

LM Studio settings for the workshop

You don't need to read this before the workshop — it's a reference for later. LM Studio's default settings are fine for 95% of use cases. Below is what's worth knowing when you want to fine-tune things.

Model

Variant	File / source	Who it's for
Bielik 4.5B v3.0 Instruct (default, Q8_0 ~4.7 GB)	Discover → bielik → speakleash/Bielik-4.5B-v3.0-Instruct-GGUF	workshop — fits in 16 GB RAM, fully accurate
Bielik 1.5B v3.0 Instruct (Q8_0 ~1.7 GB)	speakleash/Bielik-1.5B-v3.0-Instruct-GGUF	low-back-6GB RAM / older hardware — faster, but prone to hallucination
Bielik 11B via API	bielik.ai (OpenAI-compatible)	highest quality, no RAM cost

Prompt template

ChatML — auto-detected from Bielik's GGUF metadata. Don't change it. You can check it in the right chat panel (under "Prompt"). If responses are odd or cut off, make sure it's set to ChatML.

Parameters (right chat panel)

Setting	Value	Why
Temperature	0.3 for "cleaning" tasks (OCR, anonymisation, translation), 0.7 for free-form conversation	lower = more faithful, less "creative"
Context Length (n_ctx)	4096 by default; raise to 8192 when pasting long texts	longer context = more RAM
GPU Offload	max, if LM Studio detects a GPU; 0 on CPU-only	generation speed

Suggested system prompt (optional)

In the right chat panel there's a "System Prompt" field. You can paste:

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.

(Leave the prompt in Polish — it instructs the model to stay grounded in the source text and respond in Polish.)

This sets the model for the entire chat — you don't need to repeat "don't add anything of your own" in every prompt.

Common issues

- Slow generation on CPU is normal (5–15 tok/s). Shorten the text you paste, or use the [bielik.ai](#) API.
- Responses that cut off mid-answer: raise Context Length, or just ask "continue".

- Model hallucinates, loops, or spits out strange tokens (`<|...|>`): lower Temperature to 0.3, add the System Prompt above, and retry. Tends to happen with weaker builds; the 1.5B Instruct with ChatML (what the workshop uses) is normally fine.