Kacper Wikieł

AI / Machine Learning Engineer | GenAI, RAG, Computer Vision, MLOps

Warszawa | k.wikiel@gmail.com | +48 512 648 623 | LinkedIn | kacperwikiel.com

Profil zawodowy

AI / ML Engineer z doświadczeniem w budowie systemów GenAI od koncepcji i eksperymentu po wdrożenie, monitoring i optymalizację kosztów. Specjalizuję się w LLM, RAG, systemach agentowych, OCR/computer vision oraz pipeline'ach danych. Pracowałem przy rozwiązaniach enterprise dla PwC i Procter & Gamble, systemach diagnostyki przemysłowej dla operatora infrastruktury krytycznej oraz produktach własnych związanych z benchmarkami AI i przetwarzaniem dokumentów.

Łączę profil inżynierski z umiejętnością pracy z biznesem: przekładam wymagania biznesowe i regulacyjne na architekturę, metryki jakości, backlog techniczny i działające wdrożenia. Dbam o mierzalne efekty: koszt inferencji, latency, jakość odpowiedzi, stabilność pipeline'u, traceability i ograniczanie halucynacji.

Kluczowe obszary

LLM / RAG	Azure OpenAI, OpenAI API, LangChain, LlamaIndex, grounding, retrieval, chunking, compression, structured outputs
Ewaluacja GenAI	Golden datasety, testy regresyjne, LLM-as-a-judge, analiza trace'ów, metryki jakości, RCA, Langfuse
Systemy agentowe	Tool calling, workflow agentowe, orchestration, failover, ograniczanie halucynacji, prompt-as-code
Bezpieczeństwo LLM	Projektowanie pod PII, ryzyka prompt injection, audytowalność, compliance, walidacja odpowiedzi i źródeł
Computer Vision/OCR	OpenCV, PyTorch, OCR pipelines, ekstrakcja tabel i pól, przetwarzanie skanów i zdjęć dokumentów
MLOps / cloud	Python, FastAPI, Docker, CI/CD, GCP, Azure, monitoring, koszt/latency, pipeline'y produkcyjne

Wybrane osiągnięcia i projekty

- **CodeSOTA** – twórca portalu z benchmarkami AI/ML i praktycznymi porównaniami modeli; publikacje o OCR, modelach LLM i trendach rynkowych. Projekt przekroczył ok. 20k użytkowników i służy do wyboru modeli pod konkretne zadania, nie tylko do teoretycznych rankingów.
- **Benchmark OCR** – przetestowanie 47 rozwiązań OCR na realnych dokumentach: fakturach, receptach, zdjęciach z telefonu i trudnych skanach. Porównanie jakości, kosztu, odporności na błędy i użyteczności produkcyjnej.
- **Enterprise GenAI** – udział w budowie platformy GenAI dla globalnego klienta CPG, używanej w środowisku enterprise przez szeroką organizację; praca nad RAG, prompt engineeringiem, ingestion i integracją z procesami biznesowymi.
- **Cursor AI Jam** – założyciel hackathonu i inicjatywy społecznościowej wokół praktycznego użycia narzędzi AI w budowie produktów, automatyzacji i prototypowaniu.

Doświadczenie

Freelance / projekty własne Czerwiec 2024 – obecnie

ML / AI Engineer, konsultant GenAI

- Projektowanie i wdrażanie rozwiązań GenAI, RAG, OCR oraz computer vision dla klientów i własnych produktów; praca od discovery technicznego po działający prototyp lub produkcyjny serwis.
- Dobór modeli i architektury pod wymagania biznesowe: koszt, latency, jakość, prywatność danych, możliwość monitoringu i utrzymania.
- Budowa narzędzi demonstracyjnych, benchmarków, pipeline'ów przetwarzania dokumentów oraz raportów technicznych pomagających w decyzjach zakupowych i wdrożeniowych.

Vercly Grudzień 2025 – obecnie

AI Engineer / Generative AI Engineer

- Rozwój platformy AI Vercly PEP+RCA automatyzującej czynności wynikające z polskiego i europejskiego ustawodawstwa AML/CFT dla instytucji obowiązanych.
- Projektowanie workflow LLM dla analizy źródeł, dokumentów, powiązań i decyzji compliance: grounding, strukturyzowane outputy, kontrola halucynacji i audytowalność odpowiedzi.
- Debugowanie trace'ów, analiza failure modes i root cause analysis dla scenariuszy agentowych oraz pipeline'ów OCR/RAG.
- Projektowanie z uwzględnieniem ryzyk LLM: PII, prompt injection, jakość źródeł, powtarzalność wyników i zgodność z wymaganiami compliance.
- Praca w stacku Python, Azure, ML/LLM, OCR/RAG, MLOps; łączenie prototypowania z wymaganiami dotyczącymi jakości, zgodności i powtarzalności wyników.

PwC Czerwiec 2025 – Listopad 2025

Machine Learning Engineer (kontrakt)

- Projektowanie i implementacja rozwiązań RAG dla klientów enterprise: architektura, ingestion dokumentów, retrieval, ewaluacja i przygotowanie do wdrożenia.
- Integracja Langfuse do monitoringu, debugowania trace'ów i analizy jakości pipeline'ów LLM.
- Dobór modeli, strategii chunkingu, retrievalu, rerankingu i kompresji kontekstu pod konkretne use case'y biznesowe.
- Praca z interesariuszami nietechnicznymi: tłumaczenie ograniczeń LLM, ryzyk jakościowych i kosztów utrzymania na decyzje projektowe.

PERN S.A. Wrzesień 2024 – Maj 2025

Machine Learning / Computer Vision Engineer

- Budowa pipeline'ów przetwarzania sygnałów i obrazów dla diagnostyki przemysłowej oraz danych NDT.
- Praca nad detekcją elementów i anomalii w danych z inspekcji infrastruktury; łączenie klasycznego przetwarzania obrazów z modelami ML.
- Pipeline'y ETL dla danych sensorycznych, analiza jakości danych, przygotowanie cech i dashboardy dla inżynierów.
- Identyfikacja problemów jakościowych w danych pomiarowych i przekładanie obserwacji modelu na użyteczne wnioski inżynierskie.

DS STREAM Grudzień 2022 – Czerwiec 2024

Software Engineer (Generative AI) – projekty dla Procter & Gamble

- Udział w budowie produkcyjnej platformy GenAI dla globalnego klienta CPG, wspierającej wewnętrzne procesy NLP i wyszukiwania wiedzy w organizacji 30k+ użytkowników.
- Implementacja komponentów FastAPI, LangChain i LlamaIndex: ingestion dokumentów, parsowanie, indeksowanie, RAG i integracja z modelami OpenAI/Hugging Face.
- Prompt engineering dla wewnętrznych akceleratorów NLP; projektowanie promptów modułowych, szablonów odpowiedzi i mechanizmów walidacji wyników.
- Wdrożenia mikroserwisów w środowisku chmurowym: Docker, CI/CD, konfiguracja środowisk, modularność kodu i utrzymanie produkcyjnych pipeline'ów.
- Implementacja frameworka do równoległego uruchamiania zadań ML, rozdzielającego workflow na wiele procesów i jobów w celu zwiększenia przepustowości.

Igoria Trade Marzec 2021 – Listopad 2022

Python Developer

- Udział w projekcie R&D o wartości 9,3 mln PLN dotyczącym blockchainowych mechanizmów anti-fraud dla Open Banking / PSD2.
- Projektowanie i implementacja backendu w Pythonie: FastAPI, Pandas, SQLAlchemy, integracje API, przetwarzanie asynchroniczne i logika transakcyjna.
- Prace nad custom blockchainem opartym o Substrate/Rust, mechanizmami Proof of Existence oraz kryptograficznym kotwiczeniem dokumentów.
- Implementacja elementów KYC/AML i logiki compliance; przygotowywanie dokumentacji technicznej, analiz badawczych i opisów architektury.
- Analiza wydajności, identyfikacja wąskich gardeł oraz wsparcie integracji z aplikacjami webowymi w React.js.

Igoria Trade Styczeń 2021 – Marzec 2021

Software Engineer

- Rozwój komponentów backendowych i integracyjnych poprzedzający późniejszą pracę na roli Python Developer.
- Praca przy systemach transakcyjnych, integracjach API oraz automatyzacji przetwarzania danych.

Altkom Akademia Grudzień 2019 – Czerwiec 2020

Instruktor Python

- Prowadzenie ponad 40 godzin zajęć z Pythona: składnia, struktury danych, OOP, wyjątki, biblioteka standardowa, debugowanie i dobre praktyki.
- Tłumaczenie złożonych tematów technicznych osobom początkującym; przygotowanie ćwiczeń, laboratoriów i materiałów wspierających samodzielną pracę.

Trivial.co Marzec 2018 – Wrzesień 2018

Python Developer / Blockchain Data Engineer

- Projektowanie pipeline'u ETL do pobierania i normalizacji danych z blockchaina Ethereum, ze szczególnym uwzględnieniem kontraktów ERC-20.

- Rozwój mechanizmów wykrywania tokenów na podstawie bytecode'u smart contractów, zdarzeń on-chain i wzorców transakcyjnych.
- Indeksowanie zdarzeń typu Transfer/Approval oraz integracja danych zewnętrznych w celu wzbogacania bazy tokenów i analityki on-chain.

getline.in Styczeń 2015 – Luty 2018
Founder

- Uruchomienie projektu blockchain po Startup Weekend 2014 i rozwój produktu od zera bez finansowania VC.
- Praktyczne doświadczenie w budowie produktu, obsłudze użytkowników, szybkim prototypowaniu i iteracji w środowisku startupowym.

Umiejętności techniczne

Języki:	Python, SQL, podstawy Rust/Substrate, JavaScript/TypeScript
AI/ML:	Azure OpenAI, OpenAI API, PyTorch, scikit-learn, LangChain, LlamaIndex, Langfuse, Hugging Face
LLM evals:	Golden datasets, testy regresyjne, LLM-as-a-judge, metryki jakości, analiza trace'ów, prompt-as-code
Vision/OCR:	OpenCV, przetwarzanie obrazów, ekstrakcja danych z dokumentów, pipeline'y OCR, analiza jakości skanów
Backend:	FastAPI, SQLAlchemy, REST API, mikroserwisy, integracje API, przetwarzanie asynchroniczne
Dane:	Pandas, NumPy, ETL, pipeline'y danych, JSON/schema outputs, walidacja danych, raporty techniczne
Chmura/MLOps:	GCP, Azure, Docker, CI/CD, monitoring jakości, optymalizacja kosztów i opóźnień
Frontend:	React.js, budowa narzędzi demonstracyjnych i dashboardów technicznych

Edukacja i certyfikaty

PJATK – Informatyka, studia inżynierskie (2022–2026)
Uniwersytet Warszawski – Fizyka (2012–2014)
XIV LO im. Stanisława Staszica w Warszawie – profil matematyczno-fizyczny; IYPT 2012

Google Cloud Associate Engineer | Google Cloud Digital Leader | Azure AI Engineer

Języki

Polski – ojczysty | Angielski – zaawansowany

Wyrażam zgodę na przetwarzanie moich danych osobowych w celu rekrutacji zgodnie z RODO.