BudujemyAI_

MODUŁ 3: SPRAWDZAJ AI JAK EKSPERT

Halucynacje AI — jak nie dać się wpuścić w maliny

Procedura 5 kroków, która wyłapie zmyślone liczby, przepisy i źródła, zanim zobaczy je klient

JEDNA PRACA, KTÓRĄ WYKONASZ

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



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

W skrócie

DLA KOGO	Dla każdego, kto na podstawie odpowiedzi AI wysyła maile, oferty i pisma albo podejmuje decyzje w firmie.
JEDNA PRACA	**Sprawdzisz jedną ważną odpowiedź AI procedurą 5 kroków i podejmiesz decyzję z określonym poziomem pewności.**
EFEKT	Masz kartę weryfikacji, gotowe prompty i procedurę, która chroni Cię przed zmyślonymi faktami z AI.
CZAS	45 min
POZIOM	Start
KOSZT	0 zł (darmowe wersje Claude, ChatGPT i Gemini)
NARZĘDZIA	ChatGPT, Claude, Gemini, Wyszukiwarka Google, /last30days (opcjonalnie)

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/halucynacje-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 Poziom wyżej: /last30days, Buzz i rada modeli

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 korzystasz z ChatGPT, Claude albo Gemini i zdarzyło Ci się przekleić odpowiedź AI do maila, oferty albo pisma bez sprawdzania.

Jedna praca: sprawdzisz jedną ważną odpowiedź AI procedurą 5 kroków i podejmiesz decyzję z określonym poziomem pewności. Po lekcji masz kartę weryfikacji i prompty, dzięki którym zrobisz to samodzielnie przy każdej ważnej sprawie.

Czego potrzebujesz:

- komputera z przeglądarką i darmowych kont w co najmniej dwóch asystentach AI (np. ChatGPT i Claude; Gemini lub Grok jako trzeci),
- jednego prawdziwego pytania z Twojej firmy, od którego coś zależy: termin, przepis, kwota, parametr techniczny,
- 45 minut i notatnika albo arkusza na kartę weryfikacji,
- opcjonalnie, dla ambitnych: Claude Code z umiejętnością /last30days (opis w sekcji „Poziom wyżej”).

Koszt: 0 zł — wystarczą darmowe wersje modeli. **Poziom:** Start. Nie musisz znać się na technologii. Musisz tylko przestać wierzyć AI na słowo.

ROZDZIAŁ 2

Dlaczego to ważne dla Twojej firmy

AI pisze szybko, ładnie i z pełnym przekonaniem. Problem w tym, że z takim samym przekonaniem potrafi podać datę, której nie ma w ustawie, cytat, którego nikt nie powiedział, i źródło, które nie istnieje. To właśnie **halucynacja**: według definicji OpenAI — wiarygodnie brzmiące, ale fałszywe stwierdzenie wygenerowane przez model językowy.

Wyobraź sobie firmę instalacyjną, która publikuje na stronie poradnik o dofinansowaniu do pomp ciepła, napisany przez AI. Tekst wygląda profesjonalnie: kwoty, progi dochodowe, terminy naboru. Jedna kwota pochodzi jednak z poprzedniej edycji programu. Skutek: klienci przychodzą z błędnymi oczekiwaniami, handlowiec tłumaczy się z „obietnicy", a firma wygląda na niekompetentną — choć błąd zrobił chatbot.

To nie są teoretyczne ryzyka. Oto udokumentowane przypadki:

Gdzie i kiedy	Co się stało	Konsekwencja
Polska, NSA, 23.06.2026 (sygn. I FZ 104/26)	Zawodowy pełnomocnik powołał się w zażaleniu w sprawie VAT na trzy postanowienia NSA. Sygnatury istniały, ale orzeczenia miały inne daty, dotyczyły innych spraw i nie zawierały przypisanych im tez.	Zażalenie oddalone, a sąd skrytykował „bezrefleksyjne korzystanie z narzędzi AI".
USA, sprawa Mata v. Avianca, 2023	Prawnicy złożyli pismo z orzeczeniami wygenerowanymi przez ChatGPT, które nie istniały.	Grzywna 5 000 dolarów dla prawników.
Kanada, Moffatt v. Air Canada, 2024	Chatbot na stronie linii lotniczej podał klientowi błędną zasadę zwrotu za bilet. Firma broniła się, że nie odpowiada za chatbota.	Trybunał: firma odpowiada za wszystkie informacje na swojej stronie. Nakaz zapłaty 812,02 dolara kanadyjskiego.
Australia, Deloitte, 2025	W raporcie dla ministerstwa za 440 000 dolarów australijskich znaleziono nieistniejące źródła i przypisy. W poprawionej wersji firma ujawniła użycie modelu językowego.	Zwrot ostatniej raty wynagrodzenia i fala krytyki.

Skala rośnie. Baza prowadzona przez badacza Damiena Charlotina zbiera orzeczenia, w których sąd stwierdził, że strona oparła się na treściach zmyślonych przez AI. 3 października 2026 r. liczyła 2 137 pozycji z kilkudziesięciu krajów, w tym z Polski. W połowie 2025 r. było ich około 250. Modele są coraz lepsze, a wpadek przybywa, bo z AI korzysta coraz więcej osób — często bez kontroli.

Wniosek dla Twojej firmy jest prosty: za tekst wysłany klientowi, urzędowi czy kontrahentowi odpowiadasz Ty, nie chatbot. Kanadyjski trybunał powiedział to wprost. Sprawdzenie ważnej odpowiedzi zajmuje od kwadransa do 45 minut. Prostowanie błędu, który poszedł do 300 klientów,

trwa dłużej i kosztuje zaufanie.

AI wzmacnia tych, którzy umieją ocenić wynik

Ekspert widzi błąd AI w kilka sekund: księgowa od razu wyłapie zły termin, a prawnik — sygnaturę z niewłaściwego wydziału. Dlatego uważam, że AI najmocniej wzmacnia tych, którzy potrafią ocenić wynik. Dobrze pokazuje to eksperyment z udziałem 758 konsultantów Boston Consulting Group, opublikowany w „Organization Science". W zadaniach, w których AI jest mocne, osoby z AI wykonały o 12,2% więcej zadań, o 25,1% szybciej i z wyraźnie lepszą jakością. Ale w zadaniu celowo wybranym poza możliwościami AI osoby z AI o 19% rzadziej dochodziły do poprawnego rozwiązania niż te, które pracowały bez niego. To samo narzędzie raz pomaga, raz szkodzi — a różnicę robi umiejętność rozpoznania, kiedy mu nie ufać.

A co, jeśli dopiero budujesz swoją ekspertyzę? Masz dwie dźwignie. Pierwsza to **procedura**, która zastępuje doświadczenie: zamiast „czuję, że coś tu nie gra" masz listę kontrolną. Druga to **dywersyfikacja opinii**: konsultujesz wynik z innymi modelami, sprawdzasz świeże dyskusje (np. umiejętnością /last30days w Claude Code) albo sadzasz kilku agentów AI w jednym pokoju (np. w Buzz od firmy Block). Każda taka weryfikacja uczy Cię dziedziny szybciej niż samo czytanie odpowiedzi.

ROZDZIAŁ 3

Jak to działa — w prostych słowach

Model językowy nie ma w środku encyklopedii, którą przegląda. Działa raczej jak bardzo oczytany autouzupełniacz: na podstawie ogromnej ilości tekstu nauczył się przewidywać kolejne słowo. Nikt nie oznaczył mu zdań jako „prawdziwe" i „fałszywe". Widział tylko, jak wygląda płynny tekst.

Stąd podział, który dobrze opisało OpenAI we wrześniu 2025 r. Tam, gdzie są stałe wzorce — ortografia, gramatyka, styl maila — model prawie się nie myli. Tam, gdzie fakt jest przypadkowy i rzadki — data urodzin, numer orzeczenia, kwota z regulaminu konkretnego programu — wzorca nie ma. Model nie „wie", więc zgaduje. I zgaduje płynnie, bo płynności nauczył się świetnie.

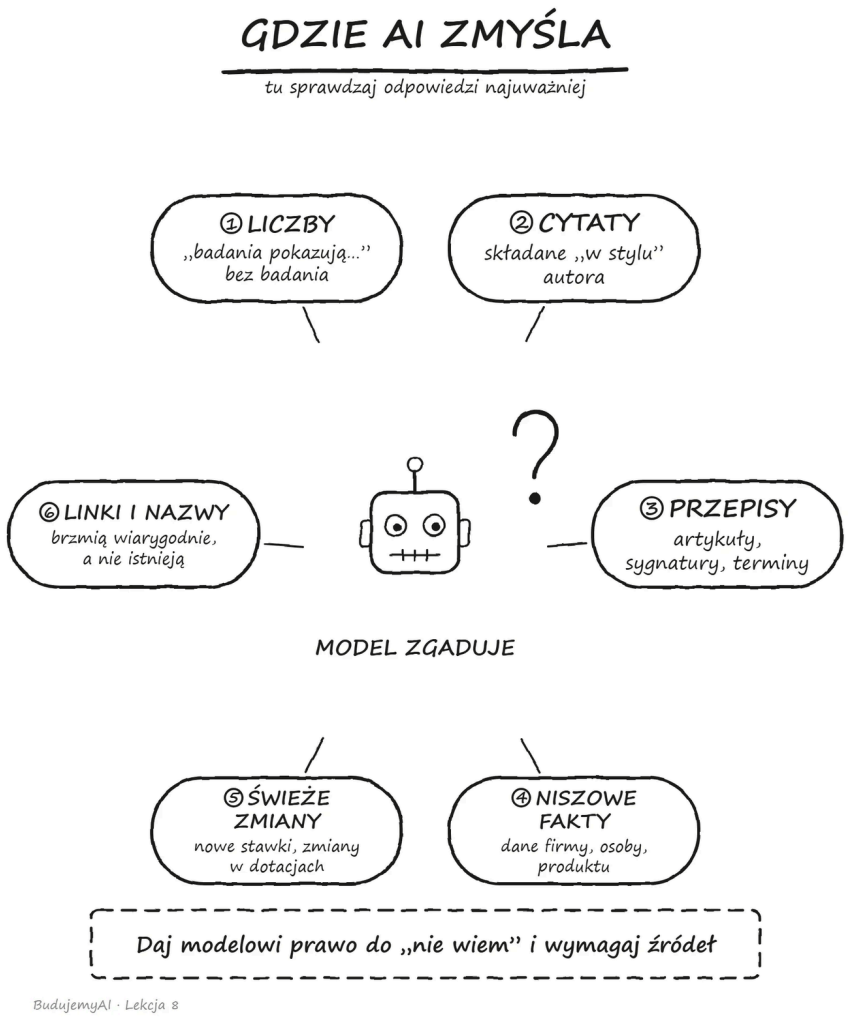
Druga przyczyna to sposób, w jaki modele się ocenia. To jak test wyboru: puste pole daje zero punktów, a strzał czasem trafia. Jeśli model nie zna czyjejś daty urodzin i strzeli, ma szansę 1 na 365. Jeśli powie „nie wiem" — ma zero. Przy tysiącach pytań „strzelający" model wygląda w rankingach lepiej niż ostrożny. Badacze OpenAI zapytali popularnego chatbota o tytuł doktoratu jednego z autorów badania i dostali trzy różne odpowiedzi — wszystkie błędne. Z datą urodzin było tak samo.

Dlaczego prawo do „nie wiem" ma znaczenie, pokazują wyniki testu SimpleQA z tej samej publikacji:

Model	Brak odpowiedzi („nie wiem")	Trafne odpowiedzi	Błędne odpowiedzi
gpt-5-thinking-mini	52%	22%	26%
OpenAI o4-mini	1%	24%	75%

Trafność podobna, ale model, który częściej mówi „nie wiem", myli się prawie trzy razy rzadziej. Wniosek dla Ciebie: **daj modelowi prawo do „nie wiem" i wymagaj źródeł**. Anthropic zaleca to samo w dokumentacji Claude. Wyszukiwanie w sieci i praca na Twoich dokumentach zmniejszają ryzyko, ale go nie usuwają: model może przekręcić źródło albo przypisać mu coś, czego tam nie ma.

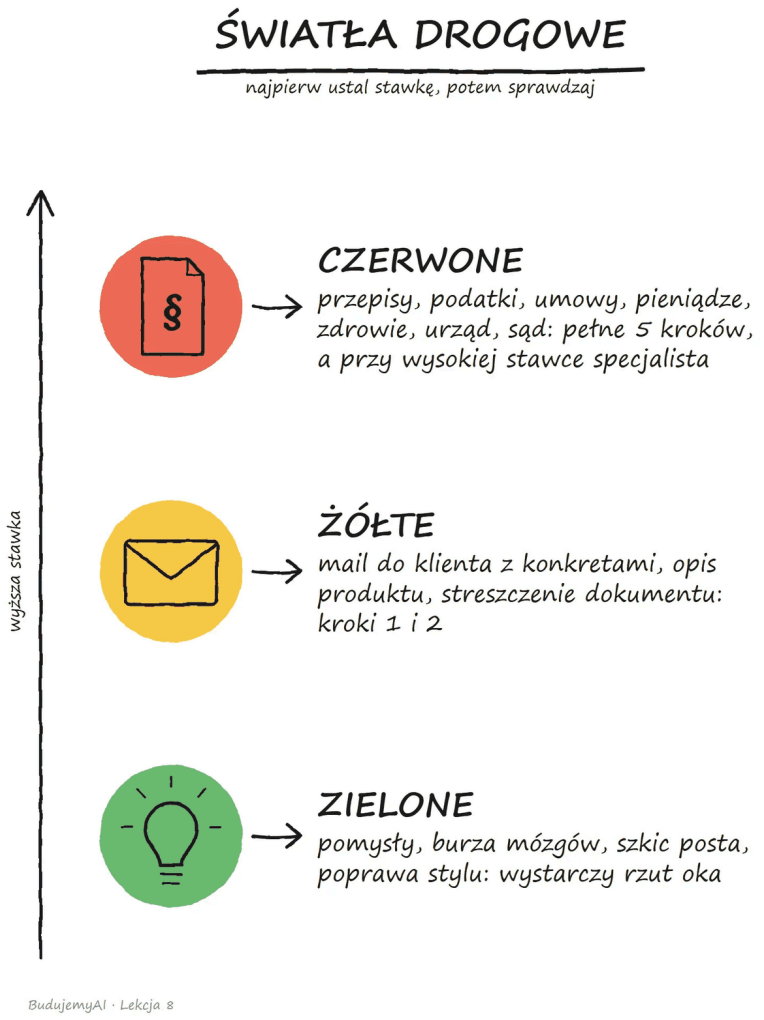
Gdzie AI zmyśla najczęściej



Obszar	Przykład z firmy	Dlaczego ryzykowne
Liczby i statystyki	„Badania pokazują, że większość klientów…"	Brzmi konkretnie, a często nie ma za tym żadnego badania
Cytaty	Wypowiedź eksperta, fragment wyroku	Model składa cytat „w stylu" autora
Przepisy i orzeczenia	Artykuł ustawy, sygnatura, termin	Rzadkie, precyzyjne i często zmieniane
Niszowe fakty	Dane konkretnej firmy, osoby, produktu	Mało tekstów, z których model mógł się uczyć
Świeże wydarzenia	Nowe stawki, zmiany w programach dotacji	Model ma wiedzę do pewnej daty
Linki i nazwy	Adres podstrony, nazwa pakietu, funkcja w aplikacji	Wygląda prawdopodobnie, a nie istnieje

Ostatni wiersz bywa groźny także technicznie. W badaniu 576 000 próbek kodu z 16 modeli komercyjne modele polecały nieistniejące pakiety oprogramowania w co najmniej 5,2% przypadków, a modele otwarte — w 21,7%. Przestępca może zarejestrować taką zmyśloną nazwę i podłożyć pod nią złośliwy kod. To zjawisko nazywa się slopsquatting i ma znaczenie, jeśli budujesz z AI stronę albo automatyzację.

Światła drogowe: kiedy sprawdzać

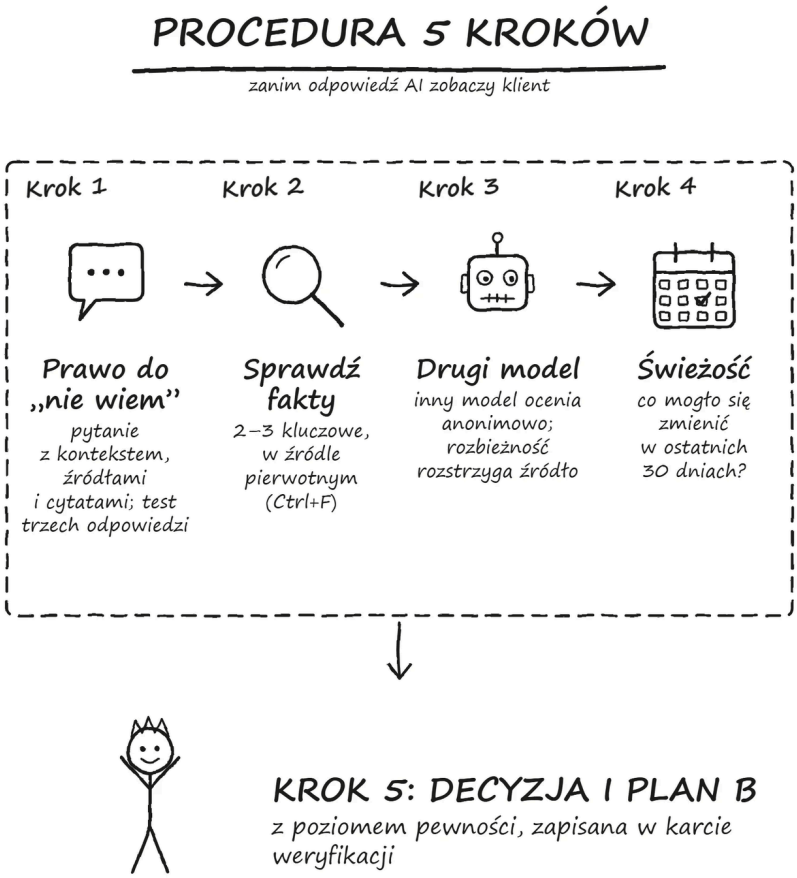


Nie każda odpowiedź AI wymaga 45 minut. Najpierw ustal stawkę:

- **Zielone** — pomysły, burza mózgów, szkic posta, poprawa stylu. Wystarczy rzut oka.
- **Żółte** — mail do klienta z konkretami, opis produktu, streszczenie dokumentu. Kroki 1 i 2.
- **Czerwone** — przepisy, podatki, umowy, pieniądze, zdrowie, bezpieczeństwo i wszystko, co trafia do urzędu albo sądu. Pełne 5 kroków, a przy wysokiej stawce także specjalista.

ROZDZIAŁ 4

Krok po kroku



Przejdziemy procedurę na jednym przykładzie. Typowa sytuacja: prowadzisz małą firmę usługową, chcesz poinformować klientów, że Twoje faktury będą trafiać do KSeF, i prosisz AI o fakty — od kiedy, z jakimi wyjątkami i co grozi za błędy. To czerwone światło: daty, progi, kary i przepisy, których harmonogram już się zmieniał.

Krok 1: Zadaj pytanie z prawem do „nie wiem” i obowiązkiem podania źródeł

Otwórz pierwszego asystenta, np. ChatGPT, i upewnij się, że korzysta z wyszukiwania w sieci (w Claude kliknij „+” w lewym dolnym rogu pola wiadomości i włącz opcję „Web search”). Nie pytaj gołym zdaniem. Daj kontekst, pozwól na niepewność i zażądaj źródeł z cytatami:

› PROMPT

```
Kontekst: prowadzę małą firmę usługową w Polsce, jestem czynnym podatnikiem VAT. Przygotowuję dla klientów krótką informację o KSeF.

Pytanie: od kiedy firma taka jak moja musi wystawiać faktury w KSeF, jakie są wyjątki i co grozi za błędy?

Zasady odpowiedzi:
1. Każde twierdzenie z datą, kwotą, progiem lub przepisem podaj z linkiem do źródła pierwotnego (ustawa, rozporządzenie, strona Ministerstwa Finansów) i z dosłownym cytatem z tego źródła.
2. Jeśli czegoś nie wiesz albo nie możesz znaleźć potwierdzenia, napisz wprost „nie wiem” lub „brak potwierdzenia”. Taka odpowiedź jest lepsza niż zgadywanie.
3. Przy każdym twierdzeniu podaj poziom pewności: wysoki, średni lub niski.
4. Na końcu wypisz, co mogło się zmienić w ostatnich miesiącach i co warto potwierdzić u księgowej lub doradcy podatkowego.
```

Zapisz odpowiedź w notatce. Przy ważnej sprawie zrób od razu **test trzech odpowiedzi**: zadaj to samo pytanie jeszcze dwa razy, w nowych oknach czatu. Jeśli daty, kwoty albo sygnatury różnią się między odpowiedziami, masz pierwszy sygnał, że model zgaduje — tak jak chatbot w eksperymencie OpenAI.

Krok 2: Wypisz twierdzenia i samodzielnie sprawdź 2–3 kluczowe

Długa, ładna odpowiedź usypia czujność. Rozbij ją na pojedyncze twierdzenia — możesz o to poprosić model w nowym oknie:

› PROMPT

```
Poniżej wklejam odpowiedź asystenta AI. Rozbij ją na listę pojedynczych twierdzeń faktograficznych (daty, kwoty, progi, przepisy, nazwy, cytaty, linki).

Dla każdego twierdzenia podaj w tabeli:
- treść twierdzenia,
- źródło podane w odpowiedzi (albo „brak”),
- ryzyko, jeśli twierdzenie jest błędne: wysokie, średnie lub niskie,
- gdzie to sprawdzić (konkretna instytucja albo rodzaj dokumentu).

Nie oceniaj, czy twierdzenia są prawdziwe. Tylko je wypisz i uporządkuj od najważniejszych dla decyzji.

[WKLEJ ODPOWIEDŹ]
```

Wybierz 2–3 twierdzenia, od których naprawdę zależy decyzja. W naszym przykładzie: datę obowiązku dla Twojej firmy, ewentualny wyjątek i kary. Sprawdź je samodzielnie w źródłach pierwotnych. Otwórz link, naciśnij Ctrl+F (na Macu Cmd+F) i wyszukaj liczbę albo kluczowe słowo z

cytatu. Szukasz odpowiedzi na dwa pytania: czy źródło istnieje i czy **mówi dokładnie to**, co twierdzi AI.

Drugie pytanie jest ważniejsze. W sprawie przed NSA sygnatury istniały — tylko że orzeczenia dotyczyły czegoś zupełnie innego. Sam fakt, że link się otwiera, niczego nie dowodzi.

Gdzie sprawdzać polskie fakty:

Co sprawdzasz	Źródło pierwotne
Ustawy i rozporządzenia	ISAP — Internetowy System Aktów Prawnych
Podatki i KSeF	podatki.gov.pl i ksef.podatki.gov.pl
Interpretacje podatkowe	EUREKA — Ministerstwo Finansów
Orzeczenia sądów administracyjnych	Centralna Baza Orzeczeń Sądów Administracyjnych
Orzeczenia sądów powszechnych	Portal Orzeczeń Sądów Powszechnych
Dane firm	KRS, CEIDG, biała lista VAT
Statystyki	GUS
Dane osobowe i RODO	UODO

Każdy wynik zapisz w karcie weryfikacji (wzór w „Zestawie do powtórzenia").

Krok 3: Daj odpowiedź do oceny drugiemu modelowi

Teraz dywersyfikacja opinii. Otwórz **inny model**: jeśli pytanie trafiło do ChatGPT, recenzentem niech będzie Claude albo Gemini. Dwa czaty tego samego modelu to wciąż jedna opinia.

Nie pisz, skąd pochodzi tekst. Modele mają skłonność do faworyzowania własnych odpowiedzi i do przytakiwania użytkownikowi (sykofancja) — więcej o tym w lekcji o słabościach AI. Wklej odpowiedź jako tekst „od asystenta" i daj recenzentowi jasne zadanie:

› PROMPT

```
Jesteś sceptycznym redaktorem sprawdzającym fakty. Poniżej jest odpowiedź asystenta AI na pytanie:
[WPISZ PYTANIE]. Nie wiesz, kto ją napisał, i nie musisz nikogo chronić.

Dla każdego twierdzenia faktograficznego:
1. Oceń je: potwierdzone, wątpliwe, prawdopodobnie fałszywe albo nie do oceny.
2. Uzasadnij ocenę jednym zdaniem i podaj źródło pierwotne z linkiem. Jeśli nie masz źródła,
napisz „brak źródła” — nie zgaduj.
3. Wskaż, czego w odpowiedzi brakuje, a co jest ważne dla decyzji.

Na końcu podaj 3 twierdzenia, które człowiek musi sprawdzić sam przed użyciem tej odpowiedzi.

[WKLEJ ODPOWIEDŹ]
```

Czytaj recenzję z tą samą podejrzliwością co oryginał — recenzent też może halucynować. Zgodność dwóch modeli to sygnał, nie dowód: modele uczą się na podobnych danych i potrafią powtórzyć ten sam błąd. Rozbieżność rozstrzyga źródło pierwotne, nie głosowanie. Jak dobrać modele do takiej „rady”, pokazuję w lekcji o wyborze modelu.

Krok 4: Sprawdź świeżość

Model ma wiedzę do pewnej daty, a przepisy, cenniki i funkcje aplikacji zmieniają się często. Zadaj sobie pytanie: czy w ostatnich 30 dniach mogło się tu coś zmienić? Trzy sposoby, od najprostszego:

1. **Wyszukiwarka z filtrem czasu.** W Google wpisz temat, kliknij „Narzędzia” i w menu czasu wybierz „Ostatni miesiąc”. Szukaj komunikatów urzędów i stron oficjalnych, nie tylko artykułów.
2. **Asystent z wyszukiwaniem** — z poniższym promptem.
3. **/last30days w Claude Code** — dla tematów, o których dyskutuje się w sieci: narzędzia, produkty, technologie. Opis w sekcji „Poziom wyżej”.

› PROMPT

```
Wyszukaj w sieci informacje z ostatnich 30 dni na temat: [TEMAT, np. terminy i zasady KSeF dla
małych firm].

Interesują mnie tylko zmiany i nowe komunikaty: nowe przepisy, projekty zmian, komunikaty urzędów,
przesunięte terminy, nowe wersje dokumentów.

Dla każdej informacji podaj datę publikacji, źródło z linkiem i jedno zdanie, co się zmieniło.
Najpierw źródła oficjalne, potem media. Jeśli nie ma nic nowego z ostatnich 30 dni, napisz to
wprost — nie uzupełniaj odpowiedzi starszymi informacjami bez podania ich daty.
```

Jeśli pojawi się zmiana, wróć do kroku 2 i sprawdź ją w źródle pierwotnym.

Krok 5: Podejmij decyzję z poziomem pewności i planem B

Weryfikacja nie kończy się na „prawda albo fałsz". Kończy się decyzją: czego używasz, na ile możesz na tym polegać i co zrobisz, jeśli kluczowy fakt okaże się błędny. Zbierz w jednym miejscu wyniki kroków 1–4 i poproś AI o uporządkowanie. Decyzja należy jednak do Ciebie.

› PROMPT

```
Pomóż mi uporządkować decyzję. Poniżej masz: (A) pytanie, (B) odpowiedź asystenta AI, (C) wyniki mojej weryfikacji w źródłach pierwotnych, (D) recenzję drugiego modelu, (E) sprawdzenie zmian z ostatnich 30 dni.

Przygotuj:
1. Tabelę: twierdzenie, czy jest potwierdzone w źródle pierwotnym (tak, nie, częściowo), poziom pewności (wysoki, średni, niski).
2. Rekomendację: czego mogę bezpiecznie użyć, co trzeba usunąć lub złagodzić, a co skonsultować ze specjalistą.
3. Plan B: co zrobię, jeśli najważniejszy fakt okaże się błędny (kto, co i do kiedy).

Opieraj się tylko na materiałach poniżej. Nie dodawaj nowych faktów.

[WKLEJ MATERIAŁY A–E]
```

W naszym przykładzie decyzja mogłaby brzmieć tak: „Data obowiązku: potwierdzona na stronie KSeF, pewność wysoka — wysyłamy. Kary: niepotwierdzone, modele podały różne informacje, pewność niska — usuwamy z tekstu i dajemy link do oficjalnej strony. Plan B: pytania o kary kierujemy do księgowej". Zapisz to w karcie weryfikacji. Po kilku tygodniach zobaczysz, w jakich tematach Twój model myli się najczęściej.

ROZDZIAŁ 5

Jak rozpoznać, że AI zrobiło dobrą robotę

W tej lekcji oceniasz dwie rzeczy: odpowiedź AI i pracę AI-recenzenta. Dobra odpowiedź:

- **ma źródła, które da się otworzyć i w których jest to, co twierdzi AI** — sprawdzone przez Ctrl+F, nie samym kliknięciem;
- **opiera się na źródłach pierwotnych i aktualnych**, a nie na blogu sprzed dwóch lat;
- **przyznaje niepewność** tam, gdzie jest naturalna: przyszłe terminy, projekty przepisów, interpretacje;
- **jest spójna w teście trzech odpowiedzi** — liczby i daty się nie zmieniają;
- **przechodzi recenzję drugiego modelu**, a zastrzeżenia są wyjaśnione w źródłach, nie przegłosowane;
- **kończy się decyzją** z poziomem pewności i planem B.

Dobra recenzja drugiego modelu ocenia każde twierdzenie osobno, podaje źródła swoich zastrzeżeń i mówi „brak źródła", gdy go nie ma. Recenzja w stylu „odpowiedź jest rzetelna i wyczerpująca" bez szczegółów nie jest nic warta.

Sygnały ostrzegawcze

- Cytat albo orzeczenie, które **idealnie** pasuje do Twojej tezy. W sprawie przed NSA zmyślone tezy pasowały do argumentacji pełnomocnika jak ulał.
- Precyzyjna liczba bez nazwy badania, autora i roku.
- Link prowadzi do strony głównej, do błędu 404 albo do dokumentu na inny temat.
- Przepisy z innego kraju wymieszane z polskimi, np. amerykańskie pojęcia w odpowiedzi o polskim prawie pracy.
- Model natychmiast zmienia zdanie, gdy dopytasz „Na pewno?", i podaje nową, równie pewną wersję.
- Brak dat przy informacjach, które się zmieniają.

Test pułapki w Twojej branży

Raz na jakiś czas zapytaj model o coś, czego **nie ma**: nieistniejący przepis, wymyśloną normę, fikcyjny produkt konkurencji. Jeśli zamiast „nie znam takiego przepisu" model zacznie go opisywać, wiesz, jak uważnie musisz mu patrzeć na ręce w swojej dziedzinie. Gotowy prompt znajdziesz w zestawie poniżej.

✓ Checklista: odpowiedź AI gotowa do użycia

- [ ] Sprawa ma przypisane światło: zielone, żółte albo czerwone.
- [ ] 2–3 kluczowe twierdzenia są sprawdzone w źródłach pierwotnych (otwarte i przeszukane Ctrl+F).
- [ ] Źródła mają datę i nie zostały zastąpione nowszą wersją.
- [ ] Test trzech odpowiedzi nie pokazał rozbieżności w liczbach i datach.
- [ ] Drugi model ocenił odpowiedź anonimowo, a rozbieżności są wyjaśnione w źródłach.
- [ ] Zmiany z ostatnich 30 dni są sprawdzone.
- [ ] Decyzja ma poziom pewności i plan B, a wynik trafił do karty weryfikacji.
- [ ] W tekście dla klienta nie ma twierdzeń z niską pewnością.

ROZDZIAŁ 6

Poziom wyżej: /last30days, Buzz i rada modeli

Procedura działa w zwykłej przeglądarce. Gdy stanie się nawykiem, te narzędzia przejmą część pracy.

/last30days — co ludzie mówią o temacie w ostatnich 30 dniach

/last30days to darmowy, otwarty skill (paczka instrukcji i narzędzi dla agenta AI) autorstwa Matta Van Horna: licencja MIT, ponad 63 tys. gwiazdek na GitHubie, wersja 3.26.0 z 1 października 2026 r. Wpisujesz `/last30days temat`, a skill przeszukuje równolegle m.in. Reddit, X, YouTube, Hacker News, Polymarket, GitHub i sieć z ostatnich 30 dni. Potem składa raport z cytatami, linkami i liczbami zaangażowania.

Instalacja w Claude Code (samą instalację Claude Code pokazuję w kursie Claude Code):

```
TEKST

/plugin marketplace add mvanhorn/last30days-skill
/plugin install last30days
```

Reddit, Hacker News, Polymarket i GitHub działają od razu, bez kluczy. Przy pierwszym uruchomieniu kreator konfiguracji pozwala dodać kolejne źródła, np. X czy YouTube — część z nich wymaga kluczy API albo zalogowanej przeglądarki. Raporty zapisują się lokalnie, na Twoim komputerze.

Jak tego używać przeciw halucynacjom: gdy AI opisuje funkcję narzędzia, cenę planu albo „nową wersję", polecenie `/last30days nazwa narzędzia` szybko pokaże, co użytkownicy piszą o tym teraz. Trzy zastrzeżenia. Posty to opinie, nie fakty — popularność nie jest dowodem. Tematy polskie i lokalne (KSeF, dotacje) mają tam słabe pokrycie, więc dla nich lepsze są wyszukiwarka i źródła oficjalne. To dodatek innej osoby, więc przed instalacją przeczytaj, co robi — jak oceniać repozytoria, pokazuję w lekcji o repozytoriach GitHub.

Buzz — agenci AI w jednym pokoju

Buzz to otwarta (licencja Apache 2.0) przestrzeń pracy dla ludzi i agentów AI. Udostępniła ją 21 lipca 2026 r. firma Block, twórca agenta goose — to nie jest produkt Anthropic. Nazywam go „Slackiem dla agentów": są kanały, wątki i wiadomości, a agenci (Claude Code, Codex, goose) dołączają do kanałów jak członkowie zespołu, z własną tożsamością i zapisem wszystkiego, co zrobili.

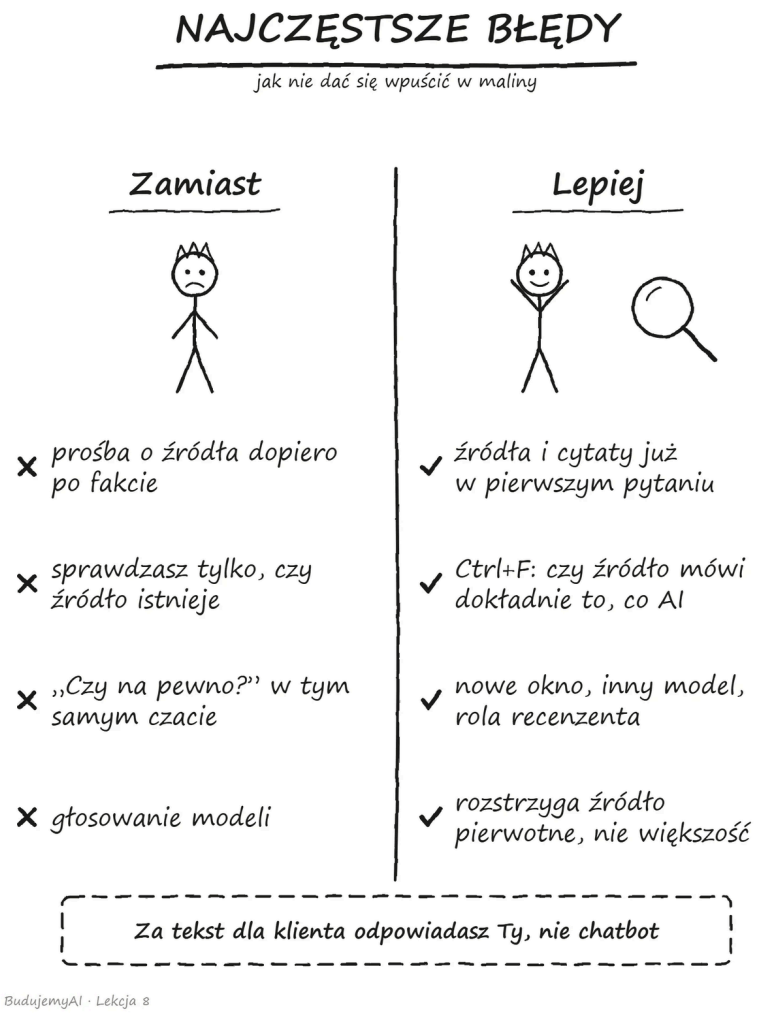
W kontekście halucynacji Buzz to sposób na „radę" w jednym miejscu: zadajesz pytanie na kanale, jeden agent odpowiada, drugi recenzuje, a cała dyskusja zostaje zapisana. Uczciwie: to wersja przed 1.0 (pod koniec września 2026 r. aplikacja desktopowa miała numer 0.5.26), aplikacje mobilne są dopiero w budowie, a wersja na Windows nie jest podpisana cyfrowo. Traktuj Buzz jako poligon dla ciekawych, nie miejsce na dane klientów ani procesy krytyczne.

Jak to robię u siebie

Na co dzień pracuję w Claude Code, a „drugie opinie" biorę z Codex CLI i Grok Build. Gdy Claude przygotuje analizę z faktami, wysyłam ją do recenzji pozostałym modelom (polecenia `codex exec` i `grok -p`) i zestawiam ich uwagi. Świeże tematy techniczne sprawdzam przez /last30days. Kluczowe fakty i tak weryfikuję sam w źródłach — narzędzia przyspieszają pracę, ale nie zdejmują ze mnie odpowiedzialności. Jak zbudować taką radę, pokazuję w lekcji o radzie doradczej AI.

ROZDZIAŁ 7

Najczęstsze błędy



1. **Prośba o źródła dopiero po fakcie.** Skutek: AI dopisuje źródła, które „pasują" do gotowego tekstu, czasem nieistniejące. Jak uniknąć: wymagaj źródeł i cytatów już w pierwszym pytaniu i otwieraj każdy link.
2. **Sprawdzanie, czy źródło istnieje, zamiast tego, co w nim jest.** Skutek: powtórka ze sprawy przed NSA — prawdziwe sygnatury, zmyślona treść. Jak uniknąć: Ctrl+F w źródle i szukanie dokładnego zdania lub liczby.
3. **„Czy na pewno?" w tym samym czacie.** Skutek: model przytakuje, przeprasza i zmienia odpowiedź albo broni błędu, bo widzi swój wcześniejszy tekst. Jak uniknąć: nowe okno, inny model, rola recenzenta i lista twierdzeń.
4. **Głosowanie modeli.** Skutek: dwa modele powtarzają ten sam błąd, a Ty uznajesz go za prawdę. Jak uniknąć: rozstrzygaj w źródle pierwotnym, nie większością głosów.
5. **Pytanie z tezą („Potwierdź, że…").** Skutek: model dopasowuje fakty do oczekiwania. Jak uniknąć: pytaj neutralnie albo proś o argumenty za i przeciw.

6. **Ślepa wiara w wyszukiwanie w sieci.** Skutek: model cytuje stary artykuł, wpis z forum albo reklamę. Jak uniknąć: sprawdzaj datę i rangę źródła — urząd i ustawa wygrywają z blogiem.
7. **Wklejanie danych klientów do kolejnych narzędzi.** Skutek: dane osobowe trafiają do usług bez umowy powierzenia, co rodzi ryzyko w świetle RODO. Jak uniknąć: anonimizuj („Klient A", bez PESEL i adresów); zasady opisuję w lekcji o Gemini i RODO. To nie jest porada prawna.
8. **Pełna procedura dla każdej drobnostki.** Skutek: po tygodniu porzucasz weryfikację. Jak uniknąć: światła drogowe — pełne 5 kroków tylko przy czerwonym.

ROZDZIAŁ 8

Zestaw do powtórzenia

Prompty do kroków 1–5 masz w sekcji „Krok po kroku”. Poniżej trzy dodatkowe prompty i szablony do druku.

Instrukcja stała przeciw halucynacjom

Wklej ją do instrukcji stałych swojego asystenta (ustawienia personalizacji w ChatGPT, preferencje profilu w Claude, instrukcje Gema lub umiejętności w Gemini) albo do projektu, w którym pracujesz:

› PROMPT

```
Zasady pracy z faktami:
1. Jeśli nie wiesz albo nie masz pewności, napisz „nie wiem” lub „brak potwierdzenia”. Wolę to niż zgadywanie.
2. Daty, kwoty, progi, przepisy, sygnatury, cytaty i statystyki podawaj tylko ze źródłem pierwotnym (link) i dosłownym cytatem.
3. Przy kluczowych twierdzeniach oznaczaj poziom pewności: wysoki, średni, niski.
4. Przy przepisach i cenach podawaj datę, której dotyczy informacja, i zaznacz, jeśli mogła się zmienić.
5. Nie wymyślaj linków, tytułów dokumentów ani nazw pakietów. Jeśli nie znasz dokładnego adresu, napisz, gdzie go szukać.
6. Gdy pytam o polskie realia, nie mieszaj przepisów i pojęć z innych krajów.
```

Odpowiedzi tylko z Twoich dokumentów

Gdy pracujesz na własnych materiałach (regulamin, umowa, cennik), ogranicz model do nich — to tzw. ugruntowanie. Najwygodniej robi się to w Gemini Notebook (dawniej NotebookLM) — pokazuję to w lekcji o drugim mózgu.

› PROMPT

```
Odpowiadaj wyłącznie na podstawie dokumentów poniżej lub w załącznikach. Nie korzystaj z wiedzy ogólnej.

Najpierw wypisz dosłowne cytaty z dokumentów, które odpowiadają na pytanie (z numerem strony lub sekcji). Potem odpowiedz, powołując się na numery cytatów.

Jeśli w dokumentach nie ma odpowiedzi, napisz: „Brak tej informacji w dokumentach”. Usuń każde zdanie odpowiedzi, którego nie potwierdza żaden cytat.

Pytanie: [WPISZ PYTANIE]
```

Test pułapki

› PROMPT

```
Opisz szczegółowo, co reguluje [NIEISTNIEJĄCY PRZEPIS, NORMA ALBO PRODUKT — wcześniej sprawdź, że naprawdę nie istnieje] i jakie obowiązki nakłada na małe firmy w Polsce. Podaj źródło z linkiem.
```

Dobry wynik: model mówi, że nie zna takiego przepisu albo nie może go potwierdzić. Zły wynik: opisuje go ze szczegółami i podaje „źródło".

Karta weryfikacji

Pytanie: ..... Data: ..... Światło: zielone / żółte / czerwone

Twierdzenie	Źródło pierwotne	Co naprawdę jest w źródle	Ocena drugiego modelu	Pewność	Decyzja i plan B

Checklista „zanim wyślesz"

- [ ] Twierdzenia z niską pewnością są usunięte albo złagodzone.
- [ ] Każda liczba i data w tekście ma źródło w karcie weryfikacji.
- [ ] Linki w tekście działają i prowadzą do właściwych dokumentów.
- [ ] Nazwy własne, nazwiska i nazwy produktów są zapisane poprawnie.
- [ ] Tekst nie zawiera danych osobowych, których odbiorca nie powinien widzieć.
- [ ] Przy sprawach podatkowych, prawnych i finansowych tekst widział specjalista albo decyzja o pominięciu konsultacji jest świadoma.

ROZDZIAŁ 9

Twój plan działania

Dziś (45 min): weź jedną odpowiedź AI z ostatnich dni, która poszła dalej — do klienta, zespołu albo urzędu — albo stała się podstawą decyzji. Przepuść ją przez 5 kroków i wypełnij pierwszy wiersz karty weryfikacji.

W tym tygodniu: wklej instrukcję stałą do swojego asystenta. Zrób test pułapki w swojej branży na dwóch modelach i porównaj wyniki. Ustal z zespołem zasadę świateł: co jest zielone, żółte i czerwone w Waszej firmie.

W tym miesiącu: zbierz 10 wpisów w karcie weryfikacji i przejrzyj je. W jakich tematach modele mylą się najczęściej? Który model był najlepszym recenzentem? Jeśli pracujesz w Claude Code, zainstaluj /last30days i użyj go przy najbliższym zakupie narzędzia albo sprawdzaniu nowej funkcji. Odważni mogą założyć testowy kanał w Buzz z dwoma agentami — bez danych klientów.

ROZDZIAŁ 10

Słowniczek

- **Halucynacja AI** — wiarygodnie brzmiące, ale fałszywe stwierdzenie wygenerowane przez model językowy: zmyślona data, liczba, cytat albo źródło.
- **Model językowy (LLM)** — program, który nauczył się przewidywać kolejne słowa na podstawie ogromnej ilości tekstu; na nim działają ChatGPT, Claude i Gemini.
- **Źródło pierwotne** — dokument, z którego informacja pochodzi: ustawa, komunikat urzędu, orzeczenie, dokumentacja producenta. Artykuł o ustawie to źródło wtórne.
- **Ugruntowanie (grounding)** — zmuszenie modelu, by odpowiadał na podstawie konkretnych dokumentów lub wyników wyszukiwania, z cytatami.
- **Sykofancja** — skłonność modelu do przytakiwania użytkownikowi i dopasowywania odpowiedzi do jego oczekiwań.
- **Test trzech odpowiedzi** — to samo pytanie zadane w trzech nowych czatach; różne szczegóły w odpowiedziach sugerują, że model zgaduje.
- **Poziom pewności** — Twoja ocena (wysoki, średni, niski), na ile informacja jest potwierdzona; od niej zależy, czy jej użyjesz.
- **Skill (umiejętność)** — paczka instrukcji i narzędzi, którą agent AI, np. Claude Code, wczytuje, gdy jest potrzebna.
- **/last30days** — otwarty skill, który zbiera dyskusje z ostatnich 30 dni (Reddit, X, YouTube i inne) w raport z cytatami.
- **Buzz** — otwarta przestrzeń pracy od firmy Block, w której ludzie i agenci AI pracują w tych samych kanałach; wersja przed 1.0.
- **Slopsquatting** — rejestrowanie pakietów oprogramowania o nazwach, które zmyśla AI, i podkładanie w nich złośliwego kodu.

ROZDZIAŁ 11

Źródła

- OpenAI — Why language models hallucinate — dlaczego modele zgadują zamiast mówić „nie wiem"; wyniki SimpleQA (5.09.2025).
- Kalai i in. — Why Language Models Hallucinate (arXiv) — pełne badanie z przykładami daty urodzin i tytułu doktoratu.
- Anthropic — Reduce hallucinations — oficjalne techniki: prawo do „nie wiem", cytaty, weryfikacja, porównanie odpowiedzi.
- Infor.pl — postanowienie NSA I FZ 104/26 — polski przypadek zmyślonych tez orzeczeń w zażaleniu.
- AI Hallucination Cases — baza Damiena Charlotina — orzeczenia, w których sądy stwierdziły treści zmyślone przez AI.
- CourtListener — Mata v. Avianca, Inc. — akta sprawy z fikcyjnymi orzeczeniami z ChatGPT.
- Civil Resolution Tribunal — Moffatt v. Air Canada, 2024 BCCRT 149 — firma odpowiada za błędne informacje swojego chatbota.
- The Guardian — Deloitte i raport przygotowany z AI — nieistniejące źródła w raporcie dla rządu Australii.
- Spracklen i in. — We Have a Package for You! (USENIX Security 2025) — zmyślone nazwy pakietów w kodzie generowanym przez AI.
- Dell'Acqua i in. — Navigating the Jagged Technological Frontier (Organization Science) — eksperyment z 758 konsultantami BCG.
- mvanhorn/last30days-skill (GitHub) — repozytorium /last30days: opis, instalacja, źródła danych.
- block/buzz (GitHub) — repozytorium Buzz: co działa, a co jest w budowie.

Co dalej

NASTĘPNA LEKCJA · 9 z 21

Claude, GPT-6 Astra, Grok czy Gemini? Odpowiedź na odwieczne pytanie

Przeprowadzisz mastermind czterech modeli AI na ważnym pytaniu z Twojej firmy i dostaniesz syntezę, w której każde zdanie ma źródło.

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/halucynacje-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.