

## Obszar biznesowy

- 1. Macie świetnie rozpisaną drogę od PoC do pełnego wdrożenia – jaki był najbardziej zaskakujący, nieoczywisty use case Bielika lub Sójki, który pojawił się dopiero w trakcie pracy z użytkownikami i zmienił wasze myślenie o tych narzędziach?**

Najbardziej zaskoczyło nas to, co wydarzyło się poza językiem polskim. Bielik był budowany jako model zoptymalizowany pod polszczyznę, a tymczasem okazało się, że w benchmarkach bił niemieckie modele... w języku niemieckim i to jeszcze zanim wypuściliśmy wersję 3.0, w której celowo pracowaliśmy nad jakością dla języków obcych. To kompletnie zmieniło naszą perspektywę na to, czym Bielik może być, czyli nie tylko polskim LLM-em, ale silnikiem, który dzięki głębokiej specjalizacji językowej zyskuje przewagi w miejscach, których nikt nie planował.

- 2. Gdzie mogę znaleźć, jakie funkcjonalności posiada obecnie Bielik?**

Najlepsze źródła do sprawdzenia aktualnych funkcjonalności Bielika:

- [bielik.ai](https://bielik.ai) - strona główna projektu z przeglądem ekosystemu: rodzina modeli, Sójka, licencja, aktualności.
- [huggingface.co/speakleash](https://huggingface.co/speakleash) - wszystkie wersje modeli z kartami technicznymi opisującymi szczegółowo możliwości każdej wersji.
- [chat.bielik.ai](https://chat.bielik.ai) - interfejs webowy z trybami: standardowym, „Rozumuj”, „Wyszukaj i Rozumuj” (z dostępem do internetu) oraz „Chroń” (filtrowanie przez Sójkę). Najlepszy do testowania Bielika.

- 3. Jaki jest optymalny pierwszy krok dla małej lub średniej firmy, która nigdy wcześniej nie stosowała rozwiązań AI? Od czego zacząć wdrażanie rozwiązań opartych na sztucznej inteligencji?**

Jeśli nie korzystaliście jeszcze z żadnych narzędzi opartych na AI, zdecydowanie zachęcamy do włączenia ChatGPT, Bielika, Claude'a czy Gemini do codziennej pracy: pisanie maili, streszczanie dokumentów, burza mózgów. Jedno zastrzeżenie: nie wrzucajcie do nich plików poufnych ani zawierających dane wrażliwe. Kolejny krok to wybór jednego konkretnego, powtarzalnego procesu, który dziś zjada czas Waszego zespołu, np. przeszukiwanie wewnętrznych dokumentów (Document Intelligence & Knowledge Bot), weryfikację klauzul w umowach (Asystent Prawny) albo ręczne sortowanie i wyciąganie informacji z maili (Inteligentna analiza korespondencji). Zróbcie z tego mały PoC na realnych danych i zmierzcie efekt: ile czasu zaoszczędziliście, ile błędów uniknęliście. To daje twarde argumenty do dalszych kroków.

- 4. Jesteśmy mikro firmą (5 osób), 3 lata na rynku. Największe wyzwanie mamy z obszarami związanymi z podatkami, compliance, procedurami, innymi przepisami. Mimo korzystania z usług specjalistów z tych dziedzin - w dalszym ciągu natykamy się na niezgodności, problemy, itp. Zastanawiam się czy Bielik może pomóc mi w tych obszarach. Musiałby mieć dostęp do wielu dokumentów firmowych, danych personalnych, wrażliwych i tajnych.**

To świetne pytanie i dokładnie ten scenariusz opisujemy w naszym use case Asystenta Prawnego. Sam Bielik jest w stanie pomóc Wam z analizą dokumentów, ale prawdziwie dobre rezultaty osiąga się mając pełne dedykowane rozwiązanie oparte modelu językowym, takie jak np. nasz Asystent Działu Prawnego. Kluczowa przewaga takiego podejścia: Bielik działa lokalnie, więc Wasze dokumenty firmowe, dane personalne i wrażliwe nie opuszczają Waszej infrastruktury. Bardzo podobne wdrożenie realizujemy m.in. dla CUI Wrocław, gdzie system codziennie analizuje umowy z kontrahentami. Analiza porównawcza, wykrywanie niezgodności, sprawdzanie zgodności z regulacjami to obszary, w których modele językowe radzą sobie najlepiej. Dla mikrofirmy taki system mógłby działać jako pierwsza linia weryfikacji, wyłapując problemy zanim trafią do Waszych specjalistów zewnętrznych, co skraca liczbę iteracji i realnie obniża koszty obsługi prawnej.

- 5. Jakie są kluczowe różnice i przewagi polskiego modelu językowego Bielik AI w porównaniu do globalnych modeli (jak ChatGPT/Gemini) w kontekście specyficznych potrzeb i języka polskiego w biznesie?**

Bielik to model otwarty (Apache 2.0), więc możecie go postawić na własnej infrastrukturze i Wasze dane nigdy nie opuszczą firmy. Jest mniejszy od globalnych modeli, ale zoptymalizowany pod język polski, więc daje porównywalną jakość odpowiedzi w polszczyźnie przy znacznie niższych kosztach infrastruktury i przetwarzania. ChatGPT czy Gemini to świetne narzędzia ogólne, ale ich znajomość polskiego wynika z tłumaczenia anglojęzycznych danych. Bielik był trenowany na polskich danych od podstaw i rozumie nasze niuanse prawne, kulturowe i biznesowe.

- 6. W jakim modelu rozliczeniowym będzie można wykupić usługi Bielika Jakiej formy subskrypcji? Local/remote? Czy coś oprócz samego modelu wchodzi w skład tego projektu jakim jest Bielik?**

Sam Bielik jako model językowy jest całkowicie darmowy (open source, licencja Apache 2.0). Możecie go pobrać, uruchomić lokalnie i korzystać bez żadnych opłat za tokeny. To o czym rozmawialiśmy na webinarze to rozwiązania oparte na Bieliku, czyli gotowe systemy dostosowane do konkretnych procesów w firmie. Cena końcowa rozwiązania i sposób

rozliczenia docelowego rozwiązania zależy od use case'u i innych czynników, które chętnie omówimy z Państwem w większych konkretach indywidualnych sesjach.

### **7. Jaka zaleta Bielika wyróżnia go wśród dostępnych powszechnie LLMów jak ChatGPT czy Gemini?**

Bielik to model otwarty (Apache 2.0), więc możecie go postawić na własnej infrastrukturze i Wasze dane nigdy nie opuszczą firmy. Jest mniejszy od globalnych modeli, ale zoptymalizowany pod język polski, więc daje porównywalną jakość odpowiedzi w polszczyźnie przy znacznie niższych kosztach infrastruktury i przetwarzania. ChatGPT czy Gemini to świetne narzędzia ogólne, ale ich znajomość polskiego wynika z tłumaczenia anglojęzycznych danych. Bielik był trenowany na polskich danych od podstaw i rozumie nasze niuanse prawne, kulturowe i biznesowe.

### **8. Document Processor - jak wykrywane są struktury dokumentów, relacje, tokenizacja, budowa bazy wiedzy, benchmarki?**

Document Processor by Deviniti obsługuje pełny cykl życia dokumentu w czterech etapach. Na wejściu przyjmuje pliki w dowolnym formacie (PDF, DOCX, HTML, PPTX, XLS, TXT), a następnie przechodzi przez digitalizację (rozpoznawanie i strukturyzacja treści), wzbogacenie (uzupełnianie informacji, standaryzacja, ekstrakcja, rozpoznawanie relacji między elementami), analizę (klasyfikacja, podsumowania, rozumienie kontekstu, wykrywanie intencji) i dostarczenie wiedzy w formatach gotowych do integracji z AI. Tak przygotowana baza wiedzy zasila potem konkretne rozwiązania: chatboty i knowledge boty (np. sprzedażowe, compliance, HR), audyty zgodności i analizę ryzyk w dokumentach prawnych, inteligentną analizę załączników w komunikacji z klientem czy transformację prezentacji sprzedażowych w asystenta wiedzy. Szczegóły techniczne dotyczące tokenizacji, wykrywania struktur i budowy bazy wiedzy stanowią know-how naszego rozwiązania i nie są publicznie udostępniane.

## **Obszar Techniczny**

### **1. Czy uda się zastosować model jeśli pracuję w środowisku polskim (polscy kontrahenci) i międzynarodowym (kontrahenci kontaktujący się w języku angielskim)**

Tak, jak najbardziej. Bielik 11B v3 został zaprojektowany jako model wielojęzyczny, trenowany na korpusie obejmującym 32 języki europejskie (w tym angielski) z ponad 1,1 biliona tokenów. W benchmarkach wielojęzycznych (EuroEval, Belebele, INCLUDE) Bielik wyprzedza modele o 2 do 6 razy większej liczbie parametrów, a w testach tłumaczeniowych FLORES osiąga wyniki znacząco lepsze od poprzednich wersji, zarówno w kierunku z polskiego na angielski, jak i odwrotnie. W praktyce oznacza to, że jedno

rozwiązanie oparte na Bieliku obsłuży Wam zarówno polskich, jak i anglojęzycznych kontrahentów bez konieczności utrzymywania dwóch osobnych modeli.

## **2. Jak postawić lokalne środowisko odcięte od sieci zewnętrznej, tak aby mieć możliwość "uczyć" na bazie własnych materiałów?**

To jeden z najczęstszych scenariuszy, na które Bielik odpowiada idealnie, właśnie dlatego, że jest modelem open source, który można uruchomić w pełni lokalnie, bez połączenia z internetem. W praktyce potrzebujecie stacji roboczej z kartą graficzną o minimum 24 GB VRAM (np. Dell Pro Max z GB10, koszt ok. 16-20 tys. zł netto), na której Bielik 11B działa komfortowo. Na takiej maszynie instalujecie model, podłączacie go do swoich dokumentów przez architekturę RAG (Retrieval Augmented Generation) i całość pracuje w zamkniętej sieci. Warto podkreślić, że „uczenie na własnych materiałach” w większości przypadków nie wymaga przetrenowywania modelu od zera. Wystarczy dobrze zaprojektowana baza wiedzy i warstwa RAG, która pozwala Bielikowi odpowiadać na pytania w kontekście Waszych dokumentów. Dokładnie tak działają nasze wdrożenia w bankach i firmach ubezpieczeniowych, gdzie dane wrażliwe nie mogą opuścić infrastruktury organizacji.

## **3. AI Compliance Navigator - jaką mamy ochronę przed halucynacją na temat reguł?**

To kluczowe pytanie i traktujemy je bardzo poważnie, szczególnie w kontekście compliance, gdzie halucynacja modelu może mieć realne konsekwencje biznesowe i prawne. Samo narzędzie to jednak dopiero początek. Główna wartość, którą dostarczamy, to praca customowa nad jakością i dopasowaniem rozwiązania do wysokiej trafności odpowiedzi w konkretnym zastosowaniu. W praktyce oznacza to, że na początku każdego projektu wspólnie z klientem opracowujemy KPI, które chcemy razem osiągnąć, np. procent poprawnych klasyfikacji, odsetek odpowiedzi popartych źródłem czy czas weryfikacji dokumentu. Następnie iteracyjnie dostrajamy model, bazę wiedzy i logikę walidacji, aż osiągniemy ustalone progi jakości. Dodatkowo architektura RAG, na której opieramy nasze rozwiązania, wymusza na modelu odpowiadanie w oparciu o konkretne dokumenty źródłowe, co znacząco ogranicza ryzyko halucynacji i pozwala użytkownikowi zweryfikować każdą odpowiedź.

## **4. Uruchomiłem Bielik-11B na karcie 4 GB i odpowiedź zajmuje 2 minuty. W co muszę zainwestować, żeby osiągnąć wydajność pozwalającą na wdrożenie tego modelu w firmie?**

4 GB VRAM to zdecydowanie za mało na Bielika 11B, stąd te 2 minuty na odpowiedź. Model musi zmieścić się w pamięci karty graficznej, a przy 4 GB system przelewa obliczenia na CPU i RAM, co zabija wydajność. Dla komfortowej pracy z Bielikiem 11B w

firmie potrzebuje minimum 24 GB VRAM. Ciekawą opcją jest Dell Pro Max z chipem NVIDIA GB10 Grace Blackwell, kompaktowa stacja (waży niecałe 1,5 kg) oferująca 128 GB zunifikowanej pamięci i do 1 petafropa wydajności AI. To rozwiązanie projektowane wprost pod lokalne LLM-y, działa na NVIDIA DGX OS z pełnym stosem narzędzi AI preinstalowanym od razu. Cena startowa to ok. 4000 USD (ok. 16 tys. zł netto), w lepszej konfiguracji z większym dyskiem ok. 25 tys. zł. Co istotne, Dell i NVIDIA współpracują nad rozwiązaniem air-gapped dla klientów wymagających pełnej izolacji od sieci, więc ten sprzęt jest projektowany dokładnie pod scenariusz, o który Państwo pytacie. Zapraszamy do kontaktu, jeśli jesteście Państwo zainteresowani naszymi rozwiązaniami na tego typu urządzenia.

**5. Jakie są najlepsze praktyki w zakresie promptowania (czyli pisania poleceń) dla modeli AI, aby uzyskać precyzyjne, biznesowe odpowiedzi, na przykład w celu stworzenia scenariuszy pracy Agentów AI?**

Temat jest szeroki, ale najskuteczniejsza rada, którą możemy dać, to struktura promptu w czterech elementach.

ROLA: kim ma być AI? Np. „Jesteś doświadczonym doradcą finansowym dla małych firm”.

KONTEKST: czego dotyczy zapytanie? Np. „Klient prowadzący jednoosobową działalność pyta o optymalizację kosztów przy przejściu na spółkę z o.o.”.

INSTRUKCJA: co dokładnie AI ma zrobić? Np. „Wyjaśnij krok po kroku, jakie koszty się zmienią i na co zwrócić uwagę”.

FORMAT: w jakiej formie ma być odpowiedź? Np. „Maksymalnie 5 punktów, prosty język zrozumiały dla osoby bez wykształcenia prawniczego, ze wskazaniem źródeł”.

Im precyzyjniej opisze te cztery elementy, tym bardziej przewidywalne i biznesowo użyteczne będą odpowiedzi, a to samo podejście stosujemy projektując scenariusze pracy dla Agentów AI w naszych wdrożeniach.

**6. Jak zmniejszyć halucynacje modelu językowego? Czy zastosowanie w prompcie informacji aby model bazował na istniejących dokumentach bądź rzeczywistych informacjach jest wystarczające?**

Sam prompt „odpowiadaj tylko na podstawie dokumentów” to dobry początek, ale zdecydowanie nie wystarczy. Model językowy z natury generuje tekst statystycznie prawdopodobny, a nie faktograficznie pewny, więc sama instrukcja nie eliminuje ryzyka halucynacji. W praktyce stosujemy kilka warstw zabezpieczeń. Pierwsza to architektura RAG (Retrieval Augmented Generation), która wymusza na modelu odpowiadanie wyłącznie w oparciu o pobrane fragmenty konkretnych dokumentów, a nie wiedzę ogólną.

Druga to walidacja odpowiedzi, czyli mechanizmy sprawdzające, czy odpowiedź faktycznie wynika z podanych źródeł i czy model jest w stanie wskazać dokładny fragment, na którym się oparł. Trzecia to Sójka (Bielik Guard), która działa jako warstwa bezpieczeństwa filtrująca treści niepożądane. Czwarta, często niedoceniana, to iteracyjna praca nad jakością na realnych danych klienta, mierzenie trafności odpowiedzi i systematyczne dostrajanie całego pipeline'u. Żadne pojedyncze rozwiązanie nie eliminuje halucynacji w 100%, ale połączenie tych warstw pozwala sprowadzić ryzyko do akceptowalnego poziomu biznesowego.

### **7. Jaki komputer najlepiej kupić do lokalnego uruchamiania Bielika i ile mniej więcej kosztuje taka optymalna konfiguracja?**

Zależy od skali. Dla pojedynczego użytkownika lub małego zespołu wystarczy komputer z kartą graficzną o minimum 24 GB VRAM (np. RTX 4090), co zamyka się w kwocie ok. 15 tys. zł za gotowy zestaw. Dla bardziej wymagających zastosowań firmowych rekomendujemy Dell Pro Max z chipem NVIDIA GB10 Grace Blackwell, kompaktową stację z 128 GB zunifikowanej pamięci, pełnym stosem narzędzi AI i wydajnością do 1 petafropa, w cenie od ok. 16 do 25 tys. zł w zależności od konfiguracji. Oba warianty pozwalają uruchomić Bielika 11B lokalnie, w pełni odciętego od sieci, bez wysyłania danych na zewnętrzne serwery.

### **8. Reprezentuję firmę MŚP w obszarze IT. Jednym z obszarów naszego działania jest wdrażanie automatyzacji - stąd pytanie - czy agentów opartych na Bieliku AI można wdrażać własnym sumptem?**

Oczywiście, Bielik jest modelem w pełni otwartym (licencja Apache 2.0), więc możecie go pobrać, uruchomić na własnej infrastrukturze i budować na nim agentów AI bez żadnych opłat licencyjnych czy kosztów tokenów. Jeśli posiadacie własny zespół deweloperski i infrastrukturę z odpowiednią kartą graficzną, nic nie stoi na przeszkodzie, żeby wdrażać rozwiązania samodzielnie. Wersje 3.0 Bielika wspierają już function calling, czyli umiejętność korzystania z zewnętrznych narzędzi, co jest fundamentem budowy systemów agentowych. Całą dokumentację, modele i przykłady znajdziecie na [huggingface.co/speakleash](https://huggingface.co/speakleash). Jeśli natomiast chcielibyście przyspieszyć ten proces lub potrzebujecie wsparcia w zaprojektowaniu architektury i osiągnięciu odpowiedniej jakości odpowiedzi dla konkretnego use case'u, zapraszamy do rozmowy.

### **9. Gdzie w chmurze najlepiej hostować bielika, chodzi mi o ceny i bezpieczeństwo?**

Jeśli nie chcecie Państwo inwestować we własną infrastrukturę, jedną z naszych rekomendacji jest Beyond.pl. Wspólnie z Beyond.pl Deviniti dostarcza bezpieczne i lokalne rozwiązania dla naszych klientów. Nie mniej ważne jest, że wspólnie z Bielik.AI

Beyond.pl powołał inicjatywę „Gniazdo”, czyli gotowe środowisko do hostowania rozwiązań opartych na Bielik w modelu chmurowym. Dane przetwarzane są w Polsce, w data center w Poznaniu, pod jurysdykcją UE, więc kwestia zgodności regulacyjnej i bezpieczeństwa danych jest rozwiązana. Firma nie musi budować własnego zaplecza GPU, a jednocześnie zachowuje pełną kontrolę nad tym, co dzieje się z jej danymi.

## **Obszar danych**

### **1. Czy można wprowadzać do Bielik AI dokumenty medyczne z danymi pacjentów bez obaw o tajemnice?**

Tak, pod warunkiem że Bielik działa lokalnie na Państwa własnej infrastrukturze, dane pacjentów nie opuszczają Waszych serwerów, co rozwiązuje kwestię bezpieczeństwa technicznego. Natomiast sama kwestia prawna jest osobnym tematem. Dane dotyczące zdrowia to w RODO kategoria szczególna (art. 9), a ich przetwarzanie w celach diagnostyki, leczenia czy zarządzania opieką zdrowotną zwykle nie wymaga osobnej zgody pacjenta, bo opiera się na przestance prawnej. Jednak wykorzystanie narzędzia AI do analizy tych danych może stanowić dodatkowy cel przetwarzania, co wymaga odrębnej oceny. Rekomendujemy skonsultowanie tego ze swoim Inspektorem Ochrony Danych lub specjalistą ds. RODO, bo odpowiedź będzie zależeć od konkretnego scenariusza użycia. Technologicznie Bielik lokalny daje Wam najlepsze możliwe warunki do spełnienia wymogów, ale decyzja prawna leży po stronie placówki.

### **2. W jaki sposób MŚP powinny weryfikować poprawność i rzetelność informacji oraz danych generowanych przez modele AI, aby uniknąć błędów krytycznych dla działalności firmy?**

Kluczowa zasada: nigdy nie wdrażajcie AI „na produkcję” bez wcześniejszej walidacji na realnych danych. Zaczynajcie od PoC na ograniczonym zakresie, np. jednym procesie czy jednym typie dokumentów, i porównujcie odpowiedzi modelu z tym, jak te same zadania wykonują dziś Wasi ludzie. Ustalcie mierzalne kryteria jakości już na starcie: jaki procent poprawnych odpowiedzi jest dla Was akceptowalny, jakie błędy są krytyczne, a jakie dopuszczalne. Następnie iteracyjnie dostrajajcie rozwiązanie, bazę wiedzy i prompty, aż osiągniecie ustalone progi. Po wdrożeniu nie odpuszczajcie, regularnie przeglądajcie próbki odpowiedzi i zbierajcie feedback od użytkowników. Najlepiej myśleć o AI jak o nowym stażystce: czy zaufalibyście świeżo zatrudnionej osobie, że bez żadnego wdrożenia, nadzoru i pracy z nią nie popełni krytycznych błędów? My też nie, i dokładnie tak samo podchodzimy do modeli AI.

### **3. Czy wprowadzanie do modeli AI takich jak Bielik AI dokumentów z danymi osobowymi pacjenta i innymi danymi medycznymi takimi jak rozpoznania chorób byłoby zgodne z prawem? Czy stanowiłoby to naruszenie tajemnicy**

**medycznej? Jestem ciekawy czy można byłoby zgodnie z prawem użyć Bielik AI do syntezy informacji z wielu stron dokumentacji medycznej przedłożonej przez pacjenta. Czy jeśli byłby użyty w tym celu Bielik AI to gromadziłby takie dane i czy ktoś inny miałby do nich dostęp?**

To pytanie pojawiło się w kilku wariantach, więc odpowiadamy kompleksowo. Jeśli Bielik działa lokalnie na Państwa infrastrukturze, dane nie są nigdzie wysyłane, model nie „gromadzi” ich w żadnej zewnętrznej bazie i nikt poza Wami nie ma do nich dostępu. To nie jest chmura, to oprogramowanie działające na Waszym serwerze, tak samo jak każdy inny system IT w placówce. Synteza informacji z wielu stron dokumentacji medycznej to zresztą jeden z najlepszych use case'ów dla modeli językowych, bo robią to szybko i dokładnie. Natomiast kwestia zgodności z prawem i tajemnicą lekarską wymaga osobnej analizy. Dane medyczne to w RODO kategoria szczególna (art. 9), a ich przetwarzanie do celów diagnostyki i leczenia zwykle opiera się na przesłance prawnej, nie na zgodzie pacjenta. Jednak wykorzystanie AI jako dodatkowego narzędzia analitycznego może stanowić nowy cel przetwarzania, co wymaga oceny przez Inspektora Ochrony Danych w Państwa placówce. Technologicznie Bielik lokalny daje najlepsze możliwe warunki do zachowania tajemnicy medycznej, ale finalną decyzję prawną powinna podjąć placówka w konsultacji ze specjalistą ds. RODO. Ważną kwestią pozostaje rozróżnienie między technicznym wsparciem procesu analizy danych a diagnozą medyczną. O ile wykorzystanie AI do syntezy, strukturyzacji czy przeszukiwania dokumentacji medycznej stanowi dopuszczalne wsparcie pracy lekarza, o tyle wykorzystanie modelu do samodzielnego formułowania wniosków diagnostycznych lub terapeutycznych wymagałoby posiadania przez system odpowiednich certyfikacji prawnych właściwych dla wyrobów medycznych (zgodnie z rozporządzeniem MDR). Każda informacja wygenerowana przez AI musi zostać poddana krytycznej weryfikacji przez wykwalifikowany personel medyczny, który zachowuje pełną i wyłączną odpowiedzialność za ostateczne decyzje diagnostyczne i proces leczenia pacjenta.