

# Idarati

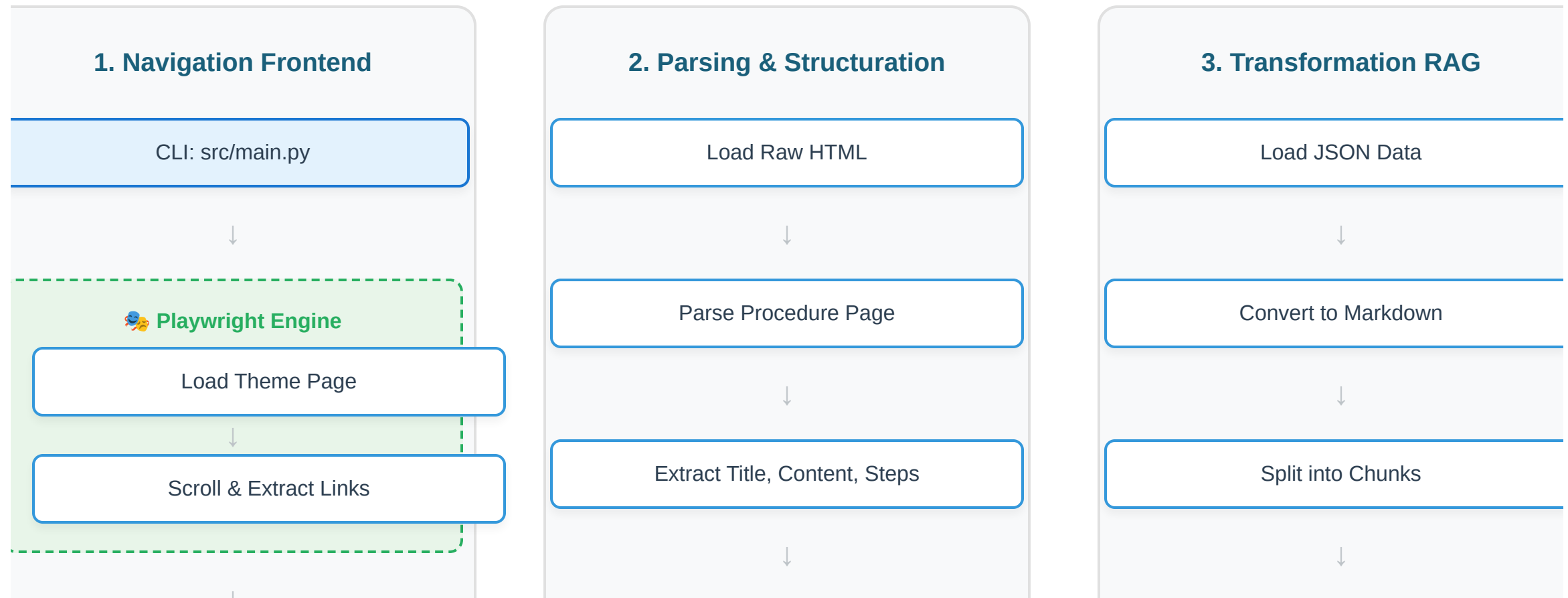
Assistant Administratif Marocain IA

RAG • QLoRA • Modal • Supabase

## Pipeline de Données (End-to-End)



## Pipeline de Collecte de Données (Scraping)

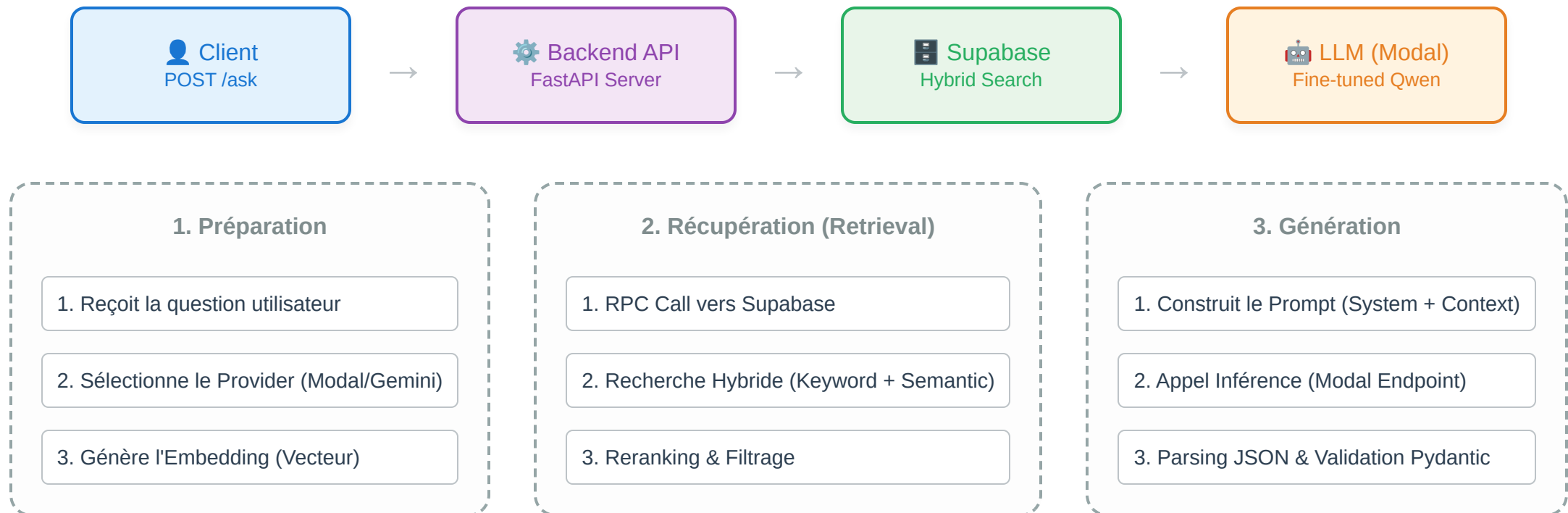


Utilisation de Playwright pour l'extraction frontend dynamique.

# Architecture RAG (Le Cerveau)

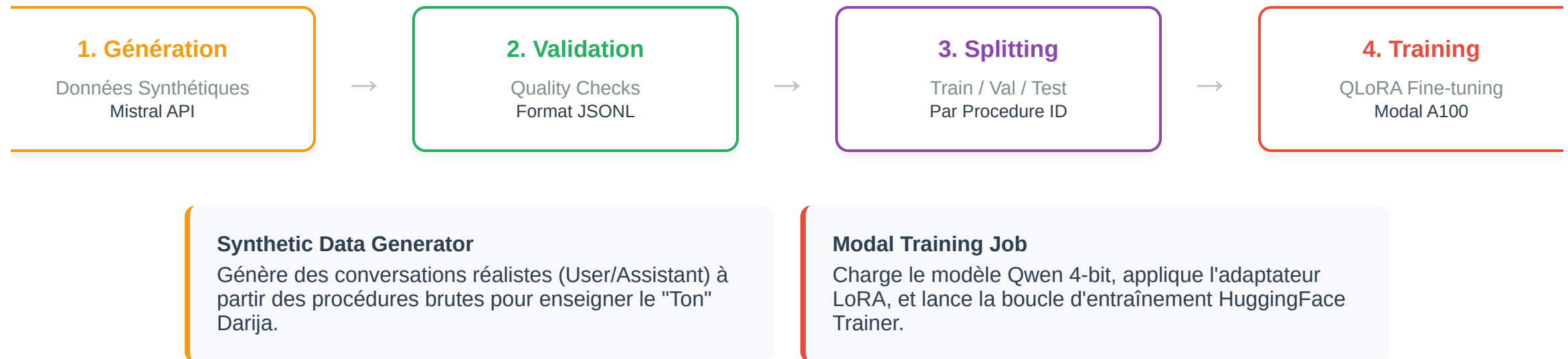
04

## Architecture RAG (Le Cerveau)



Le flux d'une question utilisateur à travers l'API et le LLM.

## Pipeline d'Entraînement (AI Engine)




## Architecture du Fine-tuning

De la donnée brute à l'adaptateur LoRA

### 1. Pipeline

 **Données (101 ex.)**  
Darija + Arabe




 **Base Model**  
Qwen2.5-7B (4-bit)



 **QLoRA Training**  
500 steps (A100)

### 2. Architecture QLoRA

 **Modèle Gelé**  
**4 Go**  
Poids figés (4-bit)



 **Adaptateur**  
**154 Mo**  
Poids entraînés

### 3. Résultats

 **Avant (Base)**

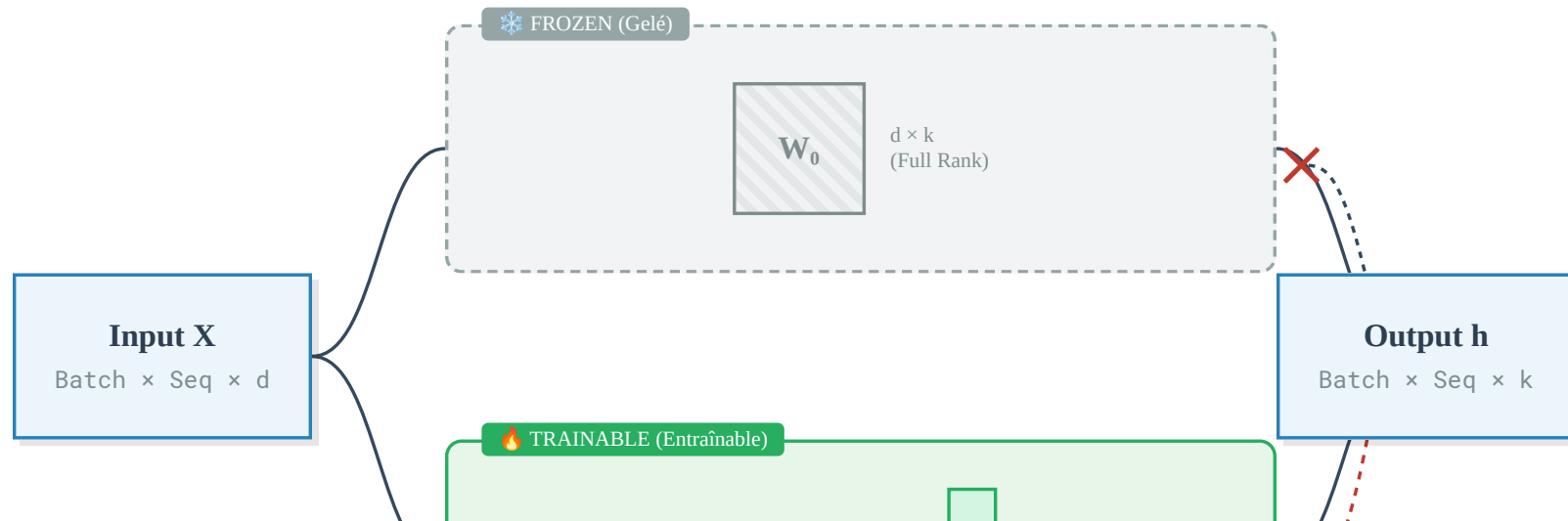
要更新您的... (Chinois/Formel)

 **Après (Idarati)**

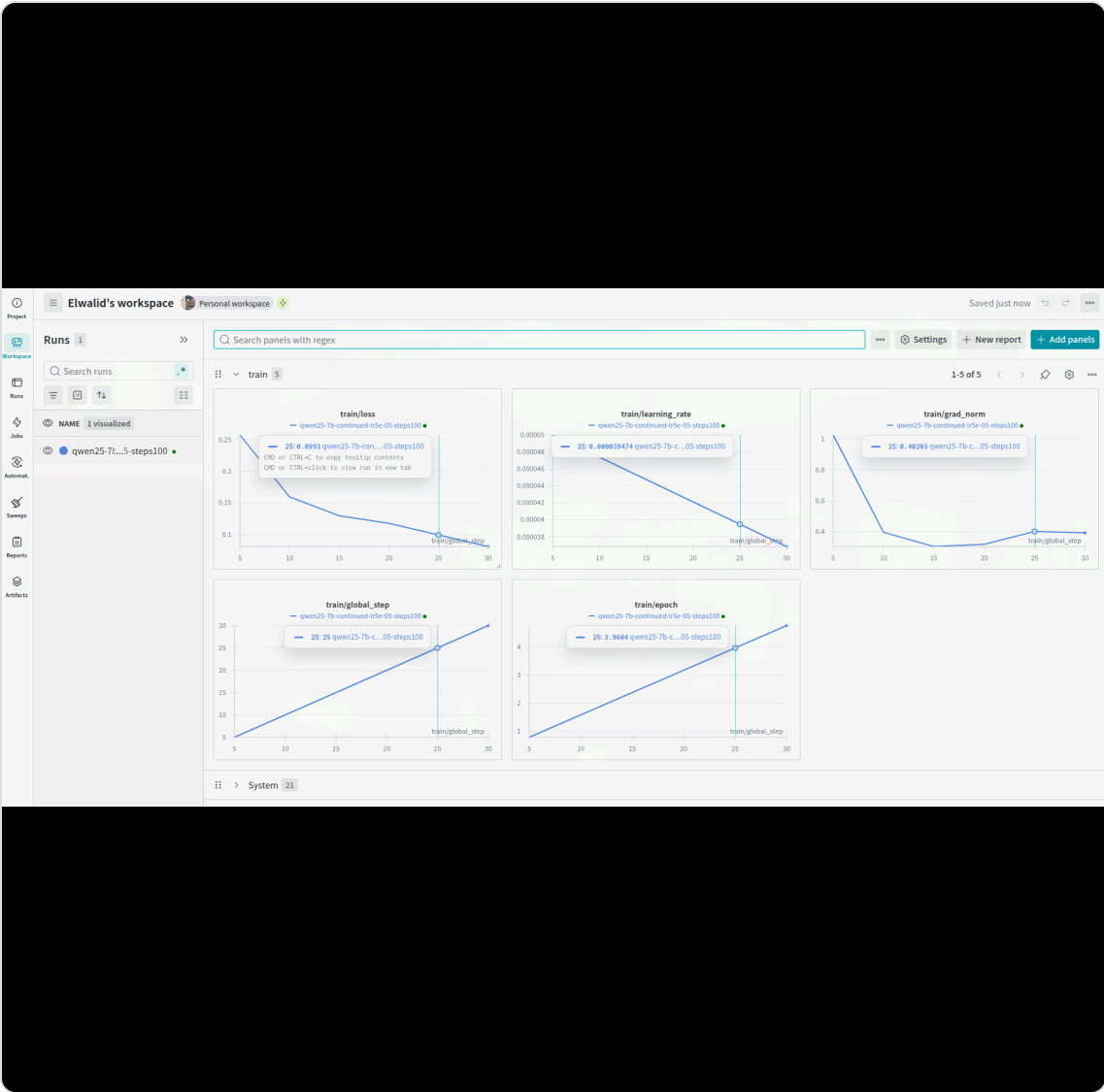
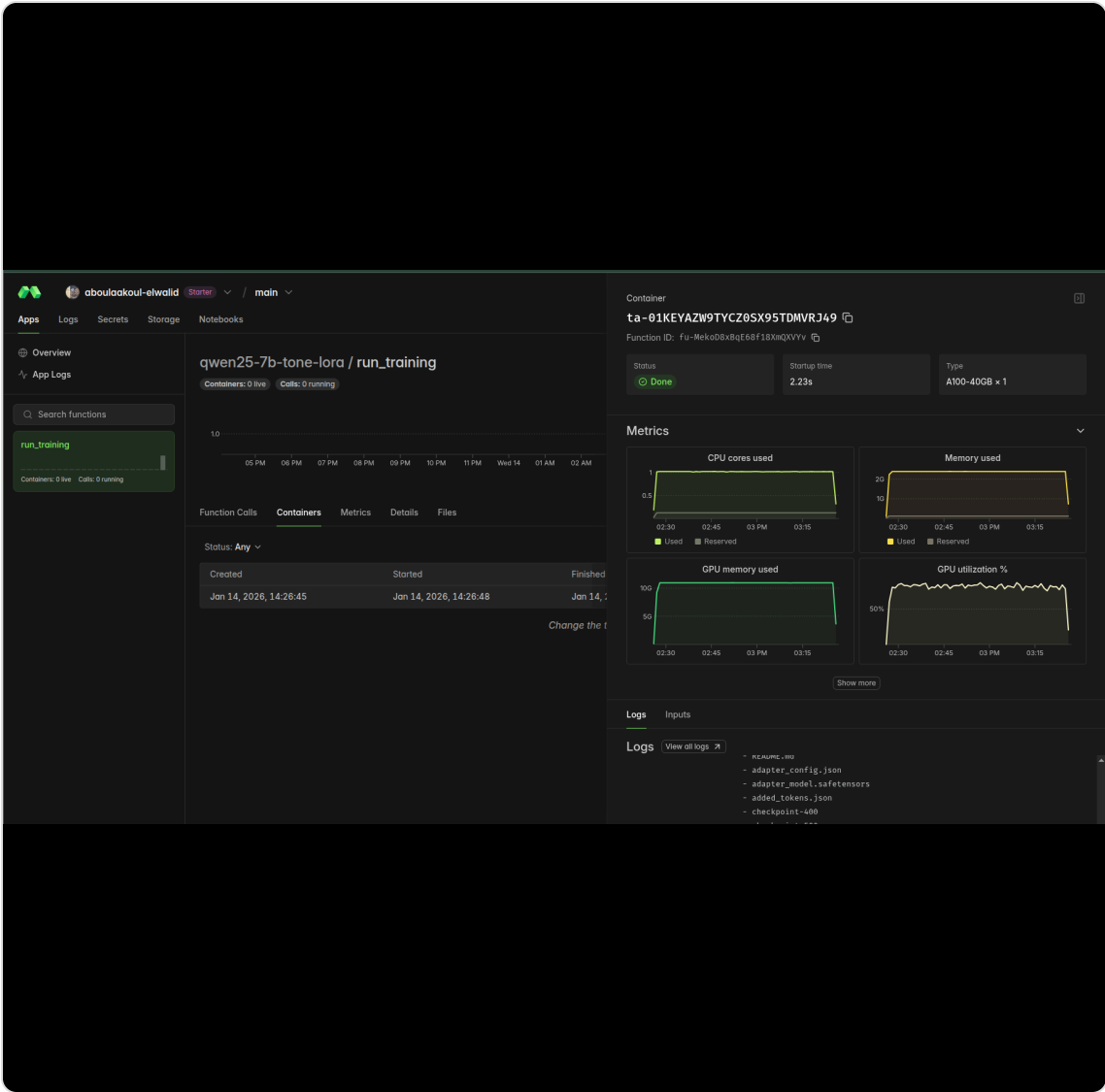
أهلاً! لتجديد البطاقة ... (Darija/Amical)

## LoRA: Low-Rank Adaptation

Schéma technique du mécanisme de fine-tuning



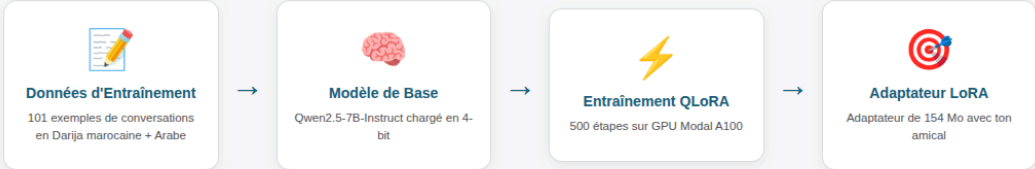
# Preuve d'Entraînement



Logs de démarrage et courbes de perte (Weights & Biases).



## Pipeline de Fine-tuning

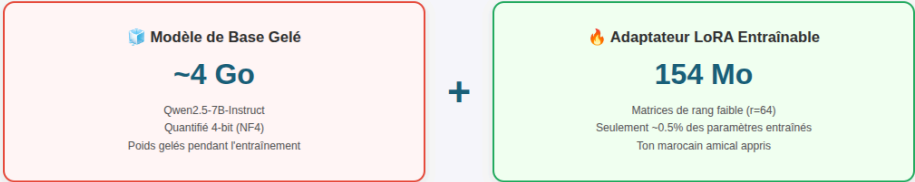


Configuration d'Entraînement	
Taux d'Apprentissage	2e-4
Taille du Batch	4
Accumulation de Gradient	4 étapes
Étapes Totales	500
Étapes de Warmup	50
Longueur Max. Séquence	512

Configuration QLoRA	
Rang LoRA (r)	64
Alpha LoRA	128
Dropout LoRA	0.05
Quantification	4-bit NF4
Type de Calcul	float16
Double Quantification	Activée

Modules Ciblés	
q_proj	✓
k_proj	✓
v_proj	✓
o_proj	✓
gate_proj	✓
up_proj / down_proj	✓

## Architecture QLoRA



Pourquoi QLoRA ?	
Efficace en Mémoire	~8 Go VRAM
Entraînement Rapide	~30 min
Petit Adaptateur	154 Mo vs 14 Go
Pas de Modification du Modèle	Sûr
Déploiement Facile	Fusion ou Échange

Infrastructure	
Plateforme	Modal
GPU	NVIDIA A100 80 Go
Temps d'Entraînement	~30 minutes
Coût	~1,50 \$
Suivi	Weights & Biases

Le modèle répond en Darija/Arabe amical (vs Chinois/Anglais avant).

### Résultats du Fine-tuning

#### ✗ Avant le Fine-tuning (Modèle de Base)

Utilisateur : كيفأجد البطاقة الوطنية؟

...要更新您的国民身份证，您需要前往

(Répond en chinois, formel, générique)

#### ✓ Après le Fine-tuning (Avec Adaptateur)

Utilisateur : كيفأجد البطاقة الوطنية؟

أهلاً لتجديد البطاقة الوطنية، تحتاج إلى التوجه للمصلحة المختصة مع الوثائق اللازمة. هل البطاقة الحالية منتهية الملاحية أم هناك سبب آخر للتجديد؟

(Arabe amical, pose des questions de clarification)

#### 🌟 Améliorations Obtenues

Langue	Arabe/Darija ✓
Ton	Amical & Chaleureux ✓
Comportement	Pose des Questions ✓
Précision	Admet l'inconnu ✓
Format	Simple & Clair ✓

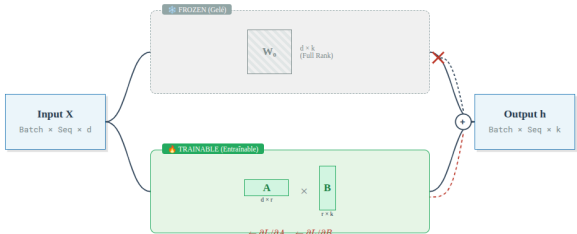
#### 📋 Prompt Système

أنت مساعد إداري مغربي ودود. تساعد المواطنين في الإجراءات الإدارية بأسلوب بسيط ومباشر. أجب بحرية بسيطة وسهلة الفهم. كن مختصراً ومفيداً. إذا لم تعرف معلومة، قل ذلك بوضوح.

(Tu es un assistant administratif marocain amical. Réponds en arabe simple. Sois concis et utile.)

### LoRA: Low-Rank Adaptation

Schéma technique du mécanisme de fine-tuning



Forward:

$$h = W_0 x + \frac{\alpha}{r} B A x$$

Update Rule:

$$A \leftarrow A - \eta \frac{\partial \mathcal{L}}{\partial A}$$

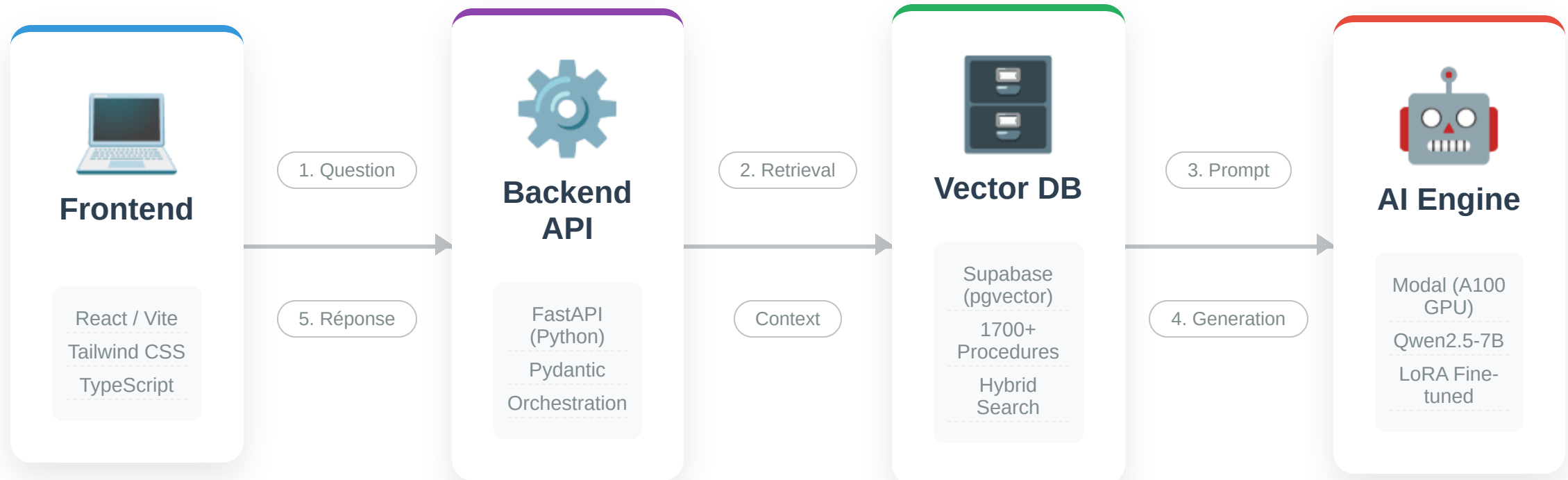
#### 🔧 Initialisation

- Matrix  $A$ : Distribution Gaussienne (Aléatoire)
- Matrix  $B$ : Zéros (Pour que  $\Delta W = 0$  au début)

#### 🌟 Avantages

- Etablissement de rendement 8.5% des paramètres
- Pas de licence supplémentaire à l'inférence (Il suffit  $W_0 = BA$ )

## Architecture Globale Idarati



# Merci

Questions ?