Small Language Models (SLM) v roce 2026

Small Language Models jsou modely umělé inteligence optimalizované pro běh s minimálními nároky na hardware. V éře „Agentic AI“ (agentní umělé inteligence) slouží SLM jako specializované mozky pro konkrétní úlohy, které nevyžadují obří znalostní databázi celého internetu.

Proč jsou SLM trendem roku 2026?

Hlavním posunem oproti předchozím letům je **kvalita trénovacích dat**. Místo kvantity se dnes sází na tzv. „učebnicová data“ a syntetická data generovaná silnějšími modely (distilace znalostí).

Klíčové výhody

- **Lokalita a soukromí:** Model běží přímo v telefonu, hodinkách nebo firemním serveru. Data nikdy neopouštějí zařízení.
- **Nulová latence:** Okamžitá reakce bez čekání na odezvu cloudu.
- **Udržitelnost:** Spotřebovávají zlomek energie (vhodné pro ekologické cíle firem).
- **Odolnost vůči halucinacím:** Díky úzkému zaměření na konkrétní doménu (např. právo, medicína) jsou fakticky přesnější než univerzální LLM.

Srovnání: LLM vs. SLM (Stav 2026)

Parametr	Large Language Models (LLM)	Small Language Models (SLM)
Příklady	GPT-5.2, Claude 5.1, Gemini 3	Phi-4, Llama 4.1 Scout, Qwen 3.5
Počet parametrů	1 bilion+	100 milionů až 15 miliard
Nasazení	Masivní datacentra	Smartphony, IoT, lokální PC
Specializace	Všeobecný génius	Expert na jednu oblast
Náklady	Vysoké (předplatné/tokeny)	Zdarma (lokální provoz)

Nejvýznamnější modely roku 2026

1. Microsoft Phi-4

Špička v oblasti „Tiny AI“. Model s méně než 4 miliardami parametrů, který v logických testech a kódování překonává starší modely typu GPT-4.

2. Llama 4.1 Scout (8B)

Nejoblíbenější model pro lokální „agenty“. Je dostatečně malý na to, aby běžel na průměrném notebooku, ale dostatečně chytrý na to, aby ovládal aplikace a automatizoval pracovní postupy.

3. Google Gemini Nano (v3)

Model integrovaný přímo do operačních systémů Android, který v reálném čase shrnuje hovory, emaily a upozorňuje na podvody (scam detection).

Praktické využití na ITWIKI

> **Pravidlo pro vývojáře:** Pokud vaše aplikace potřebuje pouze klasifikovat text, shrnout dokument nebo generovat kód v jednom jazyce, použijte **SLM**. LLM si nechte jen pro komplexní strategické plánování.

— **Související stránky:**

- Jak hostovat modely lokálně
- Agentní systémy
- Ochrana osobních údajů v AI

— **Autor:** @AI_Specialist **Aktualizováno:** ledna 2026

From:
https://www.serviceit.cz/ - **IT ENCYKLOPEDIE**

Permanent link:
https://www.serviceit.cz/doku.php?id=it:slm

Last update: **2026/01/04 16:41**