

**MINISTÈRE
DE L'INTÉRIEUR**

*Liberté
Égalité
Fraternité*

Exercice Scientifique et technique

Comment les caractéristiques des LLM influencent la qualité des réponses
selon le contexte ?

Prévot Maxime

Une école de

 CCI PARIS ILE-DE-FRANCE
EDUCATION

 **Université
Gustave Eiffel**

Sommaire

- 1 : Contexte professionnel**
- 2 : Notion de paramètres**
- 3 : Notion de quantization**
- 4 : Comparaison entre paramètres et quantification sur les performances**
- 5 : Notion d'instructions**
- 6 : Les paramètres généraux, type température**
- 7 : Notion de fenêtre de contexte → context window**
- 8 : Bilan/Préconisation par rapport aux besoins, aussi relié à mon entreprise**
- 9 : Rapport aux cours**



MINISTÈRE
DE L'INTÉRIEUR

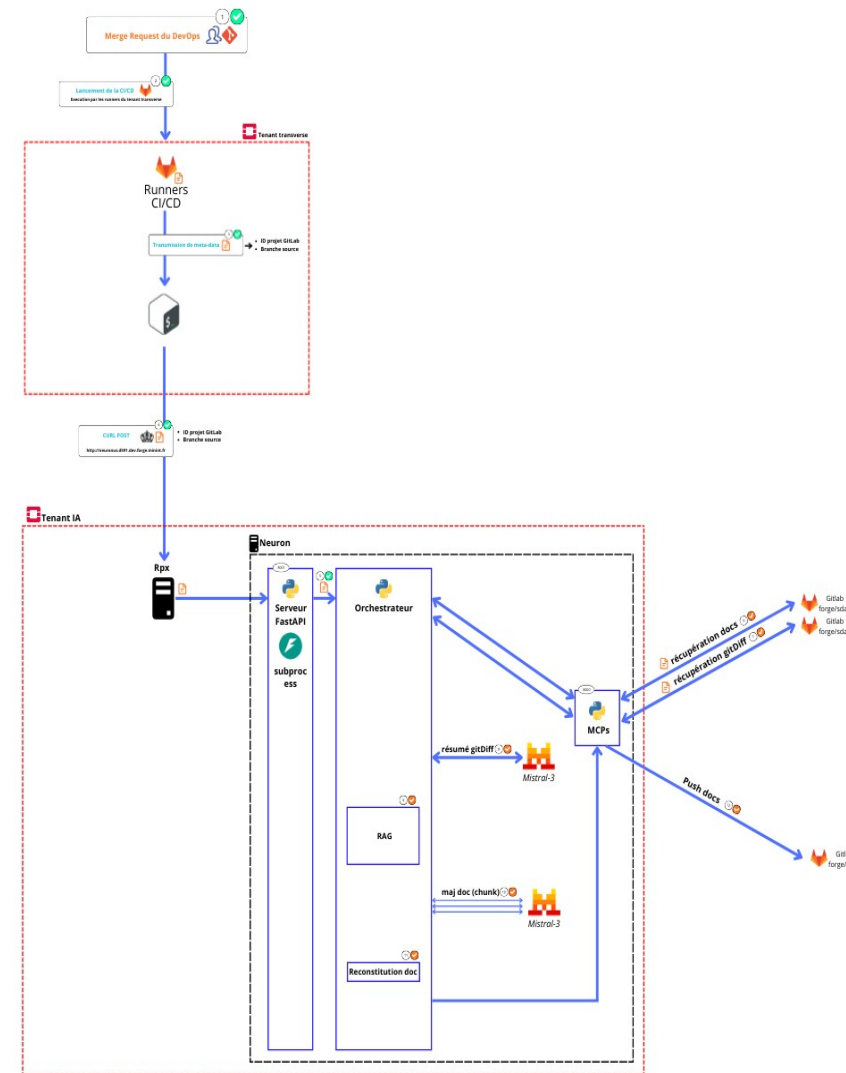
*Liberté
Égalité
Fraternité*

Contexte : Projet automatisé de la documentation assistée via LLM

1 : Détection de l'évolution d'un code (Merge Request)

2 : Mise à jour documentation par IA

3 : Nouvelle documentation pushée sur GitLab



Notion 1 : Les paramètres

Prompt de base : *Explique en 3 phrases pourquoi le ciel est bleu*

```
maxime@dell:~$ ollama run llama3.2
>>> Explique en 3 phrases pourquoi le ciel est bleu
Bien sûr ! Voici trois explications simples :

Le ciel apparaît bleu car la Terre est entourée d'atmosphère, qui scelle les rayons du soleil. L'atmosphère contient des particules appelées molécules d'azote et d'oxygène, qui réfléchissent la lumière bleue de manière plus efficace que les autres couleurs. C'est pourquoi nous voyons le ciel se teintant de bleu lors du jour.
```

```
maxime@dell:~$ ollama run mistral:7b-instruct-q4_K_M
>>> Explique en 3 phrases pourquoi le ciel est bleu
1. Rayonnement de soleil : Le ciel apparaît bleu à cause du rayonnement du soleil dans l'atmosphère. La présence d'ozone et de particules minces dans l'air filtre les ultraviolets qui sont responsables des couleurs bleues, de sorte qu'ils ne nous atteignent pas.
2. Scattering des rayons : Les rayons du soleil se dispersent dans l'atmosphère en toutes directions par les molécules d'air. Lorsqu'ils sont scatterés, ils font entrer une composante bleue à notre œil, ce qui est plus perceptible pour nos yeux que les autres colorations de la lumière.
3. Effet Tyndall : Les particules minces dans l'atmosphère scattent les rayons du soleil de manière très efficace, et cela provoque une augmentation des composantes bleues dans le spectre visible. Cet effet est connu sous le nom d'effet Tyndall.
```

```
maxime@dell:~$ ollama run minstral-3:14b-instruct-2512-q4_k_m
>>> Explique en 3 phrases pourquoi le ciel est bleu
1. **Diffusion de la lumière** : Le ciel apparaît bleu parce que l'atmosphère terrestre est composée de molécules (comme l'azote et l'oxygène) qui **diffusent (ou dispersent) la lumière blanche du Soleil**, un phénomène appelé **diffusion de Rayleigh**, plus efficace pour les **longueurs d'onde courtes** (bleu et violet).

2. **Perception humaine** : Bien que le violet soit encore plus diffusé que le bleu, nos yeux sont **moins sensibles à cette couleur**, et le Soleil émet moins de lumière violette, ce qui fait que le **bleu domine** notre perception.

3. **Absence d'obstacle** : Sans atmosphère (comme sur la Lune) ou en regardant le Soleil depuis l'espace, le ciel semble noir, car il n'y a pas de diffusion de la lumière par des molécules - le bleu n'apparaît que grâce à l'interaction entre la lumière et notre enveloppe gazeuse.
```

Llama3.2 3B

VS

Mistral 3 7B

VS

Ministral 3 14B

- Court
- Peu détaillé
- Manque de termes techniques

- Concis
- Définition des termes techniques
- Concept partiellement correcte

- Détaillé
- Compréhension du prompt exacte

Notion 2 : La quantization

Prompt de base : *Explique en 4 phrases le style d'écriture de Victor Hugo*

```
maxime@dell:~$ ollama run mistral:7b-instruct-q2_K
>>> c'est quoi le style d'écriture de Victor
... .. Hugo ? en 4 phrases stp
1. Victor Hugo écrit dans un style romantique, inspiré par la nature et les passions humaines.
2. Son travail est marqué par des langages poétiques et une forte sensibilité intellectuelle.
3. Il utilise également des métaphores et des similitudes pour créer des images vives et expressives.
4. Sa narration est souvent complexe, avec nombreux niveaux d'analyse qui invitent le lecteur à penser plus profondément sur les thèmes abordés.
```

```
maxime@dell:~$ ollama run mistral:7b-instruct-q4_K_M
>>> c'est quoi le style d'écriture de Victor
... .. Hugo ? en 4 phrases stp
1. Hugo était un poète et écrivain français, considéré l'un des plus importants auteurs du Romantisme dans la littérature française.
2. Il est connu pour ses ouvrages tels que "Les Misérables", "Notre-Dame Cathedral", et "The Legend of Ages".
3. Son style d'écriture est marqué par une grande passion, des descriptions imaginaires, et des thèmes sociaux et politiques.
4. Il était également célèbre pour sa technique de construction narrative complexe, qui mettait l'accent sur les personnages et leurs émotions.
```

```
maxime@dell:~$ ollama run mistral:7b-instruct-q8_0
>>> c'est quoi le style d'écriture de Victor
... .. Hugo ? en 4 phrases stp
1. Il utilise un style poétique lyrique, avec des métaphores et des similes qui expriment ses émotions intenses.
2. Sa prose est rythmée et harmonieuse, avec une forte présence du versification et de la musique.
3. Il emploie des vocabulaires rares et archaïques pour créer un sens d'antiquité et de profondeur historique.
4. Son écriture est aussi marquée par son grandeur d'esprit et sa capacité à aborder des questions philosophiques et religieuses.
```

Mistral 7B Q2 VS

- Texte générique

Mistral 7B Q4 VS

- Cite des œuvres
- résultat plus structuré
- Erreurs sur le nom des oeuvres

Mistral 7B Q8

- Vocabulaire riche

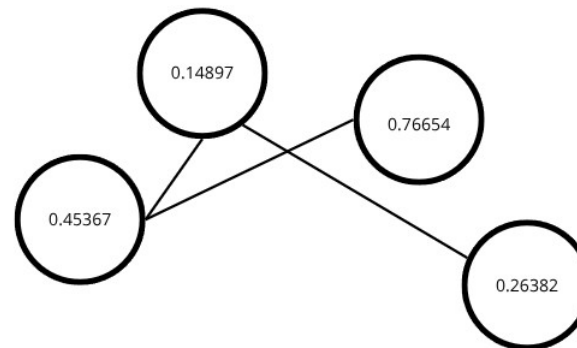
Quels types de pertes de performances ? Paramètres VS Quantization

Paramètres

- Permet la compréhension du prompt
- Suppression de connexions/poids dans le réseau de neurones

Quantization

- Permet la restitution de la réponse
- Moins de précision sur les poids



Notion 3 : Le suivi d'instructions

Prompt de base : *Voici un texte, retourne un JSON de cette X forme*

```
maxime@dell:~$ ollama run mistral:7b-text-q4_0
>>> Analyse le texte suivant et retourne uniquement un JSON valide.
...
... Contraintes :
... - aucune phrase explicative
... - aucune clé supplémentaire
... - respecter exactement les clés du schéma
... - respecter la casse
...
... Texte :
... "Marie Dupont travaille comme ingénieure logicielle à Paris. Elle développe des applications web et s'intéresse à l'intelligence artificielle."
...
... Format attendu :
... {
...   "nom_complet": "",
...   "profession": "",
...   "ville": "",
...   "interets": []
... }
```

Résultat attendu :

```
{ "nom_complet": "Marie Dupont", "profession": "ingénieure logicielle", "ville": "Paris", "interets": ["applications web", "intelligence artificielle"] }
```

```
maxime@dell:~$ ollama run mistral:7b-instruct-q4_K_M
>>> Analyse le texte suivant et retourne uniquement un JSON valide.
...
... Contraintes :
... - aucune phrase explicative
... - aucune clé supplémentaire
... - respecter exactement les clés du schéma
... - respecter la casse
...
... Texte :
... "Marie Dupont travaille comme ingénieure logicielle à Paris. Elle développe des applications web et s'intéresse à l'intelligence artificielle."
...
... Format attendu :
... {
...   "nom_complet": "",
...   "profession": "",
...   "ville": "",
...   "interets": []
... }
```json
{
 "nom_complet": "Marie Dupont",
 "profession": "ingénieure logicielle",
 "ville": "Paris",
 "interets": ["intelligence artificielle"]
}
```

### Mistral 7B (Pas Instruct)

VS

### Mistral 7B Instruct

- JSON mal construit

- Contenu illogique

- JSON Valide

## Notion 4 : La température

*Prompt de base : Écris un court paragraphe décrivant une ville sous la pluie, dans un style poétique et imagé*

```
maxime@dell:~/Bureau/docIA$ uv run python testTem
p.py
La pluie bat les vitres de la ville,
Un tapis vert et gris se forme sur le sol,
Les rues sont des rivières courantes,
Le ciel est une toile d'ombre et de bleu.
```

```
maxime@dell:~/Bureau/docIA$ uv run python testTem
p.py
Dans les rues de la ville se cache
Un monde qui se fait entourer d'eau
Les pavés glissants font que l'on s'y détache
Le pluie a cette ville donné une nouvelle hue
```

**Mistral 7B  $t = 0.2$**

**VS**

**Mistral 7B  $t = 0.7$**

- Pas de rimes
- Texte peu imagé

- Présence de rimes
- Texte plus imagé



## Notion 5 : La fenêtre de contexte – Context Window

*Taille de la Context Window de Mistral 7B : 8000 tokens → 5000 – 6000 mots (Prompt + réponse)*

*Dans le cadre de la documentation que je récupérais des projets du Ministère de l'Intérieur, il y en avait pour 50 pages/projets → 40 000 mots/projets, on dépasse largement la Context Window. Pour ne pas que l'IA hallucine, j'ai alors mis en place un RAG (même en cadre de POC).*

## Par rapport aux besoins du Ministère de l'Intérieur

*L'objectif est d'obtenir une documentation technique qualitative.*

*Pour la partie technique → Il faut un modèle avec beaucoup de paramètres pour comprendre avec précision les modifications apportés par la Merge Request.*

*Pour la partie qualitative → Il faut un modèle qui arrive à restituer avec un langage précis.*

## Par rapport aux besoins du Ministère de l'Intérieur

*Modèle qui répond le mieux au besoin :*

*Mistral 14B Q4*

*Modèle envisagé dans le projet :*

*Mistral 14B FP16*

*Il faudra juste faire attention au temps pris par le workflow complet*

X	RAM	Vitesse pour réponse complète	Compréhension	Restitution	Tests concrets pour la doc
Llama 3.2	2Go	Très rapide (2 min)	Mauvais/Moyen	Moyen	ratés
Mistral 7B Q4	6Go	12 min	Bon	Bon	presque réussis
Mistral 14B Q4	10Go	45 min	Très bon	Très bon	réussis

## Rapport aux cours → Projet FullStack

---

*Performances accrues avec GPU*

## Webographie

---

<https://mistral.ai/products/studio> → permet de connaître les caractéristiques des modèles.

<https://blent.ai/blog/a/quantization-llm> → permet de mieux comprendre la quantization et d'où elle vient.

[https://www.youtube.com/watch?v=\\_YTnZOYxSjE](https://www.youtube.com/watch?v=_YTnZOYxSjE) → permet de mieux comprendre la déterminisme dans la notion de température.