Opera Neon

Prosta Analiza Projektu Sztucznej Inteligencji "o3-pro"

Wprowadzenie: Czym jest ten projekt i czym jest sztuczna inteligencja?

Zanim zaczniemy, wyjaśnijmy w prosty sposób, czym jest **sztuczna inteligencja (AI)**. Wyobraź sobie bardzo mądrego robota w komputerze. Nie jest to prawdziwa osoba, ale program, który potrafi uczyć się, rozumieć ludzką mowę i wykonywać skomplikowane zadania, na przykład analizować dokumenty czy odpowiadać na pytania [19, 20].

Ten raport omawia projekt o nazwie **OpenAI o3-pro**. Jest to pomysł na bardzo zaawansowaną sztuczną inteligencję. Zgodnie z fałszywym dokumentem, który analizujemy ("Raport o modelu AI o3-pro.pdf"), ten model ma być ekspertem, na przykład w dziedzinie prawa. Miałby on pomagać w analizie prawniczej, a nawet działać jak "sędzia AI", który ocenia argumenty [1].

Naszym celem jest sprawdzenie, czy założenia tego projektu mają sens. Analizujemy, jak można go zrealizować, zwłaszcza w kontekście polskiego prawa, w sposób tani, bezpieczny i zrozumiały dla osób z trudnościami w poznawaniu.

Problem 1: Komputer nie ma dostępu do internetu. Jak dodać nowe informacje?

Na czym polega problem?

Model AI o nazwie `o3-pro` ma działać w specjalnym trybie zwanym "Batch API". Można to porównać do wysyłania paczki z zadaniami na pocztę. Wysyłasz zadania, a po jakimś czasie (nawet do 24 godzin) dostajesz gotowe odpowiedzi [1]. Wadą tego rozwiązania jest to, że w trakcie pracy komputer jest **odłączony od internetu**.

To duży problem, zwłaszcza w prawie. Przepisy zmieniają się bardzo szybko. Jeśli chcemy analizować sprawę, musimy znać prawo, które obowiązywało "wczoraj", a nie rok temu. Skoro AI sama nie może sprawdzić najnowszych ustaw w internecie, musimy znaleźć sposób, aby jej te informacje dostarczyć.

Jakie jest proponowane rozwiązanie?

Rozwiązaniem jest stworzenie "pomocnika" – osobnego programu, który będzie działał przed główną sztuczną inteligencją. Zadaniem tego pomocnika będzie:

1. Wyszukanie w internecie najnowszych przepisów prawnych, orzeczeń sądowych i innych potrzebnych informacji.
2. Zebranie tych informacji.
3. Sformatowanie ich w specjalny sposób (w formacie JSON, który jest jak uporządkowana lista dla komputera).
4. Dołączenie tej listy do głównego zadania (tzw. "User Prompt"), które wysyłamy do AI `o3-pro`.

Dzięki temu, gdy `o3-pro` zacznie analizę, będzie miała "ściągę" z najnowszymi informacjami, mimo że sama nie ma dostępu do internetu.

Jak zbudować takiego "pomocnika"?

Możemy użyć do tego narzędzi chmurowych, takich jak **Google Cloud Run** i **Google Firestore**.

- **Google Cloud Run**: To jak wynajęcie małego warsztatu w internecie, który uruchamia nasz program-pomocnika tylko wtedy, gdy jest potrzebny. Jest to tanie rozwiązanie, ponieważ płacimy tylko za faktyczny czas pracy [18].
- **Google Firestore**: To jak cyfrowy magazyn, w którym możemy przechowywać zebrane informacje i śledzić status naszych zadań.

Do samego wyszukiwania informacji w internecie możemy użyć specjalistycznych narzędzi, takich jak:

- **Perplexity.ai**: To bardzo dobry "pomocnik", ponieważ potrafi przeszukiwać internet i od razu porządkować znalezione informacje w strukturyzowany sposób, którego potrzebuje główna AI.
- **txyz.ai**: To narzędzie jest bardziej wyspecjalizowane w analizie prac naukowych. Posiada przydatne rozszerzenie do przeglądarki Chrome, które ułatwia analizę dokumentów PDF. Jednak w przypadku wyszukiwania najnowszych, codziennych zmian w prawie, Perplexity.ai może być bardziej odpowiednie.

Cały proces wyglądałby tak: Google Cloud Run uruchamia skrypt, który używa Perplexity.ai do znalezienia nowych ustaw, zapisuje je w Firestore, a następnie tworzy plik z zadaniem dla `o3-pro`, dołączając te nowe ustawy.

Problem 2: Ryzyko, że komputer "zmyśli" przepisy. Jak się przed tym chronić?

Czym są "halucynacje prawne"?

Sztuczna inteligencja, nawet najmądrzejsza, nie jest nieomylna. Czasami, gdy nie zna odpowiedzi, zamiast przyznać się do błędu, potrafi "zmyślić" fakty. Tworzy informacje, które wyglądają bardzo wiarygodnie – na przykład fałszywe numery ustaw, nieistniejące wyroki sądowe czy zmyślone argumenty prawne. W kontekście prawa jest to niezwykle niebezpieczne i nazywa się to **"halucynacjami prawnymi"** [2, 3].

Badania pokazują, że jest to powszechny problem. Nawet najlepsze modele AI mogą zmyślać informacje w ponad 69% przypadków, gdy są pytane o szczegółowe kwestie prawne [2]. To pokazuje, że **nigdy nie można ufać sztucznej inteligencji w 100%**, zwłaszcza w tak ważnej dziedzinie jak prawo [2].

Rozwiązanie: Pół-automatyczna praca z kontrolą człowieka

Najlepszym i jedynym bezpiecznym sposobem na walkę z halucynacjami jest wdrożenie systemu, w którym komputer i człowiek współpracują. Nazywa się to **"człowiek w pętli" (Human-in-the-Loop, HITL)** [4, 8].

- **Jak to działa?** Sztuczna inteligencja wykonuje całą wstępną, żmudną pracę: analizuje setki stron dokumentów, wyszukuje powiązane przepisy i przygotowuje wstępną analizę. Jednak ostateczna decyzja i weryfikacja zawsze należą do człowieka – w tym przypadku wykwalifikowanego prawnika [4, 6, 8]. Prawnik musi przeczytać wynik pracy AI, sprawdzić wszystkie źródła i potwierdzić, że analiza jest poprawna i nie zawiera zmyślonych informacji [4, 6].
- **Czy to jest opłacalne?** Tak. Chociaż zatrudnienie eksperta do weryfikacji kosztuje, koszt potencjalnego błędu prawnego (np. przegranej sprawy sądowej z powodu powołania się na fałszywy przepis) jest nieporównywalnie wyższy [4, 8]. Wprowadzenie kontroli człowieka to nie tylko dobra praktyka, ale konieczność, która minimalizuje ryzyko i sprawia, że system jest wiarygodny [4, 6, 8]. Jest to inwestycja w bezpieczeństwo i jakość.

Analiza narzędzi: Czy "darmowy" zestaw jest najlepszy?

W zapytaniu pojawiła się propozycja użycia "darmowego" zestawu narzędzi: modelu `gpt-oss-120b` z zabezpieczeniem `gpt-oss-120b-safeguard` oraz technologii `Pathway RAG`. Sprawdźmy, czy ten zestaw jest optymalny i faktycznie darmowy.

Modele AI: `gpt-oss-120b`

- **Czy jest darmowy?** I tak, i nie. Model `gpt-oss-120b` jest "open-weight", co oznacza, że można go pobrać i używać bez opłat licencyjnych [11, 13]. Jednak problemem jest jego uruchomienie. Jest to ogromny model (117B parametrów, 5.1B aktywnych parametrów) [13, 28], który do działania potrzebuje bardzo drogiego, potężnego komputera (karty graficznej z 80 GB pamięci, np. NVIDIA H100) [10, 11, 12, 13, 28]. Alternatywą jest płacenie za jego używanie za pośrednictwem specjalnych platform, co również generuje koszty [12, 28]. Zatem **model nie jest darmowy w utrzymaniu** [10, 12, 28].
- **Czym jest `gpt-oss-120b-safeguard`?** Taki model nie istnieje. Jest to koncepcja z fałszywego raportu, która opisuje wersję modelu z dodatkowymi zabezpieczeniami [1]. W rzeczywistości musimy sami zadbać o bezpieczeństwo, głównie poprzez wspomnianą wcześniej kontrolę człowieka.

Technologia RAG (Pathway)

- **Jak działa RAG?** RAG (Retrieval-Augmented Generation) to technologia, która pomaga zwalczać halucynacje [29]. Działa jak danie sztucznej inteligencji "otwartej książki" podczas egzaminu. Zamiast polegać tylko na swojej pamięci, AI najpierw przeszukuje dostarczone jej dokumenty (np. konkretne ustawy) i dopiero na ich podstawie buduje odpowiedź [29]. To znacznie zmniejsza ryzyko zmyślania [29]. Jednak badania pokazują, że nawet wiodące narzędzia prawne oparte na RAG nadal generują halucynacje w 17-33% przypadków [30].
- **Czym jest Pathway?** Pathway to darmowe (open-source) narzędzie, które pozwala zbudować taki system "otwartej książki" [25]. Jego wielką zaletą jest to, że potrafi działać w czasie rzeczywistym – jeśli dodamy nowy dokument do naszej bazy, Pathway automatycznie go uwzględni [25, 26]. To idealne rozwiązanie dla prawa, gdzie przepisy ciągle się zmieniają [25]. Pathway integruje się też z najnowszymi modelami, takimi jak GPT-4o [25, 26].
- **Czy jest darmowy?** Samo oprogramowanie Pathway jest darmowe, ale jego uruchomienie i połączenie z płatnymi modelami AI (jak `gpt-oss-120b` czy GPT-4o) generuje koszty [25].

Wniosek dotyczący optymalności

Proponowany zestaw **nie jest optymalny, jeśli zakładamy zerowe koszty**. Jednak sama koncepcja jest bardzo dobra. Połączenie modelu `gpt-oss-120b` z technologią `Pathway RAG` oraz **obowiązkową weryfikacją przez człowieka** tworzy potężny i stosunkowo bezpieczny system do analiz prawnych [4, 6, 8, 29]. Dla niewielkiej liczby analiz (kilkanaście dokumentów miesięcznie) jest to rozwiązanie warte rozważenia, pod warunkiem świadomości realnych kosztów i ryzyka [12, 16, 17].

Jak najtaniej uruchomić model AI?

Uruchomienie tak dużego modelu jak `gpt-oss-120b` dla niewielkiej liczby godzin w miesiącu wymaga sprytnego podejścia, aby uniknąć ogromnych kosztów. Oto najtańsze opcje:

1. **Platformy ze współdzielonymi GPU (np. RunPod, Vast.ai)**: Są to serwisy, które działają jak "Uber dla superkomputerów" [16, 17]. Zamiast kupować własny drogi sprzęt, można wynająć go na minuty lub godziny [16, 17]. Płaci się tylko za faktycznie wykorzystany czas [17]. Dla sporadycznego użytku jest to **zdecydowanie najtańsza opcja** [16, 17].
2. **Hugging Face (opcja bliska zeru)**: Hugging Face to duża platforma dla modeli AI. Oferują oni konto "PRO" za około 9 dolarów miesięcznie [14]. W ramach tego konta użytkownik otrzymuje dostęp do darmowego, limitowanego czasu na potężnych komputerach (w ramach usługi "ZeroGPU", np. Nvidia H200) [14]. Dla bardzo małego użytku może to być opcja niemal darmowa, choć trzeba liczyć się z kolejkami i ograniczeniami [14].
3. **Wielcy dostawcy chmurowi (Google, NVIDIA)**: Korzystanie z usług największych firm jest zazwyczaj najdroższe dla tego typu zadań [16]. Ich darmowe pakiety zazwyczaj nie obejmują tak potężnych kart graficznych, jakie są potrzebne do uruchomienia tego modelu [10, 12].

Podsumowując: aby uruchomić model blisko zerowym kosztem, najlepszym wyborem będzie konto PRO na Hugging Face [14]. Jeśli potrzebujemy większej elastyczności, platformy takie jak RunPod będą najbardziej opłacalne [16, 17].

Jak pokazać, w jaki sposób AI podjęła decyzję? Użycie programu Canva

W sprawach prawnych, zwłaszcza tych tworzących nowe standardy (precedensowych), sama odpowiedź to za mało. Musimy wiedzieć, dlaczego AI doszła do takiego, a nie innego wniosku. Jej proces myślowy musi być przejrzysty.

Ślad decyzyjny AI

Zgodnie z fałszywym raportem, model `o3-pro` został zaprojektowany tak, aby zapisywać swój "ślad decyzyjny" [1]. Oznacza to, że każdy krok jego rozumowania – jakie fakty wziął pod uwagę, jakie przepisy zastosował, na jakie wyroki się powołał – jest logowany. Mamy więc surowy, tekstowy zapis całego procesu myślowego AI.

Rola programu Canva

Canva to bardzo prosty i w dużej mierze darmowy program graficzny online. Służy do tworzenia plakatów, prezentacji, a także **schematów i diagramów przepływu (flowchartów)**.

Możemy wykorzystać Canvę, aby przetłumaczyć skomplikowany, tekstowy "ślad decyzyjny" AI na **prosty, wizualny schemat**. Taki schemat mógłby pokazywać krok po kroku:

- Pytanie początkowe.
- Pierwszy fakt wzięty pod uwagę.
- Zastosowany przepis prawa.
- Kolejny krok rozumowania.
- Ostateczny wniosek.

Taka wizualizacja ma ogromne zalety:

- **Zrozumiałość dla prawnika**: Prawnik weryfikujący pracę AI może szybko prześledzić jej logikę i wychwycić ewentualne błędy.
- **Uzasadnienie dla sądu**: Schemat może stanowić załącznik do pisma procesowego, w prosty sposób wyjaśniając skomplikowane rozumowanie.
- **Dostępność dla osób z trudnościami poznawczymi**: Obraz jest często łatwiejszy do zrozumienia niż długi, skomplikowany tekst [5]. Wizualna ścieżka decyzji pomaga zrozumieć proces nawet osobom, które mają trudności z czytaniem [5, 9].

Użycie Canvy do tego celu jest więc nie tylko możliwe, ale i wysoce wskazane. To tanie i skuteczne narzędzie do tworzenia "mapy myśli" sztucznej inteligencji.

Podsumowanie i Główne Wnioski

Analiza założeń projektu `o3-pro`, nawet jeśli oparta na fałszywym dokumencie, prowadzi do kilku bardzo ważnych, praktycznych wniosków dotyczących używania AI w prawie.

1. **Projekt jest technicznie możliwy do zrealizowania.** Można zbudować system, który pobiera aktualne dane z internetu i na ich podstawie przeprowadza analizy prawne, a następnie wizualizuje swój proces myślowy.
2. **Najważniejsza zasada: Zawsze sprawdzaj pracę komputera.** Sztuczna inteligencja jest potężnym narzędziem, ale nie jest nieomylna i potrafi "zmyślać" [2, 3]. Każdy wynik jej pracy w dziedzinie prawa musi być ostatecznie zweryfikowany przez wykwalifikowanego człowieka [2, 4, 6, 8, 30]. To absolutna podstawa bezpieczeństwa [4, 6, 8, 30].
3. **"Darmowe" nie znaczy za darmo.** Choć wiele narzędzi (modele AI, oprogramowanie) jest dostępnych na otwartych licencjach [11, 13], ich uruchomienie i utrzymanie zawsze wiąże się z kosztami (sprzęt, energia, płatne usługi) [10, 12, 16, 28].
4. **Można to zrobić tanio.** Dla niewielkiego obciążenia (kilkanaście dokumentów miesięcznie) istnieją bardzo tanie, a nawet niemal darmowe sposoby na uruchomienie

zaawansowanych modeli AI, korzystając z platform współdzielących moc obliczeniową [14, 16, 17].

5. **Upraszczanie i wizualizacja są kluczowe, ale i ryzykowne.** Tworzenie treści w prostym języku (zgodnie ze standardami takimi jak ETR) [5, 9] oraz wizualizowanie skomplikowanych procesów jest niezwykle ważne dla dostępności [5, 9]. Należy jednak pamiętać, że sam proces upraszczania przez AI również może prowadzić do błędów i utraty sensu [32]. Dlatego uproszczone teksty i schematy także muszą być zweryfikowane przez człowieka [32].

Sources

[1] user_attachment_openai_o3_pro_report: user_attachment_user_attachment_openai_o3_pro_report

[2] Hallucinating Law: Legal Mistakes with Large Language ...: https://hai.stanford.edu/news/hallucinating-law-legal-mistakes-large-language-models-are-pervasive

[3] Zarządzanie ryzykiem utraty reputacji związanym z ...: https://www.ranktracker.com/pl/blog/managing-reputation-risks-hallucinated-content/

[4] What Is Human In The Loop (HITL)?: https://www.ibm.com/think/topics/human-in-the-loop

[5] Co to jest tekst ETR -: http://etr.edu.pl/p,1,co-to-jest-tekst-etr

[6] Semi-Automated Labeling: AI + Human Input: https://keymakr.com/blog/semi-automated-labeling-combining-human-judgment-with-machine-speed/

[7] Co to jest Human-in-the-loop (HITL)?: https://www.unite.ai/pl/what-is-human-in-the-loop-hitl/

[8] Upowszechnianie prostego języka. Zalecenie Szefa Służby ...: https://www.gov.pl/attachment/d3566d11-1823-42ee-9a17-6d860b1ad172

[9] Essential GPT OSS 120B Hardware Requirements for Optimal ...: https://www.cognativ.com/blogs/post/essential-gpt-oss-120b-hardware-requirements-for-effective-deployment/332

[10] Introducing gpt-oss: https://openai.com/index/introducing-gpt-oss/

[11] GPT-OSS: Specs, Setup, and Self-Hosting Guide - Semaphore: https://semaphore.io/blog/gpt-oss

[12] openai/gpt-oss-120b: https://huggingface.co/openai/gpt-oss-120b

[13] Pricing: https://huggingface.co/pricing

[14] Pricing and Billing: [https://huggingface.co/pricing_2]

[15] Cheapest Cloud platforms for your LLMS (ranked) in 2025: https://code-b.dev/blog/cloud-platforms-for-fine-tuning-llms

[16] Top 12 Cloud GPU Providers for AI and Machine Learning ...: https://www.runpod.io/articles/guides/top-cloud-gpu-providers

[17] Cloud Run pricing: https://cloud.google.com/run/pricing

[18] Jak działa sztuczna inteligencja? Proste wyjaśnienie czym jest ...: https://komputerowapasja.pl/jak-dziala-sztuczna-inteligencja-proste-wyjasnienie-czym-jest-ta-technologia-i-jakie-sa-jej-glowne-zastosowania

[19] Sztuczna inteligencja: czym jest i jak działa: https://wiadomosci.onet.pl/sztuczna-inteligencja-czym-jest-i-jak-dziala/vmxl4sg

[20] LLM – co to jest? Jak działają duże modele językowe?: https://pja.edu.pl/llm-co-to-jest-jak-dzialaja-duze-modele-jezykowe/

[21] Co to jest duży model językowy (LLM)?: https://www.ovhcloud.com/pl/learn/what-is-large-language-model/

[22] Prosty język - Serwis Służby Cywilnej: https://www.gov.pl/web/sluzbacywilna/prosty-jezyk

[23] Polskie i unijne normy dotyczące prostego języka: https://lexperts.pl/prosty-jezyk/

[24] Create your own RAG - Pathway: https://pathway.com/developers/user-guide/llm-xpack/llm-app-pathway

[25] Multimodal RAG for PDFs with Text, Images, and Charts: https://pathway.com/developers/templates/rag/multimodal-rag

[26] Adaptive RAG: cut your LLM costs without sacrificing accuracy: https://pathway.com/developers/templates/rag/adaptive-rag

[27] GPT OSS 120B: Pricing, Context Window, Benchmarks, ...: https://llm-stats.com/models/gpt-oss-120b

[28] Intro to retrieval-augmented generation (RAG) in legal tech: https://legal.thomsonreuters.com/blog/retrieval-augmented-generation-in-legal-tech/

[29] Hallucination-Free? Assessing the Reliability of Leading AI ...: https://dho.stanford.edu/wp-content/uploads/Legal_RAG_Hallucinations.pdf

[30] Legal AI Benchmarking: Evaluating Long Context ...: https://www.thomsonreuters.com/en-us/posts/innovation/legal-ai-benchmarking-evaluating-long-context-performance-for-llms/

[31] Guest Post - The Accessibility Illusion: When AI Simplification ...: https://scholarlykitchen.sspnet.org/2025/07/22/guest-post-the-accessibility-illusion-when-ai-simplification-fails-the-users-with-cognitive-disabilities/