

Ustawienia LM Studio pod warsztat

Nie musisz tego czytać przed warsztatem — to ściągą na potem. Domyślne ustawienia LM Studio są dobre dla 95% przypadków. Poniżej to, co warto wiedzieć, gdy chcesz coś dostroić.

Model

Wariant	Plik / źródło	Dla kogo
Bielik 4.5B v3.0 Instruct (domyślny, Q8_0 ~4,7 GB)	Discover → bielik → speakeash/Bielik-4.5B-v3.0-Instruct-GGUF	warsztat — mieści się w 16 GB RAM, czysto i wiernie
Bielik 1.5B v3.0 Instruct (Q8_0 ~1,7 GB)	speakeash/Bielik-1.5B-v3.0-Instruct-GGUF	mało RAM / stary sprzęt — szybszy, ale potrafi halucynować
Bielik 11B przez API	bielik.ai (OpenAI-compatible)	najwyższa jakość, bez kosztu RAM

Prompt template

ChatML — wykrywa się automatycznie z metadanych GGUF Bielika. Nie zmieniaj. Sprawdzisz w prawym panelu czatu (sekcja „Prompt”). Jeśli odpowiedzi są dziwne/ucięte, upewnij się, że to ChatML.

Parametry (prawy panel czatu)

Ustawienie	Wartość	Po co
Temperature	0.3 do zadań „czyszczących” (OCR, anonimizacja, tłumaczenie), 0.7 do swobodnej rozmowy	niższa = wierniej, mniej „fantazji”
Context Length (n_ctx)	4096 domyślnie; podnieś do 8192, gdy wklejasz długie teksty	dłuższy kontekst = więcej RAM
GPU Offload	maks., jeśli LM Studio wykryje GPU; 0 na CPU-only	szybkość generacji

Sugerowany System Prompt (opcjonalnie)

W prawym panelu czatu jest pole „System Prompt”. Możesz wkleić:

Jesteś dokładnym asystentem do pracy z polskim tekstem naukowym i archiwalnym. Trzymasz się ściśle treści, którą dostajesz. Nie zmyślasz faktów, dat ani źródeł. Jeśli czegoś nie ma w tekście - mówisz, że tego nie ma. Odpowiadasz po polsku.

To ustawia model na cały czat — nie musisz powtarzać „nie dodawaj nic od siebie” w każdym promncie.

Najczęstsze problemy

- **Wolno generuje** → to normalne na CPU (5–15 tok/s). Skróć wklejany tekst albo użyj **bielik.ai API**.
- **Odpowiedź ucięta** → podnieś Context Length albo poproś „dokończ”.

- Model „gada od siebie”, zapętla się albo wypisuje dziwne znaczniki ($\langle | \dots | \rangle$) \rightarrow obniż Temperature do 0.3, dodaj System Prompt powyżej, w razie potrzeby ponów. Zdarza się na słabszych/innych buildach; wariant **1.5B Instruct z szablonem ChatML** (warsztatowy) zwykle jest czysty.