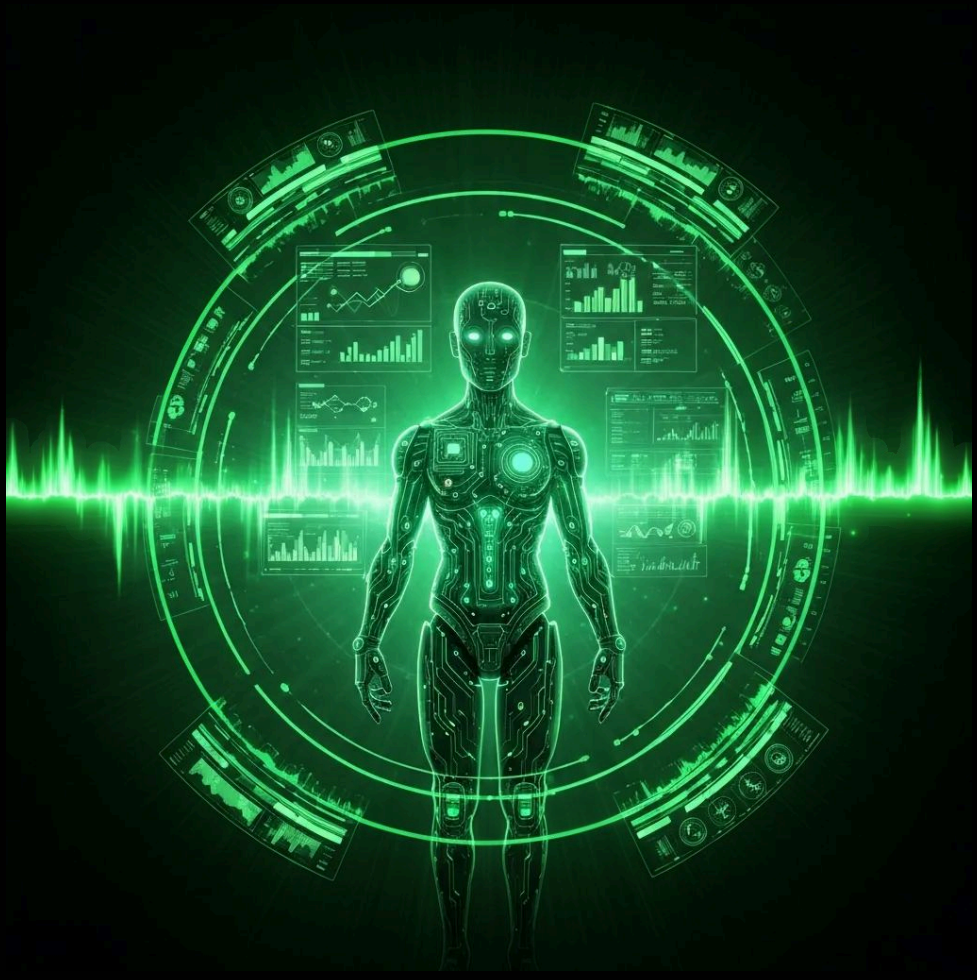
Jak mierzyć sukces

ASYSTENTA AI



Praktyczny przewodnik po metrykach, które naprawdę mają znaczenie

SYZYGY

Spis treści

Wstęp

Dlaczego analityka w projektach AI to konieczność, a nie dodatek?

W klasycznych projektach IT wszystko jest stosunkowo proste:
aplikacja działa albo nie działa. Jeśli coś się psuje — widzimy błąd, crash lub brak reakcji systemu.

W przypadku asystentów AI sytuacja wygląda zupełnie inaczej.

Asystent może:

- odpowiadać płynnie i „naturalnie”,
- brzmieć przekonująco,
- a jednocześnie wprowadzać użytkownika w błąd, frustrować go lub prowadzić do błędnych decyzji.

Dlatego w projektach AI analityka nie jest dodatkiem po wdrożeniu, ale podstawowym mechanizmem bezpieczeństwa — zarówno dla użytkowników, jak i dla biznesu.

Ta checklista PDF powstała po to, aby:

- uporządkować podejście do mierzenia sukcesu asystenta AI,
- pokazać konkretne, praktyczne metryki,
- dać gotowe instrukcje ich mierzenia — zrozumiałe również dla osób nietechnicznych.

Trójkąt wartości Asystenta AI

Aby nie zgadywać, czy asystent „działa”, warto spojrzeć na niego z trzech perspektyw, które razem dają pełny obraz jego skuteczności.

1. Efektywność i jakość

Czy asystent faktycznie rozwiązuje problemy użytkowników?
Czy odpowiedzi są trafne, kompletne i stabilne w czasie?

2. Satysfakcja użytkownika

Czy korzystanie z asystenta jest łatwe i przyjemne?
Czy użytkownicy nie muszą się powtarzać, dopytywać i frustrować?

3. Wpływ biznesowy

Czy asystent realnie odciąża zespoły, obniża koszty lub generuje przychód?
Czy inwestycja w AI ma sens z perspektywy organizacji?

Jak mierzyć metryki Asystenta AI?

FILAR I: Efektywność i jakość

1. Task Completion Rate (TCR)

Co dokładnie mierzymy?

Odsetek rozmów, w których użytkownik osiągnął swój cel, np.:

- uzyskał potrzebną informację,
- zakończył proces (zwrot, reklamacja, umówienie wizyty),
- dostał odpowiedź wystarczającą, by nie szukać dalej.

Jak mierzyć – metoda bezpośrednia (najprostsza i najbardziej wiarygodna)

Krok 1: Zdefiniuj moment „zakończenia zadania”

To musi być konkretny punkt w rozmowie, np.:

- po wyświetleniu statusu zamówienia,
- po wygenerowaniu instrukcji,
- po potwierdzeniu wykonania akcji.

Krok 2: Wyświetl pytanie w czacie

Automatycznie, bez dodatkowych ekranów:

„Czy te informacje rozwiązały Twój problem?”

Odpowiedzi:

- Tak
- Nie

Krok 3: Zapisz odpowiedź jako dane

Każda odpowiedź powinna zawierać:

- ID rozmowy,
- odpowiedź (Tak / Nie),
- timestamp.

Krok 4: Oblicz wskaźnik

Task Completion Rate (TCR) =

(Liczba odpowiedzi „Tak” / Liczba wszystkich odpowiedzi) × 100%

Jak mierzyć – metoda pośrednia (gdy użytkownicy nie klikają ankiet)

Uznaj rozmowę za „rozwiązaną”, jeśli jednocześnie:

1. użytkownik zamknął okno czatu po odpowiedzi AI
2. w ciągu 10–15 minut:
    - nie skontaktował się z konsultantem,
    - nie otworzył innego kanału (mail, formularz, infolinia) w tej samej sprawie

Ta metoda wymaga integracji danych, ale jest bardzo miarodajna.

Jak interpretować TCR

- wysoki TCR ➜ AI spełnia swoją funkcję
- niski TCR ➜ AI nie rozwiązuje realnych problemów użytkowników

2. Human Escalation Rate

Co dokładnie mierzymy?

Jak często rozmowa z AI musi zostać przekazana do człowieka, ponieważ:

- użytkownik tego chce,
- AI sobie nie radzi,
- procedura wymaga konsultanta.

Jak mierzyć – krok po kroku

Krok 1: Oznacz eskalację jako event

Każde przekazanie rozmowy do człowieka musi być zapisane w systemie jako zdarzenie.

Krok 2: Przypisz typ eskalacji

Każdej eskalacji przypisz jedną z kategorii:

- eskalacja na życzenie użytkownika
  (np. „połącz mnie z konsultantem”)
- eskalacja z powodu błędu AI
  (np. 3 odpowiedzi „nie rozumiem”)
- eskalacja proceduralna
  (sprawy, które zawsze wymagają człowieka)

Krok 3: Zlicz rozmowy

- liczba wszystkich rozmów z AI,
- liczba rozmów zakończonych eskalacją.

Krok 4: Oblicz wskaźnik

Human Escalation Rate = (Liczba rozmów przekazanych do konsultanta / Liczba wszystkich rozmów) × 100%

Jak interpretować wynik?

- dominują eskalacje proceduralne ➜ OK
- dużo eskalacji z powodu błędu AI ➜ problem z jakością lub zakresem wiedzy
- dużo eskalacji na żądanie użytkownika ➜ brak zaufania do asystenta

3. Dead-End Rate (Ślepy zaułek)

Co dokładnie mierzymy?

Ile rozmów kończy się tym, że:

- AI odpowiada komunikatem typu „nie rozumiem”,
- użytkownik rezygnuje i kończy czat.

Jak mierzyć – metoda automatyczna

Zdefiniuj regułę „ślepego zaułka”

Oznacz rozmowę jako Dead-End, jeśli:

1. ostatnia wiadomość pochodzi od AI
2. treść zawiera jedną z fraz błędu
   („nie rozumiem”, „nie mogę pomóc”)
3. użytkownik:
   - nie odpowiada,
   - zamyka czat lub sesja wygasa

Rozmowa otrzymuje flagę:

dead_end = true

Jak mierzyć – metoda ręczna (na start)

1. Raz w tygodniu wybierz 20–30 zakończonych rozmów
2. Sprawdź, ile spełnia warunki Dead-End
3. Policz procent

Dlaczego to ważne?

To czysty wskaźnik frustracji użytkownika i jedno z najlepszych źródeł insightów do poprawy AI.

4. LLM-as-Judge (zautomatyzowana ocena jakości)

Co dokładnie mierzymy?

Jakość odpowiedzi AI ocenianą przez inny model językowy, działający jako niezależny „sędzia".

Jak mierzyć – dokładna procedura

Krok 1: Zdefiniuj kartę oceny

Ustal jasne kryteria, np.:

- zgodność z faktami,
- trafność odpowiedzi,
- ton i styl zgodny z marką.

Krok 2: Ustal skalę ocen

Np. skala 1–5, gdzie:

- 1 = bardzo słabo
- 5 = bardzo dobrze

Krok 3: Wybierz próbkę rozmów

Np.:

- rozmowy testowe (development),
- rozmowy produkcyjne,
- rozmowy po zmianach promptów lub modelu.

Krok 4: Przepuść odpowiedzi przez model oceniający

Każda odpowiedź otrzymuje ocenę według ustalonych kryteriów.

Krok 5: Zapisz wyniki

Dla każdej odpowiedzi zapisz:

- ocenę,
- kryterium,
- datę.

Dlaczego to ważne?

Pozwala:

- monitorować jakość w czasie,
- wykrywać degradację odpowiedzi,
- prowadzić testy regresyjne po każdej zmianie.

5. Helpfulness Degradation

Co dokładnie mierzymy?

Czy jakość odpowiedzi spada wraz z długością rozmowy.

Jak mierzyć – krok po kroku

1. Każdą odpowiedź w rozmowie oceniaj osobno metodą LLM-as-Judge
2. Przypisz ocenę do numeru kroku rozmowy
   (np. krok 1, krok 2, krok 3...)
3. Narysuj wykres:
   - oś X: numer interakcji
   - oś Y: ocena jakości

Jak interpretować wynik?

- spadek jakości ➜ AI gubi kontekst lub kumuluje błędy
- stabilna jakość ➜ dobra kontrola procesu
- gwałtowne spadki ➜ problem w logice lub narzędziach

6. Monitoring zdrowia asystenta AI

Co mierzyć regularnie?

- aktualność źródeł danych,
- czas odpowiedzi,
- częstotliwość błędów,
- stabilność jakości w czasie.

Jak mierzyć?

- dashboard dzienny lub tygodniowy,
- alerty przy przekroczeniu progów.

FILAR II: Satysfakcja użytkownika

1. Sentiment Analysis (ocena nastroju użytkownika)

Co dokładnie mierzymy?

Emocjonalny ton wypowiedzi użytkownika w trakcie rozmowy z asystentem:

- pozytywny,
- neutralny,
- negatywny
  oraz bardziej szczegółowe emocje (np. frustracja, irytacja, sarkazm).

Jak mierzyć – dokładna procedura

Krok 1: Analizuj każdą wiadomość użytkownika

Każda wypowiedź użytkownika powinna być:

- automatycznie analizowana,
- przypisana do jednej z kategorii sentymentu.

Najczęściej stosowana skala:

- –1 ➜ negatywny
- 0 ➜ neutralny
- +1 ➜ pozytywny

W rozwiązaniach opartych o LLM:

- dodatkowe etykiety emocji (np. frustracja, złość, zniecierpliwienie).

Krok 2: Zapisuj sentyment w czasie

Dla każdej wiadomości zapisz:

- ID rozmowy,
- numer kroku rozmowy,
- ocenę sentymentu,
- timestamp.

Krok 3: Analizuj zmiany sentymentu w rozmowie

Sprawdź:

- czy sentyment pogarsza się w konkretnym momencie,
- czy powtarza się ten sam moment spadku (np. pytanie o numer zamówienia).

Jak interpretować wynik?

- stabilny lub rosnący sentyment ➜ dobra jakość doświadczenia
- powtarzalne spadki sentymentu ➜ konkretny fragment procesu do poprawy

Dlaczego to ważne?

Sentiment Analysis pozwala zrozumieć nie tylko że użytkownik jest niezadowolony, ale dokładnie dlaczego i w którym momencie.

2. Customer Effort Score (CES)

Co dokładnie mierzymy?

Jak łatwe lub trudne było dla użytkownika załatwienie swojej sprawy przy pomocy asystenta AI.

Jak mierzyć – krok po kroku

Krok 1: Zadaj jedno pytanie po zakończeniu rozmowy

Wyświetlane bezpośrednio w czacie:

„Jak łatwo było Ci załatwić swoją sprawę?”

Skala odpowiedzi:

- 1 – bardzo trudno
- 2 – trudno

- 2 – trudno
- 3 – neutralnie
- 4 – łatwo
- 5 – bardzo łatwo

Krok 2: Zapisz odpowiedź użytkownika

Każda odpowiedź powinna zawierać:

- ID rozmowy,
- wartość 1–5,
- timestamp.

Krok 3: Policz wynik

Masz dwie poprawne metody:

Opcja A – średnia:

CES = średnia wszystkich ocen

Opcja B – udział ocen pozytywnych:

CES pozytywny = (liczba ocen 4 i 5 / liczba wszystkich ocen) × 100%

Jak interpretować wynik?

- ≥80% ocen 4–5 ➜ bardzo dobre doświadczenie
- dużo ocen 1–2 ➜ użytkownik musi się „namęczyć”, nawet jeśli sprawa została rozwiązana

Dlaczego to ważne?

Użytkownicy wolą łatwe doświadczenie niż „poprawne, ale męczące”.
CES bardzo często koreluje z retencją.

3. User Rephrase Rate (URR)

Co dokładnie mierzymy?

Jak często użytkownik musi przeformułować pytanie, bo AI go nie zrozumiało.

Jak mierzyć – dokładna procedura

Krok 1: Zidentyfikuj sekwencję „rephrase”

Szukaj wzorca:

1. pytanie użytkownika
2. odpowiedź AI sygnalizująca brak zrozumienia
   (np. „nie rozumiem”, „spróbuj inaczej”)
3. bardzo podobne pytanie użytkownika chwilę później

Krok 2: Oznacz sesję jako „rephrase”

Jeśli taki wzorzec wystąpił, oznacz rozmowę flagą:

rephrase = true

Krok 3: Policz wskaźnik

User Rephrase Rate (URR) = (liczba sesji z rephrase / liczba wszystkich sesji) × 100%

Jak interpretować wynik?

- <15% ➜ akceptowalny poziom
- ≥15% ➜ AI nie radzi sobie z potocznym językiem lub synonimami

Dlaczego to ważne?

URR pokazuje miejsca, w których użytkownik musi dopasować się do AI, zamiast odwrotnie.

FILAR III: Wpływ biznesowy

1. Support Ticket Deflection

Co dokładnie mierzymy?

Na ile asystent AI zmniejsza liczbę zgłoszeń do obsługi klienta.

Jak mierzyć – krok po kroku

Krok 1: Zbierz dane historyczne

Zbierz liczbę zgłoszeń do supportu:

- z minimum 3 miesięcy przed wdrożeniem asystenta.

Krok 2: Zbierz dane po wdrożeniu

Zbierz dane:

- z analogicznego okresu po uruchomieniu AI.

Krok 3: Porównaj wyniki

Sprawdź:

- czy liczba zgłoszeń spadła,
- o ile procent.

Jak interpretować wynik?

- spadek zgłoszeń ➔ AI realnie odciąża zespół
- brak spadku ➔ AI nie przejmuje realnych problemów użytkowników

2. Wpływ na konwersję

Co dokładnie mierzymy?

Czy asystent AI pomaga użytkownikom wykonać pożądaną akcję, np.:

- zakup,
- rejestrację,
- wysłanie formularza.

Jak mierzyć – dokładna procedura

Krok 1: Dodaj linki z parametrami UTM

Każdy link podawany przez AI powinien zawierać:

- źródło = asystent AI,
- kampania = konkretna funkcja lub use case.

Krok 2: Monitoruj kliknięcia

W narzędziach analitycznych (np. GA, Mixpanel):

- sprawdzaj liczbę wejść z AI.

Krok 3: Sprawdź konwersję

Zobacz:

- ilu użytkowników wykonało pożądaną akcję po wejściu z AI,
- jaki przychód lub wartość wygenerowali.

Dlaczego to ważne?

To najmocniejszy argument biznesowy za dalszym inwestowaniem (lub zamknięciem) asystenta.

3. Cost per Resolution

Co dokładnie mierzymy?

Ile realnie kosztuje rozwiązanie jednej sprawy przez asystenta AI.

Jak mierzyć – krok po kroku

Krok 1: Monitoruj zużycie tokenów

Zbieraj:

- liczbę tokenów wejściowych,
- liczbę tokenów wyjściowych,
- koszt API dla każdej rozmowy.

Krok 2: Połącz koszt z wynikiem rozmowy

Skoreluj:

- koszt rozmowy,
- jej rezultat (np. Task Completion Rate).

Krok 3: Oblicz średni koszt

Porównaj:

- koszt rozmów zakończonych sukcesem,
- koszt rozmów zakończonych porażką.

Jak interpretować wynik?

- wysoki koszt + niski TCR ➜ nieopłacalne
- niski koszt + wysoki TCR ➜ optymalny model

4. Zadowolenie pracowników

Co dokładnie mierzymy?

Jak asystent AI wpływa na komfort i efektywność pracy zespołów wewnętrznych.

Jak mierzyć?

Metody:

- krótkie ankiety NPS lub CSAT wśród pracowników,
- analiza adopcji (kto faktycznie korzysta z AI),
- obserwacja, czy zadania są realizowane szybciej.

Dlaczego to ważne?

Jeśli pracownicy nie chcą korzystać z AI, to:

- narzędzie nie rozwiązuje ich realnych problemów,
- albo jest zbyt trudne w użyciu.

Jak pracować z metrykami, żeby miały sens?

Zbieranie metryk to dopiero początek. Same liczby niczego nie poprawią, jeśli nie będą używane do podejmowania decyzji.

1. Miej wszystko w jednym miejscu

Zbuduj jeden dashboard z kluczowymi metrykami z trzech filarów.
Dzięki temu szybciej zauważysz problemy i zależności między jakością, satysfakcją i kosztem.

2. Szukaj wzorców, nie wyjątków

Pojedyncze nieudane rozmowy są normalne.
Interesują Cię powtarzalne problemy: eskalacje, dead-endy, spadki jakości, rosnące koszty.

3. Analizuj trudne rozmowy

Najwięcej uczą rozmowy:

- drogie,
- zakończone eskalacją,
- z niską jakością lub niskim CES.

To one pokazują, co dokładnie trzeba poprawić.

4. Waliduj zmiany danymi

Każdą zmianę w AI oceniaj przez metryki „przed i po”.
Jeśli liczby się nie poprawiają — zmiana nie zadziałała.

5. Kontroluj koszty

Dobra jakość bez kontroli kosztów to nie sukces.
Monitoruj Cost per Resolution i ustaw alerty kosztowe.

Na koniec

Asystent AI to żywy produkt, nie jednorazowe wdrożenie.
Regularna praca z metrykami pozwala przestać zgadywać i zacząć podejmować decyzje oparte na danych.

Twoja firma potrzebuje sprawdzonego partnera IT?

Porozmawiajmy o potrzebach i wyzwaniach Twojej organizacji.

Michał Połetek
IT Client Partner

+48 539 524 698

Michał Połetek

SYZYGY.pl

linkedin.syzygywarsaw

SYZYGY