Advies: wat pikken we op uit Ludo's tips?

Kort: Jamal kreeg van zijn klant Ludo Baauw een reeks tips over slimmere AI-agents: een gelaagd geheugen (GBrain + Mnemosyne), Hermes als centrale hub, Mixture of Agents (MoA), en een e-mail-gateway. Dit verslag legt elk stuk uit, checkt het tegen wat Haarvisie AL heeft, en geeft een concreet stappenplan. **Conclusie: de tips zijn goed en cutting-edge, maar we slopen de werkende Conductor niet — we pikken de losse pareltjes eruit.**

---

1. Ludo's opzet vs. jouw opzet

Het belangrijkste inzicht uit het gesprek: Ludo en Jamal doen het **net anders**.

	LUDO	JAMAL (NU)
Wie staat vooraan	**Hermes** = centrale hub	**Claude Code** = Conductor
Telegram hangt aan	Hermes	Claude Code direct
Claude gekoppeld via	OAuth (Max-abo)	OAuth (Max-token)
Hermes-rol	de regisseur van alles	los hulpmiddel ernaast (Kore-stem)

Beide werken. Het verschil is puur: wie is de voordeur.

Advies: laat Claude Code je Conductor blijven. Al je geheugen, skills en Telegram-koppeling zitten diep in Claude Code ingebouwd. Hermes ervoor plakken als nieuwe voordeur = veel migratiewerk en risico op kapot maken, met weinig winst. Pik liever de losse sterke onderdelen.

---

2. De vier pareltjes

Pareltje 1 — OAuth i.p.v. API-key AL KLAAR

Ludo: koppel Claude via OAuth (Max-abo), niet via API-key — anders "loop je leeg".
Status: doen we al. Alle bots draaien op een Max-token (OAuth). API-key alleen als noodoplossing.

Pareltje 2 — GBrain + Mnemosyne (gelaagd geheugen) AL KLAAR

- **GBrain** = het brede langetermijngeheugen (de "compiled truth"): diep, doorzoekbaar op betekenis.
- **Mnemosyne** = het ultrasnelle werkgeheugen (sub-milliseconde cache): wat de agent één stap geleden deed.
- Samen = snelle cache + duurzaam langetermijngeheugen.

Status: GBrain staat live op de Conductor en is gekoppeld (MCP). Mnemosyne draait al langer. Het gelaagde geheugen staat dus.

Pareltje 3 — Mixture of Agents (MoA) via OpenRouter NIEUW

Meerdere modellen schrijven parallel een concept (de specialisten), één zwaargewicht (bijv. Claude Opus) leest alles, combineert met de diepe context en schrijft het eindantwoord (de voorzitter).

- **Hoe je de modellen aankoppelt — OpenRouter:** MoA draait op een "virtual model provider". De praktische manier hiervoor is **OpenRouter**: met **één key** krijg je toegang tot GPT-5, DeepSeek en tientallen andere modellen. Dat scheelt losse abonnementen per model.
- **Waarde:** hogere beslissingskwaliteit bij complexe analyse/code.
- ⚠ **Let op:** jaagt het tokenverbruik (en de kosten) flink omhoog.
- **Advies:** één OpenRouter-account aanmaken, MoA-preset erop, en alleen inzetten voor zware taken — niet voor dagelijks werk.

Pareltje 4 — E-mail-gateway (IMAP/SMTP) NIEUW

Een agent die je mailbox uitleest en er nuttige dingen mee doet: triage, labelen, data eruit vissen, of concept-antwoorden opstellen.

⚠ **Veiligheid is hier kritiek:**
- **Nooit** je persoonlijke hoofdmailbox — maak een apart adres (bijv. agent@...).
- Gebruik een **app-wachtwoord**, niet je echte wachtwoord.

- Zet een **whitelist**: de agent reageert alleen op vertrouwde afzenders.
- Credentials in een afgeschermd bestand (Keychain/env), nooit in plaintext.
- **Advies:** begin met alleen-lezen (triage + labelen). Pas automatisch antwoorden ná jouw akkoord.

---

3. Stappenplan (gefaseerd)

1. **Nu klaar:** OAuth + gelaagd geheugen (pareltje 1 en 2).
2. **MoA:** één OpenRouter-account aanmaken (1 key = GPT-5 + DeepSeek + meer). Dan MoA-preset aanzetten, maar alléén als provider voor zware taken. Kostenbewust.
3. **E-mail-gateway:** apart mailadres + app-wachtwoord (zet jij zelf) + whitelist. Eerst read-only triage/labelen, dan pas auto-antwoord na akkoord.
4. **Optioneel — Hermes WebUI:** Hermes heeft een web-interface (standaard localhost, of via Tailscale te benaderen). Handig om live te zien wat er draait.

---

4. Extra uit het gesprek

- **Videocall met Ludo:** hij bood aan zijn setup live te laten zien. De moeite waard — dan zie je precies hoe hij Hermes, Telegram en Claude koppelt.
- **Lokale modellen:** Ludo bouwt een lokaal cluster (Nvidia DGX Spark) om GLM/Qwen lokaal te draaien. Interessant voor later (privacy + geen tokenkosten), maar fors qua hardware. Nu geen prioriteit.
- **Geheugen/compressie-tip:** zet de context-compressie niet te agressief, en lees bij elke nieuwe sessie eerst het geheugen uit. Doen we al (start-skill + memory-uitlezen).

---

5. Advies in één zin

De architectuur van Ludo klopt en is cutting-edge, maar sluit naadloos aan op wat je al hebt. **Geen sloop — cherry-pick.** OAuth en gelaagd geheugen staan. MoA en de e-mail-gateway zijn de echte nieuwe stappen; bouw ze gefaseerd en kostenbewust.

Verslag door Conductor, 3 juli 2026.