

# Injection de connaissance terminologique et littérature grise en santé dans un LLM génératif sur infrastructure multi-GPUs

Journées Calcul et Données (JCAD) – 2025

Benoist GASTON <sup>1</sup>   Romain LELONG <sup>2</sup>   Badisse DAHAMNA <sup>2</sup>

<sup>1</sup>Centre Régional Informatique et d'Applications Numériques de Normandie

<sup>2</sup>Département de Santé Numérique (DéSaN) – CHU de Rouen

17 septembre 2025





- Statut d'Association, créée en 1992
  - ESR (ComUE Normandie Université et établissements affiliés)
  - Rectorat, établissements de santé
  - Collectivités territoriales
- Mutualisation d'équipements et services à haut niveau de performance
  - Réseau régional Syvik
  - Centre de données régional
  - Calcul intensif (HPC)
- Équipe : ~15 ETP
- Certification ISO27001 et HDS

- Réseau de 33 centres de compétence nationaux
  - Favoriser l'usage du HPC et des technologies associées (HPDA, IA & Quantum)
  - Fédérer l'écosystème
  - Développer des formations
  - Accompagner les besoins et les demandes
- Programme d'accompagnement à l'usage du HPC par les mésocentres
  - Porté par Criann & Romeo, partenaires MesoNET
  - À destination du secteur public et privé (collectivités, administrations, industrie, PME, startups)



This project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 201102063. The JU receives support from the Digital Europe Programme and Germany, Bulgaria, Austria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, Greece, Hungary, Ireland, Italy, Lithuania, Latvia, Poland, Portugal, Romania, Slovenia, Spain, Sweden, France, Netherlands, Belgium, Luxembourg, Slovakia, Norway, Türkiye, Republic of North Macedonia, Iceland, Montenegro, Serbia.

Partenaires du Centre de compétence français :



CENTRE DE COMPÉTENCE HPC.HPDA.IA



Avec la participation de MESQNET le réseau des mésocentres

# Austral – caractéristiques (septembre 2025)



	A100 80GB SXM	H200 141GB SXM
SM	108	144
Tensor Core	432	576
Bandwith	2.0 TB/s	4.8 TB/s
Interconnect (NVLink)	660GB/s	900GB/s
TFlops DP (TC)	9.7 (19,5)	34 (67)
TFlops SP (TC)	19.5 (156)	67 (987)
TFlops HP with TC	312	1979



## AUSTRAL

**UN SUPERCALCULATEUR  
POUR LA RECHERCHE  
EN NORMANDIE**

**Nœuds de calcul**

88 GPU Nvidia A100 80 Go  
16 GPU Nvidia H200 140 Go  
Interconnecté NVLink, 4 GPU par nœud de calcul

23 808 cœurs pour le calcul parallèle  
Nœuds de calcul AMD Zen4  
2x8 cœurs à 2,4 GHz et 1x8 Go FDR 6000

Un nœud à large mémoire partagée  
8x 16 MB DDR5 et 64x mémoire L3 cache

8 GPU AMD MI210  
4 GPU AMD MI210 par nœud

Interconnexion

2x InfiniBand 400Gb/s

Stockage rapide

2x NVMe 1.9x NVMe

Visualisation à distance  
Environnement interactif pour l'IA

Linux RedHat - Slurm - Lustre

**Puissance crête 2,5 PFlop/s**

MOYENS DE CALCUL PERFORMANTS  
ET SUPPORT AVANCÉ  
POUR UNE UTILISATION OPTIMALE





**CRIANN** CENTRE RÉGIONAL, INTERNATIONAL  
ET EUROPÉEN DE RECHERCHE  
EN NORMANDIE



# Contexte

Les 3 sources de « connaissances » en santé du DéSaN



Catalogue et Index des Sites Médicaux de langue Française  
([www.cismef.org](http://www.cismef.org)) – 1993

*Littérature grise en français et en santé*

- > 130 000 ressources en texte intégral



Littérature Scientifique en Santé ([www.lissa.fr](http://www.lissa.fr)) – 2014

*Banque de données bibliographiques en français et en santé*

- 1,2 M métadonnées sur articles/publications scientifiques
- > 400 K de résumés



Health Terminology/Ontology Portal ([www.hetop.eu](http://www.hetop.eu)) – 2007

*Serveur terminologique interlingue*

- 109 terminologies et ontologies en santé en 55 langues
- > 100 millions de triplets RDF

## 1. Continual Pre-training - Causal Language Modeling (CLM)

**Technique :** Le modèle apprend à prédire le prochain mot dans un texte (causal/autoregressif).

**Objectif :** Prolonger l'entraînement non supervisé avec des données spécialisées.

## 2. Fine-tuning – Low-Rank Adaptation (LoRA)

**Technique :** LoRA ajoute des matrices à faible rang dans les couches du modèle, sans modifier ses poids d'origine.

**Objectif :** Adapter le modèle à des paires instruction/réponse ou question/réponse (supervisé).

# HeTOP, une source de données structurées à linéariser.

The screenshot shows the HeTOP web interface. At the top, there is a search bar with the text 'MSH\_D\_010052' and a language dropdown set to 'français'. Below the search bar, there are checkboxes for 'Recherche sans troncature (sans wildcard)' and 'Sélection terminologies', and a checkbox for 'Ne pas chercher dans les définitions'. On the left, a sidebar titled 'Vos recherches' shows '1 ressource(s) trouvée(s) en 0,00 s' and a list of 'Meilleurs candidats' including 'ovariectomie [Descripteur MeSH]' and 'MeSH (1)'. The main content area displays the MeSH entry for 'Ovariectomie (Descripteur MeSH)'. It includes a description, tabs for 'Hiérarchies', 'Relations', and 'Pub Med / Doc CISM', and a list of associated logos: NIH NLM, Inserm, BioPortal, and RDF/XML. Below this, there is a button 'Voir toutes les langues' and a dropdown set to 'Off'. The entry details are organized into sections: 'Identifiant d'origine' (D010052), 'Libellé préféré' (ovariectomie in French and ovariectomy in English), 'Définition du MeSH' (Ablation chirurgicale d'un ou des deux ovaires. [Traduction effectuée avant 2008] and The surgical removal of one or both ovaries.), 'Synonyme MeSH' (oophorectomies, oophorectomy, ovariectomies), and 'Synonyme CISMef' (ablation ovaires, Oophorectomie).

1 concept  
=  
un ensemble de  
métadonnées/informations

# HeTOP, une source de données structurées à linéariser.

HeTOP français MSH\_D\_010052

Recherche sans troncature (sans wildcard)  
 Sélection terminologies  Ne pas chercher dans les définitions

Vos recherches

1 ressource(s) trouvée(s) en 0,00 s

Meilleurs candidats

- ovariectomie [Descripteur MeSH]
- MeSH (1)

**Ovariectomie** (Descripteur MeSH)

Description Hiérarchies Relations PubMed / DocCISMeF

NIH NLM Inserm BioPortal RDF/XML

Voir toutes les langues Off

Identifiant d'origine: D010052

Libellé préféré: ovariectomie, ovariectomy

Définition du MeSH: Ablation chirurgicale d'un ou des deux ovaires. [Traduction effectuée avant 2008], The surgical removal of one or both ovaries.

Synonyme MeSH: oophorectomies, oophorectomy, ovariectomies

Synonyme CISMef: ablation ovaires, Oophorectomie

1 concept  
=  
un ensemble de  
métadonnées/informations

métadonnées  $i_1$  →

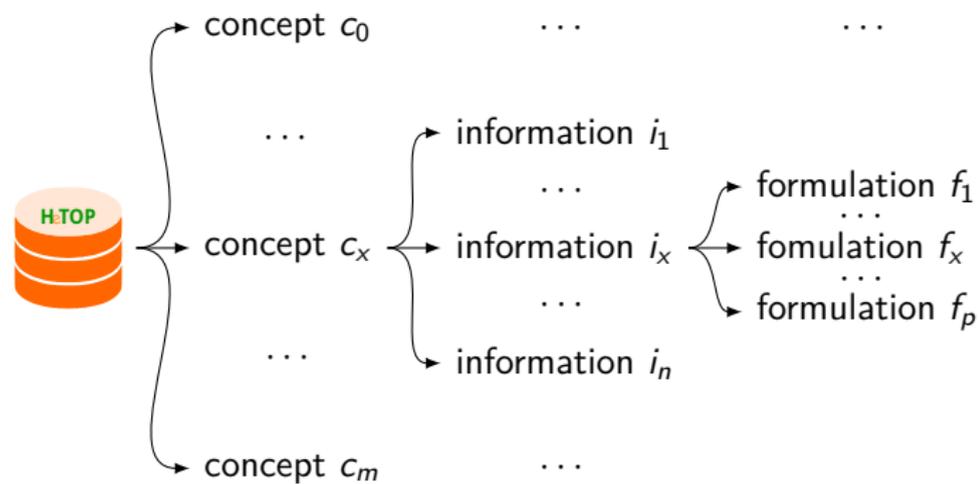
métadonnées  $i_2$  →

métadonnées  $i_3$  →

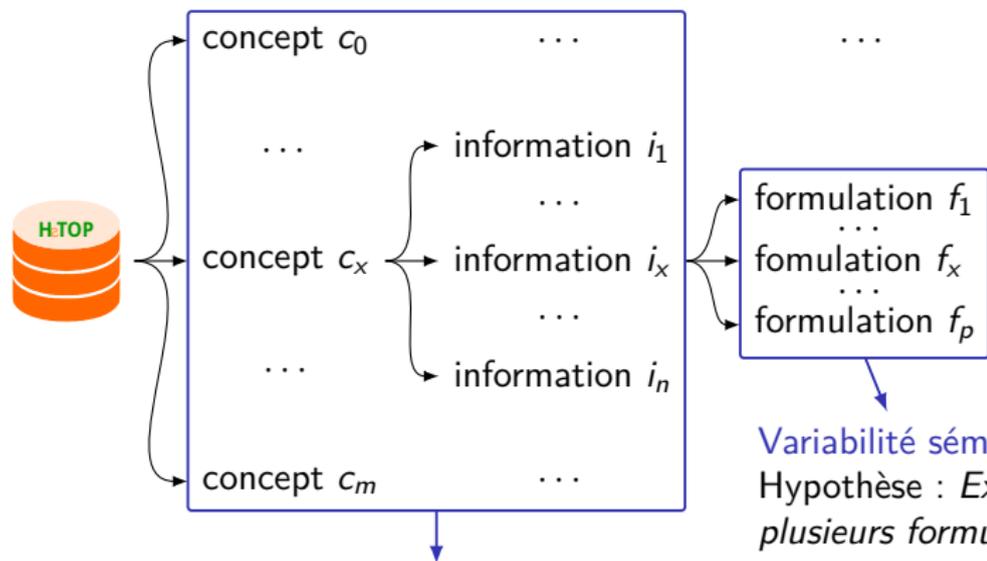
métadonnées  $i_4$  →

métadonnées  $i_5$  →

# Construction du dataset d'entraînement



# Construction du dataset d'entraînement



## Diversité des concepts

Hypothèse : *répéter un même type d'information sur plusieurs concepts aide le modèle à généraliser la structure de l'information.*

## Variabilité sémantique

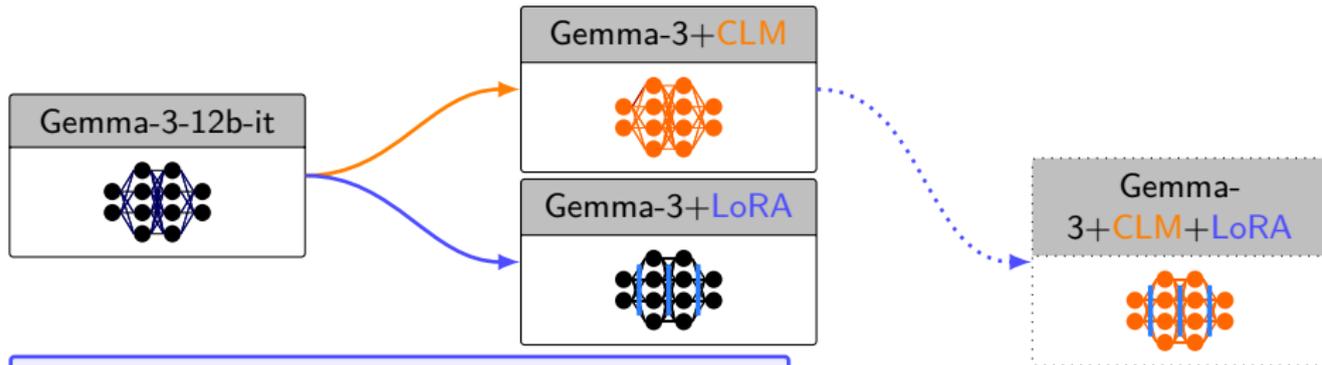
Hypothèse : *Exposer plusieurs formulations permet d'éviter que le modèle ne sur-apprenne une tournure unique.*

# Deux approches testées en pratiques

## Causal Language Modeling (CLM)

GPUs : 4 × H200 (140GB VRAM/GPU)  
Dataset : 500 concepts MeSH × 5 variations  
⇒  $\simeq$  29 575 documents  
epochs : 3.24  
Temps : 20h 8min 3sec

**VRAM globale requise en moyenne pendant l'entraînement : 406 GB**  
(GPU load : 24.1% par GPU)



## Fine-Tuning LoRA

GPUs : 4 × H200 (140GB VRAM/GPU)  
Dataset : 1000 concepts MeSH × 3 variations  
⇒ 36 128 documents  
epochs : 2.0  
Temps : 1h 53min 11sec

**VRAM globale requise en moyenne pendant l'entraînement : 110 GB**  
(GPU load : 22.4% par GPU)

## Injection de connaissance

L'entraînement en **Causal Language Modeling (CLM)** tout comme le **fine-tuning LoRA** permettent d'injecter des connaissances aux modèles.

Exemple n° 1 :

**Prompt** : « Quel est le libellé préféré du concept HeTOP MSH\_D\_010052 ? »

Gemma-3

« Le libellé préféré pour le concept HeTOP MSH\_D\_010052 est **\*\*"Maladie de Parkinson"**. [...] »

Gemma-3+**CLM**

« La notion "ovariectomie" est présente dans HeTOP avec le code MSH\_D\_010052. [...] »

Gemma-3+**LoRA**

« Le libellé préféré est "ovariectomie". »

Exemple n° 2 :

**Prompt :** « Peux tu me donner la liste complète des synonymes CISMef du concept HeTOP MSH\_D\_018259 ? »

Gemma-3

« Le concept HeTOP MSH\_D\_018259 dans CISMef représente "Hématopoïèse". Voici la liste complète des synonymes CISMef associés à ce concept :

- \* \*\*Hématopoïèse\*\* (Terme désigné)
- \* \*\*Formation hématopoïétique\*\*
- \* \*\*Génèse hématopoïétique\*\*
- \* \*\*Hématopoèse\*\* (orthographe alternative)[...]

»

Gemma-3+CLM+LoRA

« Absolument, la liste complète des synonymes CISMef pour "herpèsvirus humain de type 1" (MSH\_D\_018259) est : "HHV-1 (Human HerpesVirus 1)", "Virus herpes simplex 1", "Herpèsvirus 1 humain", "Herpesvirus simplex de type 1", "Virus herpès simplex de type 1", "HSV-1", "Herpes simplex virus 1", "HSV-1 (Herpes simplex virus 1)", "virus de l'herpès simplex de type 1", "Human herpesvirus 1". »

Les deux approches se sont montrées efficaces pour l'intégration de connaissances.

Cependant, elles nécessitent un ajustement du corpus d'entraînement pour garantir l'exactitude et la complétude des résultats finaux.

L'approche **CLM**, présente un risque élevé de surajustement (overfitting) avec 3.24 époques  $\Rightarrow$  compromission des capacités génératives et de généralisation initiales et hyperspécialisation sur les données d'entraînement.

Le **Fine-Tuning LoRa** spécialise le modèle sur des tâches précises tout en préservant ses compétences génériques de base.

L'approche **CLM**, reste compliquée à grande échelle sur ce hardware (> 20h d'entraînement pour 500 concepts ( $\simeq 30k$  docs), alors qu'HeTOP en compte > 3M sur 100 terminologies. ).

Le **Fine-Tuning LoRa**, plus réaliste et plus scalable pour l'ensemble des données HeTOP.

Autre pistes possibles :

**Approche RAG** : utiliser un contexte externe riche en connaissances pour guider l'inférence du modèle, plutôt que d'intégrer ces connaissances dans son architecture.

**Hybridation avec l'IA symbolique** : intégration de connaissances sous forme de données ontologiques (triplets RDF) lors de l'entraînement puis formation du modèle à exploiter efficacement cette connaissance formalisée.