Overzicht vlam.ai - 21-10-2025

Dit document bevat antwoorden op veelgestelde vragen over vlam.ai. Omdat we agile werken, kunnen we snel leveren, maar dit betekent ook dat details regelmatig veranderen. De titel vermeldt wanneer dit document voor het laatst is bijgewerkt.

Voor de meest actuele informatie kun je terecht in de samenwerkingsruimte, waar de laatste versie van dit document beschikbaar is.

Wat is vlam.ai?

Vlam.ai kan het beste worden gezien als AI-as-a-Service (AIaaS). AI-as-a-Service bij SSC-ICT betekent dat we geavanceerde AI-modellen, waaronder tekstgeneratie, transcriptie en beeldherkenning, beschikbaar stellen en hosten vanuit ons eigen GPU-cluster. Momenteel doen we dit vanuit de Koningskade, maar in de toekomst vanuit ODC Rijswijk. De dienstverlening van vlam.ai bestaat uit twee hoofdcategorieën:

1. **AI-modellen via een API:** Gebruikers krijgen via gestandaardiseerde API's toegang tot AI-modellen, waardoor ze eenvoudig AI-functionaliteiten kunnen toevoegen aan hun applicaties. Hiervan is het de intentie dat we dit Rijksbreed kunnen aanbieden. Hierbij bieden we modellen in twee soorten categorieën:
    a. **standaardmodellen:** dit zijn modellen die we te allen tijde beschikbaar hebben. Hierdoor zijn ze goedkoop en is er veel capaciteit voor beschikbaar.
    b. **custom-modellen:** dit zijn maatoplossingen, wanneer een afnemer een eigen model wilt laten hosten voor een applicatie. [Dit wordt op dit moment nog niet ondersteund]
2. **Applicaties met geïntegreerde AI-capaciteiten:** Direct bruikbare toepassingen met ingebouwde AI-functies, ontworpen voor directe inzet door eindgebruikers. De intentie is om dit alleen binnen ons verzorgingsgebied aan te bieden. Hier bij richten we ons in eerste instantie op twee concrete applicaties:
    a. **vlam-chat:** een veilig alternatief voor ChatGPT, waarmee gebruikers vertrouwelijke en gevoelige informatie veilig kunnen verwerken met behulp van AI. Met vlam-chat kun je bijvoorbeeld eenvoudig teksten samenvatten, vertalen of tekstvoorstellen schrijven.
    b. **vlam-search:** een intelligente zoekoplossing om grote hoeveelheden niet-gevoelige data (rubricering TLP Clear of TLP Green) snel en efficiënt bevraagbaar te maken.

> vlam.ai ontwikkelt geen custom AI-modellen en biedt ook niet de mogelijkheid om AI-modellen te (her)trainen. Verder ontwikkelt vlam.ai geen applicaties anders dan vlam-chat en vlam-search en is vlam.ai geen aanbieder van een containerplatform. We kunnen dus ook geen klant-specifieke applicaties hosten, hiervoor zal je bij SSC-ICT of je eigen IT-aanbieder moeten aankloppen.

Handige links:

- Status pagina van server en diensten.
- SSC ICT AI Team / vlam.ai Team: sscictai@minbzk.nl
- vlam-chat (inschrijving via SSO is veriest, check onder hoe je dat regelt): https://chat.demo.vlam.ai/

Planning

Aan de onderstaande planning zijn geen rechten te ontlenen. We doen ons best om alles zo snel mogelijk uit te rollen.

Wat	Wanneer
• Demo-release van vlam-chat (totaal 50-100 gebruikers) • Demo-release van vlam-chat functionaliteit: ○ Vraag & Antwoord • vlam-search voor POC's en Pilots, beperkte capaciteit • API's naar standaardmodellen voor POC's	Q3 2025

• Verdere uitbreiding van vlam-chat (totaal 10000+ gebruikers)
  ○ functionaliteiten uitbreiden met:
  ▪ Documenten uploaden
• API's naar standaardmodellen voor POC's & Pilots
	Q4 2025

• Lancering dienstverlening (chat & API) voor afnemers die een financiële commitment hebben afgegeven
• Verdere uitbreiding van vlam-chat
  ○ functionaliteiten uitbreiden met:
  ▪ Input guardrails
  ▪ Standaard prompts
• GPU cluster gebruiken vanuit ODC Rijswijk
• SOC-monitoring
• BBN2 / Dep-V
	Q1 2026

• Verdere uitbreiding van vlam-chat:
  ○ functionaliteiten uitbreiden met:
  ▪ Follow-up vragen
  ▪ Speech-to-text
  ▪ Gebruikershistorie
	Q2 2026

• Verdere uitbreiding van vlam-chat:
  ○ functionaliteiten uitbreiden met:
  ▪ Web-search
	Q3 2026

• Verdere uitbreiding van vlam-chat:
  ○ functionaliteiten uitbreiden met:
  ▪ Toegankelijkheid
	Q4 2026

Kosten

Deelname aan de vlam.ai is momenteel kosteloos, maar we hebben wel beperkte capaciteit en moeten daarom prioriteren. (dit word intern nog geëvalueerd, hier kunnen dus nog dingen wijzigen!)

De prijzen voor de diensten van vlam.ai wanneer deze in productie gaat zijn nu nog onbekend.

Prioritering

We hebben slechts beperkte capaciteit en moeten flinke investeringen doen om deze op te schroeven. We willen zoveel mogelijk partijen helpen, ook buiten ons verzorgingsgebied. Bij de uitrol, zowel van de MVP als productie, stellen we dan ook de volgende prioriteiten:

1. Afnemers die een **financiële commitment** hebben afgegeven voor de dienstverlening in 2026.
2. Afnemers zonder financiële commitment **binnen** het verzorgingsgebied van SSC-ICT.
3. Afnemers zonder financiële commitment **buiten** het verzorgingsgebied van SSC-ICT.

Deelnemen

Tot 31 december kunnen partijen zich bij ons aanmelden om te helpen testen van vlam-api en vlam-chat. Daarna zullen we de diensten nog steeds beschikbaar stellen als een product. Het doel van deelname voor ons is voornamelijk om ervaring op te doen voor productie in 2026.

De deelnameprocedure verschilt per ministerie. De meeste potentiële gebruikers kunnen rechtstreeks contact met ons opnemen. Echter, voor de onderstaande ministeries zijn er andere afspraken gemaakt:

Ministerie van Binnenlandse Zaken:

1. **vlam-api:**
    - Je kunt rechtstreeks contact opnemen met postbusteamai@minbzk.nl om toegang te krijgen tot vlam-api. Ze zullen je aanvraag zo snel mogelijk in behandeling nemen en wijzen je op de stappen die je moet zetten.
    - Zij kunnen naast onze standaard aansluitvoorwaarden aanvullende eisen stellen.
2. **vlam-chat**
    - Om toegang te krijgen tot vlam-chat, moet je contact opnemen met postbusteamai@minbzk.nl. Zij beslissen of jullie onderdeel kunnen worden van de BZK-pilot voor vlam-chat.
3. **vlam-search (Momenteel accepteren we geen aanvragen meer)**
    - Je kunt hiervoor ook rechtstreeks contact opnemen met postbusteamai@minbzk.nl. Ze zullen je aanvraag zo snel mogelijk in behandeling nemen.
    - Tijdens het eerste verkennende gesprek kan een vertegenwoordiger van ons aanwezig zijn.
    - Na afloop van dit gesprek kan het BZK AI-team beslissen of jullie casus daadwerkelijk wordt geïmplementeerd en geven ze dat aan ons door.

POCs, Pilots en Productie

In dit document worden de termen POCs, Pilots en Productie gebruikt. Hieronder volgen de definities die wij hanteren voor deze termen:

- Een Proof-of-Concept (POC) is een project dat nog in de ontwikkelfase zit en geen gebruikers heeft. Je zal het ongetwijfeld wel testen maar je draait geen pilot met gebruikers.
- Een Pilot is een project met een beperkt aantal gebruikers, maximaal 25. Het doel van een Pilot is om voor te bereiden op de productiefase.
- Productie verwijst naar een project dat op grote schaal wordt uitgevoerd met een significant aantal gebruikers. Wij verwachten pas vanaf 2026 in staat te zijn om projecten in productie te faciliteren.

Registeren tenzij, bij Pilots en Productie

Een cruciaal aspect van vlam.ai is het ondersteunen van compliance. Om dit te bereiken, volgen wij een "*registreren tenzij*" beleid voor zowel Pilots als Productie. Dit houdt in dat wanneer je een project uitvoert dat zich in de Pilot- of Productiefase bevindt, wij verwachten dat je het registreert in het algoritmeregister, tenzij er valide redenen zijn om hiervan af te zien.

Wij zijn vastbesloten om jullie zo veel mogelijk te ondersteunen bij dit registratieproces, om ervoor te zorgen dat jullie projecten soepel en compliant verlopen.

vlam-api

Assortiment standaardmodellen

Op het moment bieden we alleen api's aan uit ons assortiment standaardmodellen. Pas vanaf Q1 2026 verwachten we ook custom-modellen te kunnen draaien, mits er voldoende capaciteit beschikbaar is.

We bieden momenteel twee standaard API-functionaliteiten aan: chat en transcribe. Onder de motorkap draaien daar nu respectievelijk Mistral Medium 3 en WhisperX voor — maar dat zijn **niet** vaststaande keuzes.

Het vlam.ai team kiest zelf welke onderliggende modellen we gebruiken, en we wisselen die waar nodig uit om de kwaliteit van onze chat- en transcriptiefunctionaliteit zo goed mogelijk te houden. Je kunt er dus **niet** op rekenen dat het altijd precies dezelfde modellen blijven.

Waar je wél op kunt rekenen:

- **Het API-format blijft hetzelfde**. Dus hoe je een request doet en wat je terugkrijgt verandert niet.
- **De kwaliteit blijft gelijk of beter voor dat specifieke doel.** Dit betekent ook dat als je bijvoorbeeld chat gebruikt voor iets anders dan een chatinterface om de gemiddelde ambtenaar mee te ondersteunen, bijvoorbeeld door het te integreren in een specifieke workflow, dat we geen garanties bieden dat die workflow blijft werken nadat wij het model veranderen.

Door de tijd heen zullen we het assortiment standaardmodellen verder uitbreiden, maar tot die tijd is dit belangrijk om dit in gedachte te houden.

We zijn in ieder geval van plan de volgende modellen via API te gaan bieden met de tijd heen (hieraan kan geen rechten worden ontleend). Wanneer deze beschikbaar zijn zullen we dit bekend maken op de Samenwerkruimte:

- Mistral embed model
- Mistral OCR model
- Mistral Small model

Huidige Beschikbare Modellen via de API

Chat: taalmodel bedoeld voor chatinterfaces gericht op de gemiddelde ambtenaar		
Specificaties	**Goed in**	**Minder goed in**
• Momenteel: **Mistral Medium 3.1** • OpenAI compatible API-format • Maximale context: 120.000 tokens	• Nederlands en Engels • Creatieve schrijftaken • Samenvatten • Vertalen	• Codegeneratie • Feitelijke kennis

Transcribe: transcriptiemodel		
Specificaties	**Goed in**	**Minder goed in**
• Momenteel: **WhisperX**	• Transcriberen • Tijdscodering per woord • Speaker diarization (Speakers herkenning)	• Audiobestanden met lage geluidskwaliteit • Speaker diarization bij veel sprekers, overlappende spraak of meerdere sprekers met hetzelfde toon

Voorwaarden voor deelname

- We bieden op het moment alleen API's aan voor POC's en Pilots **die voldoen aan de aansluitvoorwaarden.**
- We hanteren een **Fair Use Beleid:**

  Je krijgt toegang met de volgende basisspecificaties (gebaseerd op Mistral Medium 3):

  - Maximum context window: 120.000 tokens
  - Maximaal gelijktijdige verzoeken: 5
  - Maximale queue size: 100
  - Tokens per dag: Geen limiet

  Deze waarden zijn ruimschoots genoeg voor de meeste POC's en testomgevingen.
- Bij beperkte capaciteit of overmatig gebruik zullen we capaciteit voor het project inperken. Denk aan minder verzoeken per seconde of kortere context window.
- We garanderen een uptime van minimaal 95%. Als we naar productie gaan is dit minimaal 98%.
- API-keys verlopen automatisch na 6 maanden tenzij anders verzocht met een goed onderbouwen reden.

Stappenplan voor deelname vlam-api

Om deel te nemen aan vlam-api, moet je de volgende stappen doorlopen:

Stap 1: Inschrijven

- Download het Word-document "Inschrijfformulier vlam-api" uit de samenwerkingsruimte.
- Vul het formulier volledig in, hiervoor is een privacy en security advies vereist. Meer uitleg daarover vind je in de aansluitvoorwaarden in de samenwerkingsruimte.
- Zet je handtekening en stuur het naar sscictai@minbzk.nl
- Door het ondertekenen van het inschrijfformulier ga je akkoord met de aansluitvoorwaarden. Deze voorwaarden zijn ook beschikbaar in de samenwerkingsruimte. Er zijn twee versies van de aansluitvoorwaarden: een voor de POC-fase met minder strenge eisen en een voor de Pilot-fase. Echter zijn we momenteel bezig met het finaliseren van de aansluitvoorwaarden voor de Pilot-fase, deze volgt binnenkort.

Stap 2: API-key ontvangen

- Voor deelname moeten de stappen hiernaast doorlopen worden.

- Na ontvangst en verwerking van je inschrijfformulier, ontvangen de door jou opgegeven personen de API-sleutel per e-mail.
- Bij de API-sleutel wordt een notebook geleverd met voorbeelden van het gebruik van vlam-api.

vlam-chat

- Vlam-chat is ontwikkeld als een veilig alternatief voor ChatGPT.
- Momenteel draait Mistral Medium onder de motorkap.
- Vlam-chat is te bereiken via chat.demo.vlam.ai. We hebben de intentie om alle AI-coördinatoren toegang te verlenen tot de applicatie. Hoewel de volledige uitrol nog in ontwikkeling is, biedt dit de mogelijkheid om vlam-chat te verkennen.
- We willen vlam-chat gecoördineerd testen in een Pilot per ministerie. Bij individuele verzoeken voor toegang zullen we die doorverwijzen naar de AI-coordinators van het betreffende ministerie. Zij kunnen vervolgens beslissen om jullie wel of niet te betrekken bij de Pilot.
- Een deelnameformulier met bijbehorende gebruiksvoorwaarden volgt binnenkort.

vlam-search

- Vlam-search is een zoekoplossing gebaseerd op Retrieval Augmented Generation (RAG). Deze zoekoplossing stelt de gebruiker in staat om een dataset met een grote hoeveelheid documenten te bevragen met AI.
- We hopen tot Q1 2026 3-4 casussen te implementeren, die ons het meeste inzicht geven in deze technologie.
- Bij deelname hebben we volgende vereisten:
    - **Data Classificatie:** Je dataset dient TLP Clear of TLP Green geclassificeerd te zijn. Meer informatie hierover is te vinden op: Traffic Light Protocol (TLP) | Aansluiten en samenwerken | Nationaal Cyber Security Centrum
    - **Testset:** Voor deelname is een testset van 50 vragen nodig, specifiek voor jouw dataset.
    - **Evaluatie:** Je hebt 5-10 domein experts nodig die de antwoorden van vlam-search op de 50 testvragen kunnen beoordelen. Dit proces kan iteratief zijn om de kwaliteit te waarborgen. Zorg ervoor dat de testers inhoudelijk deskundig zijn.
    - **Vlam-chat account:** Alle gebruikers die vlam-search willen gebruiken in de pilot/productie, hebben een actief vlam-chat account nodig. Voor de evaluatie fase is dat niet nodig.
- In de toekomst wanneer Embedding en OCR modellen ook beschikbaar zijn via de API is er mogelijkheid om zelf een RAG pipeline te bouwen.

Stappenplan voor deelname vlam-search [Momenteel inschrijving is dicht]

Gezien het grote aantal verzoeken, zullen we deze eerst verzamelen tot 1 juli 2025, voordat we de definitieve selectie maken. Om deel te nemen aan vlam-search moet je de volgende stappen doorlopen:

- Download het Word-document "Inschrijfformulier vlam-search" uit de samenwerkingsruimte.
- Vul het formulier in en stuur het naar sscictai@minbzk.nl.
- Wij zullen bevestigen dat we het hebben ontvangen.