24.09.2025, Gdańsk
Hubert Pietroń
Piotr Zieliński

Zespół AI HUB

AI Hub is here to help on Hapag-Lloyd's AI journey

Vision

We are convinced that the **smart application of AI will become a necessity** and even a differentiating factor over the next 5 years - especially in terms of cost and quality.

This is why by 2030 **100% of our customers and employees** will be positively impacted by AI.

Specifically, we commit that Customers will benefit from **AI supporting 2 Quality Promises every year** from 2025.

We aim to be of help through:

Enablement
of AI at Hapag-Lloyd

Consulting
to help navigate the AI landscape

Delivery
of AI solutions

Hapag-Lloyd

Zakres prezentacji

Kontekst biznesowy

Podłoże teoretyczne

Architektura wstępnego rozwiązania

Pierwsze wnioski I błędy

Kolejne kroki

Konkluzje & QA

Kontekst biznesowy

Hapag-Lloyd - lider w branży transportu kontenerów

Globalna obecność: 17,100 pracowników i 400 biur w 140 krajach - Hapag-Lloyd jest jedną z najbardziej rozpoznawalnych marek w branży żeglugi liniowej

Flota: 313 nowoczesnych statków kontenerowych o łącznej pojemności 2,5 mln TEU - jedna z największych flot na świecie

Usługi liniowe: 133 usługi liniowe na całym świecie, zapewniające szybkie i niezawodne połączenia pomiędzy ponad 600 portami na wszystkich kontynentach - umożliwiające klientom Hapag-Lloyd transport towarów w dowolne miejsce na świecie

Wprowadzenie do projektu i zakresu

Asysta w rozwiązywaniu spraw klientów

Obsługa klienta w Hapag-Lloyd w liczbach:

- 3200 agentów pracujących
- 20 milionów spraw rocznie
- 34 podtypy spraw z 230 różnymi scenariuszami biznesowymi

Benefity wykorzystaniu AI:

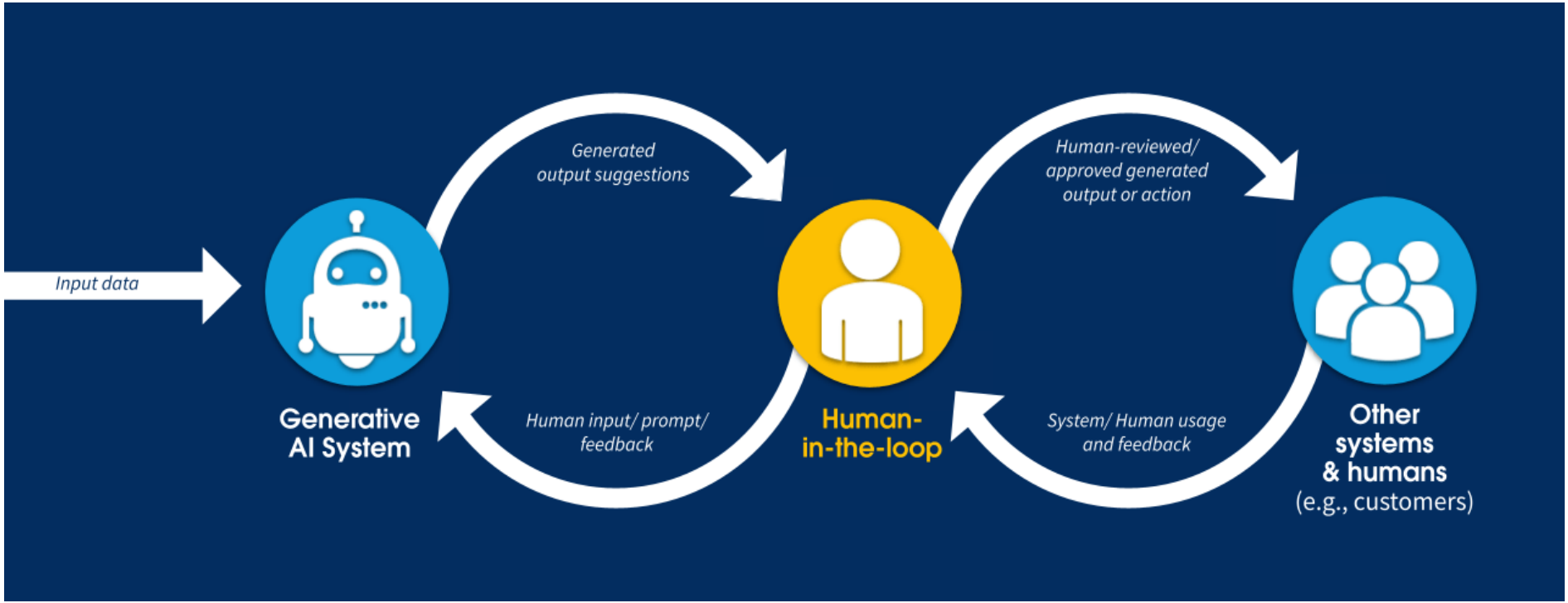
- zwiększyć efektywność
- poprawić jakość obsługi
- Automatyzacja powtarzalnych caseów

Human in the loop

Asysta w rozwiązywaniu spraw klientów

Human in the Loop to podejście, w którym człowiek jest aktywnie włączony w proces automatyzacji opartej na AI, takim jak generowanie odpowiedzi przez modele językowe (LLM). W kontekście obsługi klienta, np. w Hapag-Lloyd, AI (jak LLM) automatycznie streszcza zapytania i proponuje odpowiedzi, ale agent obsługi klienta weryfikuje, edytuje lub zatwierdza je przed wysłaniem. To minimalizuje błędy, zapewnia zgodność z regulacjami i poprawia jakość, łącząc efektywność AI z ludzkim nadzorem

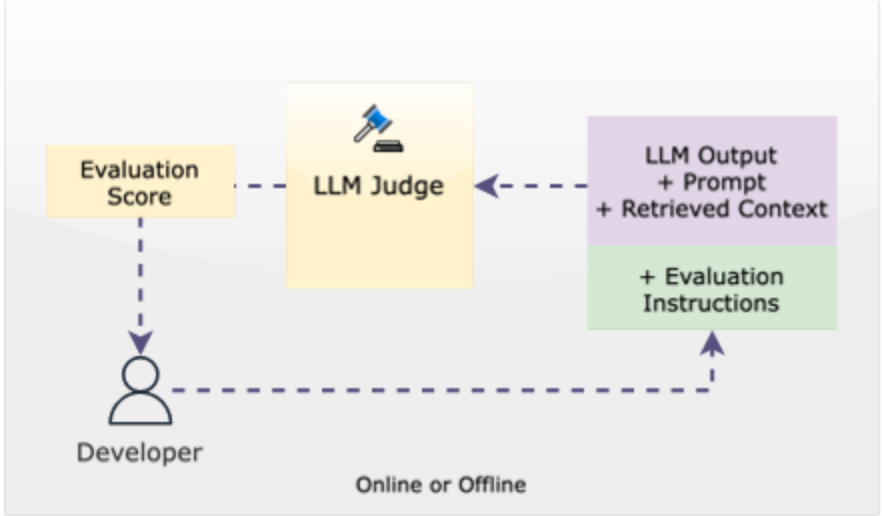
Salesforce Gen-AI



Hapag-Lloyd

LLM as a Judge

LLM as a Judge



Automatyczna ewaluacja – LLM zastępując człowieka pełni rolę sędziego, oceniając odpowiedzi innych modeli

Ocena absolutna – LLM wystawia ocenę liczbową (np w skali 1-10) albo jakościową (zła, średnia, dobra)

Redukcja kosztów/skalowalność -
zmniejsza potrzebe ręcznej oceny przez ludzi

Jak wypadają LLM jako sędziowie?

Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena

Porównanie LLM as a Judge w dwóch ustawieniach:

- MT-Bench – wieloturowe prompty, w których LLMs są proszone o ocenę odpowiedzi chatbotów (przydatność, spójność, poprawność)
- Chatbot Arena – platforma, na której ludzie głosują między dwiema anonimowymi odpowiedziami chatbotów, dostarczając danych o preferencjach na dużą skalę

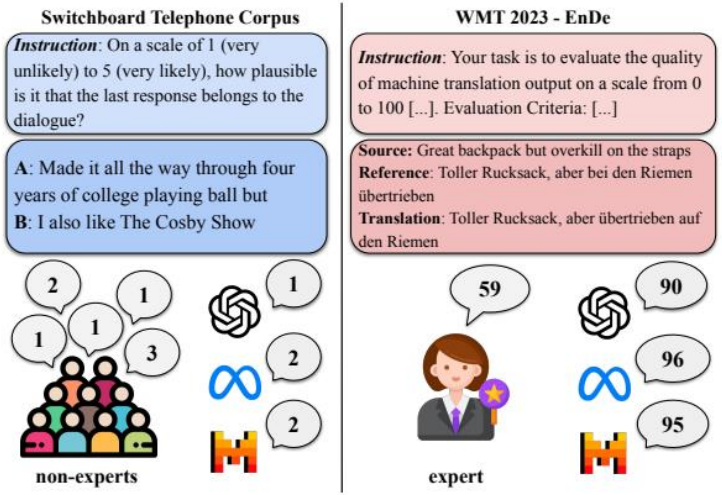
GPT-4 jako sędzia wykazał ~ 80% zgodności z ludzkimi ocenami

Wykryto pewne tendencyjności – preferowanie modeli ze swojej rodziny oraz dłuższych odpowiedzi

Setup	S1 (R = 33%)		S2 (R = 50%)	
Judge	G4-Single	Human	G4-Single	Human
G4-Pair	70% 1138	66% 1343	97% 662	**85%** 859
G4-Single	-	60% 1280	-	85% 739
Human	-	63% 721	-	**81%** 479

(a) First Turn

Jak wypadają LLM jako sędziowie?



LLMs instead of Human Judges? A Large-Scale Empirical Study across 20 NLP Evaluation tasks

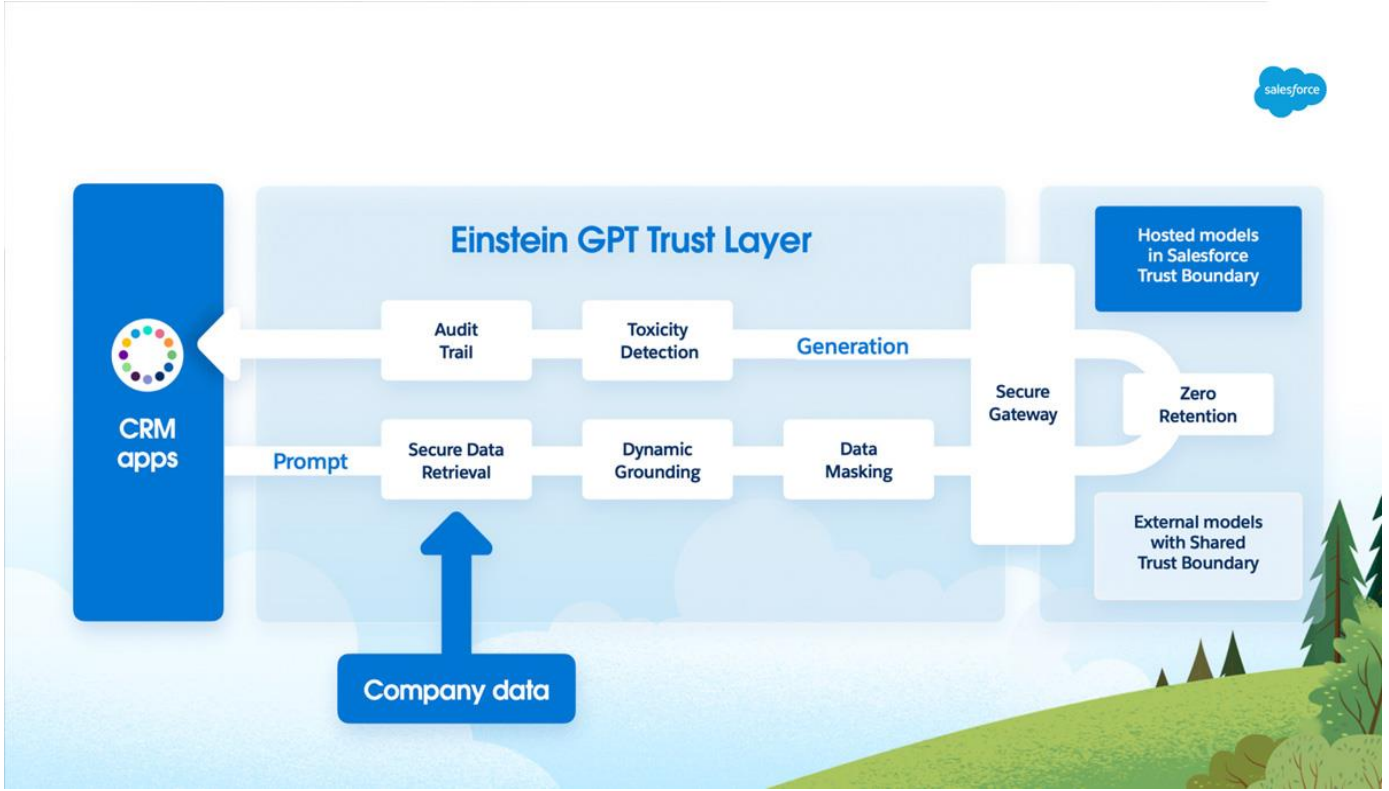
11 modeli LLM zostało ocenionych, zarówno modeli komercyjnych (GPT-4o), jak i modeli o otwarto źródłowych (Mixtral)

- Modele LLM wykazują obiecującą zgodność z ludzkimi wyborami w przypadku niektórych zadań, jednak ich wydajność jest bardzo zróżnicowana.
- Żaden pojedynczy model nie jest uniwersalnym substytutem dla ludzi. Autorzy zalecają walidacje per task dla danego zadania przed przyjęciem ewaluatorów LLM do swoich procesów.

Hapag-Lloyd

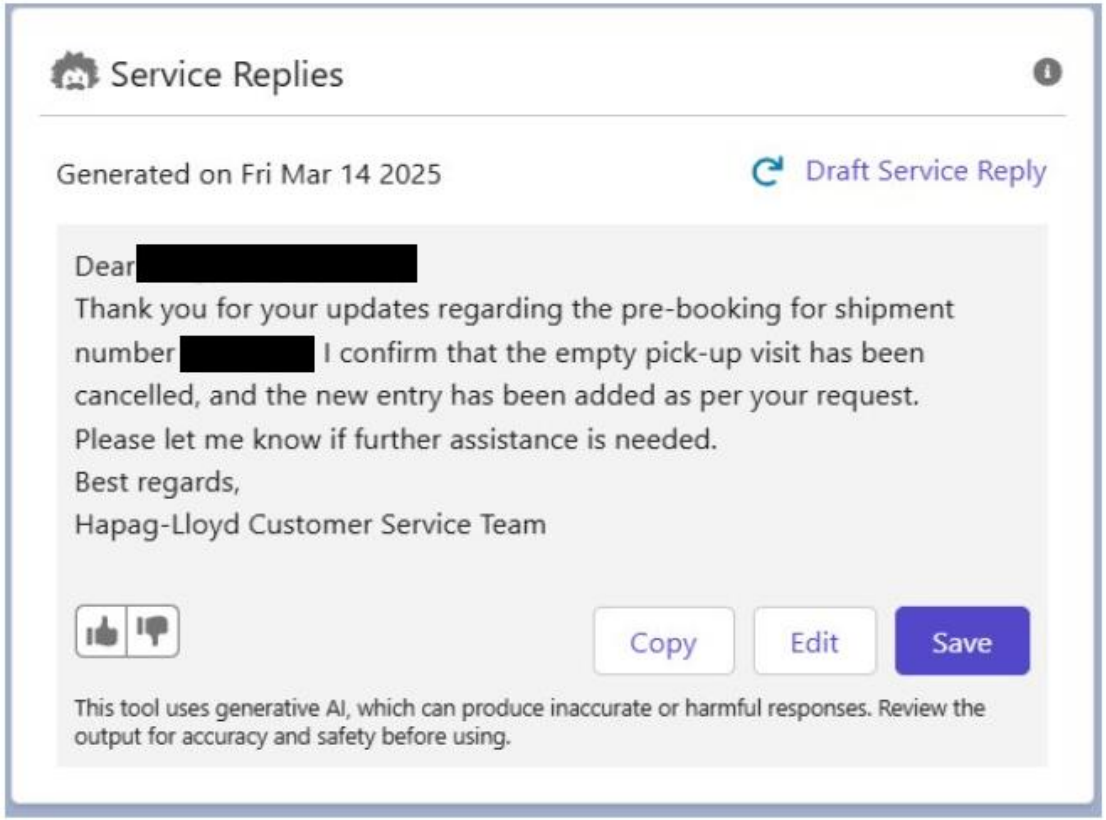
Ekosystem projektu

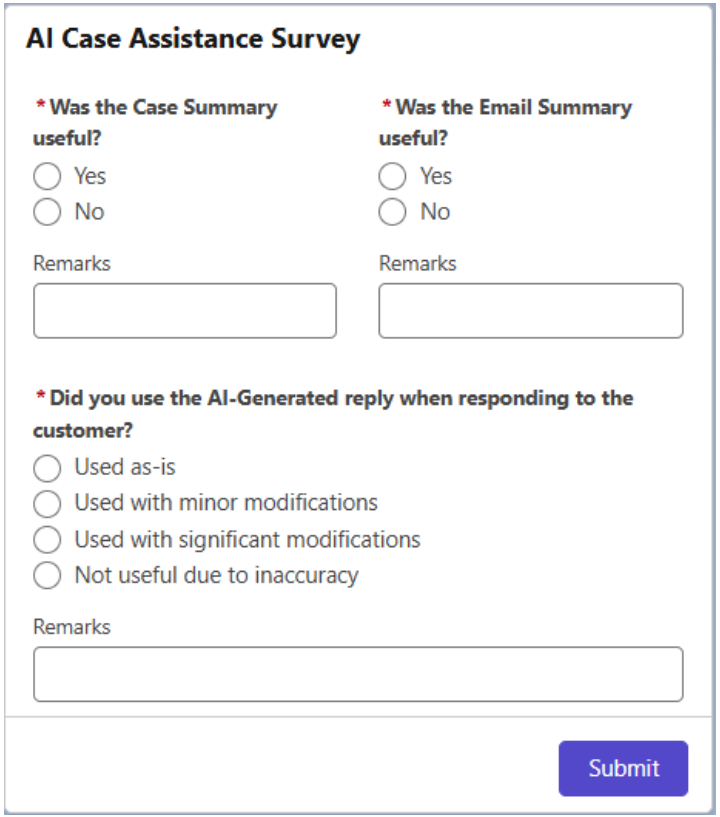
Salesforce Einstein GPT



Hapag-Lloyd

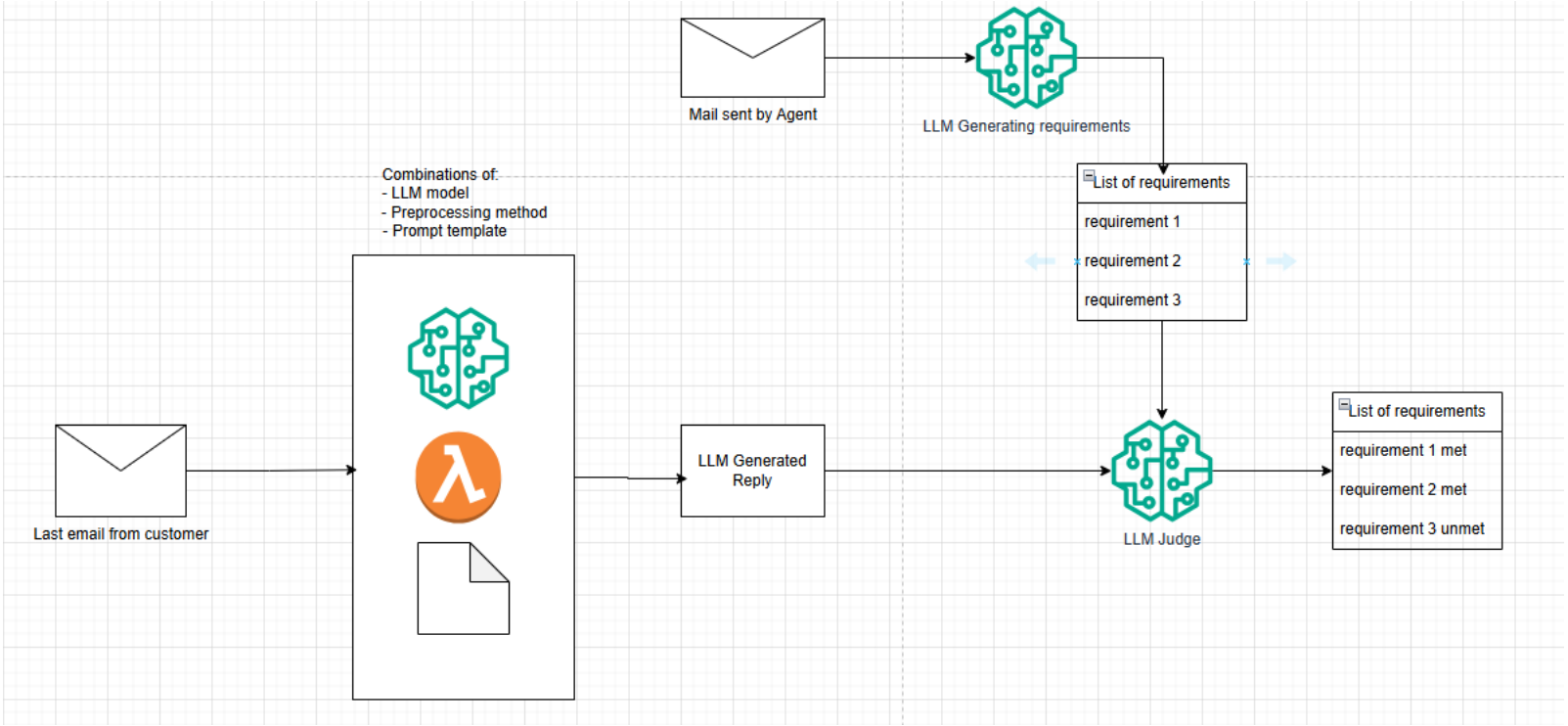
Elementy aplikacji CRM





Hapag-Lloyd

Evaluation framework 1.0



Hapag-Lloyd

Przykładowy output


```
{
  "case_number": 12345,
  "requirement": "The service agent provides an estimated timeframe for responding to the customer's request",
  "requirement_id": 1,
  "fullfiled": false,
  "justification": "The generated response only informs you that we will do it as soon as possible"
}
```

Przykładowy output


```
{
  "case_number": 12345,
  "requirement": "The reply must acknowledge receipt of the customer's message and express gratitude for their support.",
  "requirement_id": 2,
  "fullfiled": true,
  "justification": "The generated response acknowledges the customer's message by stating 'Thank you for reaching out to us' and expresses
  gratitude by saying 'We appreciate your support'."
}
```

Pierwsze sprinty – co się sprawdziło

Wprowadzone zmiany

Z pomocą frameworku, udało nam się wprowadzić zmiany, których wynikiem było:

- Priorytetyzacja najnowszej wiadomości e-mail

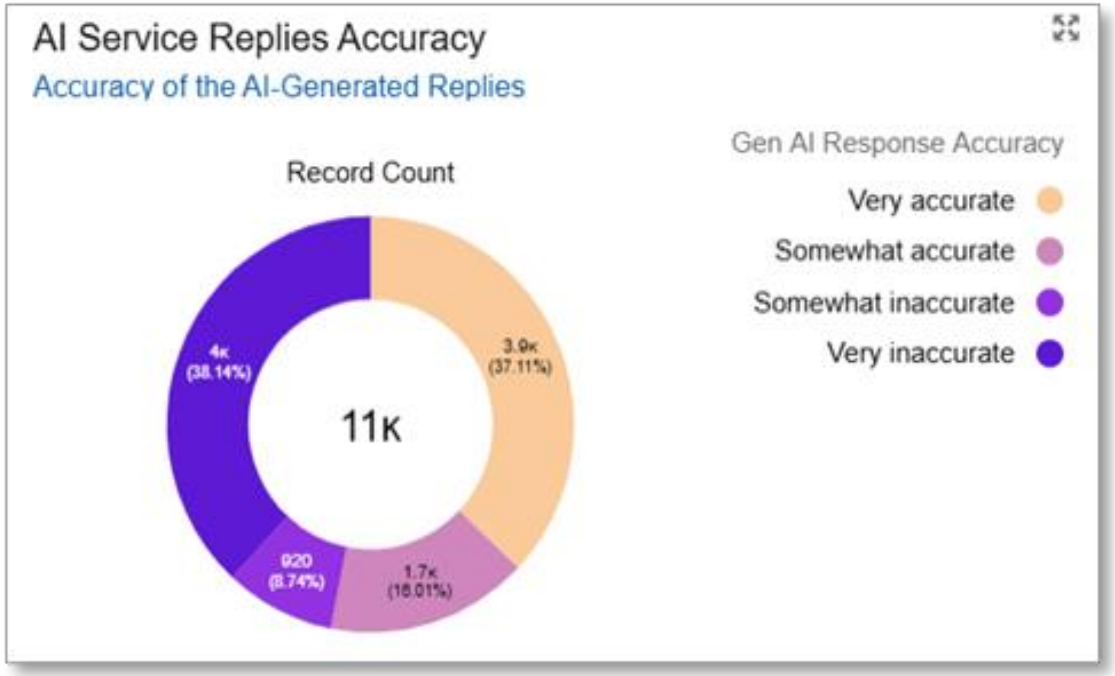
- Zmniejszenie nadmiernej szczegółowości i zbytki grzecznościowych sformułowań

- Adresowanie odpowiedzi do nadawcy ostatniej wiadomości e-mail lub alternatywnie „Szanowny Kliencie” jeśli nie można tego określić

Wprowadzone zmiany

CHANGE	OLDER VERSION	NEWER VERSION
1. Additional examples of excessive politeness markers	Avoid excessive politeness markers (e.g. multiple apologies or multiple expressions of gratitude).	Avoid excessive politeness markers (e.g. multiple apologies or multiple expressions of gratitude, such as "We sincerely apologize for the inconvenience..", "Thank you very much for your understanding and patience...").
2. Change in sign-off format	End with a friendly closing and sign off the email with "Hapag-Lloyd Customer Service Team" followed by a line break and {!$Input:Case.HLLastWorkedOnByLookup__r.FirstName}.	End with a friendly closing and sign off the email with "Hapag-Lloyd Customer Service Team" followed by a line break and {!$User.Name}.
3. Handling unknown sender	If you work on a customer Case address the recipient of your draft reply is sender of the latest email.	If you work on a customer Case the recipient of your draft reply is sender of the latest email. If the sender of the latest email cannot be determined, use "Dear Customer".
4. Conditional use of data fields	Use the following data in your draft reply if they are available...	Use the following data in your draft reply: {if {!$Input:Case.Field} not null then {!$Input:Case.Field}...
5. Removal of reference to "the original cause for the Case"	{!$Input:Case.Subject} - The original cause for the Case	No explicit mention of this description in the detailed instructions. Still available under ##Data as "Original issue: {!$Input:Case.Subject}, {!$Input:Case.HLShortDescription__c}".

Ankiety jako wyznacznik kierunku projektu



Przeprowadzenie ankiety jako ważnej części wstępnego POC pozwoliło na szybkie zmierzenie dokładności i otworzyło możliwość analizy obszarów, które nie działały zgodnie z oczekiwaniami

Wykorzystując uwagi z ankiety, przedstawiciele biznesu i LLM zdefiniowaliśmy 15 kategorii odpowiedzi

Hapag-Lloyd

Analiza dodatkowych uwag

Category	% of cases (overall)	First 4 weeks (14.02-15.03)	Last 4 weeks (16.03-11.04)	Trend	Feedback type	Explanation if needed
TOTAL:	*100% (5698)*	*100% (2889)*	*100% (2809)*			
Positive Feedback	23% (1311)	16% (475)	30% (836)	✅ +14%	1. Positive Feedback	
Answer accuracy	41% (2397)	40% (1145)	44% (1252)	❌ +4%	2. FIS Assessment Contradiction	
					3. Internal Guidelines contradiction	
					4. Critical Info Overlooked by LLM	
					5. Inaccuracy / Misinformation	E.g. using only part of customer's name or company name as the full one
					6. Hallucinations	E.g. making up vessel name
					7. Specific Action Required	specific TODO action required, currently impossible for Einstein (e.g. sending SWB)
Language	3% (158)	2% (54)	4% (104)	❌ +2%	8. Wordy/Vague or Too Detailed	
					9. Tone/Formality Issues	
Relevance and Focus	15% (867)	22% (643)	8% (224)	✅ -14%	10. Focus not on Latest Email	the auto-generated reply did not take the last msg into account
					11. No Reply Required	case did not require additional reply to the customer
					12. Paraphrasing	just paraphrasing customer's message
					13. Placeholder / Filler Reply	e.g. just providing "we'll get back to you" message
					14. Irrelevant Response	Generate reply completely off
Other	17% (965)	20% (572)	14% (393)	✅ -6%	15. Other	

Wykorzystując LLM z konsultacją z osobami bliżej biznesu udało nam sę stworzyć kategorie uwag

Dzięki kategoryzacji mogliśmy ustalić priorytety zmian w kolejnych sprintach

Większość problemów dotyczyła braku informacji, czego można było oczekiwać w przypadku modelu z tylko ogólno dostępną wiedzą, ale teraz byliśmy w stanie zmierzyć skalę tego zjawiska

Hapag-Lloyd

Pierwsze sprinty – zidentyfikowane problemy

Evaluation framework i grupowanie wymagań

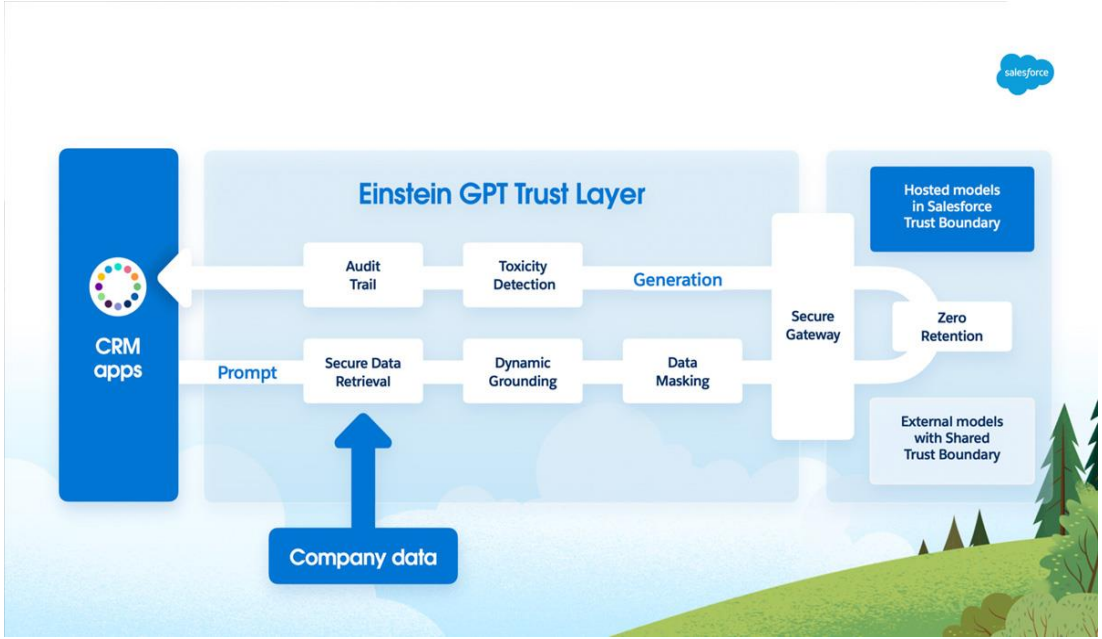
Document sent
Final copies have been sent to the customer's inbox
Mention that final rated sea waybills were sent as per the customer's request
The OBLs have been issued to the web for printing
The reply must inform the recipient that the draft and final rated SWB copies have been sent to their inbox.
The customer's request to print OBLs at destination has been submitted to the agents for handling
The request to print OBLs at destination has been submitted to the agents for handling
The reply mentions that the final seaway copies have been shared as requested
Final copies of the documents have been sent to the customer's inbox
The request to print OBLs at destination has been submitted for handling
Notification that the OBLs have been issued to the web for printing
Mention of an unrated SWB copy being sent for most of the mentioned MTDs

Próbując wykonać ten sam zabieg jak z uwagami agentów, wygenerowaliśmy kategorie wymagań wygenerowanych przez LLM

Wymagania były bardziej złożone co prowadziło do błędnych kategoryzacji

Przykładowo
informacja o brakujących dokumentach i informacja o skorzystaniu z formularza internetowego powinny znaleźć się w tych samych kategoriach, ogólny LLM starał się tworzyc oddzielne kategorie w tym przypadku

Trust Layer wpływający na porównanie



Złożony system, który chroni dane firmy

Korzystając w frameworku z modelu LLM hostowanego w databricksie nawet ten sam prompt finalnie trafiał w innej postaci w Salesforce

Modele LLM OpenAI otrzymywały różne dane, prowadząc do tworzenia metryk, którym ciężko było zaufać

Hapag-Lloyd

Następne kroki & cele dla projektu

Generyczne odpowiedzi nie nadają sie do większości spraw klientów

Przykładowa wiadomość od klienta:

"Potrzebuję kopii certyfikatu pochodzenia dla mojej przesyłki morskiej. Czy możecie mi go przesłać?"

Generyczne odpowiedzi nie nadają sie do większości spraw klientów

Przykładowa wiadomość od klienta:	"Potrzebuję kopii certyfikatu pochodzenia dla mojej przesyłki morskiej. Czy możecie mi go przesłać?"
Przykładowa odpowiedź generyczna AI:	"Przepraszamy za niedogodności. Prosimy o kontakt z naszym działem dokumentacji, aby uzyskać kopię certyfikatu pochodzenia."

Generyczne odpowiedzi nie nadają sie do większości spraw klientów

Przykładowa wiadomość od klienta:	"Potrzebuję kopii certyfikatu pochodzenia dla mojej przesyłki morskiej. Czy możecie mi go przesłać?"
Przykładowa odpowiedź generyczna AI:	"Przepraszamy za niedogodności. Prosimy o kontakt z naszym działem dokumentacji, aby uzyskać kopię certyfikatu pochodzenia."
Najbardziej odpowiednia odpowiedź	"Znalazłem Twój certyfikat pochodzenia dla numeru konosamenta HL001234. Przesyłam go Ci mailem."

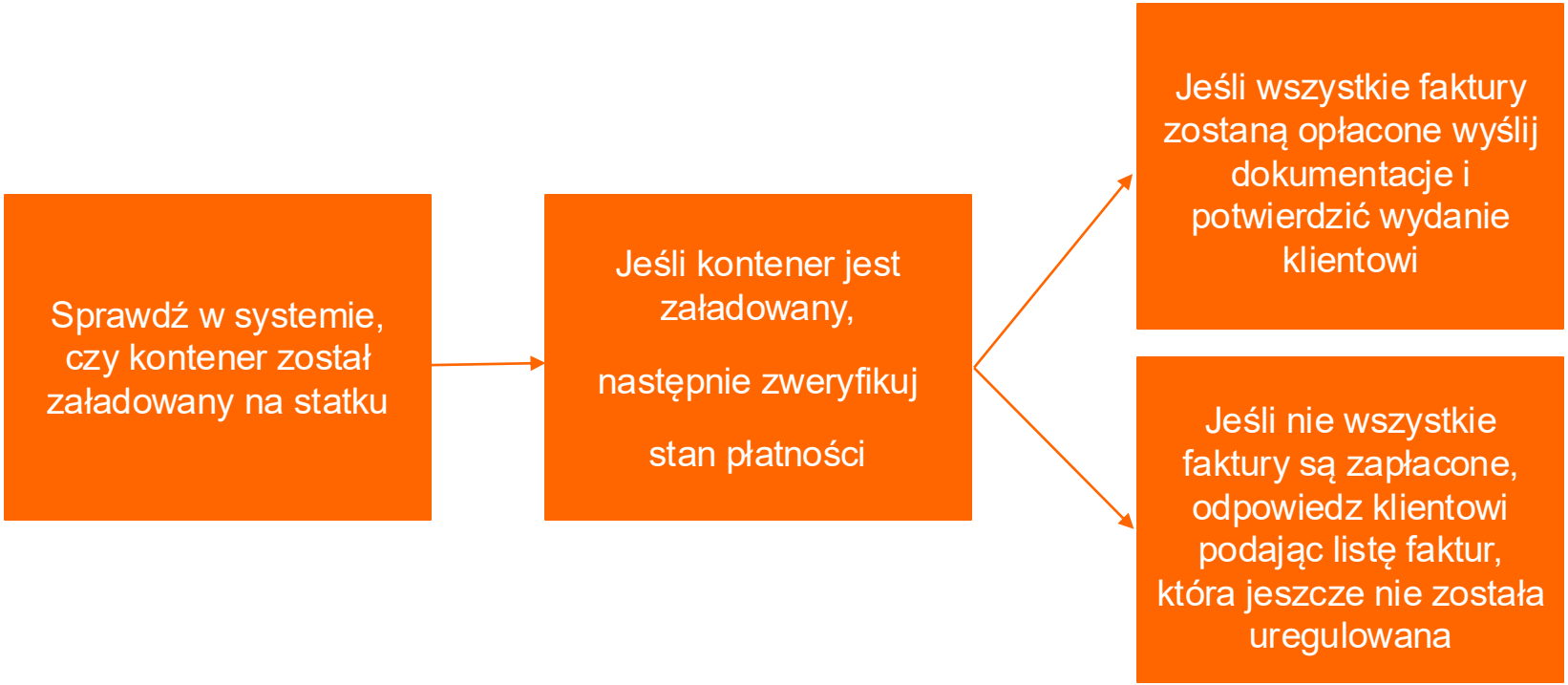
Poprawa procesu generowania podpowiedzi

Standardowa Procedura Operacyjna eng. (SOP)

Dokument opisujący szczegółowe, pisemne instrukcje, jak krok po kroku wykonać określone zadanie lub proces w organizacji, aby zapewnić spójność, jakość i wydajność działania.

Dokumentacja może być zamiennie tworzona I interpretowana przez ludzi lub AI.

Standardowa Procedura Operacyjna eng. (SOP)



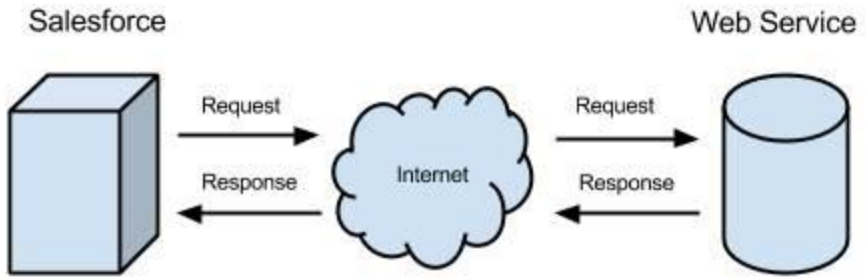
Hapag-Lloyd

Dodatkowe źródła danych

Aby sztuczna inteligencja mogła trafnie odpowiedzieć klientowi, musimy najpierw dostarczyć jej niezbędne dane.

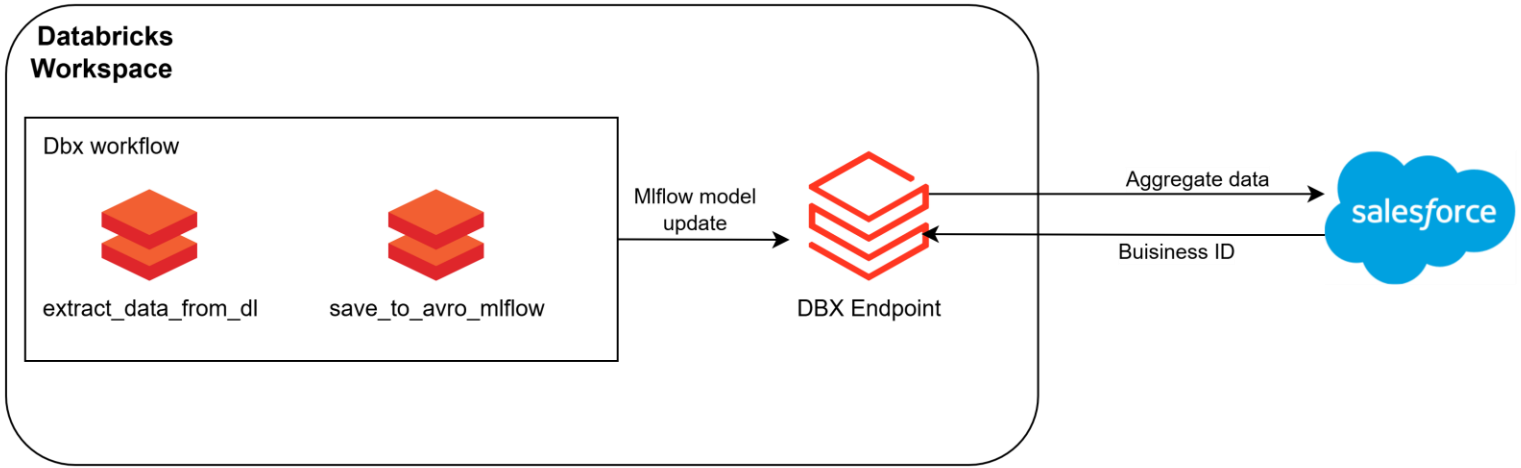
Ważne, aby dane, które są wstrzykiwane do promptu były tez widoczne dla agenta.

Do tego celu dobrze się nada REST API, dla naszego projektu stworzyliśmy Databricks Endpoint w oparciu o dane z data lake.



Hapag-Lloyd

Procesowanie I serwowanie danych do Salesforce



- Procesowanie pełnych danych
- Szybki setup – databricks zapewnia bezpieczne połaczenie, zarządzanie infrastrukturą

Wprowadzanie danych do promptu

Wprowadzanie danych do AI można zrobić na wiele sposobów, my zaczęliśmy od prostego rozwiązania z wprowadzaniem zmiennych w szablonie promptu:

Save Copy

```
1  Zweryfikuj dane klienta dotyczące przypadku {case_number}.
2  Poniższe dane powinny być sprawdzone:
3  - Czy adres klienta jest poprawny? {is_address_correct}
4  - Czy dane kontaktowe klienta są aktualne? {are_contact_details_up_to_date}
5  - Czy zamówienie zostało zrealizowane? {is_order_fulfilled}
6
7  Na podstawie powyższych danych, podejmij decyzję zgodnie z procedurą SOP.
8  Jeśli {is_address_correct} jest fałszem, skieruj klienta do działu adresowego.
9  Jeśli {are_contact_details_up_to_date} jest fałszem, poproś klienta o aktualizację danych kontaktowyc
10 Jeśli {is_order_fulfilled} jest prawdą, potwierdź zrealizowanie zamówienia.
11
12 Odpowiedź powinna zawierać:
13 - Potwierdzenie lub odrzucenie weryfikacji danych
14 - Ewentualne dodatkowe instrukcje dla klienta
```

Dedykowane prompty

Z powodu tego, że SOP-y są często dość długie, odpowiedź byłaby mniej dokładna, gdyby składała się z więcej niż jednego scenariusza biznesowego

Przed zastosowaniem konkretnego promptu dla danego scenariusza biznesowego, musimy go najpierw zidentyfikować. W naszym przypadku istnieje osobny klasyfikator, a możemy także dopasowywać do pewnych słów kluczowych

Poprawa evaluation framework

Dedykowane prompty ewaluacyjne

Ewaluacja odpowiedzi również musi być wykonywana biorąc pod uwagę SOPy

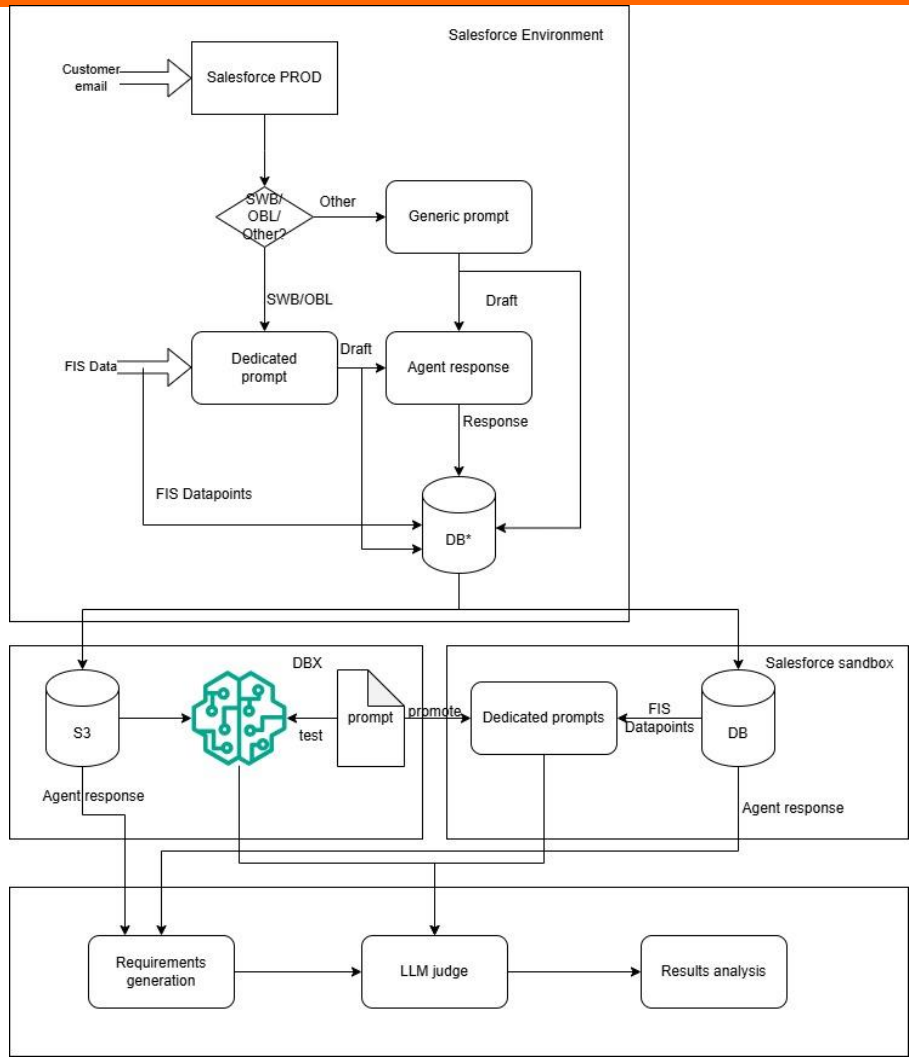
Przykładowe punkty ewaluacyjne w prompcie:

- Czy odpowiedzi są zgodne co do intencji (aprobata/odmowa)
- Czy w odpowiedzi są wymienione powody (odmowy)
- Czy w odpowiedzi są dodane instrukcje
- Czy odpowiedź jest uprzejma

Evaluation framework 2.0

Nowe evaluation framework po wnioskach:

- Salesforce endpoint
- Dedykowane prompty
- Zapisywanie danych w trakcie odpowiedzi



Resultaty I nastepne kroki

Jeden scenariusz biznesowy

Tworzenie I aktualizowanie SOPow to wymagający process

Dokładność podpowiedzi poprawiona z ~60% do 80%, celujemy w 90-95%.

Konkluzje

Protokół ewaluacji - posiłkuj się ankietami i ground truth do tworzenia metryk

Regularne audyty - Human in the loop aby zachować wiarygodność

Walidacja per task - w przypadku naszego projektu per scenariusz biznesowy

Thank you

for your attention!