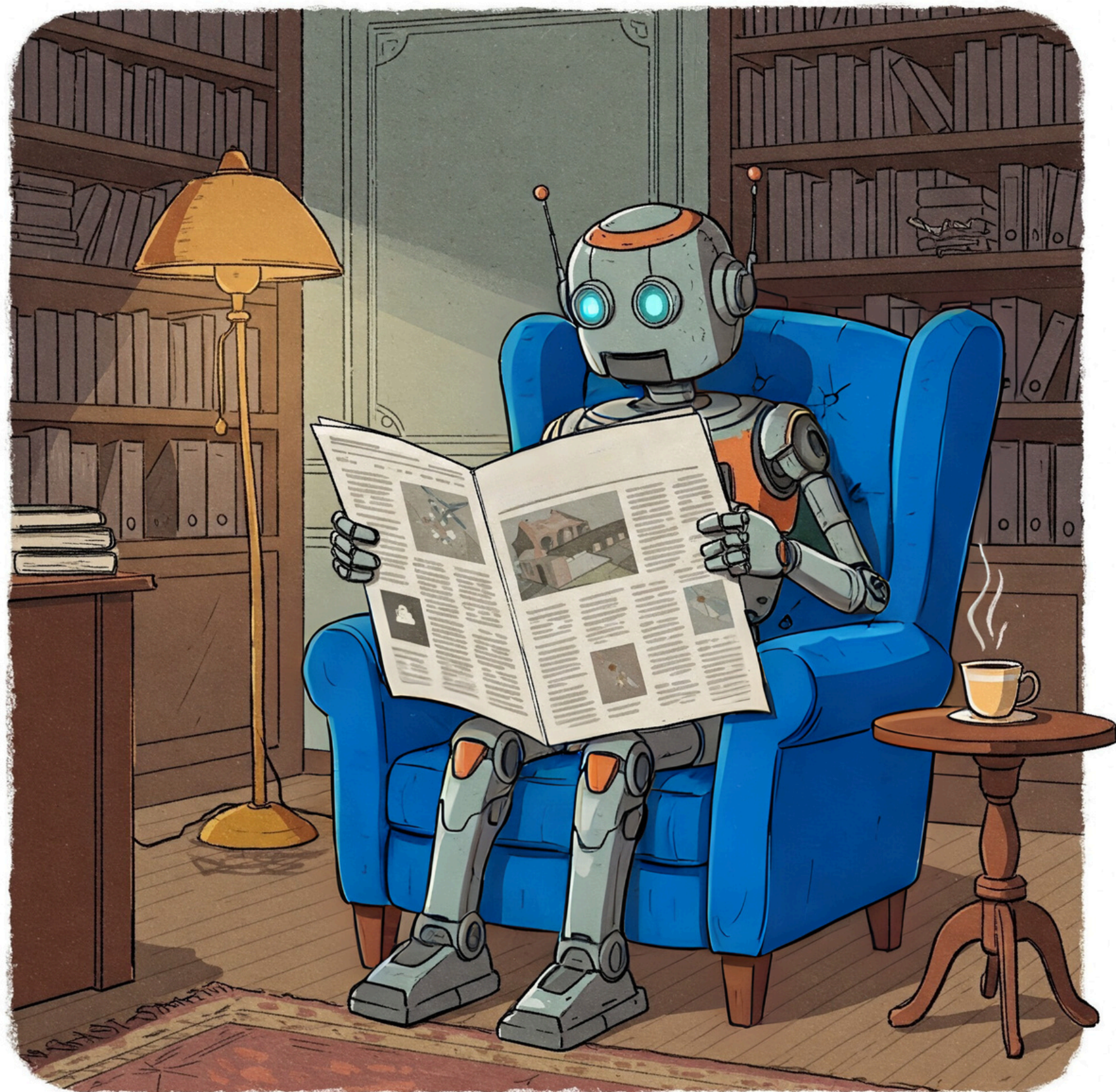




PRZEGLĄD NOWOŚCI W BRANŻY AI

Styczeń/Luty 2025



Grafika wygenerowana w aplikacji Gemini (Imagen 3)

Styczeń

2025

Poniedziałek	Wtorek	Środa	Czwartek	Piątek	Sobota	Niedziela
		1	2	3	4	5
6	7	8	9	10 <i>Premiera Asystenta AI (model DeepSeek-V3)</i>	11	12
13	14	15	16	17	18	19
20 <i>Premiera DeepSeek R1</i>	21	22 <i>Premiera Doubao 1.5 Pro</i>	23	24	25	26
27 <i>Premiera Janus-Pro</i>	28 <i>Premiera Qwen 2.5-Max</i>	29	30 <i>Gemini 2.0 Flash i Imagen 3 w aplikacji Gemini</i>	31 <i>Premiera o3-mini</i>	1	2

Poniedziałek	Wtorek	Środa	Czwartek	Piątek	Sobota	Niedziela
					1	2 Deep Research od Open AI
3	4	5 Nowości Gemini 2.0, GPT Search dla wszystkich	6	7	8	9
10	11	12	13 Porozumienie Google z PFR	14	15	16
17 Potwierdzenie inwestycji Microsoftu	18 Premiera Grok 3	19	20	21	22	23
24 Premiera Claude 3.7 Sonnet, Claude Code i PPLuM	25 Premiera QwQ- Max-Preview oraz Wan 2.1	26	27 Premiera GPT 4.5	28 Sora dostępna w Europie		

TECHNOLOGICZNI GIGANCI INWESTUJĄ W POLSCE



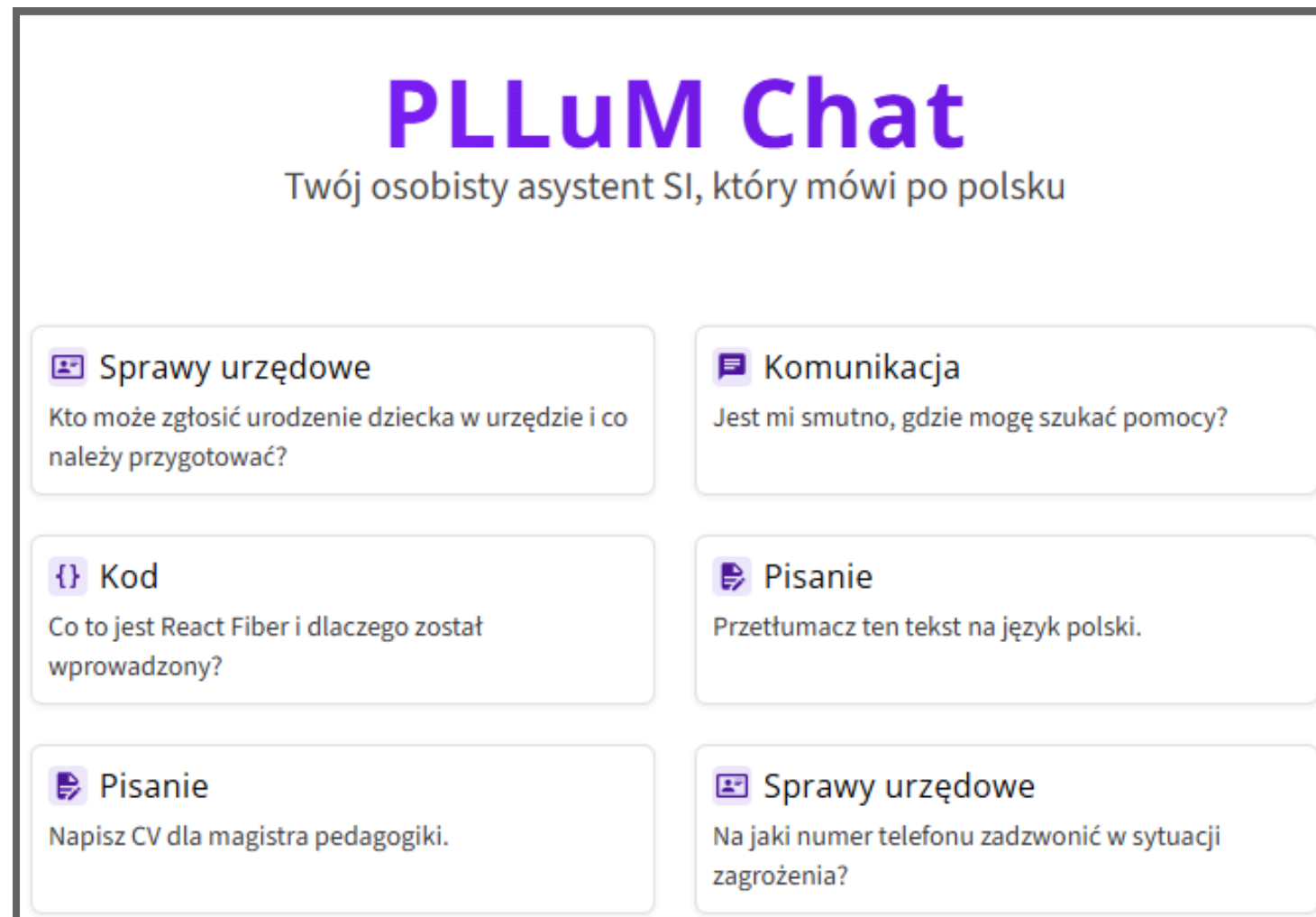
Grafika wygenerowana w aplikacji Gemini (Imagen 3)

W social mediach (a przynajmniej na LinkedInie) zawrzało po informacji, że Google chce przekazać **5 mln dolarów na przeszkolenie 1 mln Polaków w dziedzinie sztucznej inteligencji** – czyli 5 dolarów na osobę.

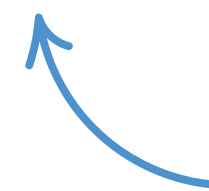
Okazuje się jednak, że mają być to miliardy (ale raczej złotych). Finalna kwota ma zostać określona dopiero po wstępnym rozpoznaniu potrzeb w administracji publicznej i biznesie.

Kilka dni później Microsoft ogłosił, iż przeznaczy **700 milionów dolarów do 2026 roku na wzrost mocy obliczeniowej w Polsce.**

POLACY NIE GĘSI SWÓJ LLM MAJĄ!



Źródło: <https://pllum.clarin-pl.eu>



czat jest dostępny pod tym linkiem

24 lutego miała miejsce premiera polskiej rodziny modeli sztucznej inteligencji PLLuM (Polish Language Large Model).

PLLuM to kolejny po Bieliku polski model językowy. Powstał na zlecenie Ministerstwa Cyfryzacji, głównie na potrzeby administracji publicznej. Model ma wspierać obywateli jako wirtualny asystent w aplikacji mObywatel oraz inteligentny asystent urzędniczy, a także ułatwić rozwój aplikacji edukacyjnych.

Czat aktualnie nie ma dostępu do internetu, a więc bazuje tylko na wyuczonym zbiorze danych. Oznacza to, że nie odpowie poprawnie na pytania o najświeższe informacje oraz nie poda informacji spoza zbioru.

PRZEGLĄD NOWOŚCI AI

DEEPSEEK

Wydanie Asystenta AI DeepSeek bazującego na modelu DeepSeek-V3 wywołało spore poruszenie w branży ze względu na niski koszt „wytworzenia” technologii. Niecały miesiąc później firma wydała kolejny model językowy DeepSeek-R1, który w testach twórców osiągnął wyniki porównywalne do openAI-o1. Model udostępniono bezpłatnie i opublikowano jego kod źródłowy. DeepSeek-R1 również jest miażdżąco tańszym modelem w porównaniu do amerykańskich rozwiązań.

DOUBAO 1.5 PRO

Doubao-1.5-pro to model firmy ByteDance – chińskiego przedsiębiorstwa, do którego należy m.in. TikTok. Doubao to model MoE (Mixture of Experts), a jego wydajność w testach przewyższyła GPT-4o czy DeepSeek-V3. Model wypuszczony z opcją „Deep Think” obsługuje też funkcje multimodalne, jednak nie jest dostępny jako open-source.

PRZEGLĄD NOWOŚCI AI

JANUS-PRO

Janus-Pro to aktualnie najbardziej zaawansowany model do generowania grafik z rodziny Janus należącej do DeepSeek. Poprawiono w nim wydajność rozumienia multimodalnego, jak i generowania obrazu z tekstu. Janus-Pro to kolejny po DeepSeek-V3 i R1 (premiery w styczniu 2025 roku) model, który firma udostępniła do ogólnego użytku.

QWEN 2.5-MAX

Qwen 2.5-Max to model MoE, który należy do firmy Alibaba – tak, tego chińskiego e-commerce’a. Qwen 2.5-Max, podobnie jak większość aktualnie pojawiających się nowości potrafi generować kod, tekst, przeszukiwać internet, a nawet tworzyć obrazy i wideo. Model jest dostępny w Qwen Chat oraz za pośrednictwem Alibaba Cloud (interfejs API). Informacja o nowości na blogu Qwen pojawiła się 28 stycznia.

PRZEGLĄD NOWOŚCI AI

DEEP RESEARCH

Deep Research to agent AI (Open AI), który znajdzie setki źródeł online, przeanalizuje je, podsumuje i dostarczy Ci w formie kompleksowego raportu. Głębokie badanie danego tematu może potrwać od 5 do 30 minut. Funkcja dostępna jest w czacie GPT w wersji Pro – od 5 lutego również w Europejskim Obszarze Gospodarczym, Szwajcarii i Wielkiej Brytanii.

GEMINI 2.0

Premiera Gemini 2.0 miała miejsce w grudniu, ale od niedawna modele z tej rodziny są ogólnodostępne. Najpierw 30 stycznia Gemini 2.0 Flash i Imagen 3 pojawiły się w aplikacji Gemini. Następnie 5 lutego Google wypuścił eksperymentalne modele Gemini 2.0 Pro i Gemini 2.0 Flash Thinking również w aplikacji Gemini, a za pośrednictwem API udostępnił Gemini Flash 2.0 i ogłosił wprowadzenie na rynek Gemini 2.0 Flash-Lite w wersji zapoznawczej.

PRZEGLĄD NOWOŚCI AI

O3-MINI

OpenAI o3-mini to pierwszy mały model rozumowania dostępny w Chat GPT dla bezpłatnych użytkowników (funkcja „Rozumuj”). Model stworzony przede wszystkim z myślą o programistach umożliwia dobór metody rozumowania (3-stopniowy), dzięki czemu użytkownicy mogą wybierać między szybkością a precyzją (w czacie dostępny jest poziom medium). Dodatkowo o3-mini korzysta z wyszukiwania i podaje linki do źródeł internetowych.

GROK 3

Grok 3 to model rozumujący z możliwością wyszukiwania informacji w sieci. Poza tym pojawił się też model Grok 3-mini, który odpowiada szybciej, ale oczywiście kosztem nieco mniejszej dokładności. Wyniki testów rozumowania dla Grok 3 i Grok 3-mini wypadły lepiej w porównaniu np. do nowego modelu OpenAI o3-mini(high). Nowość należy do firmy xAI, którą założył Elon Musk.

PRZEGLĄD NOWOŚCI AI

SONNET 3.7

Claude 3.7 Sonnet to, jak piszą twórcy, pierwszy hybrydowy model rozumowania na rynku. Ma połączony ze sobą w jednym narzędziu „zwykły” LLM i model rozumowania (zwykle są one dostępne oddzielnie). Pomiedzy „prostym myśleniem” a „rozszerzonym rozumowaniem” można się przełączać. Nowy model jest dostępny we wszystkich planach.

CLAUDE CODE

Claude Code to natomiast dedykowane narzędzie do kodowania (na ten moment w wersji beta), które może edytować pliki i naprawiać błędy, odpowiadać na pytania o architekturę i logikę kodu, uruchamiać testy. Dostępna jest też integracja z GitHub.

PRZEGLĄD NOWOŚCI AI

PLLuM

PLLuM to już wspomniana pierwsza polska rodzina modeli językowych, składająca się na razie z 18 udostępnionych modeli różniących się parametrami i funkcjami (część na licencjach w pełni otwartych, a część ograniczona np. do celów naukowych). Największą zaletą jest oczywiście dostosowanie do języka polskiego. Twórcy podkreślają etyczny i zgodny z prawem sposób pozyskania danych, a do trenowania PLLuM wykorzystano 150 miliardów tokenów w korpusie języka polskiego.

QWQ-MAX

QwQ-Max-Preview to najnowszy model z serii Qwen (Alibaba), zbudowanej na niedawno przedstawionym modelu Qwen 2.5-Max. Firma planuje udostępnić kod źródłowy QwQ-Max w ramach Apache 2.0 i mniejsze modele rozumowania na licencji open-source do użytku lokalnego. QWQ-Max jest dostępny w czacie Qwen w ramach opcji „Thinking (QWQ)”.

PRZEGLĄD NOWOŚCI AI

WAN 2.1

25 lutego chiński lider e-commerce Alibaba poinformował, że udostępni nowy model Wan 2.1 (służący do generowania wideo i obrazów) na licencji open-source. Model obsługuje kilka funkcji: Text-to-Video, Image-to-Video, Text-to-Image oraz Video-to-Audio, a do tego pozwala edytować filmy.

GPT 4.5

GPT 4.5 rozumuje i uczy się bez nadzoru. Został wytrenowany na superkomputerach Microsoft Azure AI, co dało mu głębsze zrozumienie świata i szerszą wiedzę. Ma to prowadzić do zmniejszenia halucynacji modelu. Sama rozmowa z czatem ma być też bardziej naturalna i ludzka. Aktualnie model nie obsługuje funkcji multimodalnych, takich jak tryb głosowy, wideo i udostępnianie ekranu.

ŹRÓDŁA INFORMACJI

1. Open AI o3-mini: <https://openai.com/index/openai-o3-mini/>
2. PLLuM: <https://www.gov.pl/web/cyfryzacja/polska-buduje-wlasna-sztuczna-inteligencje--pllum-gotowy-do-dzialania>
3. PLLuM: <https://www.gov.pl/web/cyfryzacja/czesc-jestem-pllum-jak-powstaje-polski-ekosystem-modeli-jezykowych>
4. Janus-Pro: <https://github.com/deepseek-ai/Janus>
5. Chat GPT Search: <https://openai.com/index/introducing-chatgpt-search/>
6. DeepSeek-R1: <https://github.com/deepseek-ai/DeepSeek-R1>
7. DeepSeek-R1: <https://api-docs.deepseek.com/news/news250120>
8. QWQ-Max-Preview: <https://qwenlm.github.io/blog/qwq-max-preview/>
9. Wan 2.1: <https://github.com/Wan-Video/Wan2.1>
10. Wan 2.1: <https://wan21ai.com/>
11. Gemini 2.0: <https://blog.google/technology/google-deepmind/gemini-model-updates-february-2025/>
12. Gemini 2.0: <https://blog.google/feed/gemini-app-experimental-models/>
13. Gemini i Imagen: <https://blog.google/feed/gemini-app-model-update-january-2025/>
14. Claude Sonnet 3.7 i Claude Code: <https://www.anthropic.com/news/claude-3-7-sonnet>
15. Grok 3: <https://x.com/i/broadcasts/1gqGvjeBlj0GB>
16. GPT 4.5: <https://openai.com/index/introducing-gpt-4-5/>
17. Sora: <https://x.com/OpenAI/status/1895353525817008447>
18. Qwen 2.5-Max: <https://qwenlm.github.io/blog/qwen2.5-max/>
19. Google w Polsce: <https://www.gov.pl/web/premier/miliardowe-inwestycje-google-w-polsce>
20. Microsoft w Polsce: <https://www.pap.pl/aktualnosci/wiceprezes-i-prezydent-microsoftu-w-przyszlym-roku-mozliwa-kolejna-inwestycja-w-polsce>
21. Doubaou-1.5-pro: <https://www.aibase.com/tool/35837>

ŚLEDŹ NAS W SOCIAL MEDIACH

FACEBOOK @semstormtool

<https://www.facebook.com/semstormtool/>

LINKEDIN @semstorm

<https://www.linkedin.com/company/semstorm/>

INSTAGRAM @semstormtool

<https://www.instagram.com/semstormtool/>