SPIES CREATIONS · LEAD-MAGNET · 2026

AI op uw eigen server

Een checklist voor MKB-directies en IT-managers — wanneer een klein taalmodel genoeg is

Door Stephan Spies · Spies Creations

Een AI-toepassing in een Amerikaans datacenter is niet vanzelf fout. Maar ook niet vanzelf goed. Voor afgebakende MKB-taken — samenvatten, classificeren, zoeken in een eigen kennisbank — is een klein taalmodel op de eigen server een serieus alternatief.

Eén A4'tje per kant, een eerlijke nulmeting, geen verkooppraatje. Loop de zeven onderdelen rustig langs en haal er één concrete vervolgactie uit.

1 Past een klein taalmodel bij uw taak?

Te weinig vinkjes: blijf voorlopig in de cloud.

- [ ] De taak is **afgebakend** (samenvatten, classificeren, vraag-antwoord boven een eigen kennisbank) — geen frontier-werk.
- [ ] U weet welk **kwaliteitsniveau** acceptabel is en hoe u dat gaat meten — niet "het voelt goed".
- [ ] De **doorlooptijd** mag binnen een redelijke marge liggen (seconden, niet milliseconden).
- [ ] Het **volume** is voorspelbaar — geen pieken van duizenden gelijktijdige verzoeken.

2 Welk model heeft u op tafel?

Eén keer bewust kiezen, niet impliciet.

- [ ] **Minimaal twee modellen** vergeleken op uw eigen taak (Phi-4, Llama 3 in de 8B-klasse, Gemma 2, Mistral Small) — geen blinde keuze.
- [ ] De **Nederlandse taalondersteuning** is gecontroleerd op uw eigen testset, niet aangenomen.
- [ ] De **licentie** staat commerciële inzet voor uw geval toe — vooral bij Llama-varianten loont het de voorwaarden door te lezen.
- [ ] De **modelgrootte** past bij uw hardware en uw kwaliteitseis — niet andersom.

3 Hardware-realiteit

Hardware bepaalt of het op de lange termijn werkbaar blijft.

- [ ] U weet of u **CPU-only**, **Apple Silicon** (sterke prijs-prestatie tot ~13B-modellen) of een **GPU** gaat draaien — gemotiveerd, niet uit voorraad.
- [ ] De **kwantisatie-keuze** is gemaakt: 4-bit voor snelheid, 8-bit voor betere kwaliteit, FP16 alleen als de hardware het rechtvaardigt.
- [ ] **Stroom, koeling en geluid** meegewogen — een werkstation in een kantoorhoek is iets anders dan een serverruimte.
- [ ] Er is een **vervangingsplan** (~3–5 jaar), met afschrijving in de begroting.

4 AVG, AI Act en data-architectuur

Lokaal draaien lost doorgifte op. De rest moet u nog zelf regelen.

- [ ] Het gebruik van het model staat in uw **register van verwerkingen**.
- [ ] **Doelbinding** en **bewaartermijnen** voor prompts, output en logs zijn vastgelegd — niet "we slaan voor de zekerheid alles op".
- [ ] U kunt **per betrokkene** inzage en wissing uitvoeren, ook op tussentijdse opslag rond het model.
- [ ] De **AI Act-classificatie** (minimaal, beperkt, hoog risico) is bepaald en de bijbehorende verplichtingen belegd.
- [ ] **Toegangsrechten** tot de machine zijn net zo strak als bij uw andere kritische systemen — geen "het staat toch lokaal".

KvK 57215782 · BTW NL001956622B49 · De Mulderij 4B, 3831 NV Leusden

5 Inferentie-stack en operations

Lokaal is geen excuus voor minder operationele discipline.

- [ ] Bewust gekozen tussen een **laagdrempelige stack** (Ollama, LM Studio) voor pilots en een **productie-stack** (vLLM, llama.cpp) – niet alles tegelijk.
- [ ] Er is **monitoring**: gebruik per dag, doorlooptijd, foutpercentage, kosten bij hybride opzet.
- [ ] Er is een **update-ritme** voor model en stack – minstens kwartaalmatig bekeken.
- [ ] Er is een **back-up** of warme stand-by – geen SLA-claims richting interne afnemers zonder.

6 Wanneer toch naar de cloud?

Lokaal is geen religie. Soms is cloud-burst de juiste keuze.

- [ ] Er is een **uitwijkregel**: welke verzoeken mogen door naar een frontier-model in de cloud, en welke nooit?
- [ ] Voor cloud-burst is waar mogelijk een **Europese cloudregio** of EU-vestiging gekozen – gemotiveerd vastgelegd.
- [ ] **Persoonsgegevens** worden vóór doorgifte gemaskeerd of vervangen door referenties, tenzij doorgifte gerechtvaardigd en gedocumenteerd is.
- [ ] U kunt **per maand** zien hoe vaak cloud-burst is gebruikt en waarom – anders is "lokaal-eerst" alleen een poster.

7 Eigenaarschap en onderhoud

Zonder eigenaar wordt elke AI-toepassing langzaam onbetrouwbaar.

- [ ] Er is een **eigenaar** – meestal de operationeel verantwoordelijke, niet alleen de bouwer.
- [ ] Er is een **vast moment** (wekelijks of maandelijks) waarop iemand naar het feitelijke gebruik kijkt, niet alleen het dashboard.
- [ ] Er is een **stop-knop**: u kunt de toepassing pauzeren zonder uw hele infrastructuur af te schakelen.
- [ ] Er is een **escalatieregel** voor kwaliteit, beschikbaarheid of beveiliging.

En nu? Tel uw vinkjes.

Twintig of meer: lokaal AI-werk past bij u en u kunt het verantwoord opzetten. **Tussen de tien en negentien:** het kan, maar er liggen onbeantwoorde vragen – pak die eerst. **Onder de tien:** blijf in een gecontroleerde cloud-omgeving en kom hier terug zodra de basisvragen zijn beantwoord.

Eén actie eruit halen. Pak het laagst genummerde onderdeel zonder vinkjes en agendeer een uur om dat in te richten. Niet alles tegelijk.

Doorpraten?

Spies Creations bouwt en begeleidt AI-toepassingen voor het Nederlandse MKB – lokaal waar het kan, cloud waar het moet, integratie waar het echt iets oplost.

Plan een uur met Stephan Spies – geen verkopers-script. spiescreations.nl/contact

Stuur een mail: stephan@spiescreations.nl

Vraag een korte review aan: wij lopen één AI-toepassing met u door op deze zeven punten.