Trzy bramki, zanim wydasz budżet

Jak w dwa tygodnie sprawdzić, czy pomysł na AI się spina

Max Małecki · Tech Lead @ Insly

Znacie tę historię startupu

Temat grzeje inwestorów.

Więc MVP dostaje AI.


ZAMKNIJ SIĘ I WEŹ MOJE PIENIĄDZE

Programista zamiast trzech if-ów woła model przez LangChain.

Działa, więc zostaje.


Trzy if-y
Model
przez
LangChain
PROGRAMISTA

Runda wpada.

Teraz to oficjalnie startup AI.


RUNDA WPADA
ZBUDUJEMY Z TEGO PRODUKT

Ruch rośnie. Przychód rośnie.

Rachunek za tokeny rośnie szybciej.


PRZYCHÓD ROŚNIE
RACHUNEK ZA TOKENY ROŚNIE SZYBCIEJ

Przychód nie nadąża.

VC potrafi to policzyć szybciej niż zespół.


KOSZT TOKENÓW PRZEGONIŁ PRZYCHÓD
DORZUĆMY JESZCZE JEDEN FEATURE Z AI

Projekt nie umiera na braku klientów.

Umiera na koszcie zapytania.


STARTUP, KTÓRY NIE POLICZYŁ KOSZTU ZAPYTANIA
Pay Respects
F
Press F to Pay Respects

Cześć, jestem Max

20 lat w IT. Manager i architekt, który nie potrafi przestać kodować.

Insly.pl — Tech Lead

Dwa projekty AI na produkcji: AWS Bedrock i vLLM na k8s z GPU.

150 000 polis w PDF miesięcznie. **7 000 USD miesięcznie** rachunku.

Zespół 15 programistów, który po transformacji, którą przeprowadziłem, dowozi **cztery razy więcej**.

Po godzinach: mmx3.pl i aienabler.tech

mmx3.pl — konsulting dla dużych firm i startupów, szkolenia, warsztaty.

aienabler.tech — mój startup edtech. Ten rachunek za tokeny płacę też sobie.

Opowiem dziś o kilku ewaluacjach

Dzięki nim Wasz ciężko wypracowany MRR nie wyląduje od razu na koncie OpenAI, Anthropic albo AWS.

Pytania wrzucajcie na Slido

Dużo materiału, mało miejsca na pytania z sali.

Piszcie w trakcie, odpowiem na końcu i przy stoliku.

`app.sli.do/event/8mA8cmhaRHc51u1amMBbSf`

Trzy bramki, zanim wydasz budżet

Jak w dwa tygodnie sprawdzić, czy pomysł na AI się spina

Max Małecki · Tech Lead @ Insly

Prototyp jest tani.

Działające demo AI zbudujecie dziś po obiedzie, za kilka złotych.

Z agentem kodującym to kwestia popołudnia, nie sprintu.

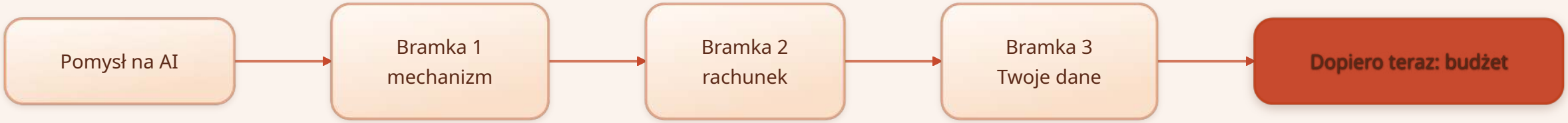
Utrzymanie nie.

Utrzymanie tego samego demo dla prawdziwych użytkowników kosztuje zupełnie inaczej.
I to ta różnica zabija projekty AI, nie jakość modelu.

Ile to kosztuje co miesiąc, kiedy już działa? Lepiej policzyć teraz niż za pół roku.

Nie powiem Wam, czy ktoś to kupi

Powiem, czy da się to zbudować, za ile i czy budżet się spina.



Dwa dni to decyzja przy bramce. **Dwa tygodnie** to prototyp, który przez nią przepuszczacie.
Bramki 1 i 2 robicie sami, na kartce. Bramka 3 to dwa dni z inżynierem.
Każda bramka kończy się decyzją: **zabij albo idź dalej**.


PZPN
NIKE
ORLEN

Kwartały do stracenia

3

Bramka 1 — czy AI jest tu właściwym mechanizmem?

Ta bramka kosztuje dwa dni. Pominięcie jej kosztuje kwartał.


NASZ STARTUP: ASYSTENT AI
NYPD
STARTUP OBOK: ASYSTENT AI

Reguły są deterministyczne?

Zwykły formularz z walidacją.

Wysoki koszt błędu i nikt nie patrzy?

Nie automatyzować. Albo human in the loop.

Bez tego cała odpowiedzialność za błąd zostaje na firmie.

Nie ma danych ani źródeł?

Najpierw zebrać dane. Model bez danych to generator zgadywanek.

Wystarczy wyszukiwanie pełnotekstowe?

Wyszukiwanie pełnotekstowe.


Zawsze było
Czekaj, to całe AI to
wyszukiwanie pełnotekstowe?

Drabina: wejdźcie na najniższy stopień, który działa

Stopień	Co to	Kiedy wystarczy
1. Prompt	jedno pytanie, jedna odpowiedź	przepisanie, klasyfikacja, generowanie

Drabina: wejdźcie na najniższy stopień, który działa

Stopień	Co to	Kiedy wystarczy
1. Prompt	jedno pytanie, jedna odpowiedź	przepisanie, klasyfikacja, generowanie
2. Prompt z kontekstem	wklejacie dokument do promptu	mało dokumentów i się nie zmieniają

Drabina: wejdźcie na najniższy stopień, który działa

Stopień	Co to	Kiedy wystarczy
1. Prompt	jedno pytanie, jedna odpowiedź	przepisanie, klasyfikacja, generowanie
2. Prompt z kontekstem	wklejacie dokument do promptu	mało dokumentów i się nie zmieniają
3. RAG	wyszukanie fragmentu, potem prompt	wiedza firmowa, duży korpus

Drabina: wejdźcie na najniższy stopień, który działa

Stopień	Co to	Kiedy wystarczy
1. Prompt	jedno pytanie, jedna odpowiedź	przepisanie, klasyfikacja, generowanie
2. Prompt z kontekstem	wklejacie dokument do promptu	mało dokumentów i się nie zmieniają
3. RAG	wyszukanie fragmentu, potem prompt	wiedza firmowa, duży korpus
4. Workflow	stałe kroki, bez „myślenia"	powtarzalny proces, mierzalny efekt

Drabina: wejdźcie na najniższy stopień, który działa

Stopień	Co to	Kiedy wystarczy
1. Prompt	jedno pytanie, jedna odpowiedź	przepisanie, klasyfikacja, generowanie
2. Prompt z kontekstem	wklejacie dokument do promptu	mało dokumentów i się nie zmieniają
3. RAG	wyszukanie fragmentu, potem prompt	wiedza firmowa, duży korpus
4. Workflow	stałe kroki, bez „myślenia"	powtarzalny proces, mierzalny efekt
5. Agent	pętla: myśl, narzędzie, ocena	trzeba dynamicznie wybierać narzędzia

Pytanie kontrolne przy każdym stopniu: **„dlaczego nie prościej?"**

150 000

polis w PDF miesięcznie

Workflow, nie agent

Kroki są znane i stałe: odbierz → wyciągnij dane → zwaliduj → zapisz.

Agent dołożyłby tylko koszt i nieprzewidywalność, dokładnie tam, gdzie potrzebujemy powtarzalności.

Rachunek: 5 500 USD miesięcznie

Claude Haiku 4.5 na AWS Bedrock, region EU. 98,5% skuteczności ekstrakcji.

To ~3,7 centa za polisę. Tanio, bo ekstrakcja wsadowa tanim modelem raz na dokument, a nie czat z długim kontekstem przy każdym pytaniu.

Trzy pytania, które ustawiają stopień. I czwarte, najważniejsze.

1. Czy kroki są znane i stałe?

Jeśli tak, to prompt, RAG albo workflow. Nie agent.

2. Czy potrzebna jest pamięć między krokami?

Jeśli tak, to workflow ze stanem. Nadal nie agent.

3. Czy trzeba dynamicznie wybierać narzędzia albo źródła?

Dopiero to uzasadnia agenta.


WELCOME, MR. ANDERSON

4. Jaki jest koszt błędu i gdzie zostaje człowiek?

Im wyższy koszt błędu, tym wcześniej człowiek zatwierdza:
przed skutkiem, nie w audycie tydzień później.


HUMAN IN THE LOOP
OSTATNI, KTÓRY MOŻE POWIEDZIEĆ NIE

Model, który halucynuje, jest gorszy niż formularz, który działa.

Bramka 2 — czy budżet się spina?

Ta bramka kosztuje dwa dni. Pominięcie jej kosztuje kwartał.

Kwartały do stracenia

2

Rachunek na kartce

```
użytkownicy × zapytania × koszt zapytania = koszt miesięczny
```

Przykład: 200 użytkowników, po 20 zapytań miesięcznie. To **4000 zapytań**.

Kiedy budżet się spina

Spina się, gdy **koszt zapytania jest niższy niż to, co za nie weźmiecie albo zaoszczędzicie**. To Wasze założenie, nie moje, ale musi być na kartce.

Na najdroższym API: $70 / 200 użytkowników = **35 centów na głowę miesięcznie**. Ile bierzecie?

Jeśli nie spina się na kartce, **nie zepnie się w produkcji**.

Ta sama funkcja: od $0,60 do $942 miesięcznie

4000 zapytań miesięcznie, zapytanie ≈ 2000 tokenów wejścia + 300 wyjścia. Ceny w USD:

Opcja	Koszt miesięczny
Model open hostowany w EU (gpt-oss na OVH)	**~$0,60**
Mistral Large 3 (EU)	**~$6**
Gemini Flash (US)	**~$10**
Opus 5 (US, frontier)	**~$70**
GPU w chmurze 24/7 (L4)	**$942** — płacicie za dostępność, nie za użycie

Między API: **ponad 100×** za tę samą funkcję. GPU to inna oś: czas, nie użycie.

Ten rachunek ma datę ważności: **te liczby zmieniły się dwa razy w trzy tygodnie**.

EU przestało być kompromisem cenowym

Model z Europy jest tańszy niż najtańszy amerykański.

Rok temu tak nie było.

Porozmawiajmy o ukrytych kosztach

Tych, których nie ma w cenniku modelu.


PSSS... MŁODY
MAM DLA CIEBIE FAKTURY

Ukryty koszt #1: długi kontekst

Płacicie za niego przy każdym wywołaniu, nie raz.

Najczęstszy błąd, jaki widziałem: prompt systemowy na dwie strony, doklejany do każdego zapytania.

Ukryty koszt #2: retry po błędzie

Nieudana odpowiedź też jest rozliczona.

Ukryty koszt #3: historia rozmowy

Rośnie z każdą turą i cały czas wchodzi jako wejście.

Ukryty koszt #4: embeddingi w pętli

Wektoryzujecie cały korpus przy każdym pytaniu zamiast raz na starcie.

Ukryty koszt #5: cold start (własny model)

Model ładuje się 30–60 sekund. Użytkownik odchodzi po trzech.

Ukryty koszt #6: człowiek, który zatwierdza

4000 zapytań × 30 sekund sprawdzenia = 33 godziny miesięcznie.

Tego nie ma w cenniku modelu, a przy takim rachunku $10 czy $70 przestaje mieć znaczenie.

Ukryty koszt #7: ewaluacja po każdej zmianie promptu

Jedno zdanie dopisane do promptu systemowego potrafi zepsuć odpowiedzi w całym systemie.

Bez pomiaru dowiecie się o tym od użytkownika, nie z CI.

Te same 30 pytań (Bramka 3) po każdej zmianie promptu, modelu i wersji API.
Tyle kosztuje to, żeby produkcja nie była krucha.


ZMIENILIŚMY
JEDNO ZDANIE
W PROMPCIE
ODPALILIŚCIE
TEST PRZED
DEPLOYEM, PRAWDA?
Prawda?

Najtańsze zapytanie to takie, którego nie wysłałeś.

Bramka 3 — czy to działa na MOICH danych?

Ta bramka kosztuje dwa dni. Pominięcie jej kosztuje kwartał.


BRAMKA 3: WASZE DANE
NIE, NIE TEN DATA

Kwartały do stracenia

1

Jakość danych jest kluczowa

Najdroższy model nie uratuje słabych danych.

Śmieci na wejściu to śmieci na wyjściu, tylko policzone po cenie tokenów.


Zły model
WASZ TEST
Wasze
własne
dokumenty

Pewnie i błędnie: „Czy HDI oferuje wariant assistance PLATYNOWY?"

> **Model ogólny, bez naszych dokumentów:** „**Tak**, HDI posiada wariant Platynowy..."
> To **wymyślone**. Taki wariant ma inny ubezpieczyciel.

> **RAG na OWU HDI 2026:** „**NIE.** HDI oferuje ZŁOTY+, PODRÓŻNIK 15..." i cytuje paragraf.

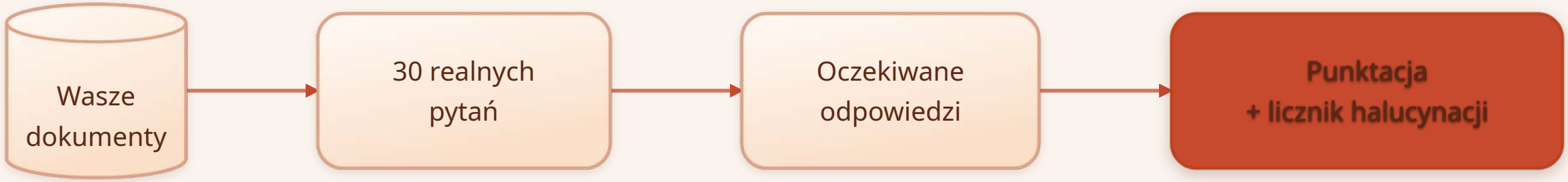
Pewnie i błędnie: odszkodowanie AC przy szkodzie całkowitej

Model ogólny: 21 000 zł. Błędna formuła: złapał liczby-wabiki z pytania.

RAG na OWU: 16 000 zł. Wartość pojazdu 20 000 minus pozostałość 4 000, z paragrafem.

Na zamkniętej domenie model bez Waszych dokumentów **halucynuje z pełnym przekonaniem**.

Jedyny wiarygodny test to Wasz własny



30 pytań, które naprawdę zadają Wasi użytkownicy. Do każdego dopisujecie odpowiedź, którą uznajecie za poprawną.
Każdy model dostaje te same pytania i punktację.

Próg zaliczenia ustalacie PRZED testem. Inaczej zawsze wyjdzie „prawie działa, jeszcze dwa tygodnie".

Dwa dni z inżynierem, nie program badawczy.

Wynik u nas: mniejszy model nie był gorszy

Ten sam test, 30 pytań o OWU:

	model 7B	model 11B
Zdane pytania	**30 / 30**	30 / 30
Idealne odpowiedzi	**26 / 30**	21 / 30
Fragmenty bez pokrycia w OWU	**5**	6
Czas odpowiedzi	**8,8 s**	13,0 s

Mniejszy nie był gorszy: **1,5× szybszy** i tańszy w hostingu.

My też wybraliśmy większy. Dopóki nie zmierzyliśmy.

Bonus: test złapał błąd, którego nie dało się wyklikać

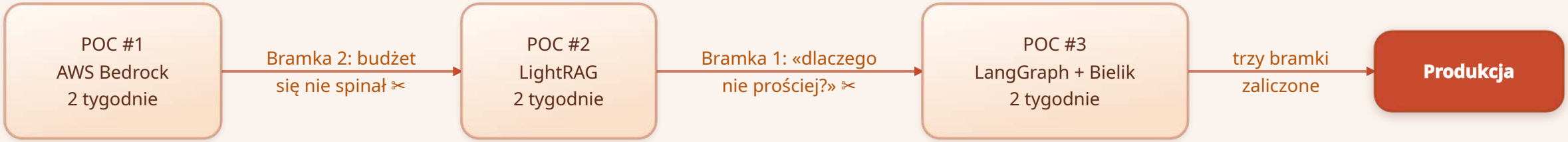
Jedna z testowanych wersji **wstrzykiwała śmieciowe znaki do każdej odpowiedzi**.

W ręcznym testowaniu niewidoczne. Prawie poszło na produkcję.
W teście widoczne od pierwszego uruchomienia.

Taki test to nie biurokracja. To ubezpieczenie.

Najmniejszy model, który przechodzi Twój test, nie największy, na jaki Cię stać.

Tak to wygląda w praktyce: trzy prototypy po dwa tygodnie



Każdy pivot to była decyzja przy bramce, nie porażka po kwartale.

Kwartały do stracenia

0


LEFT
EXIT 12
Pół roku R&D, decyzja na końcu
Zabić po dwóch tygodniach
My

Sześć tygodni zamiast pół roku

Klasyczne R&D: 6+ miesięcy, decyzja na końcu.

U nas: 6 tygodni, dwie decyzje o zabiciu, zerowy zmarnowany kwartał.

Prototyp był tani trzy razy. **Utrzymanie policzyliśmy raz i to wystarczyło, żeby dwa zabić.**

Różnica nie jest w technologii. Jest w tym, **kiedy podejmujecie decyzję**.

R&D nie kosztuje już kwartałów. Kosztują decyzje, których nie odkładasz.

Dwa pytania, które i tak padną

A RODO? Trzy drogi

	API poza EU	API w EU	GPU w chmurze 24/7
Kto	OpenAI, Google, Anthropic	CloudFerro, OVH, Mistral, PCSS	Wy, na wynajętej karcie
Koszt	grosze za zapytanie	grosze za zapytanie	~$942/mies, niezależnie od ruchu
Dane wrażliwe	wychodzą	zostają w EU	nie opuszczają Waszej instancji
Jurysdykcja	**USA, CLOUD Act**	EU	Wasza

Region EU ≠ jurysdykcja EU. Amerykański dostawca postawi Wam serwer we Frankfurcie i nadal będzie podlegał prawu USA.

Środkowa kolumna to nowość ostatniego roku i dziś jest tania.

„Postawimy własny model, będzie taniej"

GPU dostępne całą dobę: **~$942/mies, niezależnie od ruchu**.

Próg opłacalności wobec taniego modelu open hostowanego w EU ($0,06/1 mln tokenów):
~17 miliardów tokenów miesięcznie (~570 milionów dziennie przy L4 na AWS).
(Nawet na tanim EU-hostingu za $576 to wciąż ~11 miliardów tokenów).

Prawie nikt z tej sali tam nie dojdzie.

Własne GPU kupuje się dla **suwerenności i kontroli**. Cena już tego nie uzasadnia.

Reasumując

Trzy pytania na najbliższe dwa tygodnie.

1. Czy AI jest tu właściwym mechanizmem?

A może formularz, wyszukiwarka albo zwykły workflow?

2. Czy budżet się spina?

Użytkownicy × zapytania × koszt zapytania kontra to, co za to weźmiecie.

3. Czy to działa na MOICH danych?

30 pytań, próg ustalony przed testem, dwa dni z inżynierem.

Każda bramka: dwa dni.

Pominięcie: stracony kwartał.

Macie pomysł?

Złapcie mnie w kuluarach, przepuszczę go przez trzy bramki w pięć minut.

Jak zbudować taki test?

O tym będzie mówił Damian w następnej prezentacji.

A czy Ty możesz sobie pozwolić stracić trzy kwartały?

Q&A

Wasze pytania czekają na Slido — biorę je teraz.

Kto nie zdążył wrzucić, kod jest wciąż aktywny.

`app.sli.do/event/8mA8cmhaRHc51u1amMBbSf`

Dzięki!

Pytania łapcie w kuluarach, nie w kolejce do mikrofonu

Max Małecki · Tech Lead @ Insly

`linkedin.com/in/maxmalecki` · `mmx3.pl` · `aienabler.tech`