

Cairn.Info

Proposition Commerciale

Etude déploiement de LLMs en local

**Client** : Cairn.Info

**Date** : 11 octobre 2024

Version : v1

Destinataire :

|  |  |
| --- | --- |
| **Cairn.Info** | **Jean-Baptiste De Vathaire**  Courriel : jb.devathaire@cairn.info |

Auteur :

|  |  |
| --- | --- |
|  | **Vincent Nibart**  Téléphone : +33 (0)6 25 68 08 12  Courriel :[vincent.nibart@kairntech.com](mailto:vincent.nibart@kairntech.com) |

[1. RESUME EXECUTIF 4](#_Toc179548968)

[2. HYPOTHESES 4](#_Toc179548969)

[2.1. Résumé automatique 4](#_Toc179548970)

[2.2. Question-Réponse RAG 5](#_Toc179548971)

[3. PRESTATIONS DE SERVICE 5](#_Toc179548972)

[3.1. Lotissement 5](#_Toc179548973)

[3.2. Synthèse des livrables 6](#_Toc179548974)

[3.3. Planning 7](#_Toc179548975)

[3.4. Intervenant Kairntech 7](#_Toc179548976)

[4. PROPOSITION FINANCIÈRE 7](#_Toc179548977)

[5. CONDITIONS GÉNÉRALES DE VENTE 8](#_Toc179548978)

[5.1. Facturation et paiement 8](#_Toc179548979)

[5.2. Validité de la proposition 8](#_Toc179548980)

[6. RENSEIGNEMENTS JURIDIQUES ET ADMINISTRATIFS 9](#_Toc179548981)

[7. APPROBATION 10](#_Toc179548982)

# RESUME EXECUTIF

Cette proposition faite suite à la mise en production de la solution Kairntech et à la mise au point de différents chaines de traitement (pipelines) sur des problématiques d’extraction d’information et de classification de textes.

Cairn.Info souhaite à présent déployer deux nouveaux pipelines :

* un pipeline RAG de question-réponse,
* un pipeline de résumé automatique.

Ces deux nouveaux pipelines vont utiliser un LLM (éventuellement plusieurs) qui sera installé localement de manière à garantir la sécurité et la confidentialité des contenus.

Le choix du LLM va donc se porter vers l’utilisation d’un LLM en accès libre et téléchargeable qui devra être déployé dans une infrastructure garantissant de bonnes performances à un coût raisonnable.

Ce document est une proposition commerciale de prestations de service comprenant :

* l’évaluation de différents LLMs sur les deux cas d’usage mentionnés ci-dessus et la sélection de LLMs candidats,
* l’évaluation de différents frameworks pour opérer ces LLMs,
* le développement d’un nouveau composant dans la plateforme Kairntech pour se connecter au framework retenu,
* l’évaluation des performances et une recommandation pour un déploiement en production,
* l’assistance à l’installation et la configuration du LLM dans l’environnement de production.

# HYPOTHESES

Nous prenons les hypothèses suivantes avant le démarrage des travaux.

## Résumé automatique

Cairn.Info met à disposition de Kairntech un fond documentaire avec les caractéristiques suivantes :

* 100 documents au minimum
* Les documents sont représentatifs du fond documentaire à traiter
* Quelques documents représentant typiquement les documents les plus longs

Les documents et les résumés à produire sont en langue française.

Volume et flux de contenu estimé :

* 1M de documents d'archive à traiter
* 400 nouveaux documents / jour

Un document compte en moyenne 5-6000 mots, ce qui représente environ 10k token (en prenant l’hypothèse suivante en français : 1 mot = 2 token).

## Question-Réponse RAG

Cairn.Info met à disposition de Kairntech un fond documentaire pour le service de question réponse RAG :

* Les documents sont des transcriptions audio de « rencontre auteurs »
* Les documents sont représentatifs du fonds documentaires à traiter

Volume actuel :

* 40 transcriptions audio d’une durée de 1h30 environ.

Les documents et les réponses aux questions à produire sont en langue française.

Cairn.info fournit des exemples de questions, au minimum une dizaine pour mener des tests.

# PRESTATIONS DE SERVICE

## Lotissement

Lot 1 : Evaluation de LLMs

Les travaux à mener sont les suivants :

* Sélection de différents LLMs (généralistes, de tailles différentes)
* Evaluation de la qualité sur les tâches de résumé automatique et de question-réponse RAG,
* Comparaison de la qualité des différents LLMs entre eux, ajustement des prompts,
* Benchmark avec GPT4 (considéré comme étant l’état de l’art)
* Sélection de 2-3 LLMs candidats pour les tâches de résumé automatique et question-réponse RAG, en accord avec les contraintes budgétaires de Cairn.info quant au GPU nécessaire pour l’opérer

Livrable : Liste de 2-3 LLMs candidats

Lot 2 : LLMs en local - Initialisation

Les travaux à mener sont les suivants :

* Etude et sélection d'un framework open source pour opérer le(s) LLM(s) en local. Plusieurs possibilités existent (Ollama, vLLM...). Il faudra les évaluer pour identifier les avantages et inconvénients de chacun et finalement en sélectionner un.
* Développement d'un nouveau composant Kairntech pour se connecter au framework retenu
* Installation et configuration du framework retenu, des 2-3 LLMs candidats, dans un cloud de test (type cloud RunPod…)
* Tests de bon fonctionnement de la plateforme Kairntech avec LLMs en local

Livrable : un environnement de test opérationnel avec les 2-3 LLMs candidats en local

Lot 3 : LLMs en local - Tests de performance

Les travaux à mener sont les suivants :

* Campagne de tests de performance sur environnement de test avec les LLMs candidats
  + Résume automatique : disponibilité, latence, monitoring de la machine...
  + Question-Réponse RAG : disponibilité, latence, appels concurrents, monitoring de la machine...
* Itérations avec différents GPUs
* Comparaison des LLMs, sélection du LLM retenu.

Livrable : LLM retenu, recommandation de l’infrastructure cible & paramétrage

Lot 4 : LLM en local - Installation sur machine cible de production

* Accompagnement à l'installation du framework et du LLM retenu (ou des LLMs)
* Accompagnement sur réglage et optimisation des performances

## Synthèse des livrables

Les livrables sont les suivants :

1. Liste de 2-3 LLMs candidats identifiés pour répondre aux cas d’usage (résumé automatique, question-réponse RAG)
2. Un environnement de test opérationnel avec les 2-3 LLMs candidats, opérés en local
3. LLM retenu, recommandation de l’infrastructure cible avec configuration & paramétrage

## Planning

Tous les travaux inclus dans cette proposition se feront sur la dernière version Manta.

Kairntech pourra livrer ces livrables selon le calendrier ci-dessous. T0 est la date de réception de la commande.

|  |  |  |
| --- | --- | --- |
| **Livrable** | **Description** | **Date de livraison** |
| 1 | Liste de 2-3 candidats LLMs | T0+2 semaines |
| 2 | Environnement de test opérationnel | T0+6 semaines |
| 3 | LLM retenu, recommandation infrastructure | T0+10 semaines |

# PROPOSITION FINANCIÈRE

Le coût de ces prestations est indiqué dans le tableau ci-dessous. Kairntech propose de réaliser l’ensemble de ces prestations au forfait.

|  |  |
| --- | --- |
| **Description des prestations** | **Prix (€ HT)** |
| Lot 1 : Evaluation de LLMs, sélection de 2-3 candidats LLM | 3000 € HT |
| Lot 2 : Mise en place d’environnement de test opérationnel avec les 2-3 candidats LLM opérés en local | 5000 € HT |
| Lot 3 : Tests de performance, sélection du LLM retenu, recommandation | 5000 € HT |
| Lot 4 : Accompagnement installation sur plateforme cible | 2000 € HT |
| **Total** | **15 000 € HT** |

Pour le Lot 3 : ce coût inclut les frais de cloud avec machine GPU. Plafond de xxx.

# CONDITIONS GÉNÉRALES DE VENTE

## Facturation et paiement

Kairntech propose l’échéancier de paiement ci-dessous.

|  |  |  |
| --- | --- | --- |
| **Date de facturation** | **Pourcentage** | **Montant (€ HT)** |
| A la commande | 50% | 000 € HT |
| A la livraison des livrables | 50% | 000 € HT |
| **Total** | | **000 € HT** |

Tous les paiements se font à 45 jours fin de mois.

## Validité de la proposition

Cette proposition commerciale est valable jusqu’au 30 novembre 2024 inclus.

# RENSEIGNEMENTS JURIDIQUES ET ADMINISTRATIFS

Dénomination sociale Kairntech

Forme juridique SAS

Capital 170 000 Euros

Siège social 29 chemin du vieux chêne 38240 MEYLAN - FRANCE

Numéro SIRET 844 751 636 00011 RCS GRENOBLE

Code APE 5829C

TVA Intracommunautaire FR 37 844751636

Contact administratif : Anne de Chauvigny

anne.dechauvigny@kairntech.com

# APPROBATION

La signature de cette page par une personne habilitée vaut acceptation de cette proposition.

Lu et approuvé : Lu et approuvé :

Cairn.Info Kairntech SAS

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_/\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_/\_\_\_\_\_\_\_\_\_\_

Signature / Date Signature / Date

Nom : Nom :

Fonction : Fonction :