Hoe weet u of uw AI-agent het goed doet?

Observability-checklist voor MKB-directies en IT-managers

Een AI-agent zonder monitoring kan maandenlang het verkeerde antwoord geven zonder dat iemand het opmerkt. Een verkeerde garantietermijn. Een afspraak op een vrije dag. Een onderdeel dat al twee jaar uit het assortiment is.

Deze checklist helpt u in een half uur scherp te krijgen of u dat zou zien aankomen. **Eén A4'tje per blok, een eerlijke nulmeting, geen verkooppraatje.**

SECTIE 1 Welke lagen van observability heeft u eigenlijk?

Streep aan wat al ingericht is. Wat openblijft, is werk.

- ☐ **Traces.** Per gesprek of taak is het volledige spoor terug te lezen: vraag, opgehaalde documenten, model-antwoord, gereedschap-aanroepen, duur, kosten.
- ☐ **Eval-set.** Een vaste set scenario's met door uw mensen vastgesteld goed/fout-oordeel, die u periodiek tegen de agent draait.
- ☐ **Productie-signalen.** U meet ten minste afbreekpercentage, escalatie naar mens, en repeat-onderwerp per klant.
- ☐ **Regressie-check.** Bij een nieuwe modelversie of prompt-wijziging draait u eerst de eval-set, vóórdat de wijziging live gaat.

SECTIE 2 Wat meet u – en wat schoont u op?

Observability zonder beleid wordt rommel.

- ☐ Er is een lijst van **wat de traces vastleggen** (volledige conversatie, brondocumenten, model-output, tool-I/O, tokenverbruik).
- ☐ Er is een lijst van **wat juist niet in de traces komt** (betaalgegevens, BSN, gezondheidsdata – gemaskeerd vóór opslag).
- ☐ U weet **waar** de traces fysiek staan (eigen omgeving, Europese cloud, buiten-EU) en heeft dat afgestemd op uw privacybeleid.
- ☐ Er is een **bewaartermijn** per type opslag (bijv. ruwe traces 30 dagen, geaggregeerd een jaar, evals onbeperkt mits geanonimiseerd).

SECTIE 3 Kwaliteit – hoe weet u dat het antwoord klopt?

Een dashboard meet kwantiteit. Voor kwaliteit heeft u een oordeel nodig.

- ☐ Er is een **eval-set** van minimaal twintig representatieve scenario's, opgebouwd uit echte (geanonimiseerde) klantvragen.
- ☐ Per scenario is er een **door uw mensen vastgesteld goed antwoord** – geen door het model zelf bedacht referentie-antwoord.
- ☐ Voor gevallen waar één goed antwoord niet bestaat, is er een **set criteria** (feit klopt, toon klopt, geen verzonnen verwijzing, geen ongepaste belofte).
- ☐ Er is een **periodieke menselijke steekproef** op echte gesprekken – minimaal tien per week, door iemand die de inhoud kan beoordelen.

SECTIE 4 Releases & regressies – weet u nog steeds dat het werkt?

Een modelversie-bump kan stilzwijgend gedrag veranderen.

- ☐ Voor elke **modelversie-wissel** (Claude, GPT, Gemini of welke ook) draait u eerst de eval-set en vergelijkt u met de vorige run.
- ☐ Voor elke **prompt-wijziging** of kennisbank-update geldt dezelfde regel: eerst evals, dan live.
- ☐ U houdt een **logboek** van wat u wanneer aanzette, met de bijbehorende eval-score – terugkijkbaar bij een klacht.
- ☐ Er is een **rollback-route**: u kunt binnen een dag terug naar een eerdere versie zonder dat dat een crisis wordt.

SECTIE 5 AVG / GDPR-controle op traces en evals

Wat u in traces vastlegt is in veel gevallen persoonsgegeven.

- ☐ Het trace- en eval-systeem staat in uw **register van verwerkingen** of is daaraan toegevoegd.
- ☐ U kunt **per klant** binnen redelijke tijd tonen wat er over hem of haar in de traces bewaard is – overzicht of export.
- ☐ U kunt **per klant** een wisverzoek uitvoeren, ook op traces en eval-runs waarin diens gegevens voorkomen.
- ☐ Voor extern verwerkte traces (LangSmith, Braintrust, eigen Phoenix, anders) is een **verwerkersovereenkomst** afgesloten en de gegevenslocatie bekend.

SECTIE 6 Eigenaarschap & escalatie

Observability zonder eigenaar wordt observability zonder onderhoud.

- ☐ Er is een **eigenaar** binnen het bedrijf – meestal de operationeel verantwoordelijke, niet de bouwer – die beslist wat een fout is.
- ☐ Er is een **wekelijks moment** waarop iemand de traces en signalen daadwerkelijk bekijkt – niet alleen het rapport, maar de gesprekken zelf.
- ☐ Er is een **stop-knop**: u kunt de agent pauzeren of beperken tot een veiligere flow zonder het hele systeem af te schakelen.
- ☐ Er is een **escalatieregel** bij patronen – wie wordt gebeld, wie beslist over pauze of rollback.

En nu? Tel uw vinkjes.

Achttien of meer: basis op orde. Tussen de negen en zeventien: agent in productie zonder volwassen kwaliteitsspoor. Onder de negen: stel de live-gang van een nieuwe agent uit tot dit staat.

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

≥ 18 VINKJES
Basis op orde – verfijn waar nodig en houd het levend.

9 - 17 VINKJES
In productie zonder kwaliteitsspoor – risico op stille fouten.

< 9 VINKJES
Stel live-gang uit tot deze fundamenten staan.

Doorpraten met Stephan?

Eén uur, geen verkopers-script – gewoon kijken wat klopt en wat niet. Of vraag een korte review aan: wij lopen uw agent door op deze zes punten en geven een nuchtere terugkoppeling.

Contact

stephan@spiescreations.nl
spiescreations.nl/contact