MODUŁ 3: SPRAWDZAJ AI JAK EKSPERT

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

Nie szukaj jednego zwycięzcy: zbierz opinie czterech modeli i każ Claude zrobić z nich sprawdzalną syntezę

JEDNA PRACA, KTÓRĄ WYKONASZ

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



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

W skrócie

DLA KOGO	Dla właściciela lub menedżera firmy, który korzysta z AI albo wybiera, za który model zapłacić.
JEDNA PRACA	**Przeprowadzisz mastermind czterech modeli AI na ważnym pytaniu z Twojej firmy i dostaniesz syntezę, w której każde zdanie ma źródło.**
EFEKT	Masz syntezę czterech opinii z oznaczonymi źródłami, tabelę porównania, komplet promptów i wybrany model główny do codziennej pracy.
CZAS	60 min
POZIOM	Start
KOSZT	0 zł w darmowych wersjach; płatne plany flagowych modeli od ok. 20 $/mies. za usługę
NARZĘDZIA	Claude, ChatGPT, Grok, Gemini, Claude Code (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/ktory-model-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 Cztery modele w październiku 2026 — ściąga

05 Krok po kroku

06 Jak rozpoznać, że AI zrobiło dobrą robotę

07 Jak to robię u siebie

08 Ile to kosztuje

09 Bezpieczeństwo i RODO

10 Najczęstsze błędy

11 Zestaw do powtórzenia

12 Twój plan działania

13 Słowniczek

14 Źródła

15 Co dalej

ROZDZIAŁ 1

Zanim zaczniesz

„Który model AI jest najlepszy — Claude, ChatGPT z GPT-6 Astra, Grok czy Gemini?" To pytanie prędzej czy później zadaje sobie każdy, kto płaci za AI. Krótka odpowiedź: do codziennej pracy wybierz jeden model główny, który wygra Twój własny test. Do ważnych decyzji używaj wszystkich czterech naraz — i każ jednemu z nich zrobić z tego syntezę.

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

- **Dla kogo:** dla właściciela lub menedżera firmy, który korzysta z AI albo wybiera, za który model zapłacić.
- **Czego potrzebujesz:** komputera z przeglądarką, kont w Claude, ChatGPT, Groku i Gemini (darmowe wystarczą) oraz jednego pytania, od którego coś w firmie zależy.
- **Koszt:** 0 zł w wersji podstawowej. Płatne plany — w sekcji „Ile to kosztuje".
- **Czas:** 45–60 minut. Wersja pro z Claude Code: dodatkowa godzina na pierwszą instalację.
- **Poziom:** Start. Pomoże, jeśli znasz lekcję o słabościach AI i lekcję o halucynacjach.

ROZDZIAŁ 2

Dlaczego to ważne dla Twojej firmy

Spójrz na sam wrzesień 2026. OpenAI wypuściło GPT-6 Astra, xAI — Grok 4.7, Anthropic — Claude Opus 5.5 i Sonnet 5.5, a Google zapowiedział Gemini 4 Argon. Wszystko w ciągu jednego miesiąca. Ranking „najlepszego modelu" z sierpnia zestarzał się, zanim ktokolwiek zdążył go wdrożyć.

Z tego biorą się trzy kosztowne pułapki:

1. **Ślepa wierność jednemu modelowi.** Płacisz za jeden czat i bierzesz jego odpowiedzi za pewnik. Gdy model się myli, nie masz z czym porównać.
2. **Skakanie między modelami.** Co tydzień nowy „najlepszy" model, nowa subskrypcja, nowe nawyki — i nic nie zostaje wdrożone na dłużej.
3. **Jedna opinia przy ważnej decyzji.** Remontu nie zlecasz po jednej wycenie. Przy decyzji za kilka tysięcy złotych jedna odpowiedź AI to też za mało.

◆ **Typowa sytuacja (przykład ilustracyjny)**

Wyobraź sobie firmę remontową: osiem osób, kilkanaście zapytań o wycenę tygodniowo. Właścicielka pyta jeden czat: „Jak szybciej wysyłać wyceny?" i dostaje pewną siebie listę pięciu narzędzi. Przy czterech modelach zobaczyłaby, że dwa proponują zupełnie inne podejście, a jeden ostrzega przed kosztem, o którym reszta milczy. Te różnice pokazują, co sprawdzić, zanim wyda się pieniądze.

Mastermind zajmuje godzinę, a zły wybór narzędzia potrafi kosztować tygodnie. Do tego decyzja z uzasadnieniem („cztery niezależne opinie, tu zgodne, tu różne, to sprawdzone") wygląda profesjonalnie przed wspólnikiem, zespołem i klientem.

ROZDZIAŁ 3

Jak to działa — w prostych słowach

Model to silnik, aplikacja to samochód

Claude, ChatGPT, Grok i Gemini to aplikacje — jak samochody czterech marek. W środku pracują modele, czyli silniki: w Claude to m.in. Opus, Sonnet i Haiku, w ChatGPT rodzina GPT-6 z Astrą na czele, w Gemini modele Pro i Flash. W darmowej wersji jeździsz zwykle na słabszym silniku. Dlatego porównanie „darmowy Grok kontra płatny Claude" niewiele mówi.

Dlaczego modele odpowiadają różnie

Każdy model uczył się na trochę innych danych i innymi metodami. Ma też swój „styl domu": jeden pisze zwięźle, drugi szeroko, trzeci chętnie dopytuje. Do tego dochodzi przypadek — to samo pytanie zadane dwa razy temu samemu modelowi da dwie nieco inne odpowiedzi.

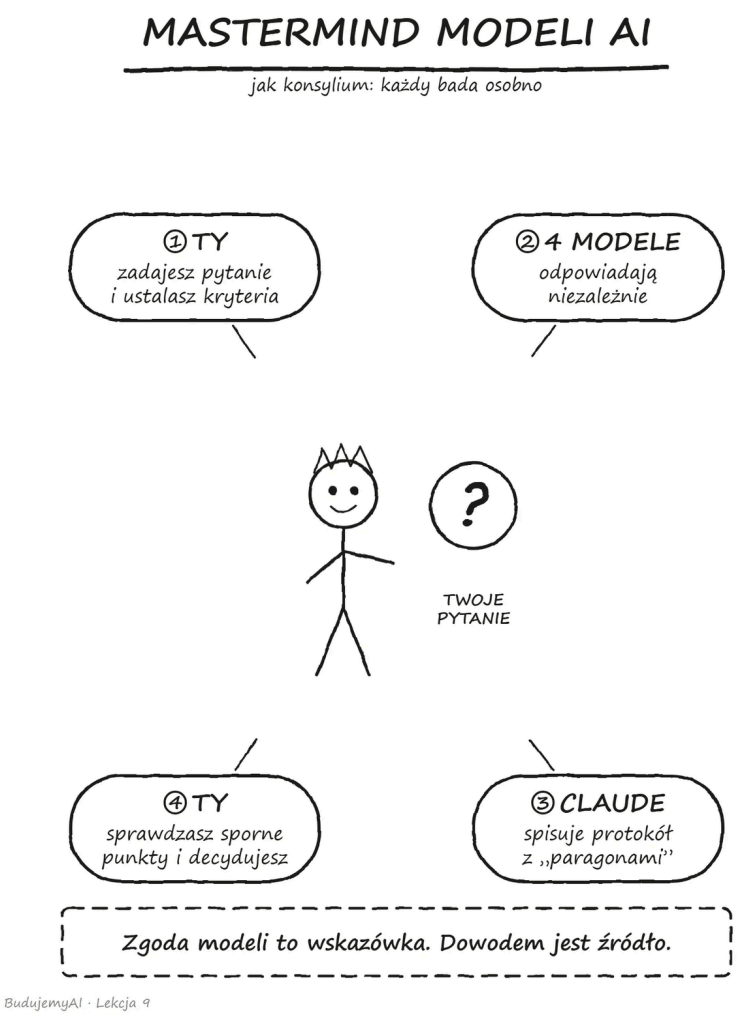
Dlaczego modele mylą się podobnie

Wszystkie cztery czytały w dużej mierze ten sam internet. To jak czterech konsultantów, którzy uczyli się z tych samych podręczników: jeśli w podręczniku jest błąd, wszyscy czterej powtórzą go z pełnym przekonaniem.

Badacze, którzy przeanalizowali ponad 350 modeli (Kim i in., ICML 2025), pokazali, że na jednym z zestawów testowych, gdy dwa modele się myliły, w 60% przypadków wybierały tę samą błędną odpowiedź. Inny zespół (Goel i in., 2025) zauważył, że im mocniejsze modele, tym bardziej podobne błędy popełniają, a model w roli oceniającego faworyzuje odpowiedzi modeli podobnych do siebie.

Wniosek: **zgoda modeli nie jest dowodem.** Jest wskazówką. Dowodem jest źródło.

Mastermind, czyli konsylium



Mastermind to narada kilku doradców nad jednym problemem. Najbliższa analogia: konsylium lekarskie. Każdy lekarz bada pacjenta osobno, potem porównują wnioski, ordynator spisuje stanowisko — a decyzję podejmuje pacjent. U nas:

- **Ty** zadajesz pytanie i ustalasz kryteria.
- **Cztery modele** odpowiadają niezależnie, nie widząc nawzajem swoich odpowiedzi.
- **Claude** jako przewodniczący spisuje protokół: co wspólne, co sporne, czego brakuje. Każde zdanie ma „paragon", czyli etykietę odpowiedzi, z której pochodzi.
- **Ty** sprawdzasz sporne punkty i podejmujesz decyzję.

Dlaczego Claude? U mnie to wybór praktyczny: pracuję w nim cały dzień, a Claude Code potrafi sam pytać pozostałe modele. Syntezę może jednak zrobić dowolny model — ważniejsze są jej zasady niż autor.

ROZDZIAŁ 4

Cztery modele w październiku 2026 — ściąga

Stan na 3 października 2026. Nazwy zmieniają się co kilka tygodni, więc przed zakupem zajrzyj na stronę producenta (linki w „Źródłach").

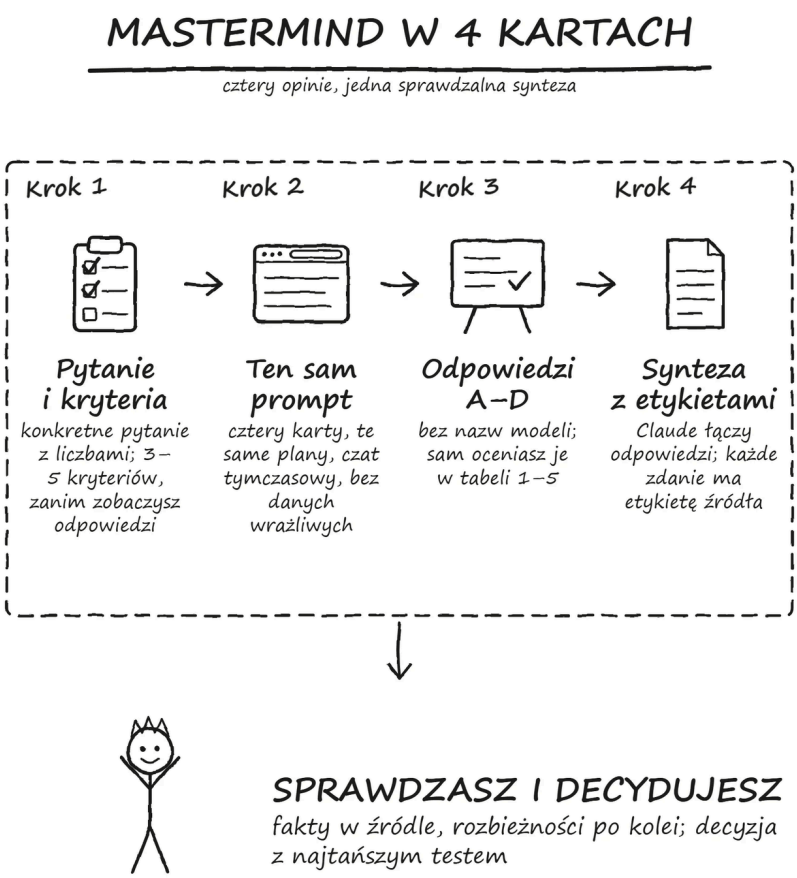
Usługa (firma)	Najmocniejsze modele	Co podkreśla sam producent	Narzędzie w terminalu
Claude (Anthropic)	Opus 5.5 (22.09), Sonnet 5.5 (28.09); najwyżej w ofercie Fable 5.1	programowanie z agentami, obsługa komputera, praca z wiedzą, bardziej naturalne pisanie	Claude Code (`claude`)
ChatGPT (OpenAI)	GPT-6 Astra (3.09) i mniejsze modele GPT-6	obsługa komputera i przeglądarki, programowanie, nauka, praca zawodowa	Codex CLI (`codex`)
Grok (xAI, na stronach jako SpaceXAI)	Grok 4.7 (21.09)	programowanie i praca z wiedzą, szybkość, niska cena	Grok Build (`grok`)
Gemini (Google)	3.1 Pro i modele Flash; Gemini 4 Argon zapowiedziany 30.09 i udostępniany etapami	złożone zadania: programowanie, dokumenty prawne i finansowe	Antigravity CLI (`agy`)

Co z tego wynika w praktyce:

- **Benchmarki z ogłoszeń to materiał producenta.** Liczby są prawdziwe, ale testy wybiera sprzedawca — jak w folderze dewelopera: zdjęcia prawdziwe, tylko słońce świeci na każdym.
- **Niezależne rankingi też się różnią.** W rankingu Arena (dawniej LMArena), gdzie ludzie głosują na ślepo, na początku października w kategorii tekstu prowadził Gemini 4 Argon (wynik wstępny), a w kategorii agentów — Claude Fable 5.1. Lider zależy od tego, co mierzysz.
- **Liczy się ekosystem.** Gdy firma żyje w Gmailu i na Dysku Google, Gemini masz pod ręką (więcej w lekcji o Gemini i RODO). ChatGPT i Claude mają integracje z Microsoft 365. Grok jest blisko serwisu X, więc szybko „wie", o czym się mówi — tylko że wpis na X to jeszcze nie źródło.
- **Gemini CLI zniknął dla kont prywatnych.** Od 18 czerwca 2026 darmowe konta Google oraz plany Google AI Pro i Ultra korzystają z Antigravity CLI. Gemini CLI został tylko w licencjach firmowych i z płatnym kluczem API.

ROZDZIAŁ 5

Krok po kroku



Krok 1: Wybierz pytanie i zapisz kryteria

Mastermind ma sens tam, gdzie błąd kosztuje: zakup narzędzia, zmiana cennika, nowa usługa, zatrudnienie, wybór dostawcy, plan marketingowy. Do poprawienia przecinków w mailu wystarczy jeden model.

Dobre pytanie ma kontekst, ograniczenia i jasny cel. Porównaj:

- **Słabe:** „Jak wykorzystać AI w firmie?"
- **Dobre:** „Firma remontowa, 8 osób, 15 zapytań o wycenę tygodniowo. Wyceny robimy w Excelu i wysyłamy mailem po 3–5 dniach. Jakie 3 usprawnienia z AI wdrożyć w pierwszym miesiącu, żeby skrócić ten czas do 48 godzin, przy budżecie do 500 zł miesięcznie? Nie zmieniamy programu do faktur."

Zanim zobaczysz jakąkolwiek odpowiedź, zapisz 3–5 kryteriów oceny, np. konkret, dopasowanie do firmy, koszt wdrożenia, ryzyka, czas do pierwszego efektu. Kryteria ustalone po fakcie zawsze dopasujesz do odpowiedzi, która Ci się spodobała (więcej w lekcji o słabościach AI).

Krok 2: Otwórz cztery karty i wyrównaj warunki

Otwórz w przeglądarce cztery karty: claude.ai, chatgpt.com, grok.com i gemini.google.com. Zaloguj się w każdej. Potem wyrównaj warunki:

- **Te same ligi.** Porównuj darmowe z darmowymi albo płatne z płatnymi. W planie płatnym wybierz w menu modeli najmocniejszy dostępny model (w Claude — Opus, w ChatGPT — GPT-6 Astra, w Gemini — Pro, w Groku — najmocniejszy tryb Twojego planu). Zapisz, co testujesz.
- **Czysta karta.** Użyj czatu tymczasowego albo incognito, jeśli usługa go ma — model nie dopasuje wtedy odpowiedzi do tego, co pamięta o Tobie.
- **Bez danych wrażliwych.** To samo pytanie trafi do czterech firm. Nazwiska zamień na „Klient A", kwoty zaokrąglij.

Krok 3: Wklej ten sam prompt do czterech modeli

Skopiuj prompt, uzupełnij nawiasy kwadratowe i wklej identyczny tekst do wszystkich kart. Nie dopisuj w jednej karcie niczego, czego nie ma w pozostałych, i nie zdradzaj, jaką odpowiedź chcesz usłyszeć.

› PROMPT

```
Jesteś doświadczonym doradcą małych i średnich firm w Polsce. Odpowiadasz samodzielnie, jako jeden
z kilku niezależnych doradców.

KONTEKST FIRMY:
[branża, liczba osób, region, główni klienci, co działa, co nie działa]

PYTANIE:
[jedno konkretne pytanie z liczbami i celem]

OGRANICZENIA:
[budżet, czas, ludzie, narzędzia, których nie zmieniamy]

ODPOWIEDZ W TYM FORMACIE:
1. Rekomendacja w 3 zdaniach.
2. Uzasadnienie: 3–5 punktów.
3. Ryzyka i warunki, w których ta rekomendacja jest zła.
4. Założenia: czego nie wiesz i co zakładasz — wprost.
5. Fakty do sprawdzenia: liczby, daty, przepisy, ceny, nazwy narzędzi — przy każdym napisz, skąd
go znasz.
6. Pewność całej odpowiedzi w skali 1–5 z jednym zdaniem uzasadnienia.

Pisz konkretnie i po polsku, bez ogólników typu „warto rozważyć”. Jeśli czegoś nie wiesz, napisz
„nie wiem”.
```

Założenia, fakty do sprawdzenia i deklarowana pewność przydadzą się w kontroli jakości. Jeśli któryś model dopyta o szczegół, dopisz odpowiedź do pytania i wyślij ją wszystkim — albo nie odpowiadaj żadnemu.

Krok 4: Zbierz odpowiedzi i zanonimizuj je

Otwórz pusty dokument (Word, Google Docs albo notatkę w Obsidianie). Wklej cztery odpowiedzi w losowej kolejności i podpisz je tylko literami: Odpowiedź A, B, C, D. Klucz („A = Grok, B = Claude...”) zapisz osobno — na kartce albo w innym pliku.

Usuń zdania, które zdradzają autora, np. „Jako Gemini...”. Po co ta ślepa ocena? Bo i Ty, i model robiący syntezę macie swoje sympatie. Model w roli oceniającego faworyzuje odpowiedzi podobne do własnych, a człowiek — marki, które lubi.

Krok 5: Oceń odpowiedzi w tabeli

Zanim ktokolwiek zrobi syntezę, oceń odpowiedzi sam według kryteriów z kroku 1. Wystarczy skala 1–5.

Kryterium (1–5)	A	B	C	D
Konkret: liczby, kroki, nazwy				

Kryterium (1–5)	A	B	C	D
Dopasowanie do mojej firmy				
Założenia i ryzyka wypisane wprost				
Uszanowane ograniczenia z pytania				
Pomysł, którego nie ma u innych				
Pewność deklarowana przez model				

Pod tabelą wypisz rozbieżności, czyli tematy, w których odpowiedzi mówią coś innego. To Twoja lista zadań na krok 7. Najwyżej oceniona odpowiedź nie wygrywa automatycznie — tabela ma pokazać, co każda wnosi.

Krok 6: Poproś Claude o syntezę z etykietami źródeł

Otwórz nowy czat w Claude — najlepiej incognito, żeby model nie korzystał z pamięci rozmów. Wklej prompt syntezy, pytanie z kroku 1 i cztery zanonimizowane odpowiedzi.

› PROMPT

```
Jesteś przewodniczącym rady doradców. Poniżej są cztery niezależne odpowiedzi na to samo pytanie,
oznaczone A, B, C i D. Nie wiesz, kto je napisał, i nie zgaduj tego.

Zadanie: zrób syntezę, która pomoże mi podjąć decyzję.

ZASADY:
- Każde zdanie syntezy zakończ etykietą źródła: [A], [B], [C], [D] albo kilkoma, np. [A, C].
- Własny wniosek, którego nie ma w żadnej odpowiedzi, oznacz [WNIOSEK] i uzasadnij.
- Nie dodawaj faktów, liczb ani nazw narzędzi, których nie ma w odpowiedziach.
- Nie uśredniaj sprzecznych opinii — pokaż je osobno.

FORMAT:
1. Konsensus — w czym zgadzają się co najmniej trzy odpowiedzi.
2. Rozbieżności — kto co twierdzi i od czego zależy, kto ma rację.
3. Głos mniejszości — pomysły z tylko jednej odpowiedzi, które warto rozważyć, i dlaczego.
4. Fakty do sprawdzenia — twierdzenia, które muszę zweryfikować w źródle przed decyzją.
5. Rekomendacja — co zrobić, pierwszy krok i prosty test na 7 dni, z pewnością 1–5.

PYTANIE:
[wklej pytanie z kroku 1]

ODPOWIEDŹ A:
[wklej]

ODPOWIEDŹ B:
[wklej]

ODPOWIEDŹ C:
[wklej]

ODPOWIEDŹ D:
[wklej]
```

Etykiety to sedno tej metody. Widzisz dzięki nim, które zdanie syntezy ma pokrycie w odpowiedziach, a które Claude dopisał od siebie. Jeśli w darmowym planie zabraknie limitu na długie odpowiedzi, skróć każdą do rekomendacji, ryzyk i faktów do sprawdzenia.

Krok 7: Sprawdź rozbieżności i podejmij decyzję

Weź listę „Fakty do sprawdzenia” z syntezy i swoją listę rozbieżności z kroku 5. Przy każdym punkcie:

1. **Fakty** (liczby, daty, ceny, przepisy) sprawdzasz w źródle: na stronie producenta, w urzędzie, w tekście przepisu. Nie w odpowiedzi innego modelu.
2. **Spór o ocenę** („warto” kontra „nie warto”) rozstrzygasz pytaniem: od czego to zależy w mojej firmie? Często obie strony mają rację, tylko w innych warunkach.

3. **Głos mniejszości** oceniasz osobno. Bywa, że to jedyna oryginalna myśl w całej naradzie.

Na koniec zapisz decyzję: co robię, po czym poznam, że działa, jaki jest najtańszy test i kiedy go sprawdzę (szablon karty decyzji jest w „Zestawie do powtórzenia"). Procedurę weryfikacji faktów opisuje lekcja o halucynacjach.

Krok 8 (wersja pro): Mastermind z terminala w Claude Code

Krok opcjonalny: to samo co kroki 2–6, ale w kilka minut. Potrzebujesz płatnego planu Claude (w darmowym nie ma Claude Code) i kont w pozostałych usługach: Codex CLI loguje się kontem ChatGPT, Grok Build — kontem Groka, Antigravity CLI — kontem Google. Każde z nich ma darmowy próg z limitami. Czym jest CLI, wyjaśnia lekcja o CLI, MCP i API.

1. Zainstaluj narzędzia — polecenia z oficjalnych stron producentów:

TERMINAL

```
# macOS / Linux
curl -fsSL https://claude.ai/install.sh | bash
curl -fsSL https://chatgpt.com/codex/install.sh | sh
curl -fsSL https://x.ai/cli/install.sh | bash
curl -fsSL https://antigravity.google/cli/install.sh | bash
```

POWERSHELL

```
# Windows (PowerShell)
irm https://claude.ai/install.ps1 | iex
powershell -ExecutionPolicy ByPass -c "irm https://chatgpt.com/codex/install.ps1 | iex"
irm https://x.ai/cli/install.ps1 | iex
irm https://antigravity.google/cli/install.ps1 | iex
```

Te polecenia pobierają i uruchamiają skrypt z internetu. Kopiuj je wyłącznie z dokumentacji producentów — Claude Code, Codex, Grok Build, Antigravity — nigdy z komentarzy i forów.

2. Zaloguj się. Uruchom w terminalu kolejno `claude`, `codex`, `grok` i `agy`. Każde narzędzie przy pierwszym starcie otworzy przeglądarkę z logowaniem.

3. Przygotuj folder. Utwórz pusty folder, np. `mastermind`, bez dokumentów firmy — narzędzia widzą pliki w folderze, w którym je uruchamiasz. Zapisz w nim pytanie z kroku 1 jako `pytanie.md`, a prompt syntezy z kroku 6 jako `synteza-prompt.md`.

4. Uruchom Claude Code w tym folderze (polecenie `claude`) i wklej:

› PROMPT

```
Zrób mastermind modeli AI na pytaniu z pliku pytanie.md. Pracuj tylko w tym folderze.

1. Najpierw zapisz własną odpowiedź w odp-claude.md, zanim zobaczysz cudze.
2. Zadaj treść pytanie.md pozostałym modelom w trybie bez okna, każdemu osobno: codex exec → odp-codex.md, grok -p → odp-grok.md, agy -p → odp-gemini.md. Nie używaj flag automatycznej zgody ani pełnego dostępu (--always-approve, --dangerously-skip-permissions, --sandbox danger-full-access).
3. Jeśli któreś narzędzie zgłosi błąd logowania albo zwróci pusty wynik, powiedz mi o tym. Nie uzupełniaj brakującej odpowiedzi sam.
4. Wymieszaj kolejność, podpisz odpowiedzi literami A–D bez nazw modeli i zapisz w odpowiedzi.md. Klucz litera → model zapisz w klucz.md.
5. Syntezę zrób w nowej sesji: claude -p z treścią synteza-prompt.md, pytaniem i odpowiedzi.md (bez klucza). Wynik zapisz w synteza.md.
6. Pokaż mi syntezę i 3 losowe jej zdania, każde z dosłownym cytatem z odpowiedzi źródłowej.
```

Claude Code poprosi o zgodę na uruchamianie poleceń. Czytaj je przed akceptacją — to Twoja pierwsza linia obrony (Codex w trybie `codex exec` domyślnie działa tylko do odczytu). Jeśli `agy -p` się zawiesi albo zwróci pusty plik — zgłaszano to w części wersji — zadaj pytanie w zwykłej sesji `agy` i wklej odpowiedź ręcznie.

Kurs Claude Code to lekcja 15, a stałą radę doradczą z modeli zbudujesz w lekcji 16.

ROZDZIAŁ 6

Jak rozpoznać, że AI zrobiło dobrą robotę



Mastermind daje więcej materiału — i więcej okazji do złudzenia pewności. Cztery zgodne odpowiedzi brzmią jak dowód, a nim nie są. Oto pięć testów do każdej syntezy.

Test 1: Paragony

Każde zdanie syntezy ma etykietę [A]–[D] albo [WNIOSEK]. Zdanie bez etykiety traktuj jak dopisek moderatora: może być trafny, ale sprawdzasz go tak samo jak nowy fakt.

Test 2: Trzy losowe zdania

Wybierz na chybił trafił trzy zdania z syntezy i znajdź je w odpowiedziach, na które się powołują. Jeśli choć jedno nie ma pokrycia albo zmienia sens oryginału („może warto" stało się „trzeba"), synteza jest niewiarygodna. Popraw prompt i zrób ją jeszcze raz.

Test 3: Rozbieżności przetrwały

Porównaj swoją listę rozbieżności z kroku 5 z sekcją „Rozbieżności" w syntezie. Jeśli w Twojej tabeli są cztery spory, a w syntezie jeden, moderator „wygładził" dyskusję. Dobra synteza nazywa spór i mówi, od czego zależy rozstrzygnięcie.

Test 4: Zgoda nie zwalnia ze sprawdzenia

Liczby, daty, ceny, przepisy i nazwy narzędzi sprawdzasz w źródle, nawet jeśli wszystkie cztery modele mówią to samo. Pamiętasz konsultantów z jednego podręcznika?

Test 5: Pułapka w ograniczeniach

Wpisz w pytanie jedno twarde ograniczenie, np. „nie zmieniamy programu do faktur". Odpowiedź, która mimo to każe go zmienić, nie przeczytała pytania uważnie — dostaje niższą ocenę w tabeli, a jej pozostałe twierdzenia sprawdzasz staranniej.

Sygnały ostrzegawcze

- Synteza wymienia narzędzia, liczby albo nazwy, których nie było w żadnej odpowiedzi.
- Sekcja „Rozbieżności" jest pusta, choć odpowiedzi wyraźnie się różniły.
- Moderator w każdym sporze przyznaje rację tej samej odpowiedzi. Sprawdź w kluczu, czy to nie jego własna.
- Same ogólniki: „warto rozważyć", „kluczowe jest podejście strategiczne".
- Wysoka deklarowana pewność przy faktach bez źródła.

✓ **Checklista kontroli jakości mastermindu**

- [ ] Kryteria oceny zapisane przed przeczytaniem odpowiedzi
- [ ] Wszystkie modele dostały identyczny prompt i te same informacje
- [ ] Odpowiedzi zanonimizowane (A–D) przed oceną i syntezą
- [ ] Każde zdanie syntezy ma etykietę [A]–[D] lub [WNIOSEK]
- [ ] 3 losowe zdania syntezy znalezione w odpowiedziach źródłowych
- [ ] Wszystkie rozbieżności z tabeli są w syntezie
- [ ] Liczby, daty, ceny i przepisy sprawdzone w oficjalnym źródle
- [ ] Decyzja ma kryterium sukcesu, test i termin

Gdzie AI pomaga, a gdzie decydujesz Ty. Modele generują opcje, wyłapują ryzyka i porządkują argumenty. Ty ustalasz kryteria, sprawdzasz fakty i odpowiadasz za decyzję. Tego podziału nie zmieni żaden nowy model.

ROZDZIAŁ 7

Jak to robię u siebie

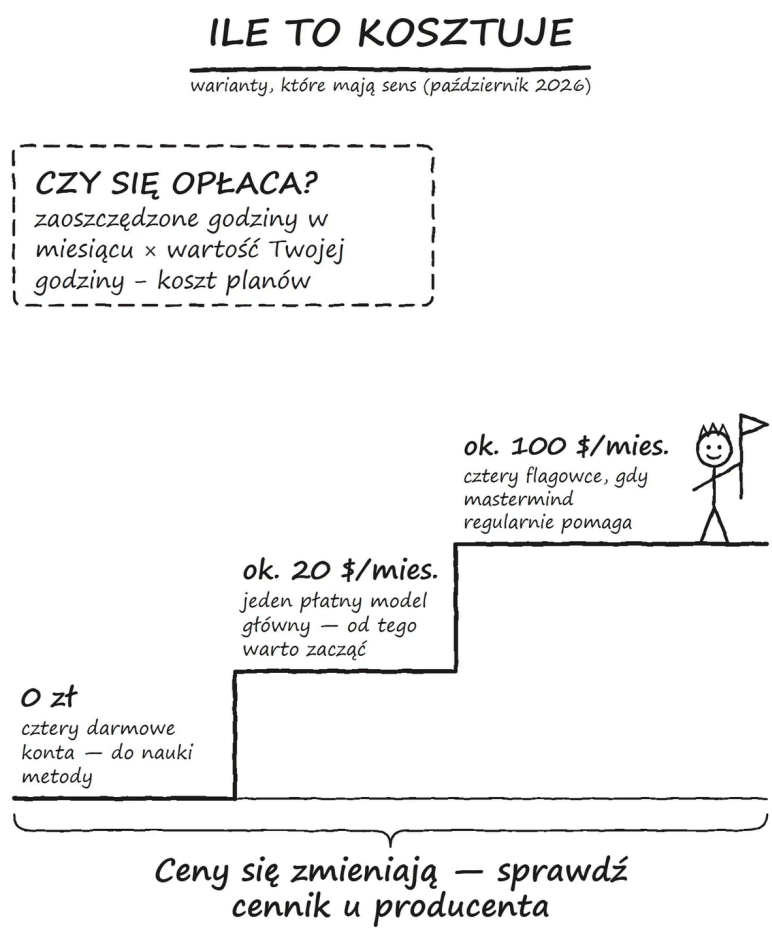
Pracuję w Claude Code na planie Max. Codex CLI (modele GPT-6) i Grok Build (grok-4.7) to moje „drugie opinie": Claude wywołuje je poleceniami `codex exec` i `grok -p`, a potem robi syntezę. Do przeglądów pracy w Claude Code używam też oficjalnej wtyczki openai/codex-plugin-cc — jej polecenie `/codex:adversarial-review` celowo podważa wybrane rozwiązanie.

Ten zestaw działa też całą dobę na moim serwerze VPS, a steruję nim z telefonu. Lekcja z praktyki: sesja logowania jednego z CLI wygasła mi kiedyś po cichu — wyszło to dopiero, gdy zamiast odpowiedzi dostałem komunikat, że nie jestem zalogowany. Od tamtej pory serwer co rano wysyła każdemu modelowi testowe pytanie i alarmuje mnie, gdy któryś milczy. W wersji pro raz w tygodniu sprawdź logowanie wszystkich narzędzi.

Mastermind odpalam przy decyzjach. Maile i codzienne teksty robię w jednym modelu — szybciej i taniej.

ROZDZIAŁ 8

Ile to kosztuje



Ceny (stan: październik 2026) zmieniają się często, więc przed zakupem sprawdź aktualny cennik u producenta. Część cenników podaje kwoty bez VAT — porównuj kwotę końcową przy płatności.

Usługa	Za darmo	Plan na co dzień	Dla intensywnych
Claude	Sonnet i Haiku, bez Claude Code	Pro: 20 $/mies. (w UE 18 € + VAT; 15 €/mies. przy płatności rocznej)	Max: 100 $ lub 200 $/mies.
ChatGPT	GPT-5.6 Luna, bez Astry (plan Go za 8 €/mies. też bez Astry)	Plus: 23 €/mies., z GPT-6 Astra	Pro: od 103 €/mies.

Usługa	Za darmo	Plan na co dzień	Dla intensywnych
Grok	z limitami	SuperGrok: ok. 30 $/mies. (wg zestawień cenowych)	SuperGrok Heavy: ok. 300 $/mies. (wg zestawień cenowych)
Gemini	3.6 Flash i ograniczony 3.1 Pro	Google AI Pro: 97,99 zł/mies. (tańszy AI Plus: 23,99 zł)	AI Ultra: od 469,99 zł/mies.

Cenę SuperGrok xAI pokazuje dopiero po zalogowaniu, więc podaję ją za niezależnymi zestawieniami.

Które warianty mają sens:

- **0 zł** — cztery darmowe konta. Wystarczą do nauki metody i prostszych pytań.
- **Ok. 20 $ miesięcznie** — jeden płatny plan dla modelu głównego, reszta za darmo. Od tego warto zacząć.
- **Ok. 100 $ miesięcznie** — cztery flagowce. Ma sens, gdy mastermind regularnie pomaga Ci w decyzjach wartych więcej.
- **Wersja pro** — minimum Claude Pro. Codex, Grok Build i Antigravity mają darmowe progi, a przy częstej pracy przydadzą się płatne plany.

Jak policzyć, czy się opłaca: (godziny zaoszczędzone w miesiącu × wartość Twojej godziny) − koszt planów. Przykład z założeniami: godzina Twojej pracy jest warta 150 zł, a plan kosztuje kilkadziesiąt złotych — wystarczy jedna zaoszczędzona godzina miesięcznie, żeby wyjść na plus. Policz to na własnych liczbach po miesiącu testu.

ROZDZIAŁ 9

Bezpieczeństwo i RODO

- **Cztery firmy, cztery serwery.** W mastermindzie to samo pytanie trafia do czterech dostawców. Dane osobowe klientów i tajemnice firmy anonimizuj zawsze, a tu podwójnie.
- **Trenowanie na Twoich rozmowach.** W planach indywidualnych Claude i ChatGPT możesz zrezygnować z wykorzystywania rozmów do trenowania modeli. U pozostałych dostawców też zajrzyj do ustawień prywatności.
- **Plany firmowe.** Do pracy na danych klientów wybieraj plany firmowe (np. Claude Team, ChatGPT Business, Gemini w Google Workspace) z umową powierzenia przetwarzania danych — przykład Google opisuje lekcja o Gemini i RODO.
- **Wersja pro.** Narzędzia w terminalu widzą pliki w folderze, w którym działają — stąd osobny, pusty folder.

> **! To nie jest porada prawna**
>
> Opisuję dobre praktyki, a nie interpretację przepisów. W sprawach umów powierzenia i danych osobowych skonsultuj się z prawnikiem albo inspektorem ochrony danych.

ROZDZIAŁ 10

Najczęstsze błędy

1. **Porównywanie różnych lig.** Darmowy model jednej firmy kontra płatny drugiej → fałszywy wniosek „X jest słabszy" → porównuj te same plany i zapisuj, który model był testowany.
2. **Ogólne pytanie.** „Jak rozwinąć firmę z AI?" → cztery zestawy ogólników, z których nic nie wynika → kontekst, liczby, ograniczenia i jedno konkretne pytanie (krok 1).
3. **Pokazywanie modelom cudzych odpowiedzi za wcześnie.** Model, który zobaczy odpowiedź innego, zaczyna się do niej odnosić zamiast myśleć samodzielnie → najpierw cztery niezależne odpowiedzi, dopiero potem synteza i krytyka.
4. **Głosowanie zamiast sprawdzania.** „Trzy modele na cztery mówią tak, więc tak" → wspólny błąd wygrywa → rozbieżność to zadanie do sprawdzenia, a fakty potwierdza źródło, nie większość.
5. **Synteza bez etykiet.** Nie wiesz, co pochodzi z odpowiedzi, a co dopisał moderator → wymuś etykiety [A]–[D] i zrób test trzech losowych zdań.
6. **Dane klientów w czterech serwisach.** Cztery razy większa ekspozycja danych → anonimizacja, a do pracy na danych plany firmowe z umową powierzenia.
7. **Mastermind do wszystkiego.** Tracisz czas i limity na sprawy, w których wystarczy jeden model → mastermind tylko przy decyzjach, w których błąd kosztuje (np. powyżej ustalonej kwoty albo dnia pracy).
8. **Wierność marce na zawsze.** „Mój model jest najlepszy" → przegapiasz, że od Twojego wyboru zmieniło się kilka generacji modeli → raz na kwartał powtórz test modelu głównego.

ROZDZIAŁ 11

Zestaw do powtórzenia

Prompty mastermindu, syntezy i dla Claude Code są w krokach 3, 6 i 8. Poniżej reszta zestawu. Do krytyki samej decyzji przyda się też prompt „adwokat diabła” z lekcji o słabościach AI.

Prompt: audyt syntezy

Najlepiej, żeby audyt zrobił inny model niż ten, który pisał syntezę.

› PROMPT

```
Poniżej są cztery odpowiedzi (A–D) i synteza, która z nich powstała. Sprawdź syntezę jak audytor.

1. Wypisz zdania syntezy bez etykiety źródła.
2. Przy każdym zdaniu z etykietą sprawdź, czy wskazana odpowiedź naprawdę to mówi. Oceń: zgodne /
zmienia sens / brak w źródle. Przy każdej ocenie podaj dosłowny cytat.
3. Wypisz rozbieżności między odpowiedziami, które synteza pominęła albo wygładziła.
4. Wypisz pomysły z tylko jednej odpowiedzi, które zniknęły z syntezy.
5. Oceń wiarygodność syntezy w skali 1–5 i napisz, co poprawić.

ODPOWIEDZI A–D:
[wklej]

SYNTEZA:
[wklej]
```

Prompt: test wyboru modelu głównego

Wklej ten sam zestaw do czterech modeli, oceń każde zadanie w skali 1–5 i wybierz model główny: ten, który wygrywa Twoje zadania, a nie ranking.

› PROMPT

```
Oto 5 typowych zadań z mojej firmy. Wykonaj każde z nich i oznacz je numerem.
1. [np. odpowiedź na mail z reklamacją — treść bez danych osobowych]
2. [np. szkic oferty dla klienta na podstawie notatek]
3. [np. streszczenie dokumentu w 5 punktach dla zespołu]
4. [np. wnioski z tabeli sprzedaży — dane bez nazwisk]
5. [np. wpis do Profilu Firmy w Google o nowej usłudze]
Na końcu wypisz, jakich informacji zabrakło Ci przy każdym zadaniu.
```

Checklista: mastermind w 10 punktach

- [ ] Pytanie dotyczy decyzji, w której błąd kosztuje

- [ ] Kontekst, liczby i ograniczenia są w pytaniu
- [ ] Kryteria oceny zapisane przed odpowiedziami
- [ ] Dane klientów zanonimizowane
- [ ] Te same plany i ten sam prompt we wszystkich modelach
- [ ] Odpowiedzi podpisane A–D, klucz zapisany osobno
- [ ] Tabela ocen i lista rozbieżności gotowe przed syntezą
- [ ] Synteza z etykietami i sekcją „Głos mniejszości”
- [ ] Audyt: 3 losowe zdania, a fakty sprawdzone w źródle
- [ ] Karta decyzji: kryterium sukcesu, test, termin

Szablon: karta decyzji

Pole	Wpis
Pytanie	
Decyzja	
Na czym się opieram (etykiety z syntezy)	
Co sprawdzone w źródle	
Kryterium sukcesu	
Najtańszy test i termin	
Kiedy wracam do tematu	

ROZDZIAŁ 12

Twój plan działania

- **Dziś (60 minut):** jeden mastermind w czterech kartach na pytaniu, które naprawdę masz na biurku. Zrób syntezę i test trzech losowych zdań.
- **W tym tygodniu:** test wyboru modelu głównego na pięciu zadaniach z firmy. Zdecyduj, za który model płacisz — albo że na razie nie płacisz za żaden.
- **W tym miesiącu:** ustal próg (np. decyzje powyżej 2 000 zł albo dnia pracy), od którego mastermind jest obowiązkowy. Jeśli korzystasz z niego często, przejdź na wersję pro i zbuduj stałą radę doradczą z lekcji 16. Za kwartał powtórz test modelu głównego.

ROZDZIAŁ 13

Słowniczek

- **Model językowy (LLM)** — program, który nauczył się języka na ogromnej ilości tekstu; „silnik" w środku czatu AI.
- **Model flagowy** — najmocniejszy, zwykle najdroższy model danej firmy, np. GPT-6 Astra albo Claude Opus.
- **Mastermind** — narada kilku niezależnych doradców nad jednym problemem; tu: kilku modeli AI.
- **Synteza** — jedno podsumowanie z kilku odpowiedzi, w którym każde zdanie ma wskazane źródło.
- **Ślepa ocena** — ocenianie odpowiedzi bez wiedzy, kto je napisał (A–D zamiast nazw).
- **Skorelowane błędy** — różne modele mylą się w ten sam sposób, bo uczyły się na podobnych danych.
- **Halucynacja** — pewnie brzmiąca, ale zmyślona informacja w odpowiedzi AI.
- **Benchmark** — standardowy test porównujący modele; mierzy tylko to, co sprawdza.
- **CLI** — narzędzie obsługiwane tekstowymi poleceniami w terminalu, np. Claude Code albo Codex CLI.
- **Tryb bez okna (headless)** — uruchomienie narzędzia AI jednym poleceniem, bez rozmowy na ekranie; wynik trafia do pliku.

ROZDZIAŁ 14

Źródła

- Anthropic — Introducing Claude Opus 5.5 — premiera 22.09.2026, mocne strony według producenta
- Claude — plany i ceny — Free, Pro, Max; Claude Code w planach płatnych
- OpenAI — GPT-6 Astra — premiera 3.09.2026, dostępność w planach ChatGPT
- ChatGPT — cennik — Free, Go, Plus, Pro; modele w planach
- xAI — Grok 4.7 — premiera 21.09.2026, deklaracje producenta
- xAI — dokumentacja Grok Build — instalacja i logowanie CLI
- Google — Gemini 4 Argon — zapowiedź 30.09.2026, etapowe udostępnianie
- Google — plany Google AI w Polsce — ceny w złotych, modele w planach
- Google Developers Blog — przejście z Gemini CLI na Antigravity CLI — zmiana od 18.06.2026
- Arena — ranking modeli — ślepe porównania, liderzy w kategoriach
- Kim i in. (2025), Correlated Errors in Large Language Models — modele często mylą się tak samo
- Goel i in. (2025), Great Models Think Alike and this Undermines AI Oversight — podobne błędy i stronniczość modelu oceniającego

Co dalej

NASTĘPNA LEKCJA · 10 z 21

CLI, MCP, API — o co z tym chodzi (+ harness, sandbox i inne pojęcia)

Podłączysz do Claude pierwszy serwer MCP tylko do odczytu i sprawdzisz, że działa bezpiecznie.

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/ktory-model-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.