Er verschijnt een nieuw vlaggenschipmodel, de benchmarkgrafiek oogt overtuigend, en je migreert voor vrijdag alles ernaartoe. Dan kruipt de latency omhoog, verdubbelt de rekening op verzoeken die nooit het slimste model nodig hadden, en gaat een stille capaciteit waarop je leunde achteruit in de update die je niet kon weigeren. Je draait productie op één model en de roadmap van één leverancier, en dat noem je eenvoud. Het is blootstelling.

De gangbare zet is: kies het beste model en standaardiseer erop. Dat faalt omdat er geen beste model bestaat, alleen een model dat het beste is voor een taak, een budget en een foutmarge, en die drie zijn het oneens zodra je verkeer echt wordt.

The Model Portfolio behandelt je modellen zoals een serieuze belegger kapitaal behandelt: als een portefeuille van bezittingen die je met opzet alloceert, niet als één weddenschap die je blijft verdubbelen. Het introduceert het PORTFOLIO-raamwerk (Profile, Options, Route, Test, Fallback en Iterate), de zetten die een stapel API-sleutels veranderen in een beheerd systeem. Het raamwerk is de structuur; het specifieke model dat je dit kwartaal verkiest, is weer en wind. Als het scorebord opnieuw kantelt, blijft je architectuur staan.

Je leert:

• Elke taak profileren op wat hij werkelijk vraagt (nauwkeurigheid, latency, kosten, context, privacy), in plaats van standaard te kiezen voor het model met de luidste lancering.

• Verzoeken routeren naar het goedkoopste model dat de lat haalt, en pas opschalen naar een sterker model wanneer het werk dat verdient.

• Modellen testen tegen je eigen evals en je eigen verkeer, niet tegen een publieke benchmark die met andermans data gewonnen is.

• Fallback en redundantie inbouwen, zodat een deprecatie, een storing of een stille terugval gecontroleerd degradeert in plaats van je plat te leggen.

• De hele portefeuille beheren als een levende positie (gemeten, geherbalanceerd en met vastgezette versies), zodat capaciteit zich opstapelt terwijl de uitgaven beheersbaar blijven.

Het resultaat is concreet: je betaalt geen frontiermodel meer te veel voor werk dat een kleiner model goed doet, je bent niet langer gegijzeld door de release notes van één leverancier, en je systeem wordt goedkoper en veerkrachtiger naarmate het veld sneller beweegt.

Draai je meer dan één model in productie en weiger je een standaardkeuze te verwarren met een beslissing, dan is dit geschreven voor de operator die capaciteit behandelt als een allocatie, niet als een gewoonte.

drs. Len P. van der Hof

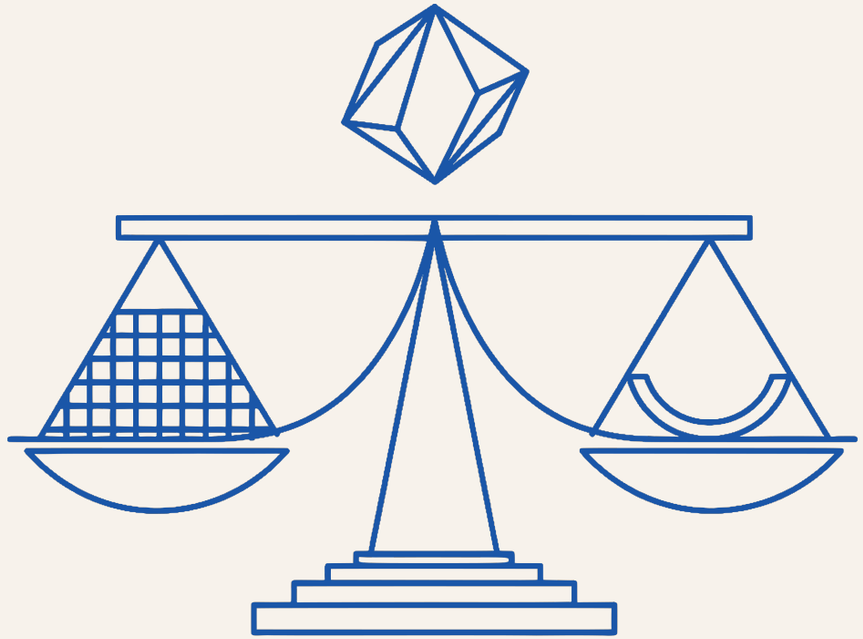
Hij bouwt besturingssystemen voor het strategische zelf. Behandelt ondernemerschap, AI en machine learning, marketing, filosofie, psychologie en gezondheidsoptimalisatie als één engineeringprobleem. MSc in Strategic Entrepreneurship, Rotterdam School of Management, Erasmus Universiteit.

Len P. van der Hof **Het Model- Portfolio**

SftSS

Het Model- *Portfolio*

Hoe je LLM's routeert, mixt en bestuurt als een strategische asset allocator



Len P. van der Hof

Systems for the Strategic Self