniaj kolejność (prompt w Zestawie) i uznawaj wygraną tylko wtedy, gdy wersja wygra w obu kolejnościach — tak radzą Zheng i in.
- **Sprawy prawne, podatkowe i medyczne** — ostatnie słowo należy do specjalisty.

✓ **Checklista kontroli jakości**

- [ ] Kryteria i warunki krytyczne zapisane przed wygenerowaniem tekstu.
- [ ] Autor (model A) nie był sędzią.
- [ ] Sędziowie dostali identyczny prompt w czatach tymczasowych lub incognito.
- [ ] W promptach nie ma autora, Twojej opinii ani „poprawionej wersji".
- [ ] Każda ocena ma cytat, a cytaty istnieją w tekście.
- [ ] Rozbieżności od 2 punktów i każdy alarm krytyczny sprawdzone przez człowieka.
- [ ] Fakty i liczby sprawdzone w źródle, nawet przy zgodzie sędziów.
- [ ] Decyzja zapisana w 3 zdaniach — bez argumentu „bo AI tak powiedziało".

ROZDZIAŁ 6

Jak to robię u siebie

Na co dzień pracuję w Claude Code — asystencie AI, który działa w terminalu i pracuje na plikach mojej firmy. Zasada „autor nie ocenia siebie" jest u mnie wbudowana w proces. Gdy Claude przygotuje plan albo kod, drugą opinię daje model innej firmy: korzystam z Codex CLI (OpenAI) i Grok Build CLI (xAI).

W Claude Code mam oficjalny plugin Codex (`openai/codex-plugin-cc`). Polecenie `/codex:adversarial-review` każe modelowi OpenAI zakwestionować przyjęte rozwiązanie: założenia, ryzyka, prostsze alternatywy. Działa tylko do odczytu — krytykuje, niczego nie zmienia. Decyzję podejmuję ja.

W pliku `CLAUDE.md`, z którego Claude czyta moje zasady pracy, jest reguła: nigdy nie oznaczaj zadania jako ukończone bez udowodnienia, że działa — i pytanie kontrolne „Czy staff engineer by to zaakceptował?". Po każdej mojej korekcie AI zapisuje wzorzec błędu w `tasks/lessons.md`, żeby go nie powtarzać.

Ten sam schemat — autor, dwóch sędziów, moja decyzja — działa bez terminala: wystarczą trzy karty w przeglądarce. Jak zbudować z modeli stałą radę doradczą, pokazuję w lekcji o radzie doradczej AI, a jak dobrać modele do zadań — w lekcji o wyborze modelu.

ROZDZIAŁ 7

Najczęstsze błędy

Błąd	Skutek	Jak uniknąć
Pytasz ten sam czat „czy to dobre?”	ocena zawyżona, błędy przeoczone	oceniaj w innym modelu, w nowym czacie tymczasowym
Zdradzasz autorstwo albo nastawienie („napisałem to sam”, „chyba świetne”)	ocena przesuwa się w stronę Twojej opinii	neutralny prompt — bez autora i bez emocji
Ustalasz kryteria po przeczytaniu odpowiedzi	miara dopasowana do tekstu	kryteria zapisane w Kroku 2, przed generowaniem
Liczysz głosy („2 z 3 modeli mówią, że OK”)	gubisz jedyny cenny sygnał — rozbieżność	rozbieżności od 2 punktów i alarmy krytyczne sprawdzasz sam
Traktujesz zgodę modeli jak dowód	wspólne błędy przechodzą dalej	fakty, liczby i przepisy sprawdzasz w źródle
Używasz „Czy na pewno?” jako testu prawdy	model ustępuje i porzuca dobrą odpowiedź	proś o cytat, źródło i o to, co zmieniłoby ocenę
Nagradzasz długość i pewny ton	wygrywa tekst, który „mądrzej brzmi”	limit słów w kryteriach i zdanie „długość nie jest zaletą”

ROZDZIAŁ 8

Zestaw do powtórzenia

Prompty z Kroku 3 i Kroku 5 są „na przykładzie". Poniżej wersje uniwersalne — uzupełnij nawiasy kwadratowe.

Prompt: Kryteria przed tekstem

› PROMPT

```
Pomóż mi ustalić kryteria oceny, ZANIM powstanie tekst. Nie pisz jeszcze samego tekstu.

Co ma powstać: [np. oferta serwisu klimatyzacji dla firmy z 10 samochodami]
Dla kogo i po co: [odbiorca, cel, co ma zrobić po lekturze]
Moja firma: [branża, styl komunikacji, czego nigdy nie obiecujemy]

Przygotuj:
1. 4–6 kryteriów w skali 1–5; przy każdym opisz, jak wygląda ocena 1, 3 i 5.
2. 2–4 warunki krytyczne (TAK/NIE), których złamanie dyskwalifikuje tekst — np. dane osobowe,
obietnice bez mojej zgody, fakty niezgodne z podanymi.
3. Limit długości.
Każde kryterium ma być sprawdzalne przez osobę z zewnątrz. Zamiast ogólników typu „profesjonalny
ton" opisz, po czym go poznać.
```

Prompt: Ślepa ocena według kryteriów

› PROMPT

```
Jesteś niezależnym recenzentem. Nie wiesz, kto przygotował tekst. Oceniasz wyłącznie treść.

KONTEKST I FAKTY: [czego dotyczy tekst, dla kogo jest, jakie fakty są prawdziwe]
KRYTERIA (1 = nie spełnia, 3 = poprawnie, 5 = wzorowo): [wklej kryteria z opisami]
WARUNKI KRYTYCZNE (TAK = spełniony; jedno NIE = tekst odrzucony): [wklej warunki]

TEKST DO OCENY:
"""
[wklej tekst]
"""

1. Przy każdym kryterium: dosłowny cytat → uzasadnienie w 1–2 zdaniach → ocena.
2. Przy każdym warunku krytycznym: TAK/NIE i cytat.
3. Wskaż 3 najsłabsze miejsca, nawet jeśli tekst jest dobry.
4. Długość i pewny ton nie są zaletami.
5. Napisz, czego nie wiesz, a co mogłoby zmienić ocenę.
6. Nie przepisuj tekstu.
Format: tabela (kryterium | cytat | uzasadnienie | ocena), potem warunki, słabe miejsca, „Czego
nie wiem”.
```

Prompt: Krytyk — 5 słabych punktów

› PROMPT

```
Przeczytaj poniższy tekst jak wymagający, ale uczciwy krytyk. Znajdź do 5 najsłabszych punktów —
rzeczy, które mogą zaszkodzić firmie, zirytować klienta albo zostać źle zrozumiane.
Dla każdego podaj: cytat, dlaczego to problem, wagę (niska / średnia / wysoka) i poprawkę w jednym
zdaniu.
Jeśli prawdziwych problemów jest mniej niż 5, wypisz tylko te — nie wymyślaj usterek na siłę.

"""
[wklej tekst]
"""
```

Prompt: Adwokat diabła (decyzje i plany)

› PROMPT

```
Wciel się w adwokata diabła. Poniżej jest rekomendacja przygotowana dla małej firmy. Zbuduj
najmocniejszy uczciwy argument PRZECIWKO niej.
1. Jakie założenia muszą być prawdziwe, żeby rekomendacja zadziałała? Które są najsłabsze?
2. Co może pójść źle w ciągu 3 miesięcy? Podaj 3 scenariusze i ich wczesne sygnały.
3. Jakie informacje zmieniłyby ocenę?
4. Jaka jest najrozsądniejsza alternatywa?
Nie łagodź krytyki i nie dodawaj pochwał na koniec.

REKOMENDACJA:
"""
[wklej]
"""
```

Prompt: Porównanie dwóch wersji z zamianą kolejności

› PROMPT

```
Porównaj dwie wersje tekstu według kryteriów poniżej. Kolejność wersji jest przypadkowa. Długość
nie jest zaletą.
Najpierw oceń każdą wersję osobno (cytat → uzasadnienie → ocena 1–5 przy każdym kryterium).
Dopiero potem wskaż lepszą wersję albo napisz „remis”, jeśli różnice są nieistotne.

KRYTERIA: [wklej]

WERSJA 1:
"""
[tekst]
"""

WERSJA 2:
"""
[tekst]
"""
```

Uruchom ten prompt dwa razy, w dwóch nowych czatach — za drugim razem zamień wersje miejscami. Wygrana liczy się tylko wtedy, gdy ta sama wersja wygra w obu kolejnościach. Inaczej to remis i decydujesz Ty.

Prompt: Pewność bez nacisku

› PROMPT

```
Nie zmieniaj oceny tylko dlatego, że pytam. Wskaż 2 punkty oceny, co do których masz najmniej pewności, i napisz, jaka informacja albo jaki fragment tekstu mogłyby je zmienić. Jeśli uzasadnienie było poprawne, podtrzymaj je.
```

Arkusz oceny (do skopiowania)

Kryterium / warunek	Ja	Sędzia B	Sędzia C	Różnica	Cytat B	Cytat C	Kto ma rację i dlaczego
K1							
K2							
K3							
K4							
K5							
W1 (TAK/NIE)							
W2 (TAK/NIE)							

Notatka decyzji: zadanie · autor (model A) · sędziowie (B, C) · największa rozbieżność · co sprawdzone poza AI · decyzja i powód · co poprawić w prompcie następnym razem.

Checklista testu krzyżowego

- [ ] Zadanie wybrane, dane osobowe zamienione na [klient], [auto], [telefon].
- [ ] 4–6 kryteriów i 2–4 warunki krytyczne zapisane przed generowaniem.
- [ ] Model A zwrócił sam tekst, bez komentarzy.
- [ ] Tekst oczyszczony ze śladów autora; własne oceny wpisane w kolumnę „Ja”.
- [ ] Sędziowie B i C dostali identyczny prompt w czatach incognito lub tymczasowych.
- [ ] Oceny w arkuszu; rozbieżności od 2 punktów i alarmy krytyczne zaznaczone.
- [ ] Cytaty sprawdzone przez Ctrl+F, fakty — w źródle.
- [ ] Decyzja zapisana, a kryteria dodane do biblioteki szablonów.

ROZDZIAŁ 9

Twój plan działania

- [ ] **Dziś (45 min):** test krzyżowy na jednym prawdziwym tekście — np. ofercie, którą masz wysłać, albo odpowiedzi na ostatnią opinię.
- [ ] **W tym tygodniu:** dwa kolejne testy na innych typach zadań, np. mail do klienta i decyzja zakupowa z promptem „Adwokat diabła". Kryteria zapisz jako szablony dla tekstów, które piszesz najczęściej.
- [ ] **W tym miesiącu:** zasada w zespole — tekst o wysokiej stawce (duża oferta, odpowiedź na reklamację, publiczna odpowiedź na opinię) przechodzi test krzyżowy. Raz zrób test podrzuconego błędu, żeby sprawdzić sędziów. Wspólną bibliotekę kryteriów i promptów zbudujesz według lekcji o współdzieleniu czatów AI.

ROZDZIAŁ 10

Słowniczek

- **Model językowy (LLM)** — program AI, który pisze tekst, przewidując kolejne słowa; tak działają Claude, ChatGPT i Gemini.
- **Test krzyżowy** — jeden model tworzy odpowiedź, a inne oceniają ją według kryteriów ustalonych wcześniej.
- **Ślepa ocena** — oceniający nie wie, kto napisał tekst ani co o nim sądzisz.
- **Rubryka** — lista kryteriów z opisem, co oznacza ocena 1, 3 i 5.
- **Warunek krytyczny** — wymóg typu TAK/NIE; jego złamanie dyskwalifikuje tekst niezależnie od punktów.
- **Faworyzowanie własnych odpowiedzi (ang. self-preference bias)** — skłonność modelu do wyższego oceniania tekstów, które sam napisał.
- **Sykofancja** — przytakiwanie: dopasowywanie odpowiedzi do opinii użytkownika zamiast do faktów.
- **AI w roli sędziego (ang. LLM-as-a-judge)** — użycie modelu do oceniania lub porównywania odpowiedzi.
- **Skrzywienie pozycji i długości** — skłonność sędziego AI do wybierania odpowiedzi pokazanej jako pierwsza albo dłuższej.
- **Kalibracja** — zgodność deklarowanej pewności z rzeczywistą trafnością.
- **Czat tymczasowy / incognito** — rozmowa, która nie trafia do historii i (zależnie od ustawień) nie korzysta z pamięci modelu.
- **Skorelowane błędy** — różne modele mylą się w ten sam sposób, więc ich zgoda nie dowodzi prawdy.

ROZDZIAŁ 11

Źródła

- Panickssery, Bowman, Feng (2024): „LLM Evaluators Recognize and Favor Their Own Generations", NeurIPS 2024 — modele rozpoznają i faworyzują własne teksty.
- Chen i in. (2025): „Do LLM Evaluators Prefer Themselves for a Reason?", preprint — kiedy faworyzowanie siebie szkodzi.
- Zheng i in. (2023): „Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena", NeurIPS 2023 — skrzywienia kolejności i długości, zamiana kolejności.
- Ye i in. (2024): „Justice or Prejudice? Quantifying Biases in LLM-as-a-Judge", ICLR 2025 — 12 skrzywień AI w roli sędziego.
- Sharma i in. (2023): „Towards Understanding Sycophancy in Language Models", ICLR 2024 — przytakiwanie; badanie Anthropic.
- OpenAI (2025): „Sycophancy in GPT-4o: what happened and what we're doing about it" — wycofana aktualizacja ChatGPT.
- Xiong i in. (2023): „Can LLMs Express Their Uncertainty?", ICLR 2024 — zawyżona deklarowana pewność.
- OpenAI (2023): „GPT-4 Technical Report" — kalibracja przed dostrajaniem i po nim.
- Sclar i in. (2023): „Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design", ICLR 2024 — wrażliwość na format promptu.
- Huang i in. (2023): „Large Language Models Cannot Self-Correct Reasoning Yet", ICLR 2024 — granice samopoprawki.
- Kim i in. (2025): „Correlated Errors in Large Language Models", ICML 2025 — wspólne błędy różnych modeli.
- Laurito i in. (2025): „AI–AI bias: Large language models favor communications generated by large language models", PNAS — modele wolą treści pisane przez AI.

Co dalej

NASTĘPNA LEKCJA · 8 z 21

Halucynacje AI — jak nie dać się wpuścić w maliny

Sprawdzisz jedną ważną odpowiedź AI procedurą 5 kroków i podejmiesz decyzję z określonym poziomem pewności.

Obejrzyj wideo

Kanał BudujemyAI na YouTube

Dołącz do Klubu BudujemyAI

Pytania, Kawa z AI co tydzień i live konsultacyjny.

klub.budujemyai.pl

Konsultacja hybrydowa

Dwa spotkania online i e-book PDF z rozwiązaniami dla Twojej firmy — 499 zł. Na start: bezpłatne 15 min.

wiedza.budujemyai.pl/konsultacja/

Pozostałe e-booki — za darmo

21 lekcji ścieżki „Jedna praca" i mapa wiedzy.

wiedza.budujemyai.pl/ebooki/

ANKIETA · 1 MINUTA

Jak mógłbym dodatkowo rozwinąć według Ciebie ten wątek, bądź co jeszcze Cię ciekawi w tym temacie?

Odpowiedz tutaj: wiedza.budujemyai.pl/baza-wiedzy/slabosci-ai/#ankieta — czytam każdą odpowiedź, a najczęstsze prośby zamieniam w kolejne lekcje.

O autorze



Daniel Baranowski — prowadzi BudujemyAI. Wdraża AI i automatyzacje w małych i średnich firmach, buduje strony www z pomocą AI (także 3D) na własnym systemie, w którym klient sam edytuje stronę, pisząc do AI po polsku.

Na co dzień pracuje z Claude Code, Codexem i Grokiem jako „radą doradczą", a całą firmę prowadzi z vaulta w Obsidianie. W tej bazie pokazuje dokładnie to, czego sam używa — prosto i konkretnie.

ai@baranowski.consulting · budujemyai.pl · wiedza.budujemyai.pl

© 2026 BudujemyAI · Daniel Baranowski. Wydanie: listopad 2026. Możesz przesłać ten e-book dalej w niezmienionej formie.

Ceny, nazwy planów i funkcje narzędzi AI zmieniają się często — przed decyzją zakupową sprawdź aktualne informacje u producenta. Treść ma charakter edukacyjny i nie jest poradą prawną ani podatkową.