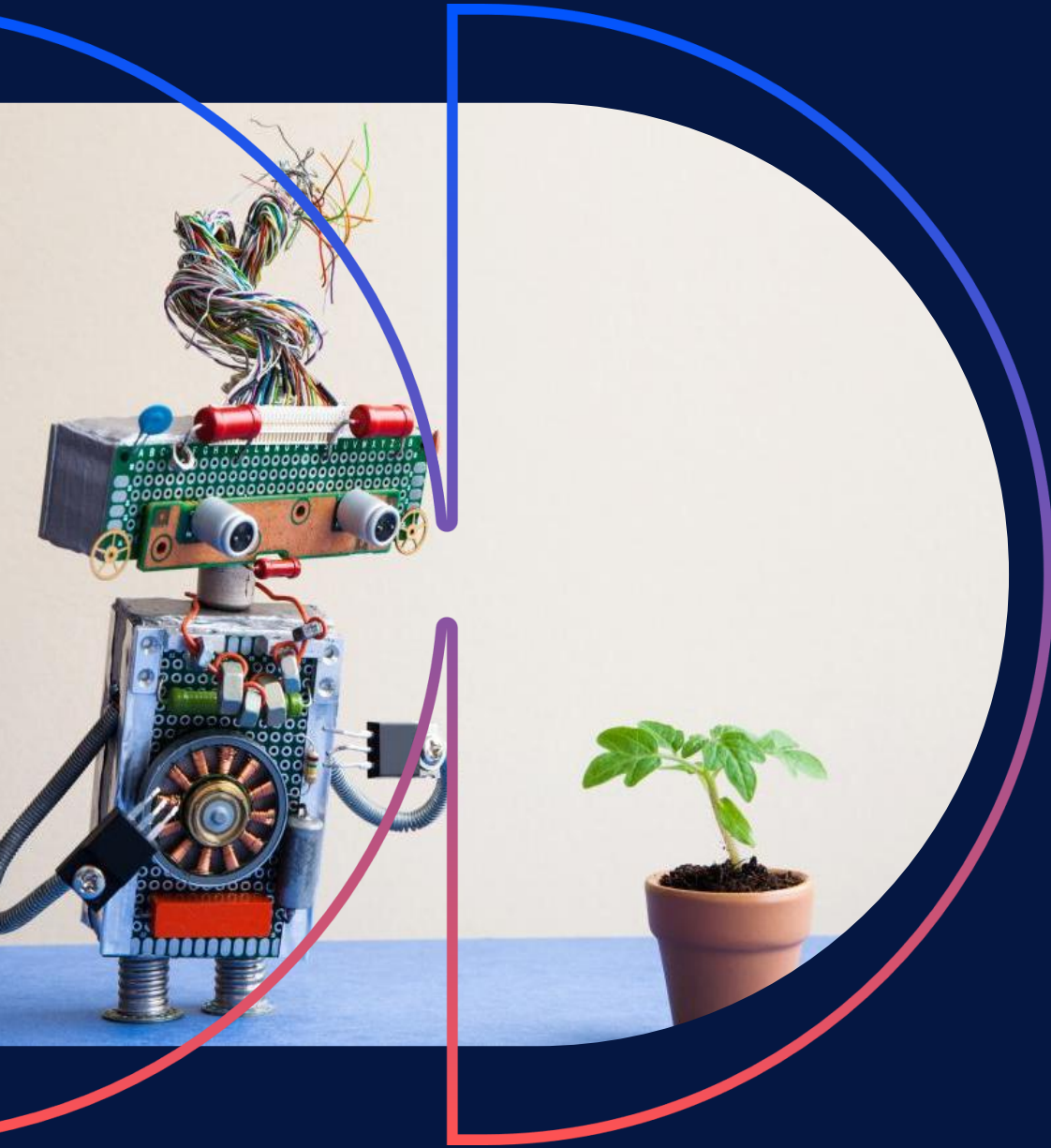




nēosoft

REX sur la création de notre IA Générative Open Source - Néogen

11/03/2026

Témoignage sur la construction de notre projet Open Source NéoGen

► ENJEUX & STRATÉGIE

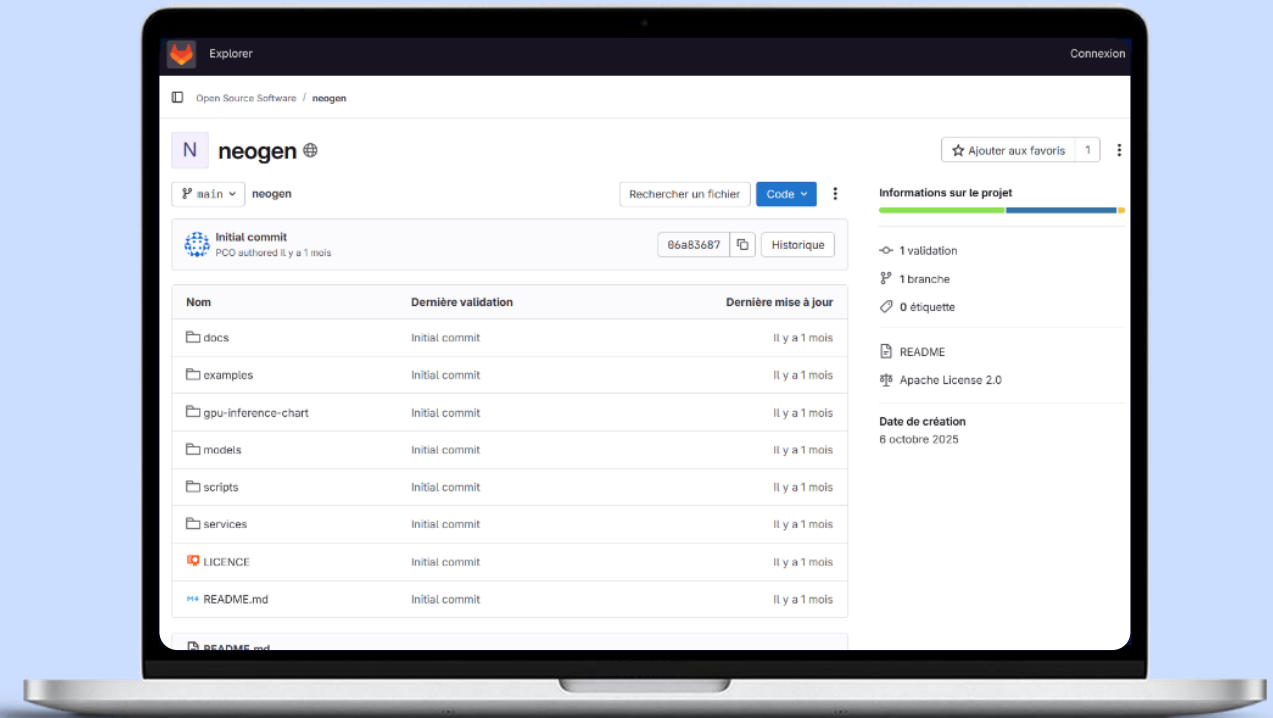
- Pourquoi disposer d'une IA Générative **privative** ?
- Quels **critères** de choix ?

► CONSTRUCTION

- Quelle **architecture** retenir ?
- Comment opter pour une **approche frugale** ?
- Quels **usages** et quels leviers d'adoption ?

► CONCLUSION

- Quel bilan en termes de **souveraineté** ?
- Pourquoi une **ouverture** communautaire ?



Si vous voyez ce signe,
c'est le moment de **voter**

Le cahier des charges de notre plateforme

nēosoft

Un projet au cœur de la souveraineté numérique

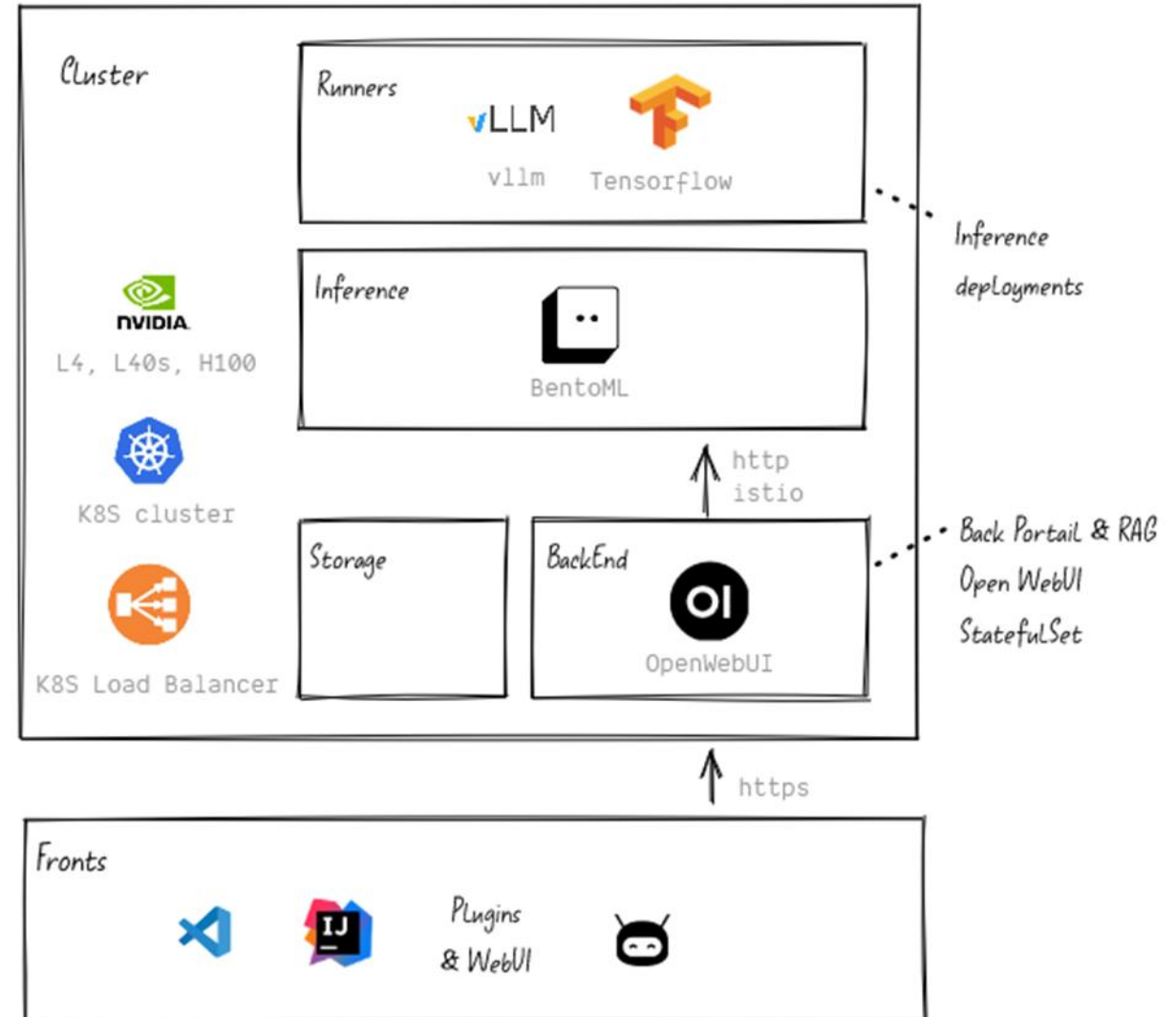


- ▮ Outil de **substitution** à ChatGPT, Claude, Gemini...
 - Interface de Chatbot + API
 - Adaptée à nos usages : génération de texte, codage, ...
- ▮ Solution **souveraine**
 - Agnostique des fournisseurs
 - A base de composants **open-source**
 - Exploitation de modèles libres, spécialisés et **frugaux**
 - Hébergé chez OVH à Gravelines sur un tenant privé
- ▮ Respectueuse de la **confidentialité**
 - Pas de conservation des données de prompt hors historique personnel
 - Pas d'exploitation des prompts

Définition de l'architecture

Une solution privée, administrée par notre équipe

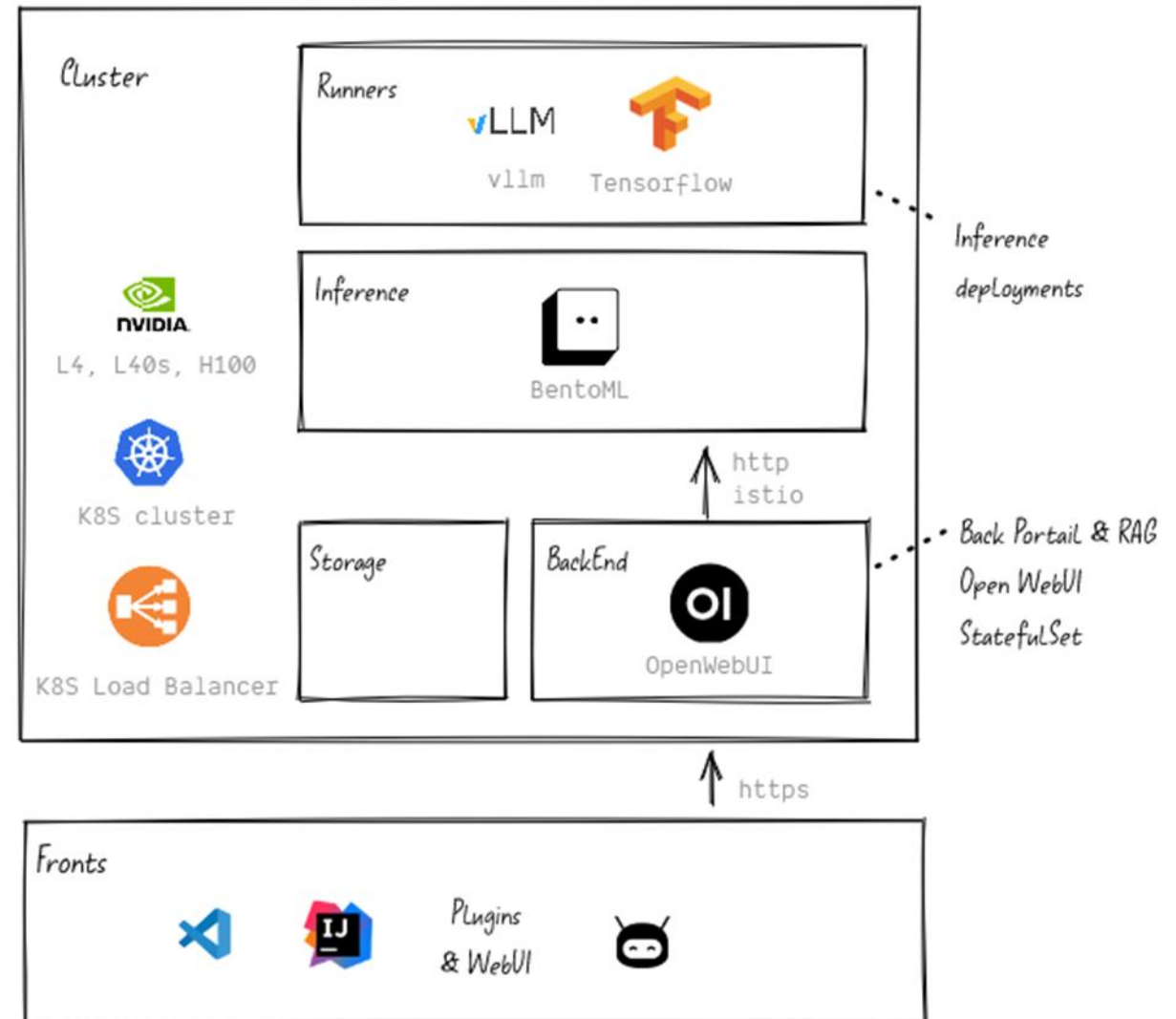
1. Hébergement sur un **tenant privé** chez OVH
capacité à monter en offre SecNumCloud
2. Infrastructure nominale sur **6 serveurs**
1 x H100-760 ; 1 x H100-380 ; 2 x L40s-90 ; 1 L4-90 ; 1 x B3-8
3. Infrastructure **modulable**
offre IaaS
4. Utilisation de puces **L40s** et **H100**
pour l'opérateur MIG (Multi-Instance GPU)
5. Activation des services en **heures ouvrées**
Extinction des nœuds d'inférence entre 19h et 7h



Définition de l'architecture

Une solution privée, administrée par notre équipe

6. Sélection de **modèles proportionnés**
(open-source ou open-weight)
7. Exposition via une **API compatible Open AI**
standardisation du marché
8. Optimisation de l'inférence (quantization)
et de la gestion de la mémoire (usage de vllm)
9. Déploiement d'un chat interne fédérant l'accès
aux différents modèles
10. Intégration aux IDE via des plugins
agnostiques du fournisseur accédant aux
modèles déployés



L'ouverture en **Open Source**

néosoft

Opter pour une approche responsable du Numérique



Bénéficier d'une **communauté** active qui porte une **innovation** continue de très haut niveau



Contribuer aux travaux communautaires pour **capitaliser** ensemble



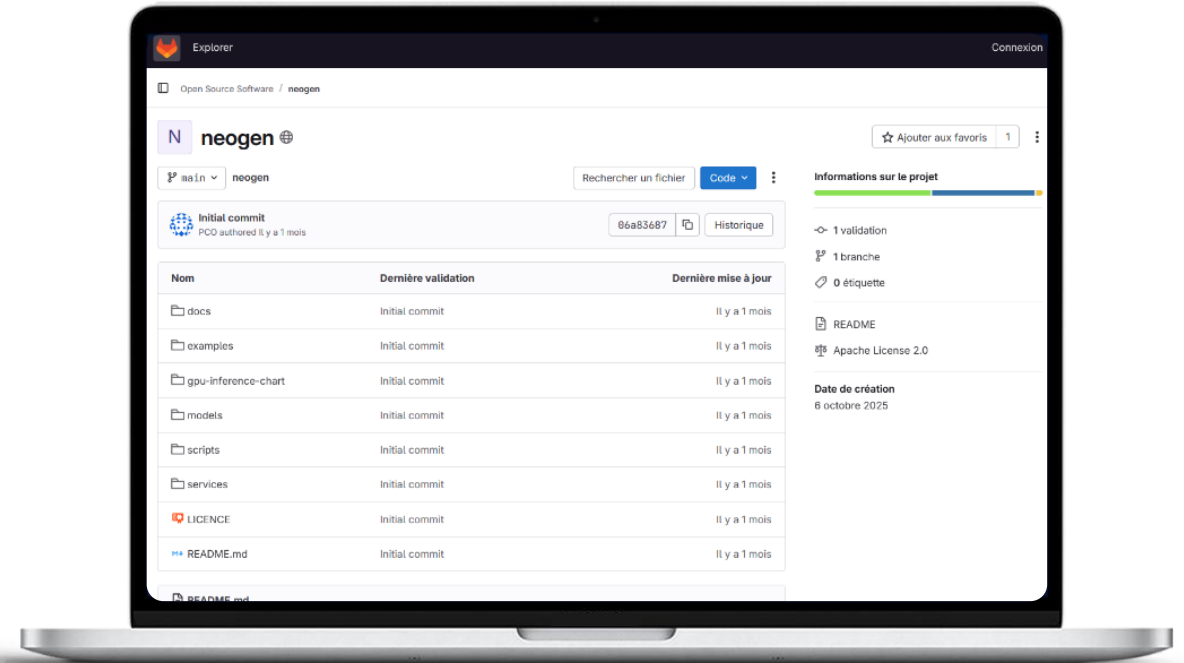
Ouvrir nos travaux pour **soutenir** la réutilisation et la création de **communs numériques**



Limiter l'impact en réutilisant plutôt que de reconstruire ou de réentraîner des modèles



Apporter une approche orientée sur le **conseil** et l'**intégration** car nous ne sommes pas éditeurs



nēosoft

Découvrez
NéoGen by Néosoft,
l'IA générative éthique,
frugale et open source

