

# Documentation du système d'extraction et de quantification des labels d'information textuelle

<b>Fonctionnalités implémentés.....</b>	<b>1</b>
Interface utilisateur.....	1
Quantification des labels.....	2
Base de données et rapports.....	2
Intégration et scalabilité.....	2
<b>Méthodes implémentées.....</b>	<b>2</b>
Factualité.....	3
Lisibilité.....	3
<b>Guide.....</b>	<b>4</b>
Importation des fichiers XLSX.....	6
Utilisation des méthodes d'extraction et d'évaluation.....	7
Exportation de l'évaluation sous XLSX.....	9

Le système à la base de ce document a été basé sur l'article "[An Information Nutritional Label for Online Documents](#)", où un label permettant d'évaluer la qualité d'un article, en s'appuyant sur les critères suivants : la factualité, la lisibilité, la viralité, l'émotion, l'opinion, la controverse, l'autorité, la crédibilité, la technicité et la topicalité. Uniquement la factualité et la lisibilité ont été implémentés dans ce système.

La factualité désigne à déterminer si l'information est factuelle du point de vue de l'auteur. Tandis que la lisibilité est le niveau de difficulté du texte, basé sur la qualité de rédaction et le public visé.

## Fonctionnalités implémentées

### Interface utilisateur

Il est possible d'importer des fichiers PDF et XSLX à évaluer, et de rédiger un texte à évaluer dans la boîte de dialogue. L'utilisateur a le choix d'extraire les valeurs des deux labels, ou d'en choisir un seul.

L'interface dispose également d'une autre boîte de dialogue permettant de visualiser les résultats des analyses précédentes.

### Quantification des labels

Les méthodes sont quantifiées et normalisées par rapport au nombre de phrases évaluées. Le score de la factualité va de 0 (Contrefait) à 1 (Fait), tandis que celui de la lisibilité varie de 0 à 5 (de A1 à C2).

### Base de données et rapports

Le système génère des rapports par texte évalué où le score de chaque phrase est affiché, permettant de comprendre le score général attribué. Il existe un historique par session, mais il est tout à fait possible d'exporter les rapports des textes évalués sous format XLSX.

### Intégration et scalabilité

Notre système est entièrement en Python, ce qui lui permet d'être facilement récupérable et adapté, notamment les méthodes des critères sélectionnés.

## Méthodes implémentées

L'approche choisie a été l'utilisation d'un modèle de Transformers, qui sont des modèles d'apprentissage profond, principalement utilisés dans le traitement du langage naturel dans le contexte des Grand Modèles de Langage, des modèles entraînés sur une grande quantité de texte.

Le modèle sélectionné pour factuality est "[XLM-RoBERTa](#)", pour sa capacité à traiter les données textuelles multilingues, du fait de son entraînement sur 100 langues avec des données d'au moins 1 million de mots par langue.

Le modèle et le tokenizer (qui est un outil qui transforme le texte en une liste structurée de morceaux compréhensibles ou tokens, tels que des mots) sont sauvegardés pour faciliter la réutilisation et le partage, avec une démonstration sur la compression des fichiers pour une distribution efficace.

Notre choix s'est porté sur du fine-tuning, qui est une technique permettant de spécialiser un modèle pré-entraîné sur une tâche spécifique, dans notre cas, une classification binaire pour Factuality. Pour la classification multiclasse de Readability, suite à des résultats de fine-tuning qui ne sont pas à la hauteur de nos attentes, nous avons décidé d'utiliser un modèle pré-fait par l'équipe readme++ disponible [ici](#).

Deux techniques de fine-tuning ont été utilisées :

- 1) PEFT (pour Parameter-Efficient Fine-Tuning) est une librairie qui a pour but d'adapter les paramètres du modèle de manière plus efficace, réduisant le sur-apprentissage, et améliore la généralisation sans un coût computationnel excessif.
- 2) LoRA (pour Low-Ranking Adaptation) permet de modifier le comportement du modèle avec une augmentation minimale des paramètres, conservant ainsi l'efficacité tout en améliorant la performance sur la tâche spécifiée.

## Factualité

Le modèle a été entraîné sur l'ensemble de données [SUBJ](#) présenté dans cet [article](#), conçu pour la subjectivité et l'objectivité du document. Cet ensemble de données n'est disponible qu'en anglais, mais il a été traduit afin de permettre un entraînement sur les langues suivantes : français, italien, espagnol, et allemand.

Par la suite, nous avons converti les fichiers en fichier CSV afin d'utiliser la librairie "dataset" de Hugging Face.

Les hyperparamètres de ce modèle sont 5 Epochs et une taille de batch de 5.

Dans la conception de cette méthode, nous nous sommes inspirés du modèle "[Fact-Or-Opinion](#)", qui reprend le même modèle, ensemble de données, et hyperparamètres.

Le code du fine tuning de la factuality est disponible [ici](#), les tests [ici](#), et le rapport du modèle [ici](#), l'ensemble des données traduites et sous CSV [ici](#).

## Lisibilité

L'ensemble de données utilisé est l'ensemble [ReadMe++](#), qui adapte la notation CEFR (de A1 jusqu'à C2) sur plusieurs langues (anglais, français, arabe, hindi, russe). L'ensemble de données [Merlin](#) a aussi été utilisé pour tester sur les langues (tchèque, allemand, italien)

mais ne permet pas d'obtenir des résultats convenables dû à la forme des données difficile à traiter.

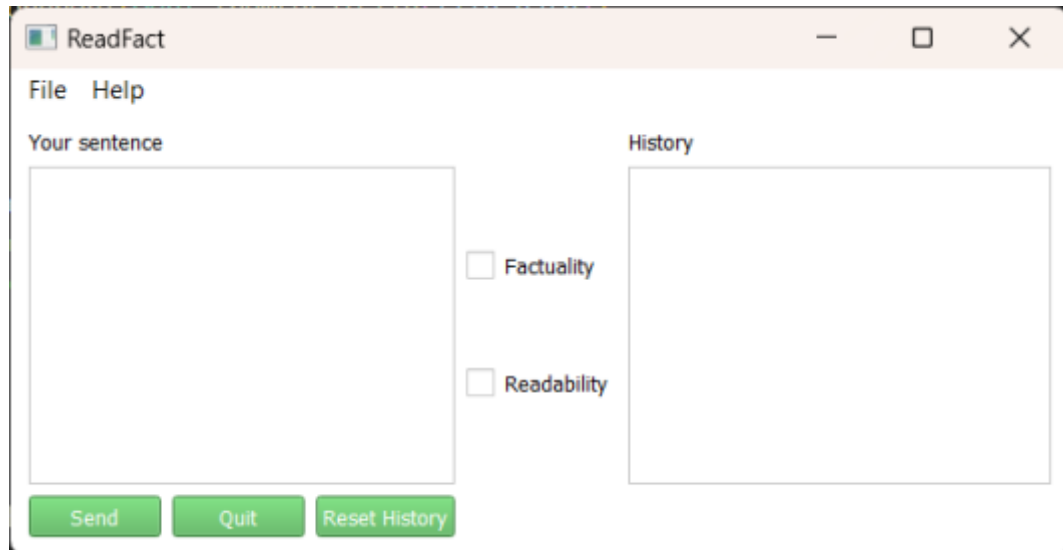
Label	Notation CEFR correspondante
1	A1
2	A2
3	B1
4	B2
5	C1
6	C2

Nous avons essayé plusieurs paramètres et différents ensemble de langues pour l'apprentissage, les résultats de fine tuning (code [ici](#)) essayés ne sont pas satisfaisants. Nous avons décidé d'utiliser le code de [ReadMe++](#) dans notre système afin d'avoir de meilleurs résultats.

Le dataset utilisé dans le code est accessible [ici](#).

# Guide

L'interface utilisateur se présente comme ce qui suit :

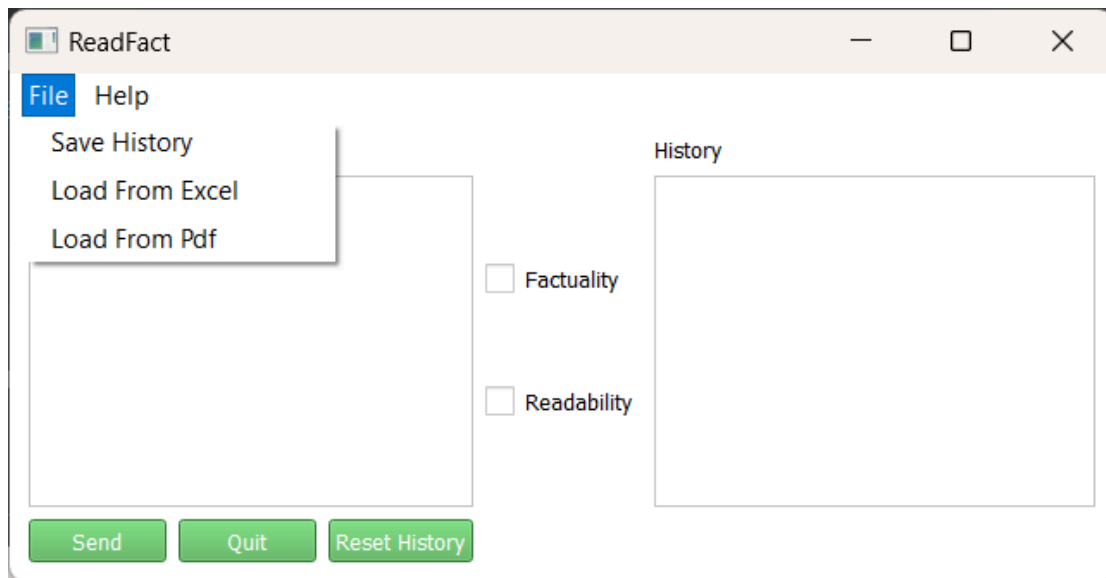


Trois boutons sont présents sous la première boîte de dialogue

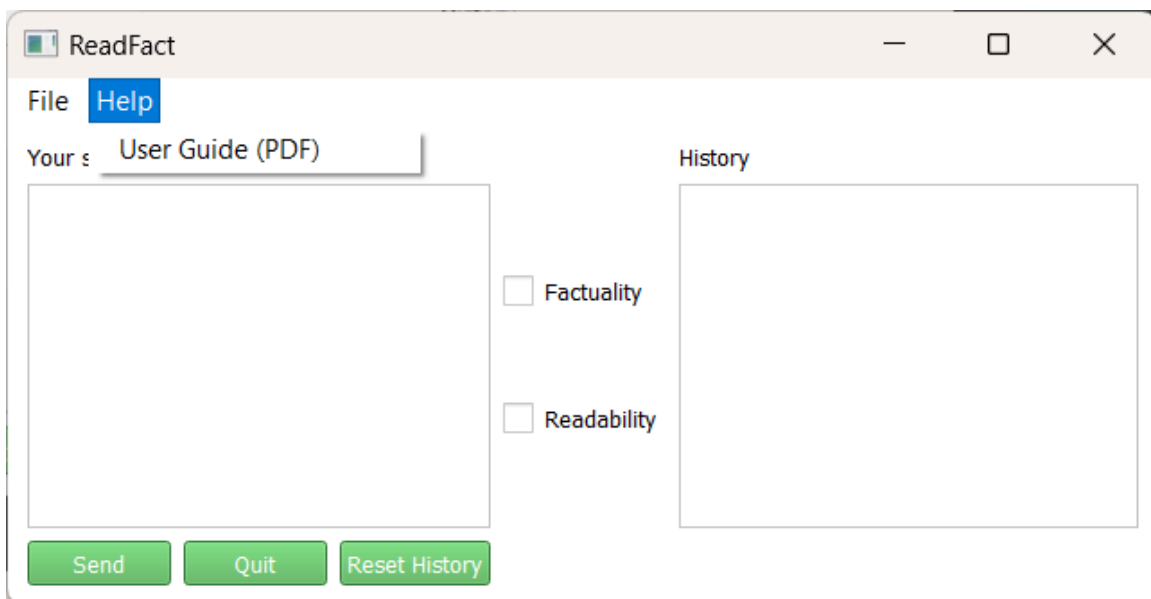
- 1) Send : permet d'évaluer le texte introduit selon les critères sélectionnés
- 2) Quit : permet de fermer le système
- 3) Reset history : nettoie l'historique, qui se situe sur la deuxième boîte de dialogue.

Deux onglets sont présentés sur la barre d'outils :

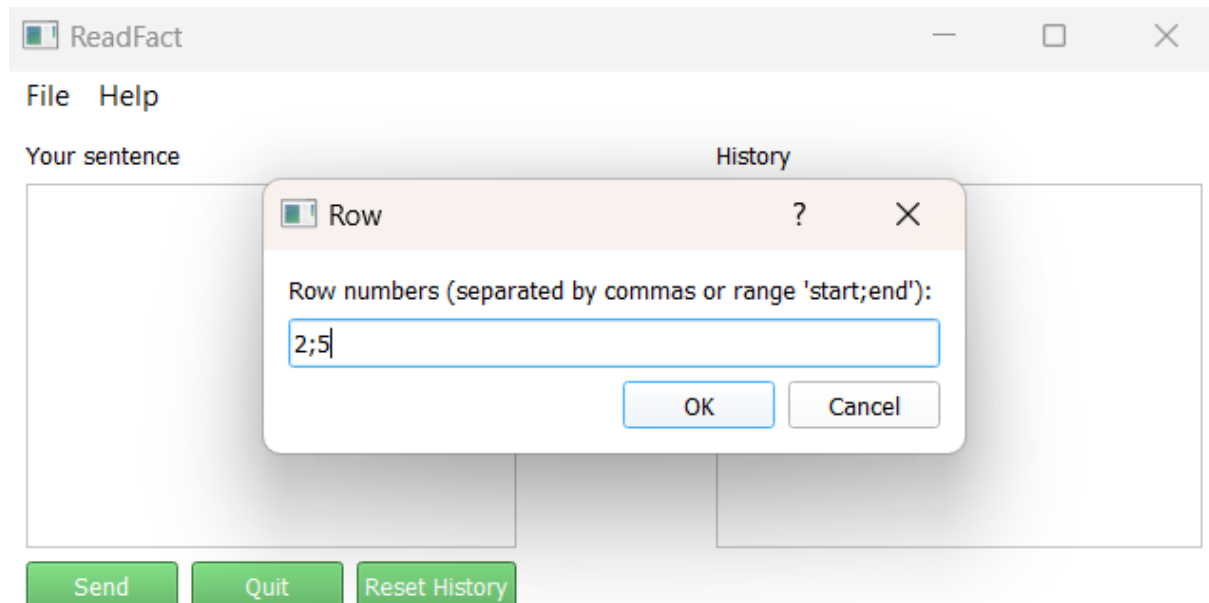
- 1) File
  - a) Save History : permet de sauvegarder l'historique des fichiers évalués sous forme de CSV.
  - b) Load from PDF: permet de charger un fichier PDF à évaluer.
  - c) Load from Excel : permet de charger un fichier XSLX à évaluer.



- 2) Help  
a) User Guide



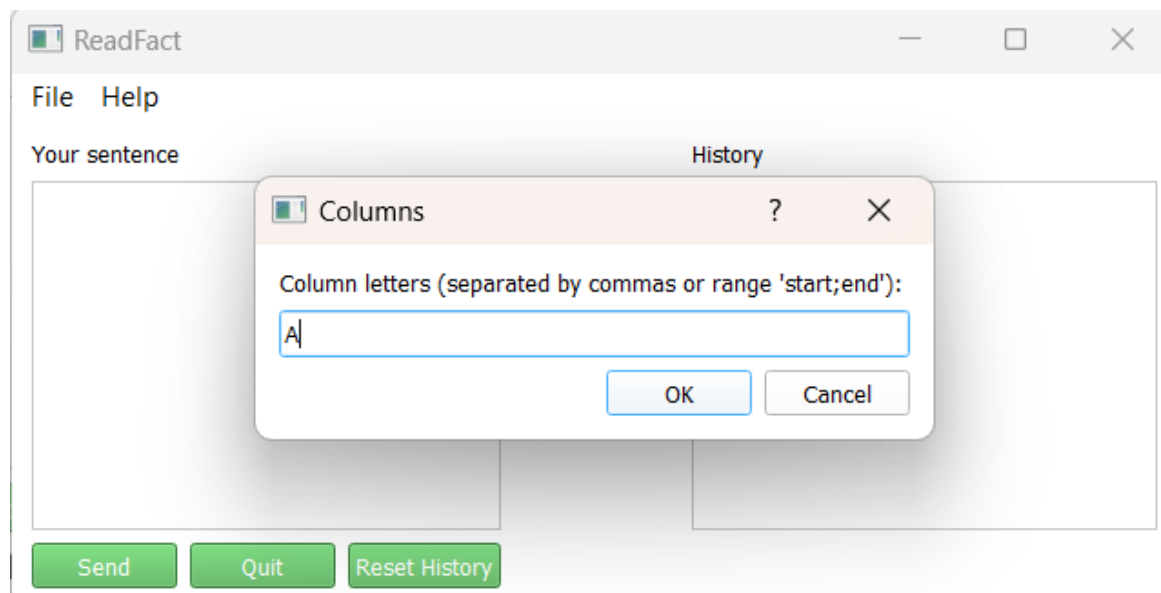
## Importation des fichiers XLSX



Il est demandé à l'utilisateur de récupérer les lignes souhaités sous deux formats :

- Séparé par des virgules (ex : 1,2,5,6)
- une rangée de lignes (ex : 1;5 ce qui donne 1,2,3,4,5)

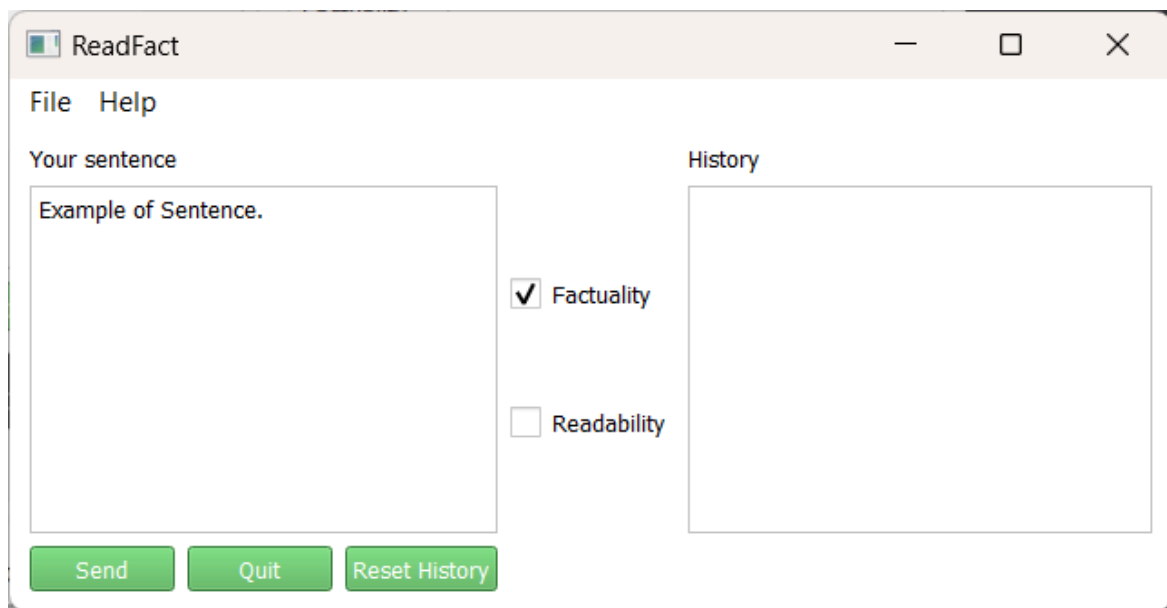
La même requête est proposée à l'utilisateur pour récupérer les colonnes souhaités



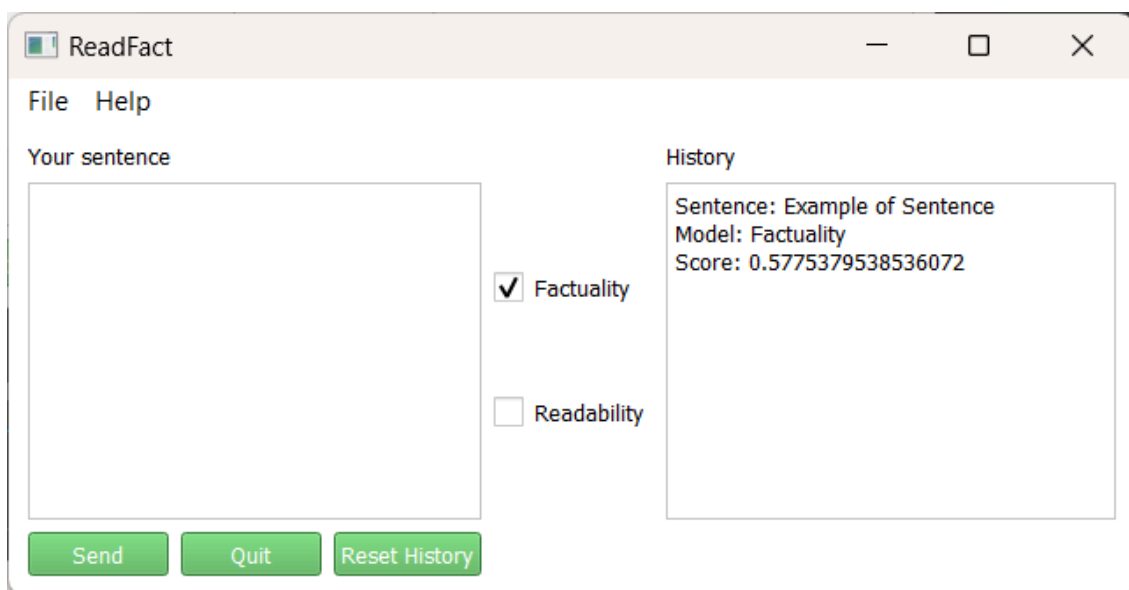
## Utilisation des méthodes d'extraction et d'évaluation

Il suffit de cocher la case "Factuality", "Readability" ou les deux selon ce qu'on désire, d'introduire le texte à évaluer (ou d'importer le document à évaluer), puis de cliquer sur "Send".

Exemple de Factuality



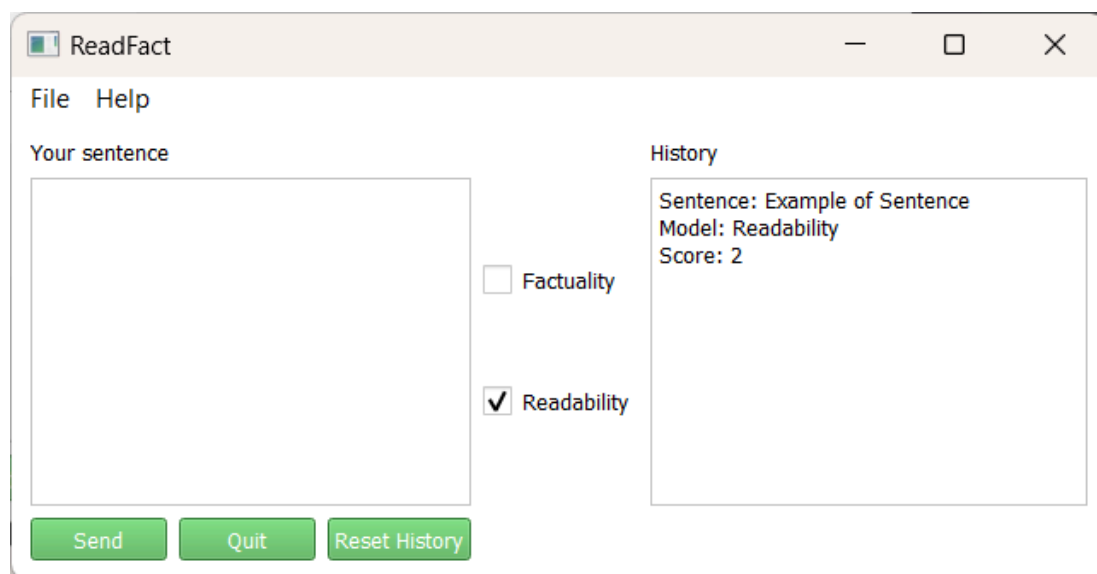
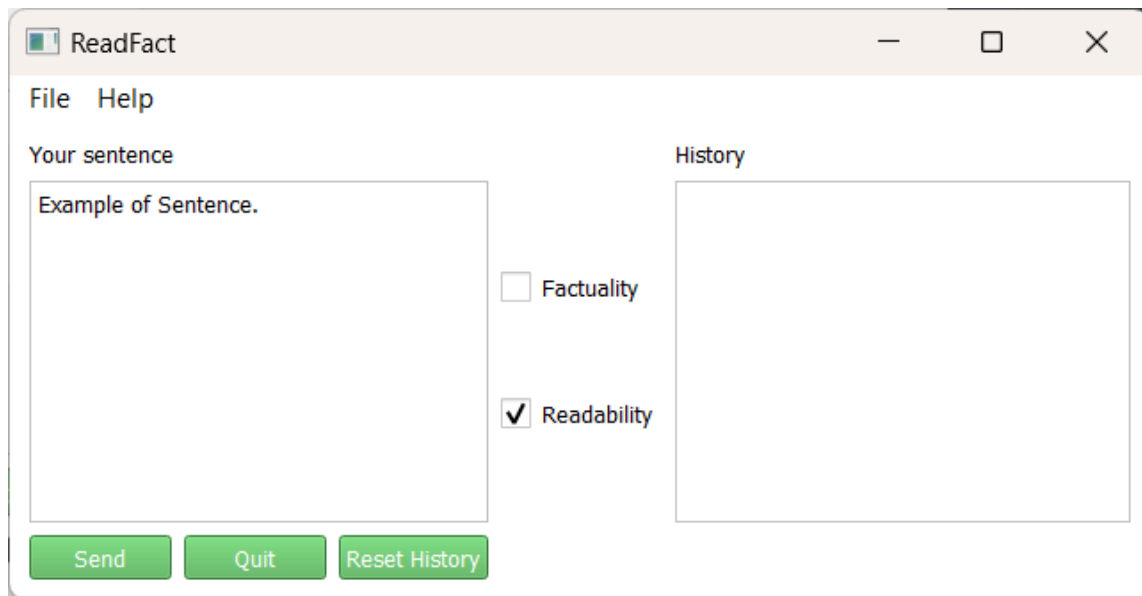
The screenshot shows the ReadFact application window. The title bar is 'ReadFact'. The menu bar has 'File' and 'Help'. The main area is divided into two sections: 'Your sentence' on the left and 'History' on the right. The 'Your sentence' section contains a text box with the text 'Example of Sentence.' and two checkboxes: 'Factuality' (checked) and 'Readability' (unchecked). Below the text box are three buttons: 'Send', 'Quit', and 'Reset History'. The 'History' section is empty.



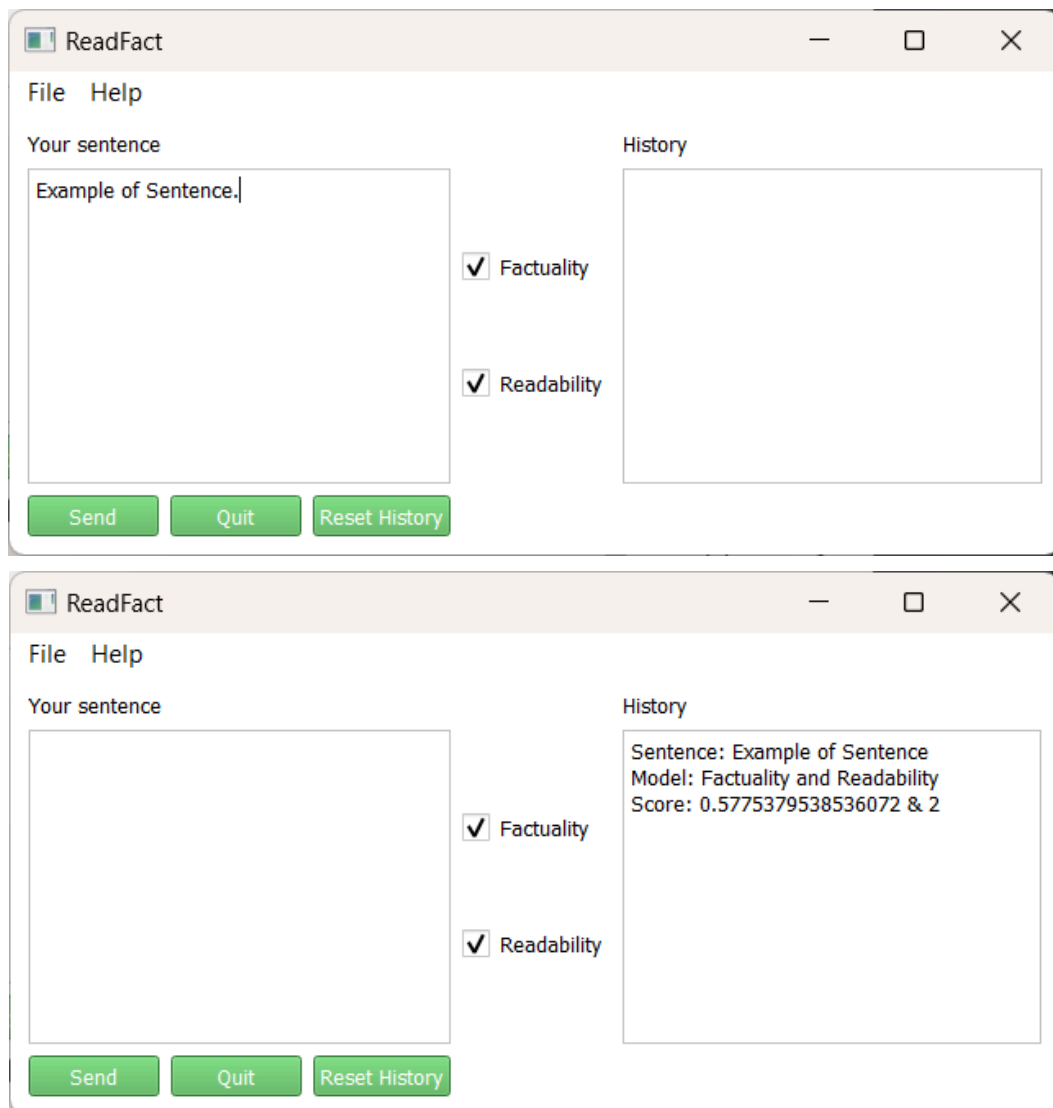
The screenshot shows the ReadFact application window after a click on the 'Send' button. The 'Your sentence' section is the same as in the previous screenshot. The 'History' section now contains the following text: 'Sentence: Example of Sentence', 'Model: Factuality', and 'Score: 0.5775379538536072'. The 'Factuality' checkbox remains checked, and the 'Readability' checkbox remains unchecked. The 'Send', 'Quit', and 'Reset History' buttons are still present.

Exemple de Readability





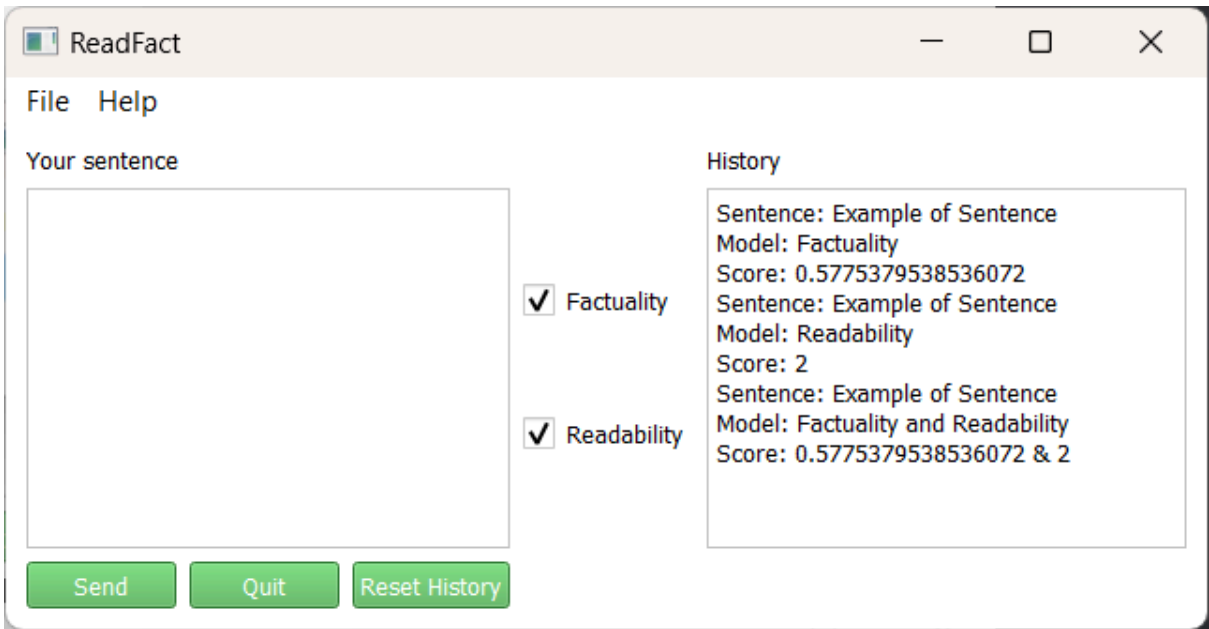
## Exemple des deux critères



La sortie "Score 0.57 & 2" signifie que le score total de la factualité est 0.57, tandis que celui de la lisibilité est 2.

# Exportation de l'évaluation sous XLSX

Une fois que le document ou le texte soumis a été évalué, il est possible d'exporter les résultats sous fichier XLSX. Voici un exemple



	A	B	C	D	E
1	Sentence	Model	Score	Average Factuality	Average Readability
2	Example of Sentence	Factuality	0.5775379538536072	0,577537953853607	2
3	Example of Sentence	Readability	2		
4	Example of Sentence	Factuality and Readab	0.5775379538536072 & 2		
5					