Wzór publicznego streszczenia treści treningowych dla modeli AI ogólnego przeznaczenia wymaganego na podstawie art. 53 ust. 1 lit. d) rozporządzenia (UE) 2024/1689 (akt w sprawie sztucznej inteligencji)

Wersja streszczenia:	*1.0*
Ostatnia aktualizacja:	*11/05/2026*

1. Informacje ogólne

1.1. Identyfikacja dostawcy

Dane osobowe i dane kontaktowe dostawcy:	Ministerstwo Cyfryzacji ul. Królewska 27 00-060 Warszawa Tel.: +48 22 250 01 10 Adres e-mail: kancelaria@cyfra.gov.pl

1.2. Identyfikacja modelu

Wersje nazwy modelu:	• PLLuM-4B-chat-2512 • Llama-PLLuM-8B-chat-2512 • PLLuM-12B-chat-2512 • Llama-PLLuM-70B-chat-2512
Zależności modelu:	• PLLuM-4B-chat-2512 jest wynikiem wychowania (ang. *alignment*) modelu PLLuM-4B-instruct-2512, który jest wynikiem dostrojenia modelu PLLuM-4B-base-2512, który z kolei jest wynikiem adaptacji językowej (kontynuacji pretreningu) modelu fundamentalnego Gemma-3-4b-pt • Llama-PLLuM-8B-chat-2512 jest wynikiem wychowania (ang. *alignment*) modelu Llama-PLLuM-8B-instruct-2512, który jest wynikiem dostrojenia modelu Llama-PLLuM-8B-base-2512, który z kolei jest wynikiem adaptacji językowej (kontynuacji pretreningu) modelu fundamentalnego Llama-3.1-8B • PLLuM-12B-chat-2512 jest wynikiem wychowania (ang. *alignment*) modelu PLLuM-12B-instruct-2512, który jest wynikiem dostrojenia modelu PLLuM-12B-base-2512, który z kolei jest wynikiem adaptacji językowej (kontynuacji pretreningu) modelu fundamentalnego Mistral-Nemo-12B • Llama-PLLuM-70B-chat-2512 jest wynikiem wychowania (ang. *alignment*) modelu Llama-PLLuM-70B-instruct-2512, który jest wynikiem dostrojenia modelu Llama-PLLuM-70B-base-2508, który z kolei jest wynikiem adaptacji językowej (kontynuacji pretreningu) modelu fundamentalnego Llama3.1-70B
Data wprowadzenia modelu do obrotu w Unii:	*21/05/2026*

1.3 Formy, ogólna wielkość danych treningowych i inne cechy charakterystyczne

Forma	*Wielkość danych treningowych*	*Rodzaj treści*
x Tekst	☐ mniej niż 1 mld tokenów x 1 mld do 10 bln tokenów ☐ ponad 10 bln tokenów Alternatywnie należy określić przybliżoną wielkość w innej jednostce miary: _____	Działania ukierunkowane były na możliwie szerokie spektrum typów tekstów i obszarów tematycznych. Wyraźnie zwiększony był jedynie udział danych pochodzących z domeny urzędowej. Przykładowe typy tekstów wchodzące do danych treningowych: ● urzędowe ● naukowe ● media społecznościowe i teksty potoczne ● publikacje prasowe ● artystyczne/ retoryczne ● kody źródłowe

Najpóźniejsza data danych: nabycie/zgromadzenie na potrzeby trenowania modelu:

12/2025
Model nie jest stale trenowany z wykorzystaniem nowych lub dynamicznych danych po tej dacie.

Opis cech językowych ogólnych danych treningowych:

Dane treningowe obejmują głównie dane w języku polskim z udziałem danych w języku angielskim. Dane w języku angielskim stanowią ok. 20% danych polskich.

Inne istotne cechy ogólnych danych treningowych:

Dane treningowe obejmują teksty z różnorodnych domen i rejestrów języka polskiego z uwzględnieniem jego współczesnych norm, konwencji stylistycznych oraz aktualnego użycia w komunikacji pisemnej. Istotną cechą danych treningowych jest duży udział treści odzwierciedlających język formalny, w tym nomenklaturę urzędową, administracyjną oraz charakterystyczne struktury językowe stosowane w dokumentach oficjalnych, regulacjach oraz korespondencji i komunikacji instytucjonalnej.

2. Wykaz źródeł danych

2.1. Zbiory danych dostępne publicznie

Czy do trenowania modelu wykorzystano publicznie dostępne zbiory danych?

x tak ☐ nie

Jeżeli tak, należy określić formę(-y) treści objętych odpowiednimi zbiorami danych:

X tekst ☐obraz ☐ wideo ☐ dźwięk ☐ inna (należy określić)

Wykaz dużych publicznie dostępnych zbiorów danych:

CommonCrawl – archiwum CC-MAIN-2025-38 (https://data.commoncrawl.org/crawl-data/CC-MAIN-2025-38/index.html) z września 2025 – zostało użyte w części; zbiór przefiltrowano pod kątem języka polskiego oraz według wyników dodatkowej weryfikacji istnienia zastrzeżeń, przed dokonaniem eksploracji tekstów i danych oraz odrzucenia adresów URL i domen, według wykorzystanych środków ochrony przed treściami nielegalnymi.

Zrzut **polskojęzycznej Wikipedii** z dnia 2025-07-01 (plwiki, https://dumps.wikimedia.org/backup-index.html).

Zrzut **angielskojęzycznej Wikipedii** z dnia 2025-09-20 (enwiki, https://dumps.wikimedia.org/backup-index.html) został użyty w części; dodatkowo przefiltrowany za pomocą klasyfikatora kontekstu/kultury polskiej.

Wybrane podzbiory zbioru danych RedPajama-Data-1T (https://huggingface.co/datasets/togethercomputer/RedPajama-Data-1T – ostatnia aktualizacja zbioru danych w repozytorium HuggingFace: 17 czerwca 2024 r.):

- podzbiór **arxiv** został użyty w części; przefiltrowany według lat 2020+, do treningu wykorzystano wyłącznie artykuły dostępne na wybranych otwartych licencjach; przefiltrowane dane zostały dodatkowo ograniczone ilościowo
- podzbiór **github** został użyty w części, dane zostały ograniczone ilościowo
- podzbiór **stackexchange** został użyty w części; przefiltrowany według lat 2020+ i dodatkowo ograniczony ilościowo.

Podzbiór **FineWeb-Edu** (https://huggingface.co/datasets/HuggingFaceFW/fineweb-edu) sample-10BT (2013 – czerwiec 2025) został użyty w części; przefiltrowany według wyników dodatkowej weryfikacji istnienia zastrzeżeń, przed dokonaniem eksploracji tekstów i danych oraz odrzucenia adresów URL i domen, według wykorzystanych środków ochrony przed treściami nielegalnymi, dodatkowo ograniczony ilościowo.

Podzbiór **FineWeb** (https://huggingface.co/datasets/HuggingFaceFW/fineweb) sample-10BT (lato 2013 – czerwiec 2025) został użyty w części; przefiltrowany według wyników dodatkowej weryfikacji istnienia zastrzeżeń, przed dokonaniem eksploracji tekstów i danych oraz odrzucenia adresów URL i domen, według wykorzystanych środków ochrony przed treściami nielegalnymi, dodatkowo ograniczony ilościowo.

Ogólny opis innych publicznie dostępnych zbiorów danych niewymienionych powyżej:

Słowosieć (http://plwordnet.pwr.wroc.pl/wordnet/, 2005–2025) – zbiór danych użyty w części (wykorzystano jednostki leksykalne, ich definicje oraz przykłady użycia) i przekształcony w zasób tekstowy w typie słownikowym z odfiltrowanymi elementami potencjalnie wzmacniającymi tendencje modeli do generowania treści toksycznych.

Mniejsze zbiory danych – użyte w części zasoby korpusowe i leksykalne w języku polskim, o zróżnicowanym charakterze, w szczególności anotowane zasoby opracowane na przestrzeni w przybliżeniu ostatnich 20 lat przez twórców modeli PLLuM w ramach prac naukowych, skonwertowane do formatu instrukcyjnego. Zbiory weryfikowane, zależnie od przypadku, pod kątem permisywności licencji.

2.2 Prywatne niedostępne publicznie zbiory danych pozyskane od osób trzecich

2.2.1. Zbiory danych licencjonowane do celów handlowych przez podmioty uprawnione lub ich przedstawicieli

Czy została zawarta transakcyjna umowa licencyjna lub transakcyjne umowy licencyjne z podmiotami uprawnionymi lub ich przedstawicielami?	X tak ☐ nie
Jeżeli tak, należy określić formę(-y) treści objętych odpowiednimi zbiorami danych:	X tekst ☐ obraz ☐ wideo ☐ dźwięk ☐ inne

2.2.2. Prywatne zbiory danych uzyskane od innych osób trzecich

Czy pozyskano od osób trzecich prywatne zbiory danych, które nie są objęte licencjonowaniem, o którym mowa w sekcji 2.2.1, takie jak dane uzyskane od dostawców prywatnych baz danych lub od pośredników w zakresie danych?	X tak ☐ nie
Jeżeli tak, należy określić formę(-y) treści objętych odpowiednimi zbiorami danych:	X tekst ☐ obraz ☐ wideo ☐ dźwięk ☐ inna
Jeżeli takie prywatne zbiory danych uzyskane od innych osób trzecich są publicznie znane, należy je wskazać:	—
Ogólny opis nieznanych publicznie prywatnych zbiorów danych uzyskanych od osób trzecich	Wybrane elementy tekstowe z podzbioru interakcji użytkowników z modelami PLLuM za pośrednictwem publicznie dostępnego okna dialogowego PLLuM Chat. Zasoby dobrano i wykorzystano wyłącznie do analizy luk kompetencyjnych modeli i jako materiał poglądowy na potrzeby generowania zbiorów instrukcji do dostrajania i preferencji do wychowania modeli. Dane przekazane przez polskie urzędy centralne i samorządowe, obejmujące teksty w języku polskim przed i po uproszczeniu zgodnie z zasadami Prostego Języka.

2.3 Dane zebrane w procesach automatycznego przeszukania i pozyskania z internetu

Czy przez dostawcę lub w jego imieniu były wykorzystywane roboty indeksujące?	X tak ☐ nie

Jeżeli tak, należy podać nazwę (nazwy) robota indeksującego/jego identyfikator:	Użyto dedykowanych rozwiązań (narzędzi/ skryptów), które nie miały określonej nazwy formalnej.
Cele robota indeksującego (robotów indeksujących):	Celem narzędzi/ skryptów było pozyskanie tematycznie zróżnicowanych danych z publicznie dostępnych dokumentów według ustalonych rodzajów treści i źródeł.
Ogólny opis zachowania robota indeksującego:	Rozwiązania operowały na wcześniej wyselekcjonowanej liście adresów internetowych. Nie podejmowały prób omijania mechanizmów CAPTCHA, sekcji chronionych hasłem ani paywalli. Pomiędzy zapytaniami wprowadzono opóźnienia, aby uniknąć nadmiernego obciążania serwerów źródłowych.
Okres zbierania danych:	Od 03/2025 do 11/2025
Wyczerpujący opis rodzaju treści i źródeł internetowych, w odniesieniu do których zastosowano robota indeksującego:	Treści obejmują główne dokumenty urzędowe wyodrębnione z załączników (.pdf, .doc, .docx itp.) znajdujących się na polskich portalach organów władzy publicznej i jednostek podległych/nadzorowanych (gov.pl) i w zasobach Biuletynu Informacji Publicznej (BIP). Pozyskiwano wyłącznie treści w języku polskim. Źródła weryfikowano pod kątem zastrzeżeń związanych z eksploracją tekstów i danych.
Rodzaj formy:	X tekst ☐ obraz ☐ wideo ☐ dźwięk ☐ inna
Streszczenie najistotniejszych nazw domen, w odniesieniu do których zastosowano robota indeksującego:	Poszczególne subdomeny w domenie .gov.pl oraz domeny adresujące zasoby Biuletynu Informacji Publicznej (BIP).
Dodatkowe uwagi:	W sposób automatyczny pozyskiwano również treści z platformy Biblioteka Nauki (bibliotekanauki.pl), teksty z Dziennika Ustaw Rzeczypospolitej Polskiej i Monitora Polskiego oraz Dziennika Urzędowego Unii Europejskiej (eur-lex.europa.eu). Treści (artykuły i książki naukowe) z platformy Biblioteka Nauki oraz teksty z Dziennika Ustaw Rzeczypospolitej Polskiej i Monitora Polskiego były pozyskiwane w sposób automatyczny, ale bez użycia robotów indeksujących – w powyższych przypadkach wykorzystano oficjalne API tych serwisów. W przypadku Dziennika Urzędowego Unii Europejskiej skorzystano z opcji pobierania zrzutu danych dla posiadaczy kont UE (datadump.publications.europa.eu). Status tych danych sprawdzano cyklicznie, z zastosowaniem mechanizmu automatycznego, a kolejne dane pobierano już poprzez API.

2.4 Dane użytkownika

Czy do trenowania modelu wykorzystano dane pochodzące z interakcji użytkownika z modelem sztucznej inteligencji (np. prompty i dane wprowadzane przez użytkownika)?	☐ tak X nie

Czy do trenowania modelu wykorzystano dane zgromadzone w wyniku interakcji użytkowników z innymi usługami lub produktami dostawcy?	☐ tak X nie
Jeżeli tak, należy przedstawić ogólny opis usług lub produktów dostawcy, które wykorzystano do gromadzenia danych użytkownika:	—

2.5 Dane syntetyczne

Czy w celu trenowania modelu zostały stworzone przez dostawcę lub w jego imieniu dane syntetyczne wygenerowane przez sztuczną inteligencję?	X tak ☐ nie
Jeżeli tak, jaką formę mają dane syntetyczne:	X tekst ☐ obraz ☐ wideo ☐ dźwięk ☐ inna
Jeżeli tak, należy określić model lub modele AI ogólnego przeznaczenia wykorzystywane do generowania danych syntetycznych, jeżeli są one udostępnione na rynku:	*DeepSeek-V3 model (https://huggingface.co/deepseek-ai/DeepSeek-V3), nie ustalono udostępnienia streszczenia.* *Mixtral-8x22B-Instruct-v0.1 (https://huggingface.co/mistralai/Mixtral-8x22B-Instruct-v0.1), nie ustalono udostępnienia streszczenia.* *Qwen3-32B (https://huggingface.co/Qwen/Qwen3-32B), nie ustalono udostępnienia streszczenia.* *gpt-oss-20b (https://huggingface.co/openai/gpt-oss-20b), nie ustalono udostępnienia streszczenia.* *gpt-oss-120b (https://huggingface.co/openai/gpt-oss-120b), nie ustalono udostępnienia streszczenia.* *PLLuM-8x7B-chat (https://huggingface.co/CYFRAGOVPL/PLLuM-8x7B-chat), nie ustalono udostępnienia streszczenia.* *PLLuM-12B-chat (https://huggingface.co/CYFRAGOVPL/PLLuM-12B-chat), nie ustalono udostępnienia streszczenia.* *Bielik-11B-v2.2-Instruct (https://huggingface.co/speakleash/Bielik-11B-v2.2-Instruct), nie ustalono udostępnienia streszczenia.* *zephyr-orpo-141b-A35b-v0.1 (https://huggingface.co/HuggingFaceH4/zephyr-orpo-141b-A35b-v0.1), nie ustalono udostępnienia streszczenia.* *Mistral-7B-Instruct-v0.3 (https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3), nie ustalono udostępnienia streszczenia.*

Informacje na temat innych modeli sztucznej inteligencji, w tym własnych modeli sztucznej inteligencji dostawcy nie udostępnionych na rynku, wykorzystywanych do generowania danych syntetycznych w celu trenowania modelu, do którego to streszczenie ma zastosowanie:	Wcześniejsze wersje modeli z rodziny PLLuM, trenowane na podzbiorze danych opisanych w niniejszym dokumencie. Wykorzystywane głównie do generowania wstępnych instrukcji lub preferencji, które następnie podlegały ręcznej korekcie lub dostosowaniu przez anotatorów (ludzi).

2.6 Inne źródła danych

Czy do trenowania modelu wykorzystano inne źródła danych niż opisane w sekcjach 2.1–2.5?	X tak ☐ nie
Jeżeli tak, należy przedstawić opis tych danych oraz źródeł tych danych:	Zbiór instrukcji przeznaczonych do dostrajania modeli, opracowany przez anotatorów: - tworzone ręcznie od podstaw polecenia oraz odpowiadające im odpowiedzi; - tworzone ręcznie od podstaw polecenia oraz ręcznie skorygowane i dostosowywane odpowiedzi generowane przez modele językowe udostępniane na otwartych licencjach. Zbiór preferencji przeznaczonych do wychowywania modeli (ang. *alignment*), opracowany przez anotatorów: - tworzone ręcznie od podstaw polecenia oraz odpowiadające im pary odpowiedzi wraz z ręczną oceną ich jakości; - tworzone ręcznie od podstaw polecenia oraz ręcznie skorygowane i dostosowywane pary odpowiedzi generowane przez modele językowe udostępniane na otwartych licencjach. Materiały złożone z elementów udostępnianych na szerokich otwartych licencjach, elementów niepodlegających ochronie prawnoautorskiej lub w odniesieniu do których autorskie prawa majątkowe wygasły.

3. Aspekty przetwarzania danych

3.1. Respektowanie zastrzeżenia praw wynikających z wyjątku lub ograniczenia dotyczącego eksploracji tekstów i danych

Czy są Państwo sygnatariuszami kodeksu postępowania dotyczącego modeli AI ogólnego przeznaczenia, w którym przewidziano zobowiązania do przestrzegania zastrzeżeń praw wynikających z wyjątku lub ograniczenia TDM?	☐ Tak X Nie
Należy opisać środki wdrożone przed rozpoczęciem trenowania modelu w celu przestrzegania zastrzeżeń praw wynikających z wyjątku lub ograniczenia dotyczącego eksploracji tekstów i danych przed gromadzeniem danych i w jego trakcie, w tym protokołów i rozwiązań opt-out uznanych przez dostawcę lub, w stosownych przypadkach, przez osoby trzecie, od których pozyskano zbiory danych:	W trakcie opracowywania modelu wdrożono zróżnicowane środki mające na celu uwzględnienie dokonanych wyraźnie zastrzeżeń dotyczących TDM. Z uwagi na to, że znacząca część danych wykorzystanych na potrzeby pretrenowania pochodziła z zasobów publicznie udostępnionych w taki sposób, aby każdy mógł mieć do nich dostęp w miejscu i czasie przez siebie wybranym, podejmowane działania dotyczyły zastrzeżeń dokonanych w formacie przeznaczonym do odczytu maszynowego wraz z metadanymi. W przypadku otwartego zbioru danych Common Crawl przeznaczonego do trenowania dużych modeli językowych uwzględniano dokumenty tekstowe pozyskane ze snapshotów Common Crawl zgromadzonych przez robota indeksującego, powszechnie identyfikowanego jako CCBot, zgodnie z zapewnieniem CommonCrawl działającego z poszanowaniem obowiązujących protokołów, umożliwiając prawidłowe pozyskiwanie i utrwalanie treści stron internetowych w oparciu o jasno określone reguły zawarte w pliku robots.txt. Zbiory składające się z oczyszczonych i zdeduplikowanych anglojęzycznych danych internetowych pochodzących z CommonCrawl, w tym wchodzących w skład datasetów FineWeb i FineWeb-Edu, włączono w zbiór pretreningowy na tej samej zasadzie. W ramach kolejnego wdrożonego środka przeprowadzano weryfikację, według listy domen adresujących uwzględniane zasoby, w celu potwierdzenia braku zastrzeżeń w związku ze zwielokrotnianiem zasobów. Weryfikację prowadzono z wykorzystaniem zautomatyzowanego narzędzia działającego zgodnie z protokołem REP, który nakłada na

roboty sieciowe obowiązek odczytywania i stosowania się do instrukcji określonych w pliku robots.txt.

Weryfikację pliku robots.txt stosowano każdorazowo w przypadku zwielokrotniania zasobów bezpośrednio ze źródeł publicznie udostępnionych w taki sposób, aby każdy mógł mieć do nich dostęp w miejscu i czasie przez siebie wybranym.

Wprowadzono także środek polegający na dodatkowej zautomatyzowanej analizie uwzględnianych materiałów w oparciu o dedykowany wykaz sformułowań, wykaz kluczowych fraz, zliczanie i agregowanie wyników. Automatyczna weryfikacja obejmowała zarówno pliki robots.txt, jak i szeroki zakres plików tekstowych, ze szczególnym uwzględnieniem plików pierwszego poziomu, a także plików zawierających typową dokumentację formalną stron internetowych (warunki korzystania z usług, polityki prywatności, itp.).

Ze względu na fakt, że pozyskiwanie zasobów było ukierunkowane przede wszystkim na źródła polskojęzyczne, wdrożono dodatkowy środek w postaci analiz zespołów anotatorskich, uzupełniający działanie środka automatycznego w przypadkach wymagających w szczególności rozstrzygnięcia wątpliwości natury językowej.

Z uwagi na formalną specyfikę formuły projektu dotyczącego realizacji modelu, Kodeks GPAI opublikowany na stronach Unii Europejskiej zgodnie z informacją dostępną pod adresem: https://digital-strategy.ec.europa.eu/pl/policies/contents-code-gpai („Kodeks") nie podlegał sygnowaniu.

Dodatkowe uwagi:

Niezależnie od faktu opublikowania Kodeksu w czasie, w którym prace projektowe znajdowały się na zaawansowanym etapie, według wszelkich analiz i najlepszej wiedzy, formalne założenia projektowe od początku realizacji prac uwzględniały zasady, wartości i środki zbieżne z tymi, które zostały ujęte w Kodeksie. Dotyczy to w szczególności należytej staranności w zakresie przestrzegania zastrzeżeń praw wynikających z wyjątku lub ograniczenia TDM, na poziomie nie niższym od wynikającego z przepisów prawa i zakresu

	zobowiązań ujętych w Kodeksie, w szczególności poprzez wdrożenie adekwatnych środków dedykowanych temu celowi.

3.2 Usuwanie nielegalnych treści

	W celu przeciwdziałania wykorzystaniu nielegalnych treści stosowano filtrowanie danych na poziomie adresów URL oraz domen, w oparciu o "czarną listę" utworzoną na podstawie trzech źródeł – dwóch krajowych i dodatkowo trzeciego zagranicznego (ośrodka uniwersyteckiego). Dodatkowo korzystano z zasobów publikowanych przez krajowy CERT/CSIRT NASK i konsultacji specjalistycznych. Respektowano skuteczne środki technologiczne (zgodnie z definicją zawartą w art. 6 ust. 3 Dyrektywy 2001/29/WE), mających na celu zapobieganie lub ograniczanie nieautoryzowanych działań w odniesieniu do utworów i innych przedmiotów chronionych prawami autorskimi, w szczególności poprzez przestrzeganie technologicznych blokad lub ograniczeń dostępu narzuconych przez modele subskrypcji lub paywalle.
Ogólny opis wprowadzonych środków:	Według najlepszej wiedzy nie korzystano ze stron internetowych udostępniających publicznie treści, które, były uznawane przez sądy lub organy publiczne w Unii Europejskiej i Europejskim Obszarze Gospodarczym za trwale oraz wielokrotnie naruszające prawa autorskie lub prawa pokrewne, na skalę komercyjną. Dla mitygacji ryzyka zastosowano podejście oparte na przyznaniu generalnie wyższego priorytetu zasobom pochodzącym ze źródeł, do których: - zastosowanie miały przepisy Rozporządzenia 2022/2065 (DSA), zgodnie z zakresem wynikającym z art. 2 DSA, ze szczególnym uwzględnieniem mechanizmów prawnych, o których mowa w art. 16-18, art. 20-21 i art. 35 DSA; - w zakresie wykraczającym poza obszar stosowania DSA, zastosowanie miały mechanizmy prawne o zbliżonym charakterze, w szczególności pozwalające odbiorcom usług zgłaszanie nielegalnych treści (w tym tekstów i danych udostępnianych bez wymaganej zgody podmiotów uprawnionych) i blokowanie przez podmioty udostępniające ich dostępności.

3.3. Inne informacje

Inne istotne informacje na temat przetwarzania danych:	Dodatkowo, w celu minimalizowania ryzyka generowania przez systemy AI, z którymi zintegrowany może zostać model, danych wyjściowych mogących naruszać prawa autorskie lub prawa pokrewne, wdrożono środki mające na celu zapobieganie generowaniu przez modele danych wyjściowych, powielających treści treningowe w sposób naruszający takie prawa.