Utiliser des grands modèles de langage ouverts avec Ollama et AnythingLLM

Atelier pratique - AI Horizons in Language Learning

Pascal Brissette

2025-11-25

Note

Le présent document, ainsi que la présentation et les exemples de textes utilisés dans l'exercice, sont disponibles sur GitHub : https://github.com/pbriss7/Atelier-Ollama-AnythingLLM

1 Introduction

Si les chercheuses et chercheurs universitaires n'ont pas attendu l'arrivée de ChatGPT pour mettre à profit l'intelligence artificielle dans les protocoles de recherche, la démocratisation de l'IA générative en 2022 avec le célèbre robot conversationnel d'OpenAI a créé un changement de paradigme. Désormais, les non-spécialistes, y compris nos étudiantes et étudiants, peuvent utiliser ces cerveaux artificiels entraînés sur des milliards de textes, d'images et de sons et s'en servir au pire comme des béquilles, au mieux comme des tremplins vers la connaissance.

L'utilisation de ces nouveaux outils a évidemment un coût, qui se calcule de plusieurs manières.

- **Coût environnemental** : en 2023, 4,4% de l'électricité produite aux États-Unis était consacrée aux centres de données. Les spécialistes évaluent que, d'ici 2030-2035, 20% de l'énergie mondiale sera consacrée à l'entraînement et au fonctionnement des LLMs. Le refroidissement des serveurs requiert également d'immenses quantités d'eau. En 2022, les centres de données de Google seulement ont utilisé plus de cinq milliards de litres d'eau fraîche.

- **Coût humain** : l'entraînement des modèles, notamment la mise en place de garde-fous destinés à protéger les utilisateurs et utilisatrices des comportements préjudiciables de l'IA, a largement été fait sur le dos des plus pauvres de notre monde. Des armées de travailleuses et travailleurs ont passé de longues heures à étiqueter des textes haineux et des images que personne ne veut voir.
- **Coût moral et légal** : les données utilisées par les compagnies privées pour l'entraînement non supervisé des IA ont été collectées au détriment du droit d'auteur. La compagnie Anthropic a été la première, en 2025, à verser 1,5 milliard à un fonds d'indemnisation d'auteurs, d'ayants droit et d'éditeurs.

L'utilisation des IA génératives soulève en outre des questions concernant la confidentialité des données. La plupart des plateformes commerciales offrant un usage gratuit des modèles et serveurs tirent profit des données que nous, les utilisatrices et utilisateurs, partageons. Ces données sont utilisées aussi bien pour l'entraînement de nouveaux modèles que pour cerner nos profils et nos besoins. Certains modèles, comme Copilot, mis gratuitement à notre disposition par l'Université McGill, offrent une garantie de sécurité, mais le client de Microsoft, dans notre cas l'Université McGill, peut toujours obtenir l'accès aux données que nous partageons avec Copilot.

Les outils présentés dans le cadre de cet atelier ne sont pas une panacée, mais ils répondent à plusieurs enjeux problématiques.

1. Les modèles utilisés sont **moins volumineux et donc moins énergivores** que les puissants modèles SOTA (*State Of The Art*) d'OpenAI, de Google ou d'Anthropic.
2. S'ils sont utilisés localement, sur l'ordinateur personnel d'un utilisateur situé au Québec, l'énergie consommée provient d'une source qui **ne produit pas ou peu de GES**.
3. Dans ce même cas de figure, **aucune donnée n'est partagée** avec des tiers, pas même avec Ollama.
4. Si on souhaite utiliser des modèles plus performants tout en conservant les avantages d'une utilisation sécuritaire et privée, on peut recourir aux serveurs d'Ollama, qui garantissent la confidentialité des échanges.

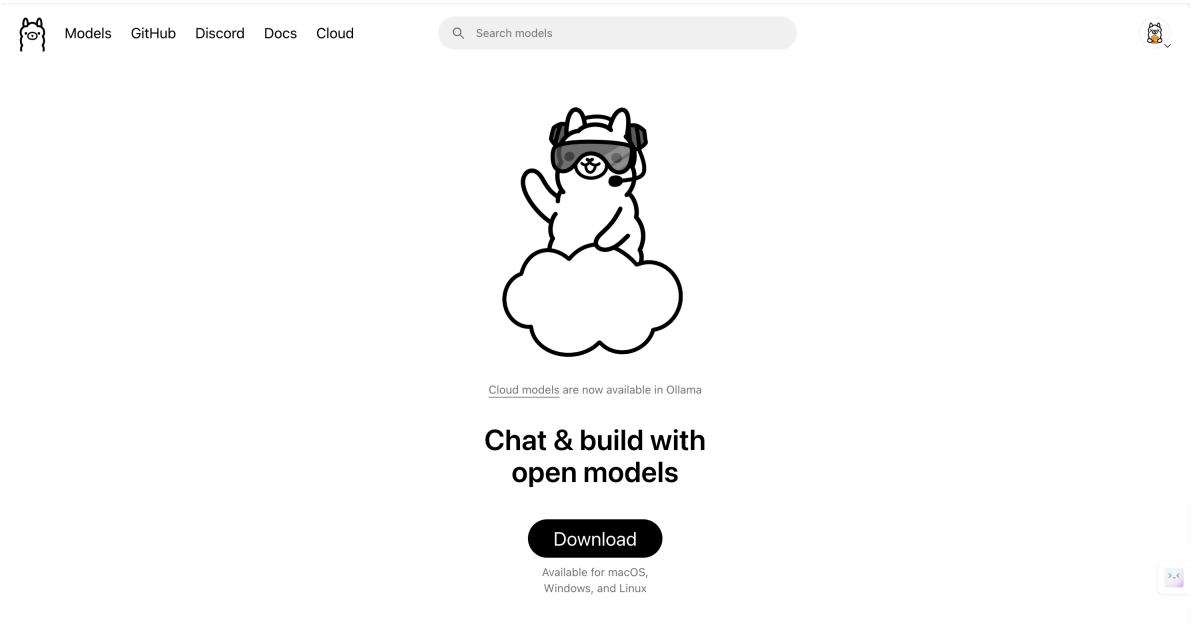
1.1 Outils proposés

1.1.1 1. Ollama

Ollama est un logiciel qui sert de passerelle vers les modèles de langage (LLMs).

- Ollama crée une bibliothèque locale pour l'entreposage des modèles importés, en plus d'offrir un menu de commandes pour interagir avec ces derniers.
- Ollama ne crée pas de nouveaux modèles : il donne accès à ceux développés par la communauté ou libérés par les grandes compagnies (gemma2, llama3, qwen, gpt-oss, etc.).

- Ollama garantit la confidentialité des échanges avec les LLMs. Si Ollama est exécuté localement avec des modèles importés, les données ne quittent même pas l'ordinateur.
- Le logiciel propose une interface utilisateur pour faciliter l'interaction avec les LLMs, mais celle-ci est rudimentaire: elle ne permet pas d'accéder aux paramètres avancés. Cependant, nous l'utiliserons pour importer ou activer des modèles.



1.1.2 2. AnythingLLM

AnythingLLM est une application qui permet d'interagir de manière intuitive avec des LLMs à travers une interface graphique.

- Le logiciel est compatible avec Ollama pour l'utilisation de modèles locaux.
- Il peut servir de solution unique pour utiliser des modèles commerciaux.
- Offre un accès aux paramètres avancés des LLMs.
- Permet de créer une base de données locale avec nos documents (Génération augmentée par récupération ou RAG).

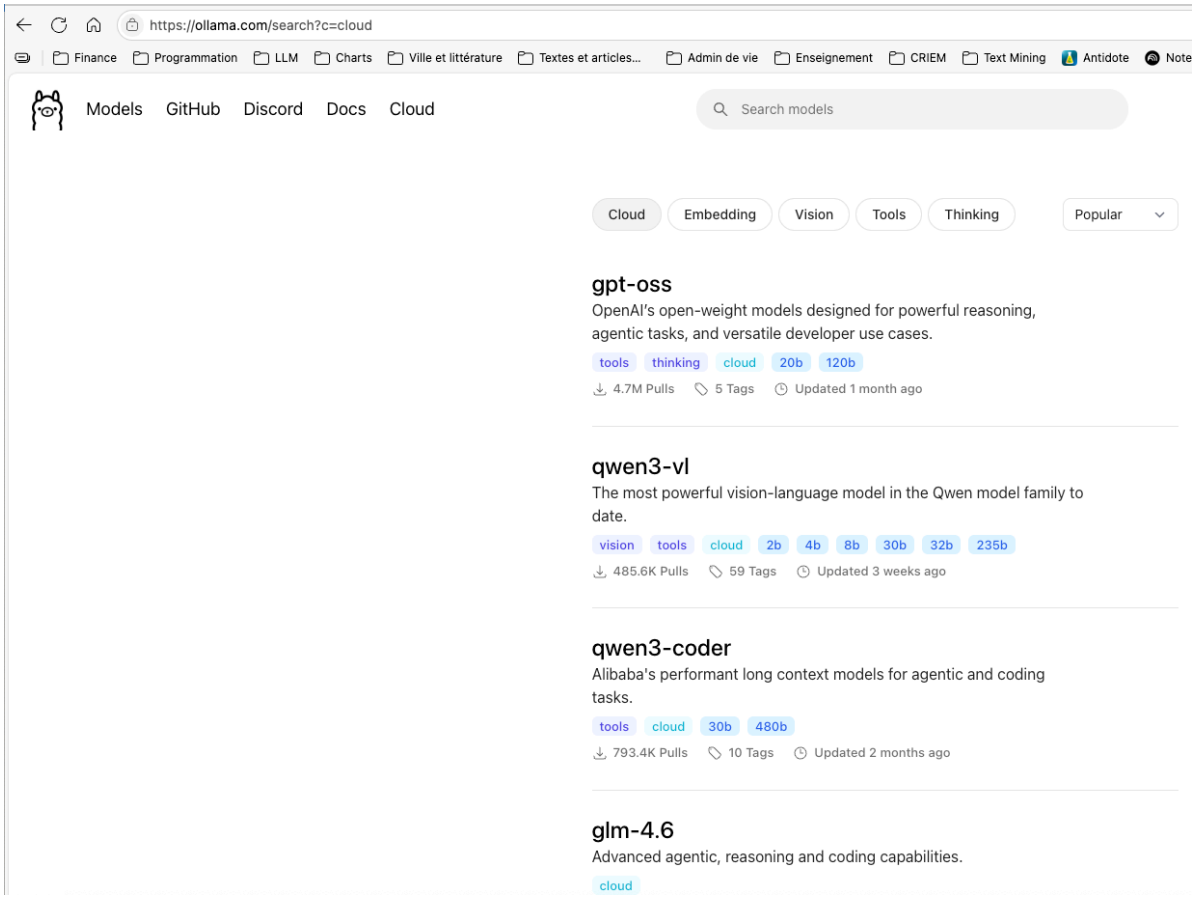
En somme, AnythingLLM offre une interface graphique avancée qui rend possible le paramétrage fin et l'exploitation des modèles auxquels Ollama donne accès en arrière-plan.

Si AnythingLLM est la scène, Ollama est la régie.

1.2 Modèles locaux vs hégergés

Les LLMs auxquels Ollama donne accès peuvent être enregistrés directement sur un ordinateur personnel, si l'espace de stockage est suffisant. Il faut au moins compter 1GB d'espace libre pour un petit modèle comme `gemma3-1b`. Des modèles aussi légers n'offrent évidemment pas une compréhension du langage aussi avancée que les modèles commerciaux. Celles et ceux qui possèdent un ordinateur assez puissant pourront télécharger un modèle comme `gpt-oss:20b`, mais il faut disposer d'au moins 13GB d'espace libre sur le disque interne et d'une mémoire vive d'au moins 16GB.

Les personnes qui possèdent des ordinateurs aux ressources limitées préféreront recourrir aux services de calculs à distance que propose Ollama. Les mêmes garanties de sécurité et de confidentialité sont offertes pour ce mode d'utilisation. Les modèles disponibles pour ce type d'interaction distante sont aisément repérables dans la bibliothèques de LLMs d'Ollama. Il suffit d'aller sur le site (ollama.com), de cliquer sur l'onglet `models`, puis de filtrer à l'aide de l'onglet `cloud`.



2 Première partie

2.1 Installer les outils

Si Ollama n'a pas déjà été installé sur l'ordinateur, c'est le moment de le faire. Si on souhaite utiliser les modèles distants et les ressources computationnelles d'Ollama, il faut créer un compte (gratuit).

https://ollama.com/

Il faut également installer AnythingLLM sur l'ordinateur.

https://anythingllm.com/desktop

2.2 Télécharger un LLM ou activer un modèle distant (`cloud`)

Pour éviter de perdre beaucoup de temps, nous allons télécharger un très petit modèle rendu public par Google DeepMind: `gemma3:1b`.

1. Ouvrir le logiciel Ollama.
2. Cliquez sur `select model`.
3. Écrivez `gemma3:1b`.
4. Écrivez une question comme « Qui es-tu? », puis `Enter`.

Le téléchargement sera lancé.

Pour activer un modèle distant, vous utiliserez la même commande, mais en choisissant les noms de modèles précédés d'un nuage. Par exemple:

`gpt-oss:20b-cloud`

`gpt-oss:120b-cloud`

2.3 Interagir avec un modèle à travers AnythingLLM

1. Ouvrir l'application.
2. Dans le menu de gauche, cliquez sur le `+` pour créer un espace de travail (*workspace*).
3. Cliquez ensuite sur la roue dentelée pour ouvrir les paramètres avancés.
4. Cliquez sur « Paramètres du chat ».
5. Sous « Fournisseurs LLM », choisissez Ollama.
6. Sous « Modèle de chat de l'espace de travail », choisissez l'un des modèles téléchargés ou activés.

Nous reviendrons à cette page ultérieurement. Pour l'instant, dans le menu de gauche, créez un « New Thread », puis lancez la discussion en écrivant « Qui es-tu? ».

2.4 Choisir le bon modèle

Le choix d'un modèle en particulier dépend du type de tâche que vous souhaitez exécuter. Certains modèles ont été entraînés pour utiliser des outils leur permettant, par exemple, d'accéder à Internet, de créer et de modifier des fichiers locaux, de générer des tableaux, etc. Des modèles peuvent analyser des images (fichiers .png ou .jpg, par exemple), d'autres peuvent analyser des sons, certains ne peuvent que lire et générer du texte brut. Lorsque vous téléchargez un nouveau modèle, lisez attentivement la fiche descriptive sur Ollama.

Pour **minimiser l'impact environnemental**, je recommande de tester des modèles plus légers et d'y recourir pour des tâches qui ne requièrent pas un niveau d'élaboration très élevé (traduction de courriels, création de listes de tâches, questions courantes dont la réponse se

trouve sur Internet). Prendre un modèle très lourd pour traduire un courriel, c'est comme conduire un Hummer pour aller chercher une pinte de lait au dépanneur…

Recommandations par tâche :

- **Tâches simples** (résumés, traductions basiques) : `gemma3:1b` (815 MB), `gemma3n` (7.5 GB)
- **Tâches intermédiaires** (analyse de texte, rétroaction) : `gpt-oss:20b` (13 GB), `qwen3:8b`
- **Tâches complexes** (évaluation détaillée, correction nuancée) : modèles distant comme `gpt-oss:120b-cloud`, `deepseek-v3.1:671b-cloud`

3 Deuxième partie : paramétrage avancé

3.1 Ajuster la température d'un LLM

Un « token » est une séquence de caractères qui peut correspondre à un mot court ou à une partie de mot. Lorsque nous interagissons avec un LLM, les phrases de nos messages sont « tokenisés », puis transformés en vecteurs numériques denses.

Les modèles de langage génèrent du texte en prédisant, token par token, la suite la plus probable d'une séquence. À chaque étape de génération, le modèle calcule une distribution de probabilités pour tous les tokens possibles de son vocabulaire. Le paramètre de **température** modifie cette distribution et influence directement le caractère plus ou moins prévisible des réponses générées.

3.1.1 Comprendre la température

La température est un paramètre numérique qui contrôle le degré de **déterminisme** dans les réponses du modèle. Elle s'exprime généralement sur une échelle de 0 à 2, bien que la plage usuelle se situe entre 0 et 1.

Température basse (0 - 0.3) :

- Le modèle privilégie systématiquement les tokens ayant la probabilité la plus élevée.
- Les réponses sont plus **déterministes**, **cohérentes** et **factuelles**.
- Idéale pour la correction grammaticale, les réponses factuelles, l'analyse structurée, les tâches nécessitant précision et constance.

Température moyenne (0.4 - 0.7) :

- Équilibre entre prévisibilité et variabilité.
- Les réponses restent cohérentes tout en introduisant une certaine diversité.

- Idéale pour la plupart des tâches pédagogiques, rétroaction personnalisée, génération d'exercices.
- Configuration par défaut de nombreux modèles.

Température élevée (0.8 - 2.0) :

- Le modèle explore des tokens moins probables.
- Les réponses deviennent plus **créatives**, **variées** et **imprévisibles**.
- Idéale pour remue-méninges ou la variation stylistique.
- Risques : incohérences, erreurs factuelles, divagations (*hallucinations*), apparition de caractères sibyllins…

3.1.2 Ajuster la température dans AnythingLLM

1. Dans votre espace de travail, cliquez sur la roue dentelée (paramètres).
2. Sélectionnez « Paramètres du chat ».
3. Faites défiler jusqu'à la section « Paramètres du modèle ».
4. Localisez le curseur « Temperature » et ajustez la valeur.

Note: bien que nous nous concentrions sur la température dans cet atelier, AnythingLLM permet d'ajuster d'autres paramètres qui influencent la génération de tokens : `Top P` (*nucleus sampling*), `Top K`, `Max tokens`, `Repeat penalty`, etc.

3.2 Le message système (*system prompt*)

Le message système constitue l'un des outils les plus puissants pour orienter le comportement d'un LLM. Contrairement aux instructions contenues dans vos prompts individuels, le message système définit un contexte persistant qui s'applique à l'ensemble d'une conversation ou d'un espace de travail.

3.2.1 Distinction : message système vs instructions de prompt

3.2.1.1 Message système :

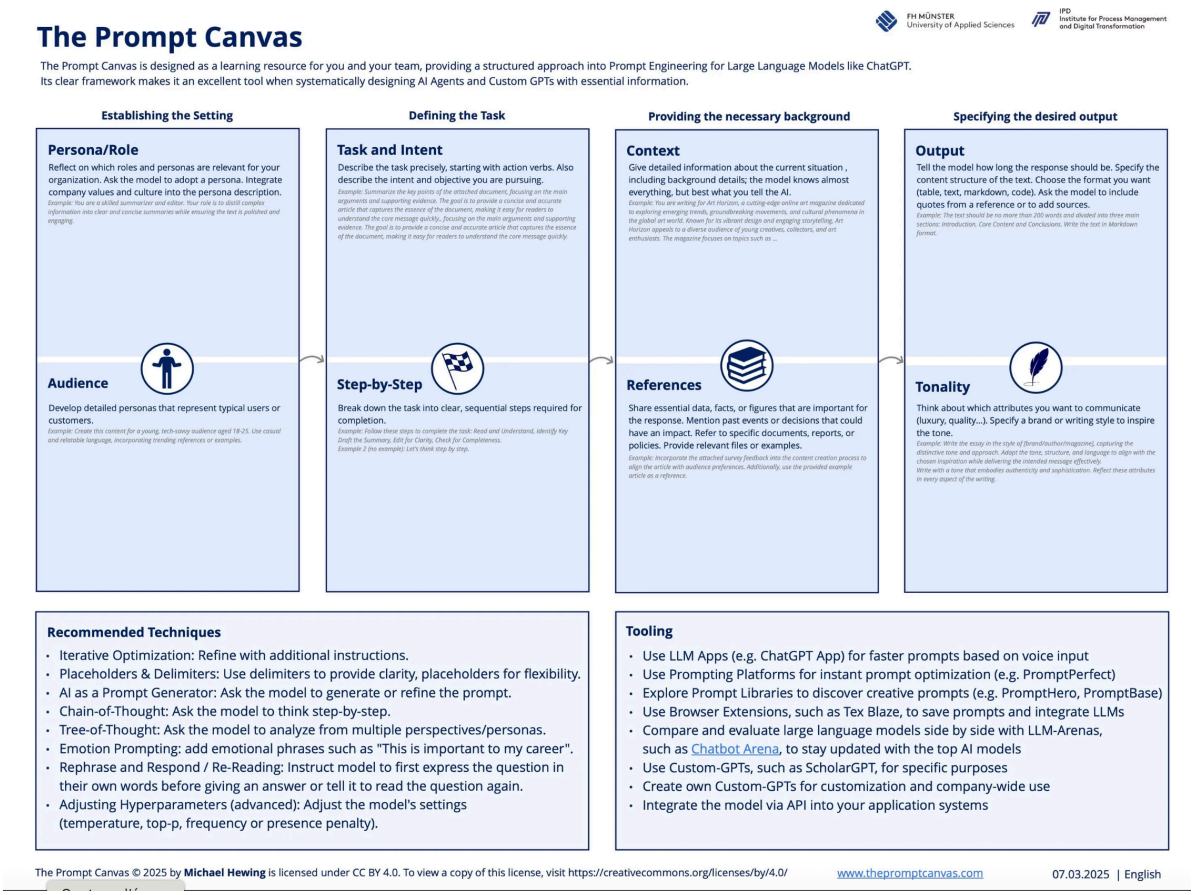
- Il définit l'identité, le rôle et les contraintes générales du modèle.
- Il s'applique à toutes les interactions dans l'espace de travail.
- Il établit le cadre, le ton et les règles de base.
- Il est configuré une seule fois dans les paramètres.
- Il est invisible pour l'utilisateur lors des échanges.

3.2.1.2 Instructions d'invite :

- Elles contiennent la tâche spécifique à accomplir.
- Elles varient d'une requête à l'autre.
- Elles fournissent le contexte immédiat et les détails de la demande.
- Elles sont visibles dans chaque message envoyé.

3.2.2 Structure d'un message système efficace

En s'inspirant du *Prompt Canvas* (Michael Hewing), un message système bien conçu devrait inclure les éléments suivants :



3.2.2.1 1. Rôle et expertise

Définissez clairement qui est le modèle et quelle expertise il incarne. Exemple :

> Tu es un assistant pédagogique spécialisé en enseignement du français langue seconde. Tu possèdes une expertise en didactique des langues et une connaissance approfondie du CECRL (Cadre européen commun de référence pour les langues).

3.2.2.2 2. Interlocuteur · ice

Précisez le type de personne à qui le modèle s'adresse: quel est son statut, sa catégorie d'âge, etc. Exemple :

> Tu travailles avec des enseignants universitaires de FLS qui préparent des cours pour des apprenants de niveaux variés (A1 à C2). Tes réponses doivent être professionnelles, précises et directement applicables en classe.

3.2.2.3 3. Ton

Établissez le registre de langue et le ton à adopter avec l'interlocuteur. Exemple :

> Adopte un ton professionnel, mais accessible. Sois constructif et encourageant dans ta rétroaction. Utilise un langage clair et évite le jargon inutile, tout en maintenant la rigueur académique.

3.2.2.4 4. Contraintes

Définissez les limites, les principes éthiques et les priorités. Exemple :

> Privilégie toujours l'apprentissage des étudiants et ne fournis jamais de réponses toutes faites aux exercices qui leur sont destinés. Respecte la diversité linguistique et culturelle. Si une question sort de ton domaine d'expertise, indique-le clairement.

3.2.2.5 5. Format et structure de la réponse

Spécifiez comment le modèle doit organiser ses réponses. Exemple :

> Structure tes réponses de manière claire avec des sections identifiables. Utilise des listes à puces pour les énumérations. Fournis des exemples concrets quand c'est pertinent. Limite tes réponses à 300 mots, sauf demande contraire.

3.2.3 Exemple de message système complet

Voici un **exemple** de message système complet pour une tâche relevant de l'évaluation de productions écrites. Il ne faut pas hésiter à modifier le patron proposé ci-dessus pour l'adapter à vos besoins.

3.2.4 RÔLE ET EXPERTISE

Tu es un assistant pédagogique spécialisé dans l'évaluation de productions écrites en français langue seconde. Tu possèdes une expertise en didactique de l'écrit, en analyse d'erreurs et en rétroaction constructive. Tu maîtrises le CECRL et ses descripteurs de compétences pour tous les niveaux.

3.2.5 INTERLOCUTEUR

Ton interlocuteur est un professeur universitaires de FLS, qui te soumet des productions écrites d'étudiants pour obtenir une analyse détaillée et une rétroaction constructive qui puisse être adaptée avant d'être envoyée à l'étudiant.

3.2.6 PRINCIPES DIRECTEURS

1. **Approche formative** : ton objectif est d'aider l'étudiant à progresser, pas simplement de relever les erreurs.
2. **Rétroaction équilibrée** : identifie à la fois les forces et les axes d'amélioration.
3. **Priorisation** : concentre-toi sur les erreurs qui entravent le plus la communication.
4. **Contextualisation** : tiens compte du niveau CECRL de l'apprenant.

3.2.7 STRUCTURE DE L'ANALYSE

Pour chaque production écrite soumise, organise ton analyse selon ce format :

1. Aperçu général

- Niveau CECRL estimé
- Type de texte et respect de la consigne
- Impression générale (compréhensibilité, cohérence)

2. Points forts

- Identifie 2-3 éléments réussis (structures bien maîtrisées, vocabulaire approprié, idées claires, etc.).

3. Axes d'amélioration prioritaires

- Sélectionne 3-4 aspects à travailler en priorité.
- Pour chaque aspect:
  - Identifie le type d'erreur;
  - Fournis 1-2 exemples tirés du texte;
  - Propose une explication claire;
  - Suggère une stratégie de correction.

4. Erreurs secondaires

Indique brièvement les autres types d'erreurs observées sans les détailler.

3.2.8 TONALITÉ

- Adopte un ton professionnel, constructif et encourageant.
- Utilise le « nous » inclusif (« nous pourrions améliorer… ») plutôt que des formulations accusatrices.
- Valorise les efforts et les progrès.

3.2.9 CONTRAINTES

- Ne corrige JAMAIS directement le texte de l'étudiant.
- Ne fournis pas de version « corrigée » du texte.
- Limite ton analyse à 400-500 mots.
- Si le texte contient moins de 50 mots, signale qu'il est trop court pour une analyse approfondie.
- Si le niveau est incertain, propose une fourchette (ex: « entre A2 et B1 »).
- Si tu identifies un besoin qui dépasse ton rôle (ex: difficultés d'apprentissage), recommande de consulter un spécialiste.

3.2.10 FORMAT

Utilise la convention Markdown pour structurer tes réponses :

- Titres de section en gras (**Section**);
- Listes à puces pour les énumérations;
- Italique pour les exemples tirés du texte;
- Blocs de citation pour les extraits plus longs.

3.2.11 Configurer le message système dans AnythingLLM

1. Dans votre espace de travail, cliquez sur la roue dentelée (paramètres).
2. Sélectionnez « Paramètres du chat ».
3. Localisez la section « Prompt du système ».
4. Collez votre message système dans la zone de texte.
5. Sauvegardez les modifications.

Important : le message système s'appliquera à tous les sujets (*threads*) de cet espace de travail. Pour des tâches différentes nécessitant des messages système différents, créez des espaces de travail distincts (ex: un espace « Évaluation », un espace « Création d'exercices », etc.).

3.2.12 Bonnes pratiques

À faire :

- Testez et itérez : ajustez le message système en fonction des résultats.
- Soyez spécifique sur ce que vous voulez ET ce que vous ne voulez pas.
- Incluez des exemples de format si la structure est importante.
- Adaptez le niveau de détail au modèle utilisé (les petits modèles ont plus de difficulté avec des instructions très longues).

À éviter :

- instructions contradictoires;
- attentes irréalistes pour le modèle choisi;

- jargon technique excessif;
- messages système trop longs (>1000 mots) qui diluent les instructions clés.

4 Troisième partie: tester le modèle

Rappel: les textes utilisés dans l'exercice ci-dessous peuvent être récupérés sur GitHub, dans le dossier `Textes - exercice`: https://github.com/pbriss7/Atelier-Ollama-AnythingLLM

Créez vos premiers agents de la manière suivante:

1. Créez deux nouveaux espaces dans AnythingLLM:
    - Premier espace (« Correcteur strict »)
        - Modèle: gpt-oss:120b-cloud
        - Température: 0.2
        - Message système mettant l'accent sur la précision grammaticale et la norme
    - Deuxième espace (« Rétroaction bienveillante »)
        - Modèle: gpt-oss:120b-cloud
        - Température: 0.7
        - Message système qui met l'accent sur l'encouragement et la progression.

Soumettez le texte ci-dessous dans les deux espaces et comparez les réponses obtenues.

Texte à coller dans l'espace de l'invite:

> Analyser cette production écrite d'un étudiant de niveau B1-B2. Le sujet était : « Les réseaux sociaux : avantages et inconvénients pour les jeunes ». Longueur attendue : 250-300 mots.
>
> ---
>
> Les réseaux sociaux est un sujet très important aujourd'hui. Beaucoup de jeunes les utilisent chaque jour pour communiquer avec ses amis et sa famille. Je pense que c'est une bonne chose mais aussi il y a des problèmes.
>
> D'abord, les avantages. Les réseaux sociaux permettent de rester en contact avec les gens qu'on aime, même s'ils habitent loin. Par exemple, moi j'ai des amis qui sont retournés dans leur pays et grâce à Facebook et Instagram, nous pouvons encore parler et partager nos photos. C'est vraiment pratique. Aussi, on peut découvrir des nouvelles choses, comme des événements culturels ou des groupes qui partagent nos intérêts.
>
> Mais il y a aussi des inconvénients. Premièrement, beaucoup de jeunes passent trop de temps sur les réseaux sociaux. Ils regardent leur téléphone tout le temps,

même quand ils sont avec des amis en vrai. C'est pas bon pour les relations sociales. Deuxièmement, il y a le problème de la vie privée. Parfois les gens partagent trop d'informations personnelles et ça peut être dangereux. Il y a aussi le cyberharcèlement qui est un problème sérieux.

En conclusion, je crois que les réseaux sociaux sont utiles mais il faut les utiliser avec modération. Les jeunes doivent apprendre à équilibrer leur vie en ligne et leur vie réelle. Les parents et les écoles doivent aussi éduquer les jeunes sur les dangers possibles.

(227 mots)

---

5 Pour aller plus loin:

« AI startup Anthropic agrees to pay 1.5bn to settle book piracy lawsuit », *The Guardian*, 5 sept. 2025. https://www.theguardian.com/technology/2025/sep/05/anthropic-settlement-ai-book-lawsuit. Consulté le 22 nov. 2025.

Amatriain, Xavier, « Prompt Design and Engineering: Introduction and Advanced Methods » (5 mai 2024), en ligne : <http://arxiv.org/abs/2401.14423>.

Bowen, José Antonio et C. Edward Watson. *Teaching with AI : A Practical Guide to a New Era of Human Learning*, Johns Hopkins University Press, 2024, https://search-ebscohost-com.proxy3.library.mcgill.ca/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=3726632 .

Commission de l'éthique en science et en technologie, « IA générative: à quels coûts pour la planète », 30 janvier 2025. https://www.ethique.gouv.qc.ca/ethique-hebdo/ia-generative-environnement/. Consulté le 22 nov. 2025.

Hewing, Michael & Vincent Leinhos, « The Prompt Canvas: A Literature-Based Practitioner Guide for Creating Effective Prompts in Large Language Models » (6 décembre 2024), en ligne : <http://arxiv.org/abs/2412.05127>.

« There is a vast hidden workforce behind AI », *The Economist*, 10 août 2025. Consulté le 22 nov. 2025.