

programista

4/2026 (125)

Cena 33.90 zł (w tym VAT 8%)

ESP-NOW

Protokół komunikacyjny dla drona

JUŻ W EMPIKU



ISSN 2084-9400

04



Komputer niebinarny. Czy to jest możliwe?

Programista wspomagany sprzętowo

Czy można zastąpić płatne AI lokalnymi?

Własny bot na Discordzie

ECS w silnikach gier i symulacji

Strefa CTF



Czy można zastąpić płatne AI lokalnymi?

Zgodnie z prawem nagłówków Betteridge’a... Stop. Tym razem odpowiedź jest jednak twierdząca. Tak, można. Nie w absolutnie wszystkim – do trudniejszych zadań wciąż używam płatnego Claude’a – ale w tym artykule przedstawię sposób, w jaki z powodzeniem wykorzystuję lokalnie serwowany model w życiu codziennym, a także jego ograniczenia.

I POSTĘP JEST NIEUNIKNIONY

Półtora roku temu w wydaniu 111 Programisty [0] popeliłem artykuł „LLM w użyciu”, gdzie przedstawiłem rzeczywisty przykład codziennego użycia dużego modelu językowego (LLM) z wykorzystaniem Mistral-7B¹. Używałem go do wstępnego odfiltrowywania wiadomości ze skrzynki pocztowej i przedstawienia streszczeń wiadomości uznanych za interesujące.

Z perspektywy czasu muszę jednak przyznać, że był to bardziej ciekawy eksperyment niż pełnoprawne narzędzie. Mistral działał, ale każdą jego odpowiedź trzeba było traktować ze znaczącą rezerwą. Potrafił ignorować polecenia z promptu systemowego, a streszczenia bywały kompletnie chybione. Ponadto dla różnych wiadomości różne treści promptu wydawały się mieć losowy efekt na jakość odpowiedzi. Można więc powiedzieć, że nadawał się on do jednego, wąsko zdefiniowanego zadania, które ciężar zapewnienia tej precyzji składało na barkach użytkownika. Na domiar złego jakość obsługi wywoływania narzędzi (ang. *tool calling*) była bardzo niska – właściwie można powiedzieć, że model tego nie wspierał w ogóle. Dlatego też trudno się dziwić, że do rozmowy, pisania czy analizy dokumentów nadal sięgałem po płatne usługi.

Od tamtej pory zmieniło się bardzo wiele – ale nie mój sprzęt. Na tym samym komputerze z kartą RTX 4090 działa dziś Qwen 3.8 27B² od Alibaby [1], opublikowany po raz pierwszy 14 sierpnia 2026 r. Ma prawie cztery razy więcej parametrów niż Mistral-7B, ale po kwantyzacji zajmuje w pamięci karty niewiele więcej miejsca. Różnica w jakości jest za to ogromna (Tabela 0).

W tabeli przedstawiono dwa warianty wielkości okna kontekstowego, ponieważ korzystam z nich w zależności od potrzeb. Domyślnie używam wersji 96k. Zostawia ona ok 4GB wolnej pamięci VRAM, co pozwala na swobodne działanie innych serwisów wymagających GPU, jak Immich lub Jellyfin. Wersję 192k stosuję, gdy potrzebuję większego okna kontekstu – w dalszej części artykułu przeszukanie skrzynki e-mailowej wymagało dużego okna kontekstowego (Rysunek 6).

	Mistral-7B-Instruct v0.2	Qwen 3.8 27B (192k)	Qwen 3.8 27B (96k)
Data publikacji wag	11 grudnia 2023	14 sierpnia 2026	14 sierpnia 2026
Liczba parametrów	7.24e9	2.73e10	2.73e10
Format wag	F16	Q4_K_M	Q4_K_M

1. W dalszej części artykułu pisząc „Mistral” będę miał na myśli konkretnie model Mistral-7B, który wcześniej opisywałem, jeśli nie zaznaczę inaczej.
2. W dalszej części artykułu, pisząc „Qwen”, będę miał na myśli Qwen 3.8 27B kwantyzowany Q4_K_M, jeśli nie zaznaczę inaczej

	Mistral-7B-Instruct v0.2	Qwen 3.8 27B (192k)	Qwen 3.8 27B (96k)
Okno kontekstu	32k	192k	96k
Pamięć karty graficznej	15.4 GB	23.2 GB	20.5 GB
Liczba tokenów tego samego tekstu	7475	5770	5770
Przetwarzanie promptu	5157 tok/s	1726 tok/s	1750 tok/s
Generowanie odpowiedzi	58.7 tok/s	64.1 tok/s	65.7 tok/s
Wywoływanie narzędzi	nie	tak	tak
Rozumienie obrazów	nie	tak	tak
„What is 17*23? Answer with the number only.”	„The answer is 391.”	„391”	„391”

Tabela 0. Mistral-7B i Qwen 3.8 27B na karcie RTX 4090 (Ollama 0.32, mediana z trzech pomiarów; Qwen generuje z dekodowaniem spekulacyjnym MTP, bez niego byłoby to około 48 tok/s)

Qwen całkiem nieźle rozmawia po polsku, rozumie zdjęcia i – co najważniejsze – sprawnie korzysta z narzędzi. Na co dzień zapisuje za mnie posiłki i pomiary wagi, przegląda pocztę, rozwiązuje moim klientom proste zadania supportowe, pomaga przy fakturach i szuka informacji w sieci. W ciągu kilku tygodni przetworzył dziesiątki milionów tokenów, a ja nie zapłaciłem za to ani złotówki abonamentu. Nie oznacza to jednak, że rezygnuję z płatnych modeli. Do programowania i innych, bardziej skomplikowanych zadań wciąż używam Claude’a, bo tam lokalny model zwyczajnie im nie dorównuje.

I MISTRAL A QWEN

Na początku na tapet wezmę odtworzenie przykładu z poprzedniego artykułu (Programista 111) i sprawdzenie jego działania z użyciem Mistrala oraz Qwena. Był to prosty program podsumowujący nowe e-maile ze skrzynki, pomijający wiadomości promocyjne i spam. Część czytelników zapewne powie, że jest to porównanie nieuczciwe, ponieważ modele te mają nieporównywalne liczby parametrów, a ponadto dzieli je prawie trzy lata rozwoju technologii, co jest w tej branży przepaścią. W istocie, jest to prawda, jednak istotą mojego porównania jest właśnie pokazanie skali zmian, szczególnie że z perspektywy użytkownika uruchamia on po prostu inny model na tym samym sprzęcie, otrzymuje bardzo zbliżone prędkości, a jednocześnie jakość jest z zupełnie innej półki.

We wspomnianym tekście przedstawiłem pięć przetworzonych wiadomości: zaproszenie na konferencję, tygodniowy raport z serwisu Backtrace, którego wcześniej byłem pracownikiem, informację o nieudanym buildzie z GitHuba (również Backtrace), przypomnienie

z Duolingo i powiadomienie o doręczeniu przesyłki przez firmę kurierską FedEx. Zaproszenie i przypomnienie zostały odrzucone jako promocyjne, pozostałe trzy – wyświetlone. Te same pięć wiadomości przepuściłem teraz przez oba modele.

Dla każdego e-maila program wywołuje model dwukrotnie: najpierw prosi o streszczenie wiadomości promptem z Listingu 0, a potem, w tej samej rozmowie, pyta, czy wiadomość jest promocyjna i niezwiązana z pracą (Listing 1). Odpowiedź „yes” oznacza pominięcie wiadomości.

W listingu 0 znajduje się pierwszy prompt dla modelu. Promptu nie zmieniałem, żeby oba modele dostały dokładnie to samo. W przypadku Qwena lub obecnie dostępnych komercyjnie modeli może on się (słusznie) wydawać bardzo niezręczny. Powstał on jednak półtora roku temu metodą prób i błędów – każda kolejna wersja miała wymusić na Mistralu odrobinę więcej posłuszeństwa.

Listing 0. Prompt z przykładu

Please summarize the following message. Be very succinct! Do not include any unnecessary data, such as company numbers. Focus on the underlying message, rather than the message format. In case of longer message threads, prefer the newest updates.

Try to limit your summary to one or two sentences.

Do not include any of the following information in your summary UNDER ANY CIRCUMSTANCES:

- copyright or trademark
- legal
- "do not reply" in any form. If there is a "do not reply" message, ignore it.
- "unsubscribe" in any form. If there is an "unsubscribe" message, ignore it.
- that the message is confidential
- that the message is automatically generated

Do not mention the same thing several times.

W Listingu 1 przedstawiono drugi prompt, decydujący o tym, by wiadomość w ogóle się pokazała.

Listing 1. Prompt decydujący o pominięciu wiadomości

Is this an automated or promotional email that's also not related to work? Important automated emails (such as from your bank or delivery service) are not promotional. Answer yes if the message is promotional and not related to work. Answer no otherwise. Start your answer with 'yes' or 'no'.

Wynik działania programu z Mistralem przedstawiono na Rysunku 0. Model przepuścił te same trzy wiadomości co półtora roku temu – choć wiadomość od FedEx pojawiła się tylko w części wywołań. Mimo wszystko można uznać, że na poziomie decyzji „pokazać czy pominąć” poradził sobie dobrze. Gorzej ze streszczeniami. Podsumowanie raportu z Backtrace kończy się informacją, kiedy raport został wygenerowany, czyli dokładnie tym, czego prompt zabrania. Streszczenie z GitHuba mówi, że część zadań się nie powiodła, ale nie mówi które – zamiast tego odsyła do GitHuba.

Najgorzej wypada FedEx. Zamiast jednego czy dwóch zdań dostajemy niemal całą treść wiadomości: nadawcę, adres doręczenia (na zrzucie zamazany), podziękowanie za skorzystanie z usług, prośbę o sprawdzenie statusu przesyłki, a na koniec zdanie „This message

was generated automatically, please do not reply”. Prompt zabrania wprost obu tych ostatnich rzeczy. Ten sam błąd opisywałem w wydaniu 111 – i nie ma w tym nic dziwnego, bo to wciąż ten sam model.

```
krzak@krzak:~/code/programista/ai_helper main*
% bundle exec ruby repro_111.rb creds_gmail.json
Model: mistral-7b-instruct-v0.2
Message: Backtrace: wtf for 01/09/2024 - 01/23/2024 (no-reply@backtrace.io)
Summary: No crashes or errors reported for krzak from January 9 to 23, 2024. Recent activity report generated on January 24, 2024.
Message: [backtrace-labs/crashpad] Run failed: CI - backtrace (bc30fe3) ("Paweł Zakrzewski" <notifications@github.com>)
Summary: The CI workflow run for the backtrace-labs/crashpad repository finished, with some jobs failing for the Android builds. The exact failure details can be viewed on GitHub.
Message: FedEx | Doręczyliśmy Twoją przesyłkę. (FedEx Express Polska <pldomestic@fedex.com>)
Summary: Your package from SPPrzedsiębiorstwa Sp. z o.o. was delivered by FedEx on January 23, 2024, to Paweł Zakrzewski at
services. Please check your package's status on FedEx Express. This message was generated automatically, please do not reply.
```

Rysunek 0. Wiadomości z wydania 111 przetworzone przez Mistral-7B

Qwen dostał te same wiadomości i ten sam prompt (Rysunek 1). Każde streszczenie mieści się w jednym zdaniu. Raport z Backtrace to „All clear for krzak: zero errors, crashes, or suspicious activity over the past two weeks”. Przy buildzie z GitHuba model podaje to, co najważniejsze: Windows, Linux i macOS przeszły, a wszystkie buildy Androida padły. Tej informacji Mistral w ogóle nie przekazał. Powiadomienie od FedEx-u skraca się do adresata, miasta i daty – bez zakazanych formułek i bez mojego adresu domowego. Qwen nie jest jednak bezbłędny: bywa, że dokłada emoji, o które nikt nie prosił. Jest to jednak wada kosmetyczna, a nie merytoryczna.

```
krzak@krzak:~/code/programista/ai_helper main*
% bundle exec ruby repro_111.rb creds_gmail.json
Model: qwen3.8-72b-96k
Message: Backtrace: wtf for 01/09/2024 - 01/23/2024 (no-reply@backtrace.io)
Summary: All clear for krzak: zero errors, crashes, or suspicious activity over the past two weeks. 🎉
Message: [backtrace-labs/crashpad] Run failed: CI - backtrace (bc30fe3) ("Paweł Zakrzewski" <notifications@github.com>)
Summary: CI for crashpad failed: all Windows, Linux, and macOS builds passed, but every Android build failed.
Message: FedEx | Doręczyliśmy Twoją przesyłkę. (FedEx Express Polska <pldomestic@fedex.com>)
Summary: FedEx delivered a package to Paweł Zakrzewski in Szczecin on 23 Jan 2024.
```

Rysunek 1. Te same wiadomości przetworzone przez Qwen 3.8 72B

Decyzje modelu program wypisuje tylko w trybie diagnostycznym (Rysunki 2 i 3). Widać w nim także dwie pominięte wiadomości. Mistral na polecenie „Start your answer with ‘yes’ or ‘no’” odpowiada pełnym zdaniem, na przykład „Yes, the email is promotional”. Qwen odpowiada samym „yes”. Przy pięciu wiadomościach nie ma to znaczenia, bo kod i tak dopasowuje wyrażenie regularne na początku odpowiedzi. Przy narzędziach, gdzie odpowiedź modelu jest wejściem kolejnego kroku, jest to różnica między programem działającym a niedziałającym.

```
krzak@krzak:~/code/programista/ai_helper main*
% make repro-mistral DEBUG=1
bundle exec ruby repro_111.rb creds_gmail.json
Model: mistral-7b-instruct-v0.2
Skipping message Darzona konfederacja dla uczestników HackYeah 🍷 ("CodeFrenzy | Aleksandra" <alek.sandra@codefrenzy.pl>)
Decision: Yes, the email is promotional.
Message: Backtrace: wtf for 01/09/2024 - 01/23/2024 (no-reply@backtrace.io)
Summary: No crash activity or errors reported for krzak from January 9 to 23, 2024. Recent activity report generated on January 24, 2024.
Decision: No, the message is not promotional and not related to work.
Message: [backtrace-labs/crashpad] Run failed: CI - backtrace (bc30fe3) ("Paweł Zakrzewski" <notifications@github.com>)
Summary: A CI workflow run for the "backtrace-labs/crashpad" repository finished, with some jobs failing on Android builds. The exact failures are not specified in the message.
Decision: No.
Skipping message Hi krzak! Keep your 483 day German streak going! (Duolingo <hello@duolingo.com>)
Decision: Yes, the email is promotional and unrelated to work.
Message: FedEx | Doręczyliśmy Twoją przesyłkę. (FedEx Express Polska <pldomestic@fedex.com>)
Summary: Your package from SPPrzedsiębiorstwa Sp. z o.o. was delivered by FedEx on January 23, 2024, to Paweł Zakrzewski at
services. This message was generated automatically, please do not reply. FedEx Express Poland Sp. z o.o. (NRS number: 0000357766) delivered the package.
Decision: No, the message is not promotional and not related to work. It is an automated email regarding a delivery.
```

Rysunek 2. Mistral-7B w trybie diagnostycznym, z decyzjami i pominiętymi wiadomościami