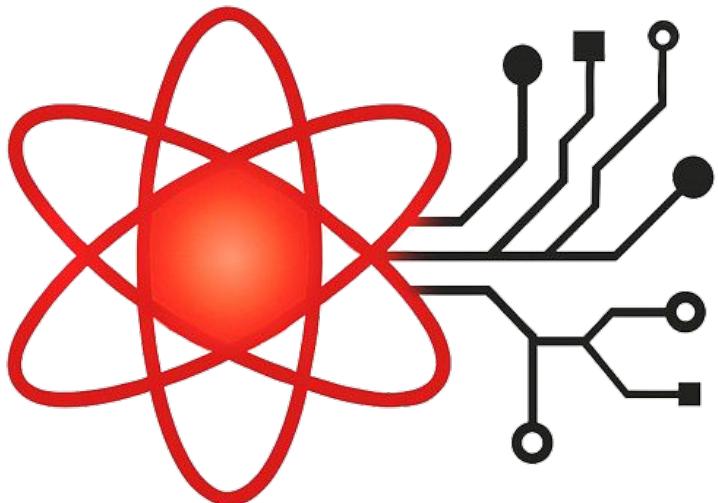




IA générative au service de la constitution d'une base de données de propriétés CLT à partir de publications scientifiques et construction d'un modèle prédictif de la vitesse d'altération initiale

Abdelhak OUANZOUGUI - Caroline CHABAL

Pierre FRUGIER – Stéphane GIN



Cadre général du stage

Contexte: Conditionnement des déchets nucléaires à haute activité à l'aide de matériaux vitreux (verres).

Phénomène: Le **contact du verre avec l'environnement** (comme l'**eau**) provoque son altération et sa dégradation à long terme.

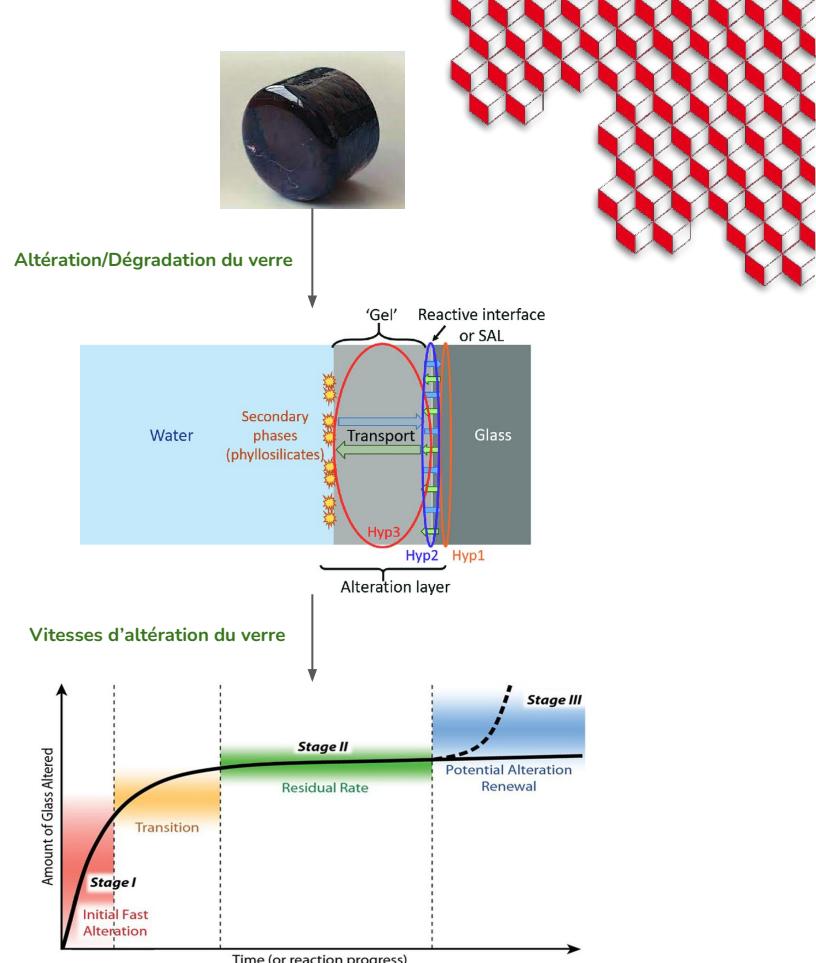
Étapes de l'altération:

1. Vitesse d'altération initiale
2. Vitesse résiduelle
3. Possible reprise d'altération

Objectif: Modéliser la vitesse d'altération initiale des verres nucléaires à l'aide d'une approche prédictive basée sur l'IA.

Problématique: Base de données non disponible.

Enjeu: Montrer la faisabilité de l'extraction automatique de données via une IA générative.



Cadre de travail :

Cas d'étude expérimental réalisé pour benchmarker les solutions d'IA générative qui seront mises à disposition des salariés du CEA.

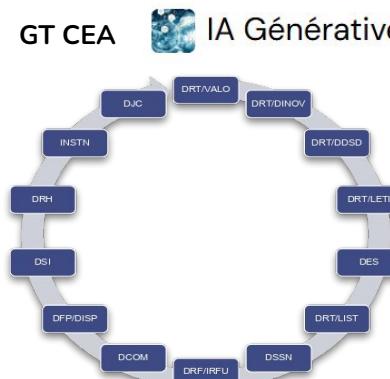


2 thématiques :

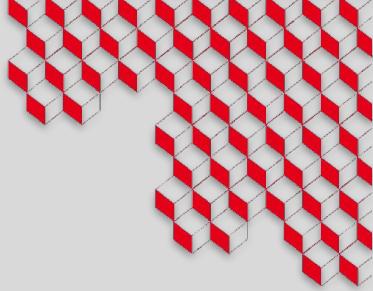
- IA générative pour les métiers CEA (Francois LE FEVRE)
 - IA générative sur les données scientifiques (Christophe CALVIN)

Limitations :

- Infrastructure limitée : 1 serveur IXIA (1 machine avec 4 GPU RTX6000)
 - Données autorisées : Diffusion limitée max
 - Utilisation de Mistral22B



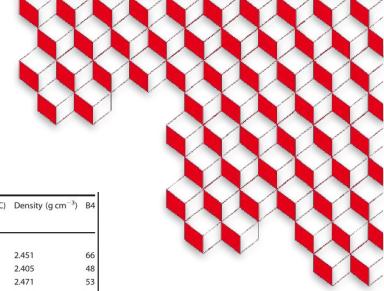
Expérimentations proposées	plateforme	qui	où
recherche de solutions techniques à des problèmes techniques afin de générer des concepts innovants	GPT4 (tbc)	Florence AUBERT, ingénieur brevets	DES/IRSENES
Utilisation de Chat GPT comme assistant de programmation sur des thèmes comme le développement de modèles d'apprentissage automatique ou l'écriture de scripts divers	GPT4	ALLINEI Pierre-Guy	DES/IRSENES/DTN /SMTA/LMN
Évaluer l'efficacité de ChatGPT à fournir des informations utiles et pertinentes en 15 secondes pour une recherche de licence sur 5 négociateurs sélectionnés	GPT4	Jean-Christophe GILLET	DRT/VALO/SAPRI /BLIC
un chatbot pour faire des recommandations à utilisateur explorant le catalogue public de formation INSTN : https://infn.cea.fr/ qui est ouvert publiquement sur Internet	GPT4 pour la partie 1 et LOOMS/IAGEN sur serveur CEA extra pour la partie 2	Minier Vincent / Moussea Olivier / Le Favre François	INSTN / DAM / DISP
demandera à l'IA, à partir de prompts avec les mot clés, de nous fourrir un état des lieux des entreprises d'un secteur donnée	GPT4	Roué Amélie / Minier Vincent	INSTN
faciliter la recherche d'art antérieur en vue du dépôt d'un brevet :	chatGPT4 (tbc)	Corinne VEDEL + Aurelie Fargier	DRT/VALO/SAPRI /BPI + SBEIM
#REF!	#REF!	#REF!	#REF!
Utilisation de l'IA pour avoir une meilleure connaissance de nos partenaires principaux et de nos futurs prospects	GPT4	Talowski	DOSP/DPMA
Aide à l'étude bibliographique : recherche de documents pertinents, synthèse, réponse à quelques questions	GPT4 et Lollms	Godin Christelle	LETI/DSYS
Aide à la rédaction de documents : rapports, articles, mails, présentations			
Aide à l'écriture de code matlab/python			
Un chatbot "vie pratique" pour tous les salariés/part de la base documentaire	LOOMS/IAGEN	ARLOT Julien	DCOM
agent conversationnel pour la mise à disposition de nouvelles procédures clarifiées, si possible simplifiées et les actions de sensibilisation/formations associées permettant aux salariés de mieux s'approprier les règles de fonctionnement et de gestion du CEA	LOOMS/IAGEN	Serge Penet	DFP
mise en production d'un agent dédié à l'aide à l'exploitation et à l'agrégation des données du tokamak WEST.	GPT4 ou LOOMS/IAGEN	Feda Almuhsen	IRFM
Fdea Almuhsen			
fournir aux porteurs de projets, engagés dans des démarches d'entrepreneuriat, une série de chatbots avancés capable de les aider:	Exp1 : GPT4 (SAAS)	PASTERNAK R. / LE FEVRE F. / DSSN	DRT/DINOV/DIR et DFP/Disp
	Exp2 : LOOMS/IAGEN	LE FEVRE Francois	DFP/Disp
expérimentations comparées de reformulation/synthèse à partir de sujets d'innovations DINOV et de rapports #VALEUR!			
générer automatiquement une BDD à partir des documents produits sur le comportement à long terme des matricées	LOOMS/IAGEN	Edith LAVIEC	DRT/DFCM/SMIR
	Langflow	Caroline Chabal	DES/SEC/OPME /
Utilisation des LLM pour mettre en place un système de recommandation de nouveaux concepts. Mais faut utiliser des LLM.			
Analyse de données multidimensionnelle pour détecter des tendances (prédictif) ou des classifications,(clusterisé) et génération de	pour les benchmark utilisation de base d'exigence open source + tests sur des LOOMS/IAGEN	Patrick Tessier	DRT/LIST/DILS/CS
#VALEUR!	Patrick BALDIT	DSI	



Partie 1 : Crédation d'une Base de Données sur l'Altération des Verres

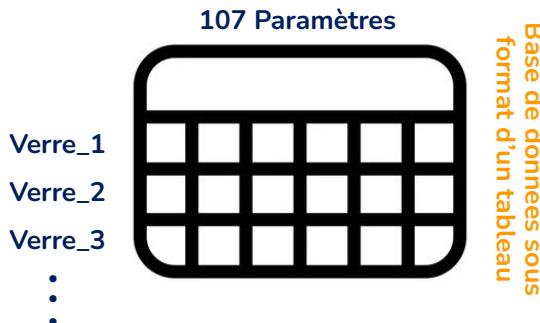
Partie 2 : Entraînement d'un réseau de neurones sur la base de données

Partie 1 : Création d'une Base de Données sur l'Altération des Verres



Partie 2 : Entraînement d'un réseau de neurones sur la base de données

La Structure de la base de données à concevoir :



Exemple de paramètres de la base de données :

- Document de référence : Titre, premier auteur, référence
- Composition du verre : Li, B, O, Na, Al ...
- Propriétés du verre : Densité, Masse, Surface, Volume, Température, ph, Vitesse d'altération initiale (Si)/(B)/(Na)...

Name	Other name in the literature	Composition (mol%)	Melting T (°C)	T _g (°C)	Density (g cm ⁻³)	B4	
		SiO ₂ B ₂ O ₃ Na ₂ O Al ₂ O ₃ CaO ZnO ₂ Li ₂ O					
<i>Borosilicate glasses</i>							
NBS14/18 ^a	CJ1 ^a	67.8 18.0 14.2	1500	570 2.451	66		
NBS17 ^a	CJ2 ^a	64.9 17.3 13.7 4.1	1450	510 2.405	48		
NBSAC ^a	CJ3 ^a	61.2 16.3 12.8 3.9 5.8	1400	510 2.471	53		
NBSACZ ^a	CJ4 ^a , ISG ^b	60.1 16.0 12.7 3.8 5.7 1.7	1400	510 2.500	48		
NBSAZ ^a	CJ7 ^a	63.8 17.0 13.4 4.0 1.8	1400	510 2.505	41		
NBSC ^c	CJ8 ^a	63.6 17.0 13.4 6.0	1400	510 2.505	66		
NBSCZ ^c	CJ9 ^a	62.5 16.7 13.1 5.9 1.8	1350	550 2.547	63		
NBS12/28	NBS17-24 ^c	60.5 27.7 11.6	1250	540 2.462	43		
NBS36/21	NBS35-19 ^c	43.2 20.7 36.2	1100	470 2.537	63		
NBS31/15	NBS29-13 ^c	54.2 15.0 30.9	1250	505 2.538	78		
<i>Aluminosilicate glasses</i>							
NSAC19 ^g		55.3 19.0 9.9 13.9 1.9	1450	610 2.591			
NSAC17		48.6 17.0 20.1 12.3 1.9	1500	580 2.569			
NSAC21 ^{g,h}		57.4 21.3 5.9 15.4	1350	580 2.592			
NSACO		57.3	5.9 36.8	1500	580 2.790		
NSAC35 ^g		58.1 34.9 6.9	1350	580 2.522			
NSA	Albite glass ^d	75.0 12.5 12.5	1650	530 2.340			

Note that the glasses NBS12/28, NBS36/21, and NBS31/15 are close but not exactly the same as those reported by Pacaud et al.²⁷. B4 represents the fraction of four fold coordinated B in the glass, determined by MAS NMR spectroscopy. Annealing was conducted at $T_0 = 20^\circ\text{C}$.

^aGin et al.³⁷

^bGin et al.³⁷

^cPacaud et al.²⁷

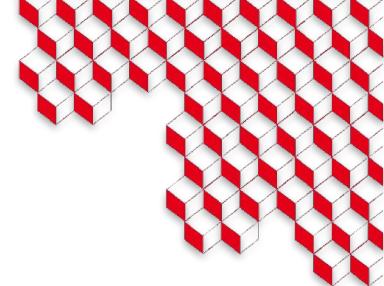
^dPerez et al.³⁸

Glass	Test #	FAG	Congruency (Na/Si)	Final pH ^{b,c}	Parameters considered for r_0 calculation	Fraction of altered glass (FAG) and pH are given at the end of the test. Uncertainty on the pH is 0.1 unit. Congruency between Na and Si is calculated at the end of the test, n.d. not determined.		
						r_0 (g m ⁻² d ⁻¹)	Ln(r ₀)	y-intercept (g m ⁻²)
NBS12/28	1	<0.01	1.1 ± 0.1	8.9	Si	192 ± 60	5.3	0.4
	1			Mass loss		223 ± 45	5.4	
NBS36/21	2	0.26		9.0	Si	202 ± 60	5.3	0.995
	1	0.24	1.2 ± 0.1	9.5	Si	47,370 ± 14,360	10.8	-31.2
NBS31/15	1	0.02	1.1 ± 0.1	9.1	Si	49,048 ± 9,909	10.8	
	1		Mass loss			2070 ± 620	7.6	3.8
NBS14/18	1	0.06	n.d.	9.0	Si	2756 ± 551	7.9	
	2	0.08	8.9	Si	17.7 ± 5.3	2.9	0.1	0.989
	3	0.07	9.0	Si	12.2 ± 3.7	2.9	0.1	0.999
NBSA	1	0.03	1.1 ± 0.1	8.9	Si	2.6 ± 0.8	0.9	-0.02
NBSAC	1	0.04	1.0 ± 0.1	8.9	Si	9.9 ± 3.0	2.3	-0.08
NBSACZ	1	0.09	1.1 ± 0.1	9.0	Si	8.2 ± 2.5	2.1	0.2
NBSAZ	1	0.02	1.0 ± 0.1	8.9	Si	2.2 ± 0.6	0.8	0.04
NBSC	1	0.24	1.0 ± 0.1	8.9	Si	52.2 ± 15.7	4.0	0.4
NBSCZ	1	0.26	1.1 ± 0.1	8.9	Si	56.4 ± 17.0	4.0	0.8
NSAC19	1	0.08		8.8	Si	5.6 ± 1.7	1.7	0.05
NSAC17	2	0.02	1.2 ± 0.1	9.1	Si	7.2 ± 2.2	2.0	-0.1
	3	0.02	1.2 ± 0.1	9.1	Si	5.5 ± 1.7	1.7	-0.12
	1	0.14		8.8	Si	9.8 ± 2.9	2.3	0.2

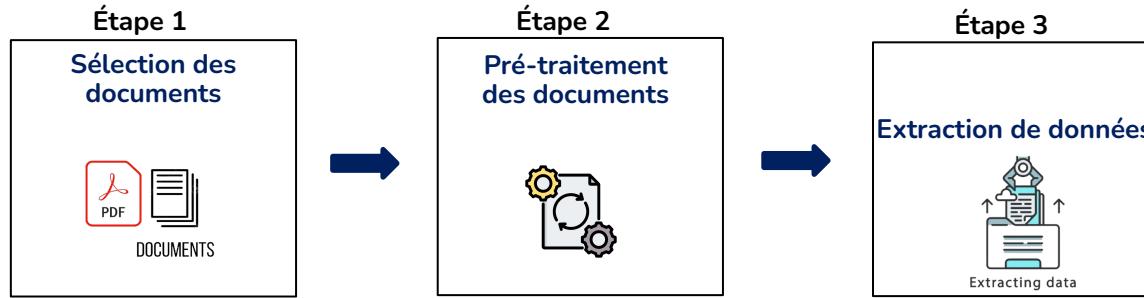
Exemple d'article sur l'altération des verres

Partie 1 : Création d'une Base de Données sur l'Altération des Verres

Partie 2 : Entraînement d'un réseau de neurones sur la base de données

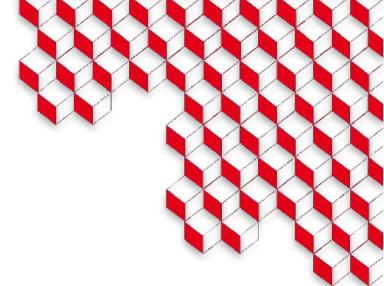


La création de la base de données s'est déroulée en trois étapes:



Partie 1 : Création d'une Base de Données sur l'Altération des Verres

Partie 2 : Entraînement d'un réseau de neurones sur la base de données

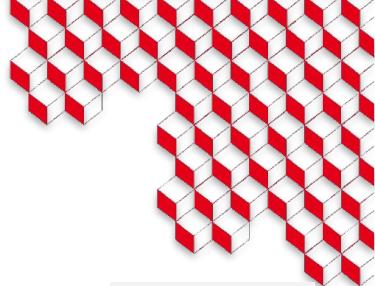


La création de la base de données s'est déroulée en trois étapes:

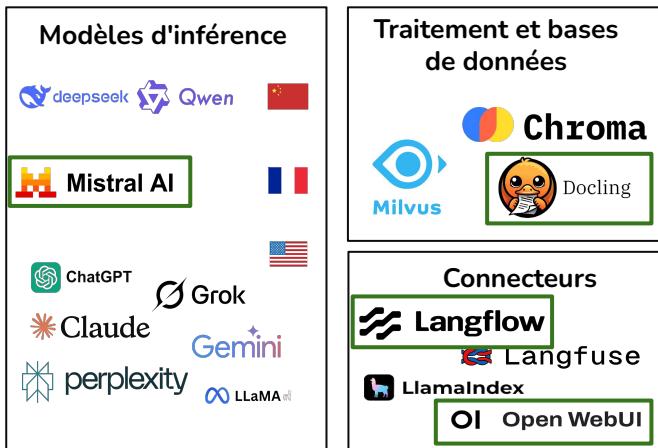


Partie 1 : Création d'une Base de Données sur l'Altération des Verres

Partie 2 : Entraînement d'un réseau de neurones sur la base de données



LLM (Large Language Model) : Modèle d'intelligence artificielle qui génère et comprend des séquences de données, comme du texte, des images ou de l'audio.



Outils de l'infrastructure CEA

Modèle d'inférence = Interface conversationnelle (par exemple, ChatGPT)

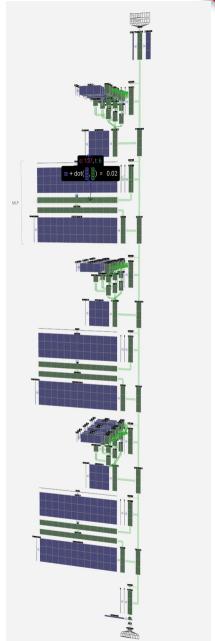
OI Open WebUI <https://openwebui-cea-exp.ixia.intra.cea.fr/>

mistralsmall-22b

Modèle d'inférence par défaut pour vos travaux au CEA, mistrail/Mistral-Small-3.1-24B- Instruct-2503

Comment puis-je vous être utile aujourd'hui ?

- + Recherche Web
- + Suggéré
- Checking grammar & typos
- Traduction en français
- Développement de Partenariats
- Traduction de Mots Techniques

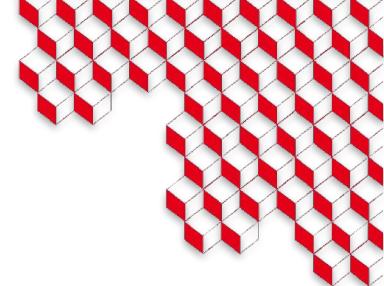


<https://bbycroft.net/llm>



Partie 1 : Création d'une Base de Données sur l'Altération des Verres

Partie 2 : Entraînement d'un réseau de neurones sur la base de données



<https://langflow-cea-exp.ixia.intra.cea.fr/>

The screenshot shows a complex workflow configuration in the Langflow platform. It starts with a 'Chat Input' component where 'Hello' is entered. This feeds into an 'OpenAI' component, which generates text using an OpenAI LLM. The 'OpenAI' component requires an 'Input' (Receiving Input), 'System Message' (Receiving Input), 'Model Name' (set to gpt-4o-mini), and an 'OpenAI API Key' (OPENAI_API_KEY). The generated text from the OpenAI component is then processed by a 'Chat Output' component, which displays the message in the playground.

Objectif :

- Créer et déployer des workflows automatisés.
- Aller au-delà des simples interactions conversationnelles pour automatiser des tâches complexes et l'analyse de données.

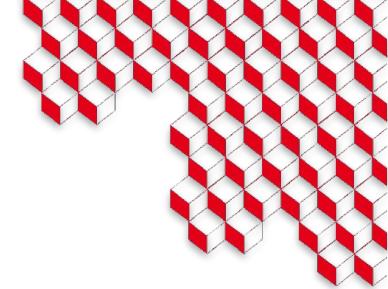
Avantages :

- Permet la conception de processus personnalisés et flexibles.
- Intègre les capacités avancées des LLM pour automatiser des tâches.
- Offre des solutions puissantes et adaptables pour la gestion des applications basées sur l'intelligence artificielle.



Partie 1 : Création d'une Base de Données sur l'Altération des Verres

Partie 2 : Entraînement d'un réseau de neurones sur la base de données

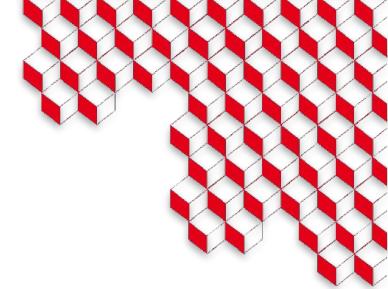


Problématique : Comment sélectionner les articles pertinents pour la base de données ?

Manuellement ?

Partie 1 : Création d'une Base de Données sur l'Altération des Verres

Partie 2 : Entraînement d'un réseau de neurones sur la base de données



Problématique : Comment sélectionner les articles pertinents pour la base de données ?

Manuellement ?

Non!

Automatiquement

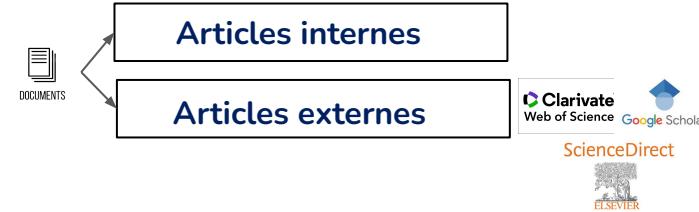
Partie 1 : Création d'une Base de Données sur l'Altération des Verres

Partie 2 : Entraînement d'un réseau de neurones sur la base de données



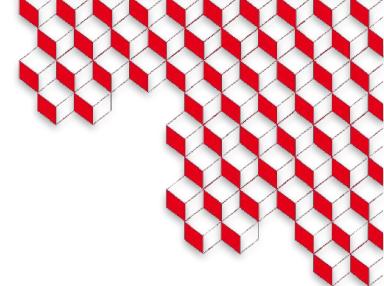
Objectif:

- Constituer un corpus d'articles potentiellement pertinents.



Partie 1 : Création d'une Base de Données sur l'Altération des Verres

Partie 2 : Entraînement d'un réseau de neurones sur la base de données



Objectif :

- Constituer un corpus d'articles potentiellement pertinents.

Remarques :

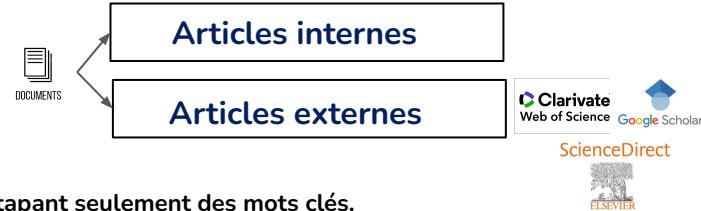
On peut récupérer **les articles externes** automatiquement en utilisant l'outil Langflow et en tapant seulement des mots clés.

The screenshot shows the arXiv search interface. The search query 'glass alteration' has been entered. The results page displays a list of papers found, with a total time of 694ms indicated at the top.



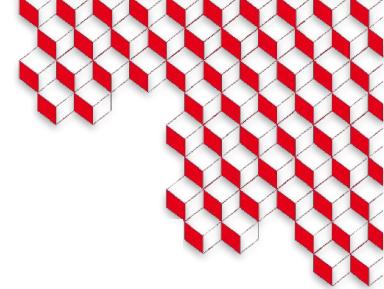
The screenshot shows the Langflow component output interface. It displays a table of search results with the following columns: id, title, summary, published, updated, authors, arxiv_url, pdf_url, comment, journal_ref, primary_category, category, and abstract. The results are listed as follows:

id	title	summary	published	updated	authors	arxiv_url	pdf_url	comment	journal_ref	primary_category	category	abstract
http://arxiv.org/abs/2408.0	Thermal conductivity in m...	The thermal conductivity o...	2024-08-01T11:11:36Z	2024-08-01T11:11:36Z	["Philip Rasmussen", "Se...	http://arxiv.org/abs/2408.0	http://arxiv.org/pdf/2408.0	The following article has b...		cond-mat.dis-nn	[cond-mat]	
http://arxiv.org/abs/1711.0	Adversarial Information F...	We propose a novel gener...	2017-11-14T14:25:09Z	2018-09-28T14:42:04Z	["Antonia Creswell", "Yu...	http://arxiv.org/abs/1711.0	http://arxiv.org/pdf/1711.0			cs.CV	[cs.CV]	
https://arxiv.org/abs/1904.1	Waste Forms for actinides:	This high level liquid radio...	2019-04-28T13:43:25Z	2019-04-30T11:29:53Z	["Bend Grambow"]	https://arxiv.org/abs/1904.1	https://arxiv.org/pdf/1904.1	21 pages	P.C. Burns, G.E. Simon, M...	cond-mat.mtrl-sci	[cond-mat]	
http://arxiv.org/abs/1904.2	Interpretation of the appar...	The temperature depende...	2020-12-14T05:24:45Z	2020-12-14T05:24:45Z	["Koush Shirai"]	http://arxiv.org/abs/1904.2	http://arxiv.org/pdf/1904.2	23 pages, 5 figures		cond-mat.dis-nn	[cond-mat]	
http://arxiv.org/abs/1811.1	Decoupling between them...	We rejuvenate well-aged q...	2018-11-07T01:31:13Z	2018-11-07T01:31:13Z	["Xiaomin Yang", "Hua To...	http://arxiv.org/abs/1811.1	http://arxiv.org/pdf/1811.1	14 pages, 6 figures		cond-mat.soft	[cond-mat]	
http://arxiv.org/abs/2411.0	Effect of Random Pinning i...	We investigate the effects...	2024-11-06T12:43:17Z	2024-11-07T03:42:38Z	["Ron Chatterjee", "Mon...	http://arxiv.org/abs/2411.0	http://arxiv.org/pdf/2411.0			cond-mat.soft	[cond-mat]	
http://arxiv.org/abs/2411.1	Nanoscale Analysis of Surf...	To investigate the effect o...	2024-11-22T10:18:52Z	2024-11-22T10:18:52Z	["Mohamad Hosseini ..."]	http://arxiv.org/abs/2411.1	http://arxiv.org/pdf/2411.1			cond-mat.soft	[cond-mat]	
http://arxiv.org/abs/2108.0	Visualizing and Quantifying...	An aqueous suspension of...	2021-08-12T06:46:16Z	2021-08-12T06:46:16Z	["Shidong Li", "Aiqi Sng", ...]	http://arxiv.org/abs/2108.0	http://arxiv.org/pdf/2108.0			physics.app-ph	[physics]	
http://arxiv.org/abs/1801.1	Theory for Swap Accelerati...	Swap algorithms can shift...	2018-01-11T14:53:18Z	2018-04-10T25:51:42	["Carolina Briz", "Edan L...	http://arxiv.org/abs/1801.1	http://arxiv.org/pdf/1801.1	8 pages, 7 figures in the m...	Phys. Rev. X, 8, 031050 (201...	cond-mat.soft	[cond-mat]	
http://arxiv.org/abs/2301.1	Electrochemically induced ...	We report the electroch...	2023-01-25T03:03:00Z	2023-01-25T15:42:07Z	["Cencon Zhao", "Shiwani ..."]	http://arxiv.org/abs/2301.1	http://arxiv.org/pdf/2301.1	31 pages, 7 Main Figures 5		cond-mat.mtrl-sci	[cond-mat]	



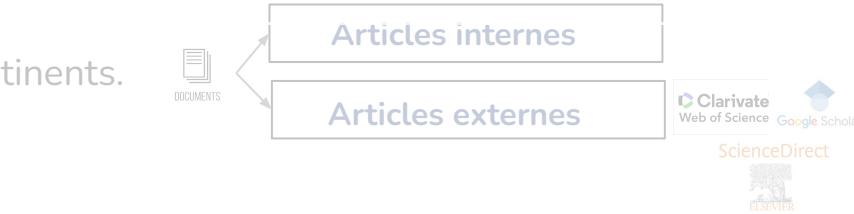
Partie 1 : Création d'une Base de Données sur l'Altération des Verres

Partie 2 : Entraînement d'un réseau de neurones sur la base de données



Objectif :

- Constituer un corpus d'articles potentiellement pertinents.
- Utiliser un outil dédié pour valider la pertinence de l'article **automatiquement** en se basant sur des critères.

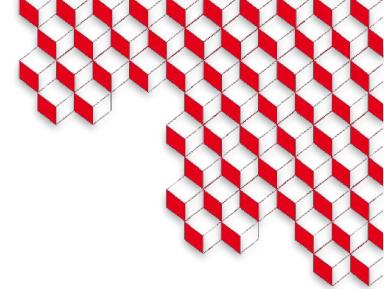


Exemple de critères de validation d'un article :

- Le document contient-il la composition du verre ?
- Le document mentionne-t-il la vitesse initiale d'altération du verre ?
- Le document inclut-il au moins un paramètre expérimental, comme le pH, la température, ou autre ?

Partie 1 : Création d'une Base de Données sur l'Altération des Verres

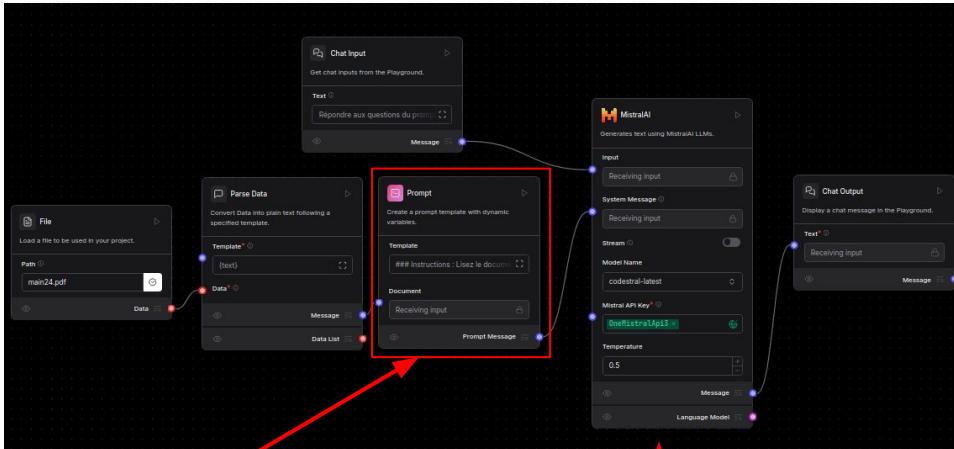
Partie 2 : Entraînement d'un réseau de neurones sur la base de données



Workflow utilisé pour la validation des documents



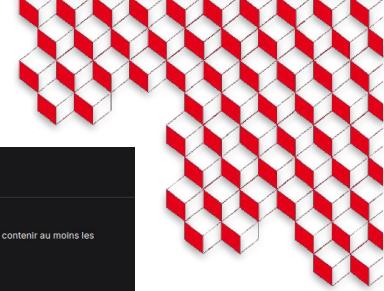
Télécharger l'article



Prompt : Définir les critères de validation

LLM (Mistral)

Partie 1 : Crédit d'une Base de Données sur l'Altération des Verres



Partie 2 : Entraînement d'un réseau de neurones sur la base de données

Étape 1

Sélection des documents
 PDF
 DOCUMENTS

Edit Prompt

Create your prompt. Prompts can help guide the behavior of a Language Model. Use curly brackets {} to introduce variables.

Instructions :

Lisez le document ci-dessous et déterminez si cet article est pertinent pour contribuer à une base de données visant à prédire la vitesse initiale d'altération des verres à l'aide de l'apprentissage automatique. Pour qu'un article soit considéré comme pertinent, il doit contenir au moins les informations suivantes pour chaque verre ou condition expérimentale évaluée :

Document: **(Document)**

1. ***Composition chimique du verre*** :
- Fournie en pourcentage (sous forme d'oxydes ou d'éléments, par exemple : SiO₂, B₂O₃, Na₂O, etc.).
- La somme des pourcentages doit être proche de 100 % (tolérance ±5 %) pour représenter la quasi-totalité de la composition. Les compositions partielles ou incomplètes ne sont pas acceptables.

2. ***Vitesse initiale d'altération explicite*** :
- Fournie directement sous forme numérique avec une unité claire (ex. : g·m⁻²·d⁻¹, µm·d⁻¹, mg·m⁻²·h⁻¹).
- Peut être globale (pour le verre entier) ou spécifique à un élément (ex. : V_xSi_yB_z), mais doit être mesurée dans les conditions initiales de dissolution (avant saturation ou formation de couches altérées).
- Doit être normalisée (par ex., à la surface géométrique ou BET) avec la méthode de normalisation indiquée.
- Les données indirectes (ex. : courbes de concentration, épaisseurs de couche altérée) ou calculables (ex. : via des équations ou pentes) ne sont pas acceptables.

3. ***Au moins un paramètre expérimental*** parmi :
- pH initial à température ambiante ou d'essai,
- Température de l'essai,
- Surface spécifique (BET ou géométrique, si poudre),
- Volume de la solution,
- Débit de la solution (pour essais dynamiques).
- Ce paramètre doit être clairement associé à l'expérience où la vitesse d'altération est mesurée.

Si l'un de ces trois critères (composition complète, vitesse explicite avec unité, et au moins un paramètre expérimental) est absent pour un verre ou une condition donnée, cet ensemble de données n'est pas pertinent. Évaluez chaque verre ou condition expérimentale séparément si l'article en présente plusieurs.

Informations demandées :

1. ***Pertinence de l'article*** : [Oui/Non]
2. ***Justification*** : [Explication concise indiquant si les critères sont remplis ou non, avec mention des éléments présents et absents pour chaque verre ou condition.]



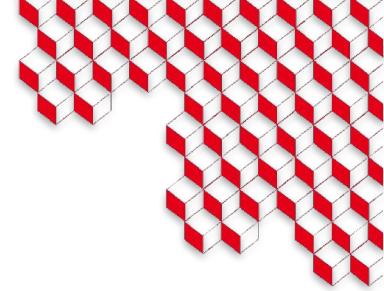
Télécharger l'article

Prompt : Définir les critères de validation

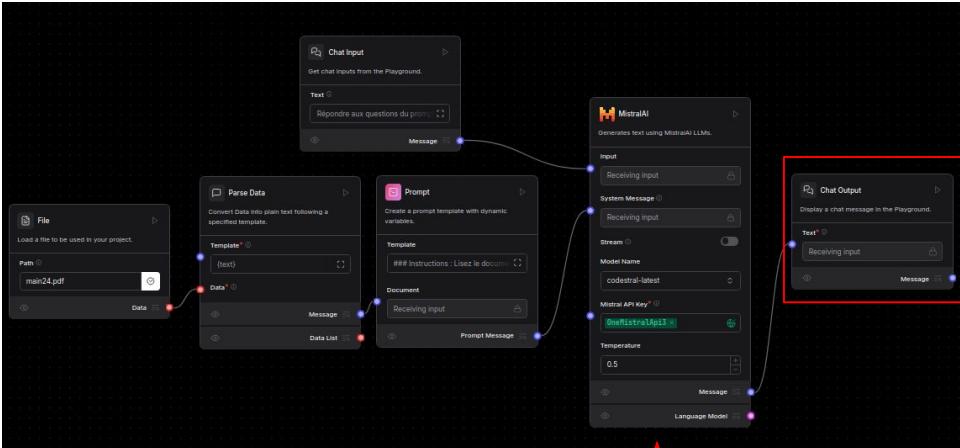
LLM (Mistral)

Partie 1 : Création d'une Base de Données sur l'Altération des Verres

Partie 2 : Entraînement d'un réseau de neurones sur la base de données



Workflow utilisé pour la validation des documents



M codestar-latest
Pertinence de l'article :
Pertinence : Oui

Justification :

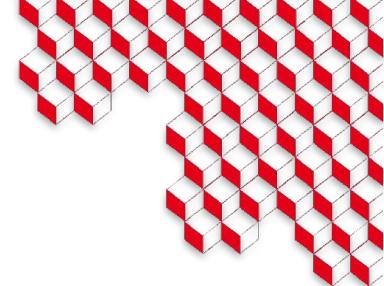
L'article est pertinent pour contribuer à une base de données visant à prédire la vitesse initiale d'altération des verres à l'aide de l'apprentissage automatique. Voici les raisons détaillées :

- Composition chimique du verre :
 - L'article fournit la composition chimique des deux verres étudiés : SON68 et ISO. Les compositions sont données en pourcentage pour chaque oxyde ou élément (par exemple, SiO₂, B₂O₃, Na₂O, etc.). La somme des pourcentages pour chaque verre est proche de 100 %, ce qui est conforme aux critères demandés.
- Vitesse initiale d'altération explicite :
 - L'article fournit plusieurs vitesses initiales d'altération pour les deux verres. Par exemple, pour le verre ISO, la vitesse initiale est de (6.10 m·l⁻²·h⁻¹) et pour le verre SON68, elle est de (4.4 m·l⁻²·h⁻¹). Ces valeurs sont explicites, numériques et normalisées à la surface géométrique (SA).
- Paramètres expérimentaux :
 - L'article fournit plusieurs paramètres expérimentaux, tels que la température (70 °C), le pH initial (12.5), la surface spécifique (BET) des verres, le volume de la solution, et la surface area to solution volume ratio (SA/V). Ces paramètres sont clairement associés aux expériences où les vitesses initiales d'altération sont mesurées.

En résumé, pour chaque verre ou condition expérimentale évaluée, l'article contient les informations nécessaires pour contribuer à la base de données. Les compositions chimiques sont complètes, les vitesses initiales d'altération sont explicites et normalisées, et les paramètres expérimentaux sont fournis.

Partie 1 : Création d'une Base de Données sur l'Altération des Verres

Partie 2 : Entraînement d'un réseau de neurones sur la base de données

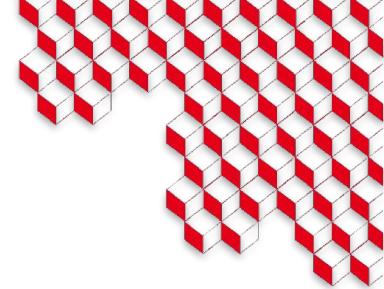


Résultats :



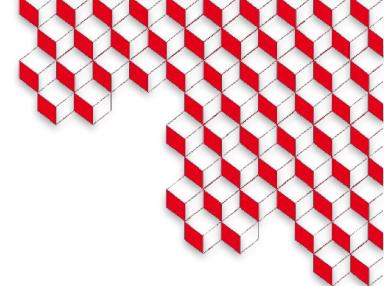
Partie 1 : Création d'une Base de Données sur l'Altération des Verres

Partie 2 : Entraînement d'un réseau de neurones sur la base de données



Partie 1 : Création d'une Base de Données sur l'Altération des Verres

Partie 2 : Entraînement d'un réseau de neurones sur la base de données



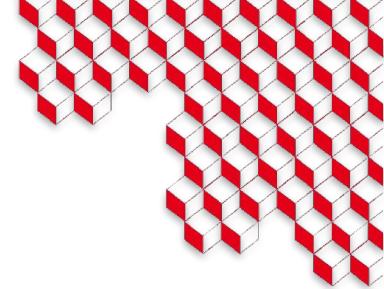
Contexte : Certaines articles données directement par les experts du domaine sont des documents scannés (image de texte) d'où la nécessité de la conversion (**OCR**) pour pouvoir en faire des textes numériques exploitables.



Océrisation : L'océrisation (ou **OCR**, Reconnaissance Optique de Caractères) est le processus de conversion d'images de texte (documents scannés, PDF, photos) en texte numérique exploitable, à l'aide d'outils comme Docling, pour extraire et traiter les informations contenues.

Partie 1 : Création d'une Base de Données sur l'Altération des Verres

Partie 2 : Entraînement d'un réseau de neurones sur la base de données

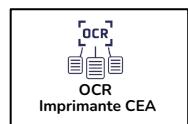


Outils testés et sélectionnés :

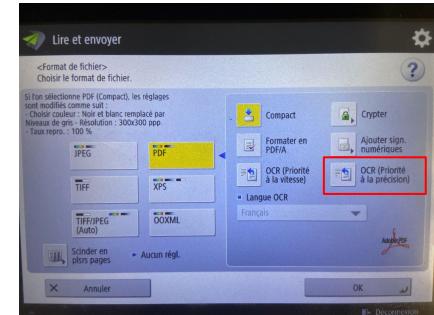
Après le test de plusieurs outils Open source d'Océrisation et la vérification manuelle de la qualité des documents convertis, deux outils permettent d'avoir des résultats très pertinentes :



Avantages		Inconvénients	
- Accepte plusieurs formats de documents.	- Rapide et précis pour les documents simples.	- Pas 100 % précis pour les documents complexes.	- Garde pas la forme originelle du document.

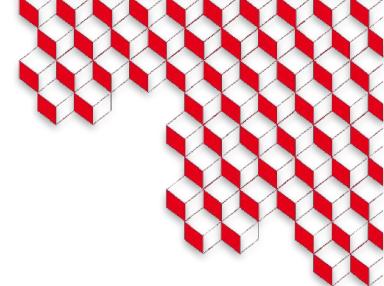


Avantages		Inconvénients	
- 100 % précis pour les documents complexes.	- Garde la forme originelle du document.	- Prend quelques minutes pour l'exécution.	- Nécessité d'imprimer le document.



Partie 1 : Création d'une Base de Données sur l'Altération des Verres

Partie 2 : Entraînement d'un réseau de neurones sur la base de données



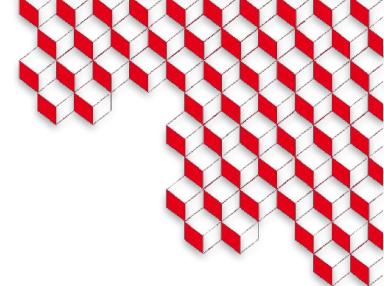
Processus de traitement :



Remarque: Les documents scannés sont donnés directement par des experts du domaine sans passer par l'outil de validation.

Partie 1 : Création d'une Base de Données sur l'Altération des Verres

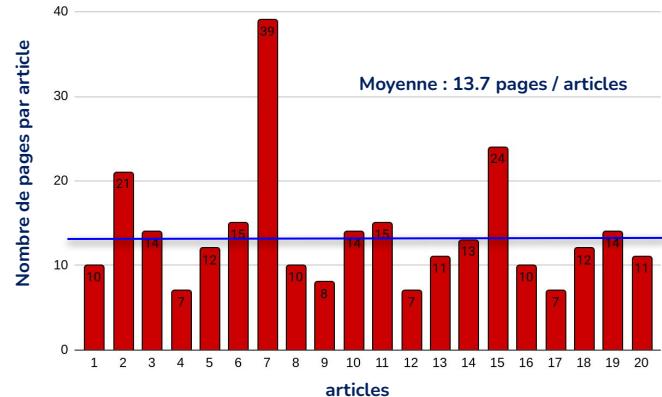
Partie 2 : Entraînement d'un réseau de neurones sur la base de données



Résultats :

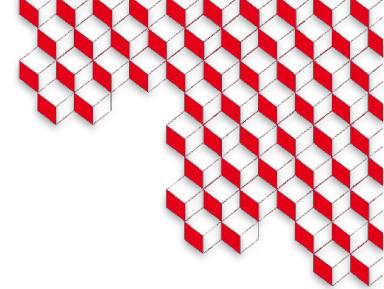


20 documents numériques



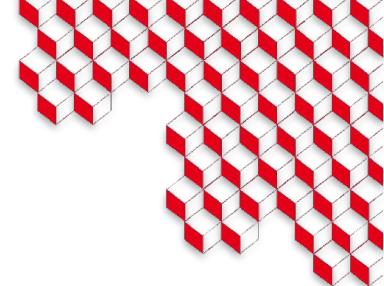
Partie 1 : Création d'une Base de Données sur l'Altération des Verres

Partie 2 : Entraînement d'un réseau de neurones sur la base de données

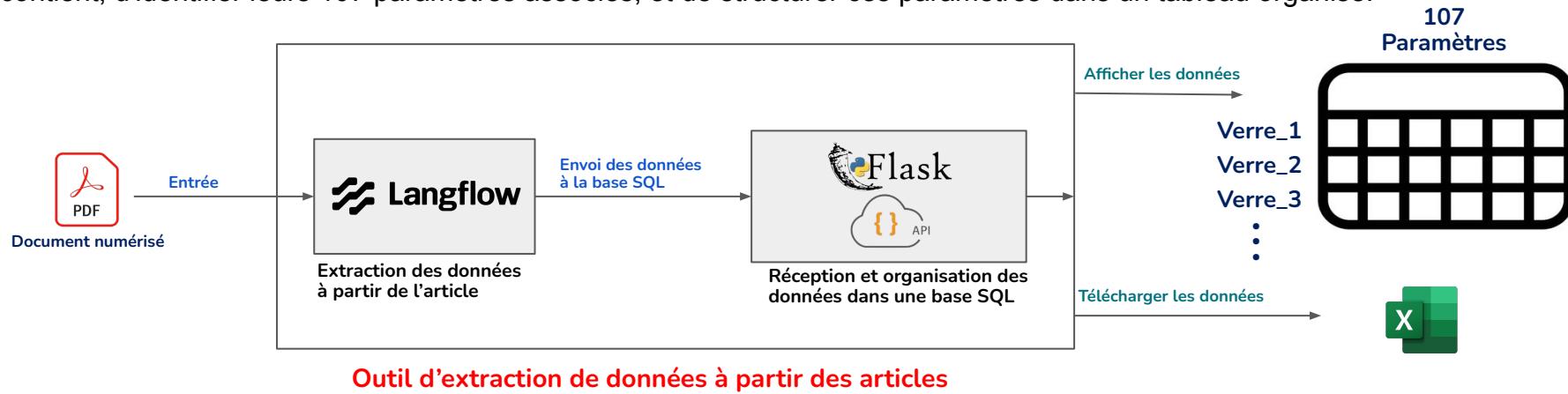


Partie 1 : Création d'une Base de Données sur l'Altération des Verres

Partie 2 : Entraînement d'un réseau de neurones sur la base de données

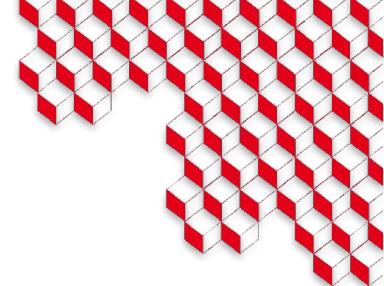


Objectif : Développer un outil capable d'analyser un document numérique en entrée, d'extraire tous les verres qu'il contient, d'identifier leurs 107 paramètres associés, et de structurer ces paramètres dans un tableau organisé.



Partie 1 : Création d'une Base de Données sur l'Altération des Verres

Partie 2 : Entraînement d'un réseau de neurones sur la base de données



Objectif : Développer un outil capable d'analyser un document numérique en entrée, d'extraire tous les verres qu'il contient, d'identifier leurs 107 paramètres associés, et de structurer ces paramètres dans un tableau organisé.



Partie 1 : Création d'une Base de Données sur l'Altération des Verres

Partie 2 : Entraînement d'un réseau de neurones sur la base de données



Problématiques identifiées :

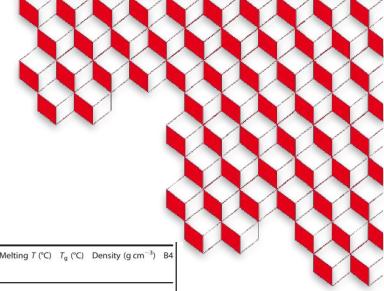
- Grand nombre de verres par article** : Environ une dizaine de verres par article, chacun nécessitant l'extraction de 107 paramètres.
- Multiplicité des tests par verre** : Chaque verre est soumis à plusieurs tests, chacun avec ses 107 paramètres à extraire.
- Variation des unités de composition** : Les compositions des verres sont exprimées en différentes unités (%mol, wt%) ou formes (élément, oxyde), nécessitant une unification des unités dans la base de données finale.
- Variation de la nomenclature des paramètres** : Les paramètres varient dans les articles (ex. V0/r0, densité/density), nécessitant un outil capable de reconnaître et de gérer ces différences d'unités et de nomenclature.

Name	Other name in the literature	Composition (mol%)						Melting T (°C)	Tg (°C)	Density (g cm ⁻³)	B4
		SiO ₂	B ₂ O ₃	Na ₂ O	Al ₂ O ₃	CaO	ZrO ₂	Li ₂ O			
Borosilicate glasses											
NBS14/18 ^a	CJ1 ^b	67.8	18.0	14.2		1.9	1450	570	2.451	66	
NBSA ^c	CJ2 ^b	64.9	17.3	13.7	4.1		1450	510	2.405	48	
NBSAC ^c	CJ3 ^b	61.2	16.3	12.8	3.9	5.8	1400	510	2.471	53	
NBSACZ ^c	CJ4 ^b , ISO ^d	60.1	16.0	12.7	3.8	5.7	1.7	1400	510	2.500	48
NBSA2 ^c	CJ7 ^b	63.8	17.0	13.4	4.0		1.8	1400	510	2.505	41
NBSC ^c	CJ8 ^b	63.6	17.0	13.4		6.0		1400	510	2.505	66
NBS2 ^c	CJ9 ^b	62.5	16.7	13.1		5.9	1.8	1350	550	2.547	63
NBS12/28	NBS17-24 ^e	60.5	27.7	11.6				1250	540	2.462	43
NBS36/21	NBS35-19 ^e	43.2	20.7	36.2				1100	470	2.537	63
NBS31/15	NBS29-13 ^e	54.2	15.0	30.9				1250	505	2.538	78
Aluminosilicate glasses											
NSAC19 ^g		55.3	19.0	9.9	13.9	1.9	1450	610	2.591		
NSAC17		48.6	17.0	20.1	12.3	1.9	1500	580	2.569		
NSAC21 ^g		57.4	21.3	5.9	15.4		1350	580	2.592		
NSAC0		57.3		5.9	36.8		1500	580	2.790		
NSAC35 ^g		58.1	34.9	6.9			1350	580	2.522		
NSA	Albite glass ^f	75.0	12.5	12.5			1650	530	2.34		

Note that the glasses NBS12/28, NBS36/21, and NBS31/15 have four-fold coordinated B in the glass, determined by XRD^g and Raman^h.
^aGin et al.ⁱ
^bPaucaut et al.^j
^cPerez et al.^k

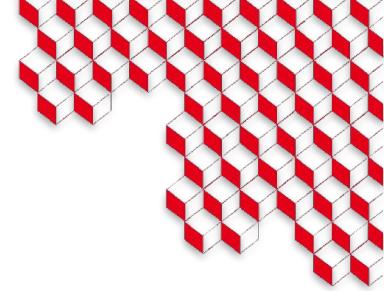
Table 2. Results of the initial dissolution rate measurements ($\text{d}i/\text{dt}$) obtained by linear regression at the end of the test. Uncertainty on the pH is 0.1 unit. Congruency between Na and Si is calculated at the end of the test, n.d. not determined.

Glass	Test #	FAG	Congruency (Na/Si)	Final pH ^h	Parameter considered for r_0 calculation	r_0 (g m ⁻² d ⁻¹)	$\ln(r_0)$	y-intercept (g m ⁻²)	R ²
NBS12/28	1	<0.01	1.1±0.1	8.9	Si	192±60	5.3	0.4	1.000
	1				Mass loss	223±45	5.4		
	2	0.26	8.9	9.0	Si	202±60	5.3	0.6	0.995
NBS36/21	1	0.24	1.2±0.1	9.5	Si	47,370±14,360	10.8	-31.2	0.986
					Mass loss	49,048±9,809	10.8		
NBS31/15	1	0.02	1.1±0.1	9.1	Si	2070±620	7.6	3.8	0.999
					Mass loss	2756±551	7.9		
NBS14/18	1	0.06	n.d.	9.0	Si	12.0±3.6	2.5	0.2	0.999
	2	0.08	8.9	Si	17.7±5.3	2.9	0.1	0.989	
	3	0.07	9.0	Si	12.2±3.7	2.5	0.1	0.999	
NBSA	1	0.03	1.1±0.1	8.9	Si	2.6±0.8	0.9	-0.02	0.997
NBSAC	1	0.04	1.0±0.1	8.9	Si	9.9±3.0	2.3	-0.08	0.986
NBSACZ	1	0.09	1.1±0.1	9.0	Si	8.2±2.5	2.1	0.2	0.990
NBSA2	1	0.02	1.0±0.1	8.9	Si	2.2±0.6	0.8	0.04	0.997
NBSC	1	0.24	1.0±0.1	8.9	Si	52.2±15.7	4.0	4.0	1.000
NBS2	1	0.26	1.1±0.1	8.9	Si	56.4±17.0	4.0	0.8	0.999
NSAC19	1	0.08		8.8	Si	5.6±1.7	1.7	0.05	0.988
	2	0.02		9.0	Si	7.2±2.2	2.0	-0.1	0.999
	3	0.02	1.2±0.1	9.1	Si	5.5±1.7	1.7	-0.12	0.984
NSAC17	1	0.14		8.8	Si	9.8±2.9	2.3	0.2	0.997
	3	0.02	1.0±0.1	9.0	Si	5.0±1.5	1.6	-0.03	0.996

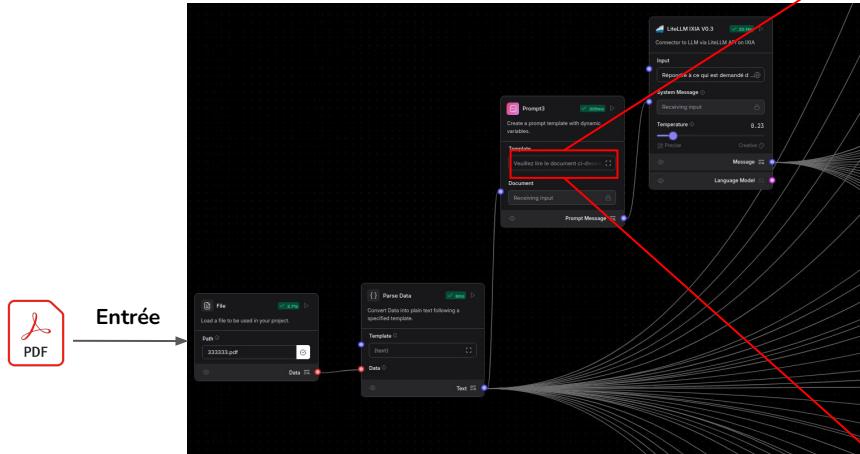


Partie 1 : Création d'une Base de Données sur l'Altération des Verres

Partie 2 : Entraînement d'un réseau de neurones sur la base de données



1. Extraire la liste des verres disponible dans le pdf :



Edit Prompt

Create your prompt. Prompts can help guide the behavior of a Language Model. Use curly brackets {} to introduce variables.

Veuillez lire le document ci-dessus et extraire uniquement les informations suivantes. Formatez les réponses clairement pour chaque élément :

Document :
(Document)

Informations demandées :
Extraire uniquement les noms officiels des verres dont la composition est explicitement donnée dans le document et pour lesquels au moins un paramètre expérimental est disponible. Ne pas extraire les secondaires ou alternatifs qui n'ont pas de caractéristiques numériques dans le document.

1. verre_1 : [Nom clair, sans texte additionnel]
2. verre_2 : [Nom clair, sans texte additionnel]
3. verre_3 : [Nom clair, sans texte additionnel]
4. verre_4 : [Nom clair, sans texte additionnel]

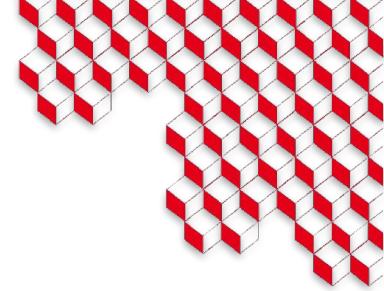
Format attendu :
1. verre_1 : (Nom)
2. verre_2 : (Nom)
3. verre_3 : (Nom)
4. verre_4 : (Nom)
...
n. verre_n : (Nom)

Instructions strictes :
- N'extraire que les verres dont la composition et au moins un paramètre expérimental sont explicitement mentionnés dans le document.
- Ne pas inclure de noms secondaires ou alternatifs sans caractéristiques numériques.
- Ne pas ajouter de commentaires supplémentaires et respecter strictement le format attendu de la réponse.
- Si un verre ne répond pas aux critères, ne pas l'inclure dans la liste.
- Si un verre a une composition théorique et une composition expérimentale, il faut prendre celui dont on a la composition expérimentale. L'insiste beaucoup sur la composition expérimentale.

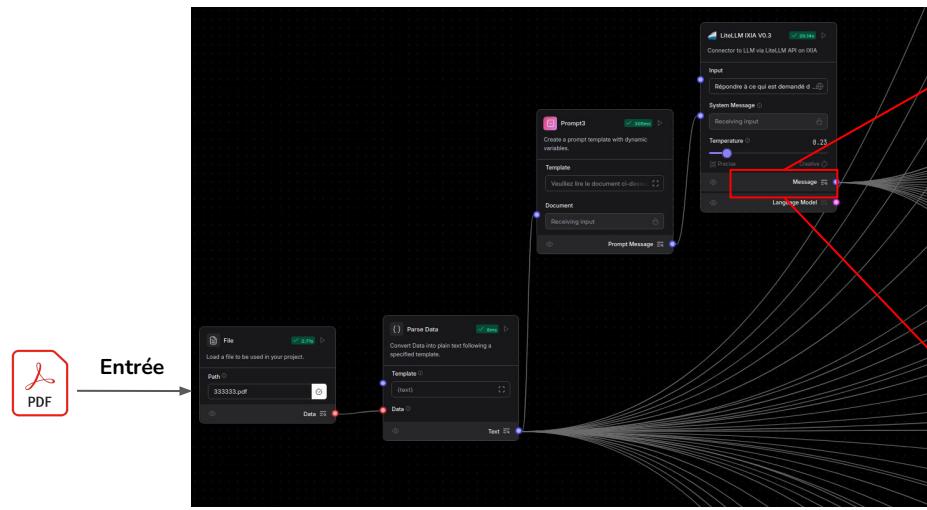
Prompt Variables: Document

Partie 1 : Création d'une Base de Données sur l'Altération des Verres

Partie 2 : Entraînement d'un réseau de neurones sur la base de données



1. Extraire la liste des verres disponible dans le pdf :



Résultats LLM:

```
Component Output
Inspect the output of the component below.

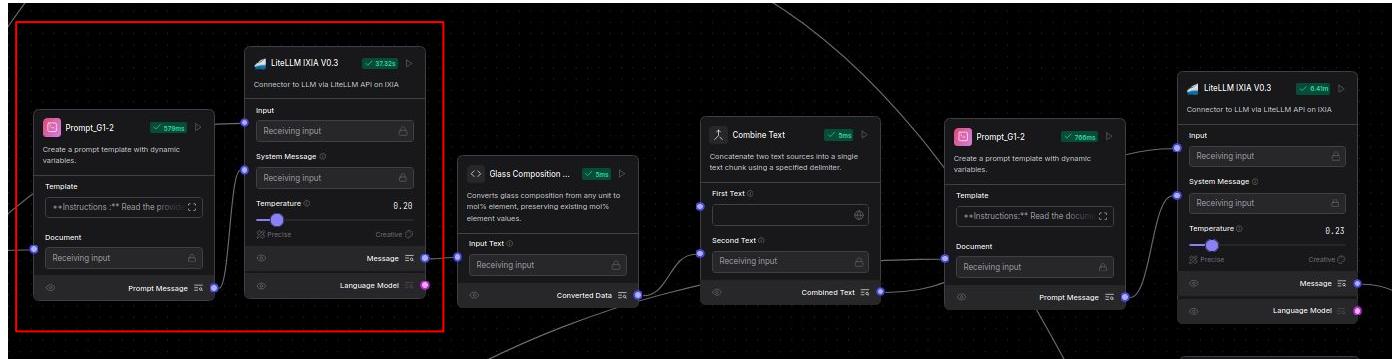
1. verre_1 : AFCI
2. verre_2 : EWC-C
3. verre_3 : IDF1-B2
4. verre_4 : IDF21-EC14
5. verre_5 : IDF7-E12
6. verre_6 : ISG
7. verre_7 : IDF16-A161
8. verre_8 : LAW-A44
9. verre_9 : LAW-ABP1
10. verre_10 : LAW-B45
11. verre_11 : LAW-C22
12. verre_12 : LD6-5412
13. verre_13 : MT25
14. verre_14 : MT30
15. verre_15 : ORLEC28
16. verre_16 : ORLEC33
17. verre_17 : ORPLG9
18. verre_18 : SON68
19. verre_19 : SRL-202
```

Partie 1 : Création d'une Base de Données sur l'Altération des Verres

Partie 2 : Entraînement d'un réseau de neurones sur la base de données

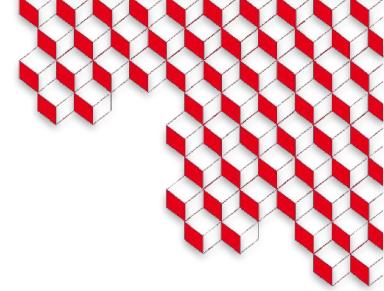


1. Extraire la composition des deux premiers verres dans la liste sous les différentes formes possibles :

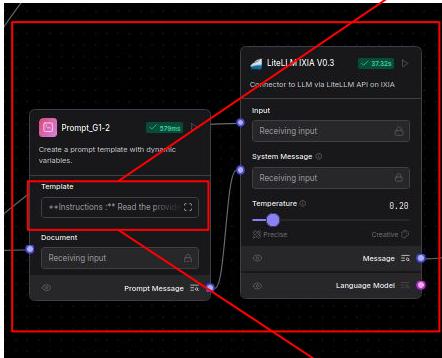


Partie 1 : Création d'une Base de Données sur l'Altération des Verres

Partie 2 : Entraînement d'un réseau de neurones sur la base de données



1. Extraire la composition des deux premiers



Edit Prompt

Create your prompt. Prompts can help guide the behavior of a Language Model. Use curly brackets {} to introduce variables.

Instructions :

Read the provided article and extract the composition of glasses 1 and 2 as listed in the user's input (verre_1 and verre_2). The composition may be expressed in oxides or elements, with units such as mol% or wt%. Present the composition strictly in the format below, adhering to the rules without deviation. If the article contains fewer than 2 glasses, assign [0] to all composition values for the missing glasses (e.g., glasses 1 and 2 must have all values set to [0]).

Document :

{Document}

Expected Format :

Glass 1:

1. [oxide][element] : [value][wt%][oxide]; [value][wt%][element]; [value][wt%][cation]; [value][mol%][oxide]; [value][mol%][element]; [value][mol%][cation]
2. [oxide][element] : [value][wt%][oxide]; [value][wt%][element]; [value][wt%][cation]; [value][mol%][oxide]; [value][mol%][element]; [value][mol%][cation]

...

Glass 2:

1. [oxide][element] : [value][wt%][oxide]; [value][wt%][element]; [value][wt%][cation]; [value][mol%][oxide]; [value][mol%][element]; [value][mol%][cation]
2. [oxide][element] : [value][wt%][oxide]; [value][wt%][element]; [value][wt%][cation]; [value][mol%][oxide]; [value][mol%][element]; [value][mol%][cation]

...

Strict Rules :

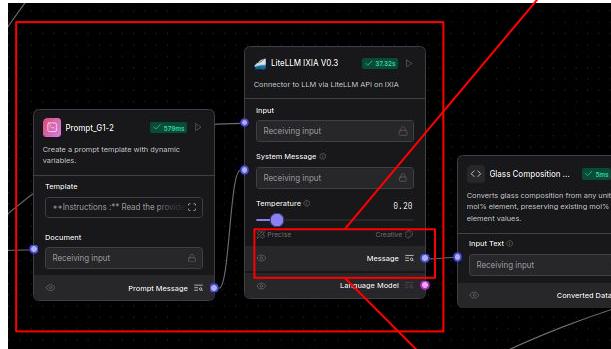
- Extract only the composition values explicitly provided in the article for glasses 1 and 2, based on their order in the document (verre_1 is the 1st glass, verre_2 is the 2nd glass).
- If the article contains fewer than 2 glasses (e.g., only 1), use [0] for all six values ([wt%][oxide], [wt%][element], [wt%][cation], [mol%][oxide], [mol%][element], [mol%][cation]) for each oxide or element (SiO₂/Si) with all six values set to [0].
- Include all oxides and elements explicitly mentioned in the article for glasses 1 and 2 (if present). For glass 2 not present, replicate the same list of oxides/elements from glass 1 (if any) and set all six corresponding values to [0].

Partie 1 : Création d'une Base de Données sur l'Altération des Verres

Partie 2 : Entraînement d'un réseau de neurones sur la base de données



1. Extraire la composition des deux premiers verres



Résultats LLM:

Component Output

Inspect the output of the component below.

Glass 1:

1. [SiO₂][Si] : [53.67][wt%][oxide]; [0][wt%][element]; [0][wt%][cation]; [58.40][mol%][oxide]; [0][mol%][element]; [0][mol%][cation]
2. [Na₂O][Na] : [7.00][wt%][oxide]; [0][wt%][element]; [0][wt%][cation]; [7.38][mol%][oxide]; [0][mol%][element]; [0][mol%][cation]
3. [B₂O₃][B] : [9.65][wt%][oxide]; [0][wt%][element]; [0][wt%][cation]; [9.06][mol%][oxide]; [0][mol%][element]; [0][mol%][cation]
4. [Al₂O₃][Al] : [19.38][wt%][oxide]; [0][wt%][element]; [0][wt%][cation]; [6.01][mol%][oxide]; [0][mol%][element]; [0][mol%][cation]
5. [CaO][Ca] : [5.00][wt%][oxide]; [0][wt%][element]; [0][wt%][cation]; [5.83][mol%][oxide]; [0][mol%][element]; [0][mol%][cation]
6. [ZrO₂][Zr] : [1.15][wt%][oxide]; [0][wt%][element]; [0][wt%][cation]; [0.61][mol%][oxide]; [0][mol%][element]; [0][mol%][cation]
7. [Li₂O][Li] : [4.50][wt%][oxide]; [0][wt%][element]; [0][wt%][cation]; [9.85][mol%][oxide]; [0][mol%][element]; [0][mol%][cation]
8. [Nd₂O₃][Nd] : [2.01][wt%][oxide]; [0][wt%][element]; [0][wt%][cation]; [0.39][mol%][oxide]; [0][mol%][element]; [0][mol%][cation]
9. [MoO₃][Mo] : [1.50][wt%][oxide]; [0][wt%][element]; [0][wt%][cation]; [0.68][mol%][oxide]; [0][mol%][element]; [0][mol%][cation]
10. [Ce₂O₃][Ce] : [1.19][wt%][oxide]; [0][wt%][element]; [0][wt%][cation]; [0.24][mol%][oxide]; [0][mol%][element]; [0][mol%][cation]
11. [Cs₂O][Cs] : [1.10][wt%][oxide]; [0][wt%][element]; [0][wt%][cation]; [0.26][mol%][oxide]; [0][mol%][element]; [0][mol%][cation]

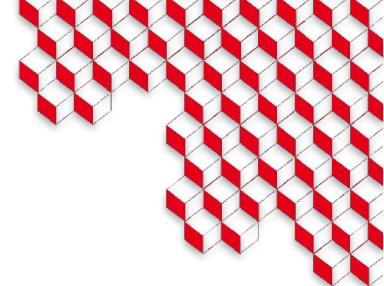
Glass 2:

1. [SiO₂][Si] : [56.20][wt%][oxide]; [0][wt%][element]; [0][wt%][cation]; [60.10][mol%][oxide]; [0][mol%][element]; [0][mol%][cation]
2. [Na₂O][Na] : [12.20][wt%][oxide]; [0][wt%][element]; [0][wt%][cation]; [12.65][mol%][oxide]; [0][mol%][element]; [0][mol%][cation]
3. [B₂O₃][B] : [17.30][wt%][oxide]; [0][wt%][element]; [0][wt%][cation]; [15.97][mol%][oxide]; [0][mol%][element]; [0][mol%][cation]
4. [Al₂O₃][Al] : [16.10][wt%][oxide]; [0][wt%][element]; [0][wt%][cation]; [3.84][mol%][oxide]; [0][mol%][element]; [0][mol%][cation]
5. [CaO][Ca] : [5.00][wt%][oxide]; [0][wt%][element]; [0][wt%][cation]; [5.73][mol%][oxide]; [0][mol%][element]; [0][mol%][cation]
6. [ZrO₂][Zr] : [3.30][wt%][oxide]; [0][wt%][element]; [0][wt%][cation]; [1.72][mol%][oxide]; [0][mol%][element]; [0][mol%][cation]
7. [Li₂O][Li] : [0][wt%][oxide]; [0][wt%][element]; [0][wt%][cation]; [0][mol%][oxide]; [0][mol%][element]; [0][mol%][cation]
8. [Nd₂O₃][Nd] : [0][wt%][oxide]; [0][wt%][element]; [0][wt%][cation]; [0][mol%][oxide]; [0][mol%][element]; [0][mol%][cation]
9. [MoO₃][Mo] : [0][wt%][oxide]; [0][wt%][element]; [0][wt%][cation]; [0][mol%][oxide]; [0][mol%][element]; [0][mol%][cation]
10. [Ce₂O₃][Ce] : [0][wt%][oxide]; [0][wt%][element]; [0][wt%][cation]; [0][mol%][oxide]; [0][mol%][element]; [0][mol%][cation]
11. [Cs₂O][Cs] : [0][wt%][oxide]; [0][wt%][element]; [0][wt%][cation]; [0][mol%][oxide]; [0][mol%][element]; [0][mol%][cation]

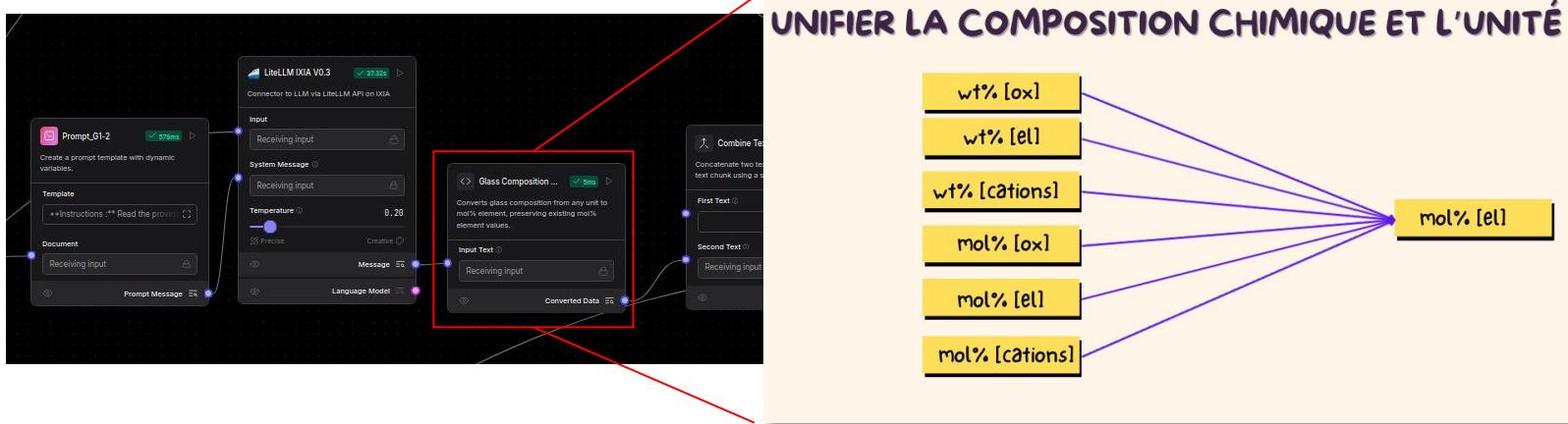
Outputs Logs

Partie 1 : Création d'une Base de Données sur l'Altération des Verres

Partie 2 : Entraînement d'un réseau de neurones sur la base de données

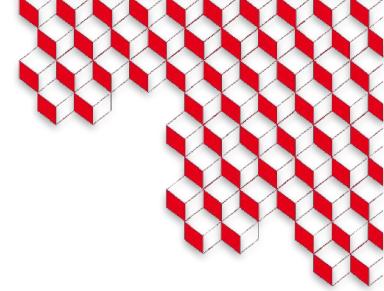


2. Conversion des unités de la composition des verres en %mol d'élément :

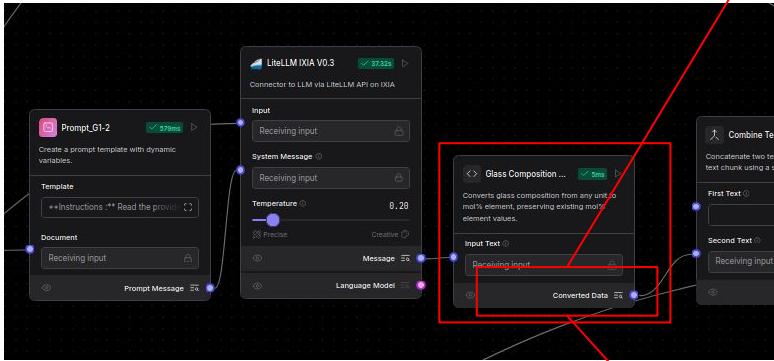


Partie 1 : Création d'une Base de Données sur l'Altération des Verres

Partie 2 : Entraînement d'un réseau de neurones sur la base de données



2. Conversion des unités de la composition des verres en %mol



Résultats LLM:

Component Output

Inspect the output of the component below.

glass 1:

```
mol% Si = 18.12
mol% O = 59.09
mol% Na = 4.58
mol% B = 5.62
mol% Al = 3.73
mol% Ca = 1.81
mol% Zr = 0.19
mol% Li = 6.11
mol% Nd = 0.24
mol% Mo = 0.21
mol% Ce = 0.15
mol% Cs = 0.16
```

glass 2:

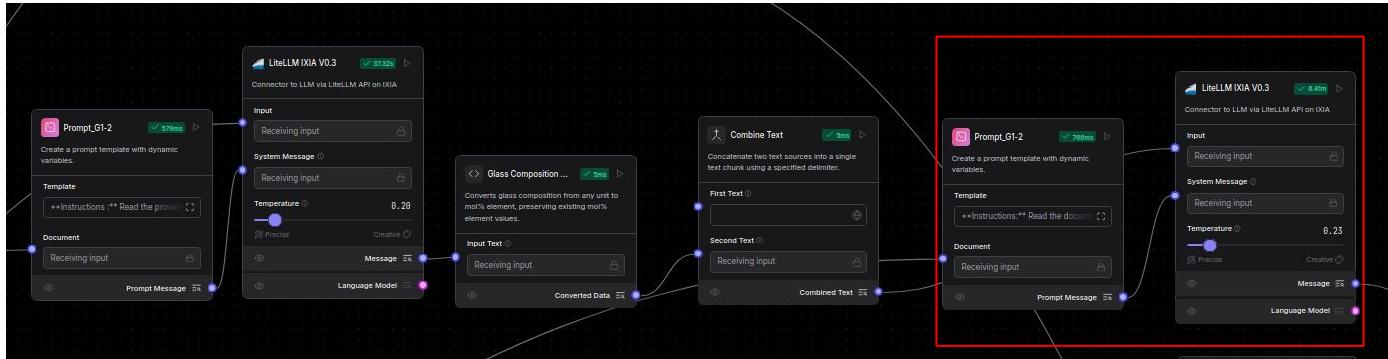
```
mol% Si = 18.0
mol% O = 60.33
mol% Na = 7.58
mol% B = 9.57
mol% Al = 2.3
mol% Ca = 1.72
mol% Zr = 0.52
```

Partie 1 : Création d'une Base de Données sur l'Altération des Verres

Partie 2 : Entraînement d'un réseau de neurones sur la base de données

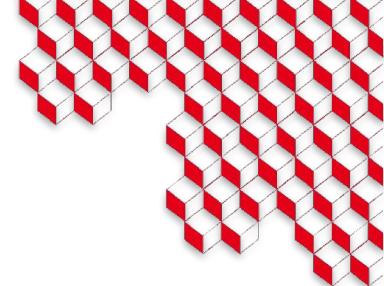


3. Extraire les autres Paramètres pour compléter les 107 paramètres par verre:



Partie 1 : Crédit d'une Base de Données sur l'Altération des Verres

Partie 2 : Entraînement d'un réseau de neurones sur la base de données



Étape 1

Étape 2

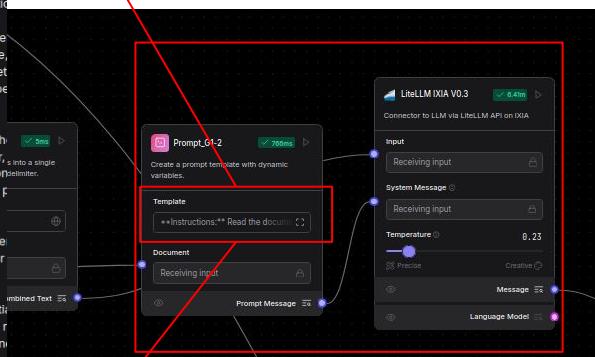
Étape 3

Edit Prompt

Create your prompt. Prompts can help guide the behavior of a Language Model. Use curly brackets {} to introduce variables.

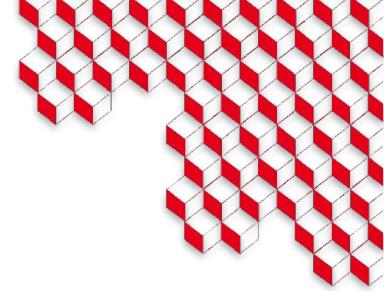
```
73. ***Np** : [List of numeric values, one per test; use 0 if not found in the input]
74. ***An** : [List of numeric values, one per test; use 0 if not found in the input]
75. ***Cr** : [List of numeric values, one per test; use 0 if not found in the input]
76. ***Stautres An*** : [List of numeric values, one per test; sum of other actinides not listed above, use 0 if not found]
77. ***Somme** : [List of numeric values, one per test; sum of the composition values above for each test; Numeric value, calculate the sum of the composition values above]
78. ***Densité** : [List of numeric values, one per test; if available, otherwise "Not available"; Numeric value]
79. ***Homogénéité** : [List of comments, one per test; if available, otherwise "Not available"; Comment]
80. ***B(IV)** : [List of numeric values, one per test; if available, otherwise "Not available"; Numeric value, fraction or percentage of tetracoordinated boron in the glass]
81. ***Irradié** : [List of "O/N" values, one per test; if available, otherwise "Not available"; Y/N, Was the glass studied for initial rate pre-irradiated?]
82. ***Caractéristiques si irradié** : [List of comments, one per test; if available, otherwise "Not available"; Comment, If the answer to the left column is yes, what are the irradiation characteristics?]
83. ***Température** : [List of numeric values, one per test; if available, otherwise "Not available"; Numeric value]
84. ***Statique/dynamique** : [List of comments, one per test; if available, otherwise "Not available"; Comment, Static = the alteration solution is not renewed. Dynamic = the alteration solution is renewed]
85. ***Plage granulométrique (si poudre)** : [List of numeric values, one per test; if available, otherwise "Not available"; Numeric value, terms that may be used: mesh, mesh size, granulometry]
86. ***Surface spécifique géométrique (si poudre)** : [List of numeric values, one per test; if available, otherwise "Not available"; Numeric value, terms that may be used: geometrical surface area]
87. ***Surface spécifique BET (si poudre)** : [List of numeric values, one per test; if available, otherwise "Not available"; Numeric value, terms that may be used: BET must be specified in square meters per gram]
88. ***Qualité de polissage (si monolithe)** : [List of comments, one per test; if available, otherwise "Not available"; Comment, terms that may be used: polishing grade]
89. ***Masse du verre*** : [List of numeric values, one per test; if available, otherwise "Not available"; Numeric value, here we seek the mass of glass placed in the reactor]
90. ***Surface du verre (S)** : [List of numeric values, one per test; if available, otherwise "Not available"; Numeric value, here we seek the surface area of the glass placed in the reactor]
91. ***Volume de la solution (V)** : [List of numeric values, one per test; if available, otherwise "Not available"; Numeric value, here we seek the volume of solution in the reactor]
92. ***Débit de la solution*** : [List of numeric values, one per test; if available, otherwise "Not available"; Numeric value, if the test is performed dynamically, we seek the solution flow rate]
93. ***pH initial (T amb)** : [List of numeric values, one per test; if available, otherwise "Not available"; Numeric value, possible terms: onset pH, starting pH, pH of the leachate, pH at ambient temperature (next field)]
94. ***pH initial (T essai)** : [List of numeric values, one per test; if available, otherwise "Not available"; Numeric value]
95. ***Composition de la solution*** : [List of comments, one per test; if available, otherwise "Not available"; Comment, here we seek to know the chemical composition of the alteration solution]
96. ***Durée de l'expérience*** : [List of numeric values, one per test; if available, otherwise "Not available"; Numeric value, convert to hours if the unit is not already in hours. For example, 1 day = 24 hours]
97. ***pH final (T amb)** : [List of numeric values, one per test; if available, otherwise "Not available"; Numeric value, same as for initial pH]
98. ***pH final (T essai)** : [List of numeric values, one per test; if available, otherwise "Not available"; Numeric value]
99. ***Normalisation de la vitesse (S0 ou SBET)** : [List of numeric values, one per test; if available, otherwise "Not available"; Numeric value, here we seek to know if the initial dissolution rate is expressed in S0 or SBET units]
100. ***Va(S) ou ra(S)** : [List of numeric values, one per test; if available, otherwise "Not available"; Numeric value, here we seek the initial rate (initial dissolution rate, forward reaction rate), provided in a different unit or in a transformed form (such as log(rf)), convert it to the specified units. For logarithmic values, assume log is base 10 unless specified otherwise, and for linear values, assume linear regression coefficient]
101. ***r2(S)** : [List of numeric values, one per test; if available, otherwise "Not available"; Numeric value, possible terms: linear regression coefficient, regression coefficient]
102. ***Ordonnée à l'origine (S)0*** : [List of numeric values, one per test; if available, otherwise "Not available"; Numeric value, terms that may be used: y-intercept]
103. ***Va(B) ou ra(B)** : [List of numeric values, one per test; if available, otherwise "Not available"; Numeric value, same as for S0]
104. ***Ordonnée à l'origine (B)** : [List of numeric values, one per test; if available, otherwise "Not available"; Numeric value, terms that may be used: y-intercept]
105. ***Va(Na) ou ra(Na)** : [List of numeric values, one per test; if available, otherwise "Not available"; Numeric value]
106. ***r2(Na)** : [List of numeric values, one per test; if available, otherwise "Not available"; Numeric value]
107. ***Ordonnée à l'origine (Na)** : [List of numeric values, one per test; if available, otherwise "Not available"; Numeric value, terms that may be used: y-intercept]
108. ***Va(LM) ou ra(LM)** : [List of numeric values, one per test; if available, otherwise "Not available"; Numeric value]
109. ***Congruence*** : [List of comments or numeric values, one per test; if available, otherwise "Not available"; Comment, sometimes congruence may be expressed as a numerical value, such as 100%]
```

tres par verre:



Partie 1 : Création d'une Base de Données sur l'Altération des Verres

Partie 2 : Entraînement d'un réseau de neurones sur la base de données



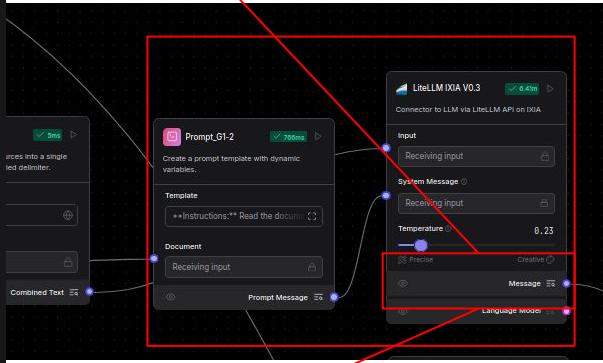
Component Output

Inspect the output of the component below.

Outputs Logs

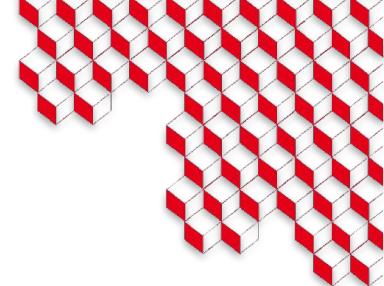
```
6. Verre_type1 : AFCI
7. Nombre de tests(Verre_type1) : 16
8. Li(Verre_type1) : [9.85, 9.85, 9.85, 9.85, 9.85, 9.85, 9.85, 9.85, 9.85, 9.85, 9.85, 9.85, 9.85, 9.85]
9. B(Verre_type1) : [9.06, 9.05, 9.06, 9.06, 9.06, 9.06, 9.06, 9.06, 9.06, 9.06, 9.06, 9.06, 9.06, 9.06, 9.06]
10. O(Verre_type1) : [59.09, 59.09, 59.09, 59.09, 59.09, 59.09, 59.09, 59.09, 59.09, 59.09, 59.09, 59.09, 59.09, 59.09, 59.09, 59.09]
11. Na(Verre_type1) : [7.38, 7.38, 7.38, 7.38, 7.38, 7.38, 7.38, 7.38, 7.38, 7.38, 7.38, 7.38, 7.38, 7.38, 7.38, 7.38]
12. Mg(Verre_type1) : [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
13. Al(Verre_type1) : [6.01, 6.01, 6.01, 6.01, 6.01, 6.01, 6.01, 6.01, 6.01, 6.01, 6.01, 6.01, 6.01, 6.01, 6.01]
14. Si(Verre_type1) : [58.4, 58.4, 58.4, 58.4, 58.4, 58.4, 58.4, 58.4, 58.4, 58.4, 58.4, 58.4, 58.4, 58.4, 58.4, 58.4]
15. P(Verre_type1) : [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
16. K(Verre_type1) : [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
17. Ca(Verre_type1) : [5.83, 5.83, 5.83, 5.83, 5.83, 5.83, 5.83, 5.83, 5.83, 5.83, 5.83, 5.83, 5.83, 5.83, 5.83, 5.83]
18. Ti(Verre_type1) : [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
19. V(Verre_type1) : [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
20. Cr(Verre_type1) : [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
21. Mn(Verre_type1) : [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
22. Fe(Verre_type1) : [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
23. Co(Verre_type1) : [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
24. Ni(Verre_type1) : [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
25. Cu(Verre_type1) : [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
26. Zn(Verre_type1) : [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
27. Ga(Verre_type1) : [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
28. Ge(Verre_type1) : [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
29. As(Verre_type1) : [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
30. Se(Verre_type1) : [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
31. Rb(Verre_type1) : [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
32. Sr(Verre_type1) : [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
33. Y(Verre_type1) : [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
34. Zr(Verre_type1) : [0.61, 0.61, 0.61, 0.61, 0.61, 0.61, 0.61, 0.61, 0.61, 0.61, 0.61, 0.61, 0.61, 0.61, 0.61, 0.61]
35. Nb(Verre_type1) : [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
36. Mo(Verre_type1) : [0.68, 0.68, 0.68, 0.68, 0.68, 0.68, 0.68, 0.68, 0.68, 0.68, 0.68, 0.68, 0.68, 0.68, 0.68, 0.68]
37. Tc(Verre_type1) : [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
```

ètres par verre:

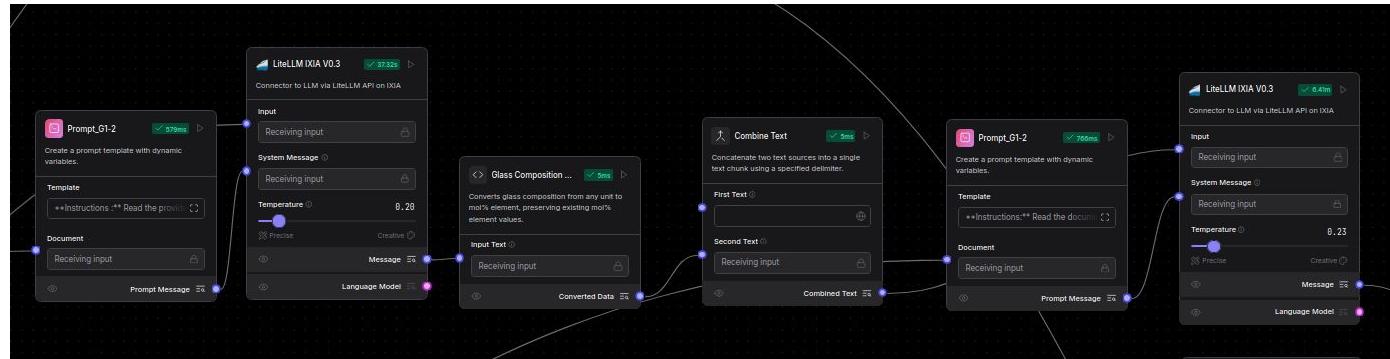


Partie 1 : Création d'une Base de Données sur l'Altération des Verres

Partie 2 : Entraînement d'un réseau de neurones sur la base de données



4. Dupliquer ce processus pour les autres verres :



X 15

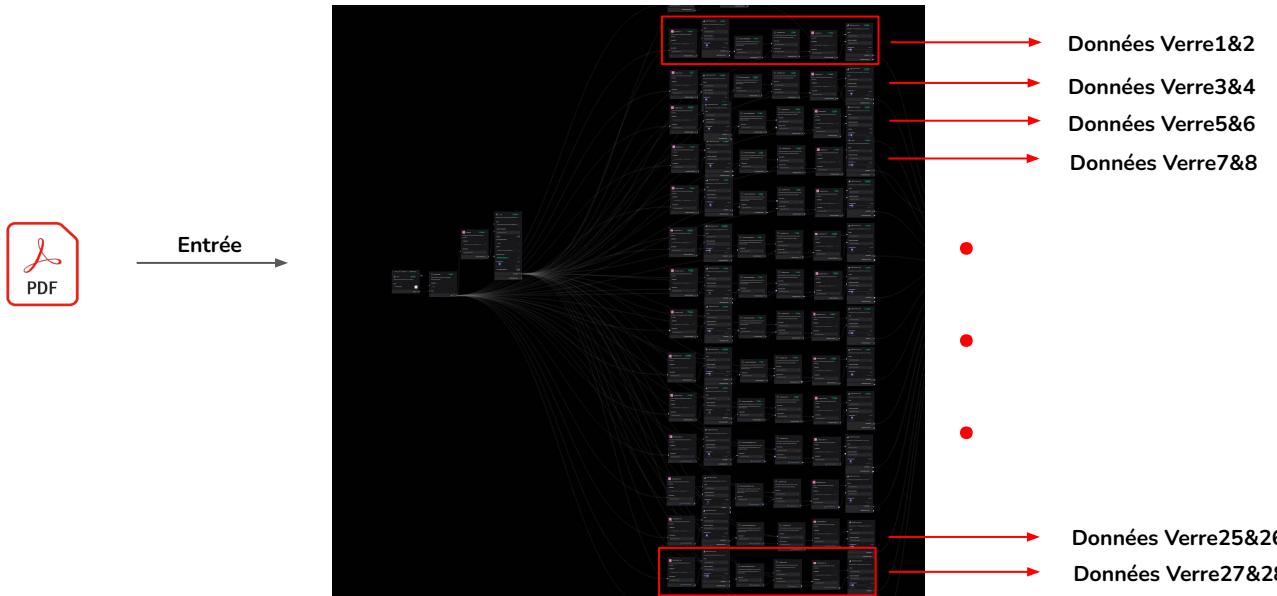
On s'est limité à 30 verres par article, car nous avons constaté que la plupart des articles ne dépassent pas ce nombre.

Partie 1 : Création d'une Base de Données sur l'Altération des Verres

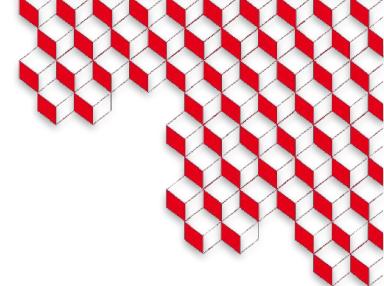
Partie 2 : Entraînement d'un réseau de neurones sur la base de données



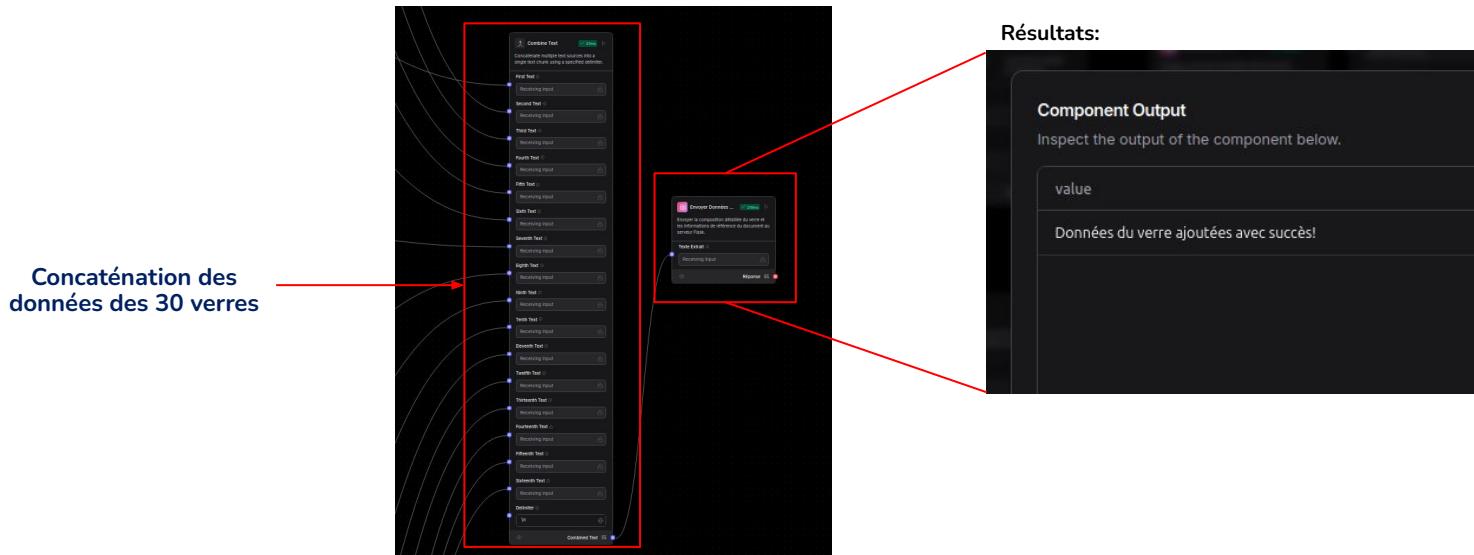
4. Dupliquer ce processus pour les autres verres :



Partie 1 : Création d'une Base de Données sur l'Altération des Verres

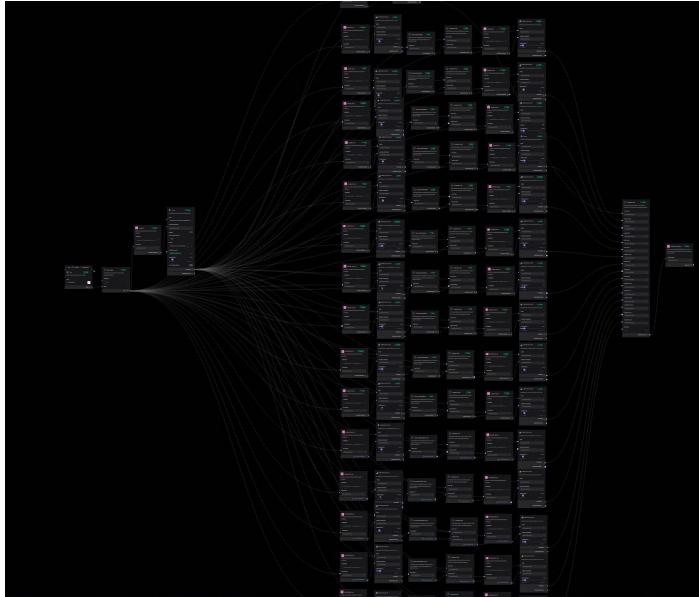
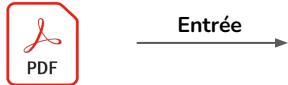


5. Concaténer toutes les résultats obtenus puis envoyer les données au tableau :



Partie 1 : Création d'une Base de Données sur l'Altération des Verres

Partie 2 : Entraînement d'un réseau de neurones sur la base de données



Architecture du workflow Langflow pour l'extraction et l'envoi des données à l'application Flask

Partie 1 : Création d'une Base de Données sur l'Altération des Verres

Partie 2 : Entraînement d'un réseau de neurones sur la base de données



Temps d'exécution : Le temps d'extraction de données à partir de chaque article dépend des éléments suivants :

- Nombre de verres présents dans l'article.
- La complexité de l'article en terme de la forme.

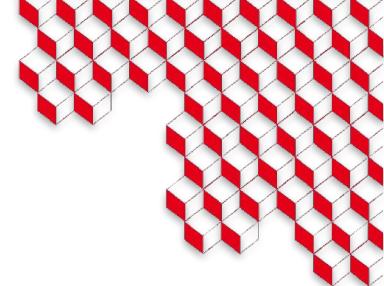
En moyenne : 2:55 minutes/article

Temps d'exécution pour chaque article (av. 2:55 minutes)



Partie 1 : Création d'une Base de Données sur l'Altération des Verres

Partie 2 : Entraînement d'un réseau de neurones sur la base de données



Ensuite on passe à la création d'une interface tableau qui va contenir les données (Frontend) :



Partie 1 : Création d'une Base de Données sur l'Altération des Verres

Partie 2 : Entraînement d'un réseau de neurones sur la base de données



Une interface capable de :

- Envoyer un document PDF à l'outil d'extraction de données.
- Recevoir les données extraites et les afficher.
- Exporter les données sous format Excel.

Premier Auteur	Type de Verre	Li	B	O	Na
Stéphane Gin	NBS14/18_test1	0	10.71	60.65	8.45
Stéphane Gin	NBS14/18_test2	0	10.71	60.65	8.45
Stéphane Gin	NBS14/18_test3	0	10.71	60.65	8.45
Stéphane Gin	NBSA	0	10.09	60.59	7.99
Stéphane Gin	NBSAC	0	9.74	60.25	7.65
Stéphane Gin	NBSACZ	0	9.58	60.32	7.61
Stéphane Gin	NBSAZ	0	9.94	60.7	7.84
Stéphane Gin	NBSC	0	10.37	60.24	8.17
Stéphane Gin	NBSCZ	0	10.2	60.37	8.0
Stéphane Gin	NBS12/28_test1	0	15.61	60.79	6.54
Stéphane Gin	NBS12/28_test2	0	15.61	60.79	6.54
Stéphane Gin	NBS36/21	0	12.12	54.05	21.19
Stéphane Gin	NBS31/15	0	9.08	55.8	18.71
Stéphane Gin	NSAC19_test1	1.24	0	57.24	12.42
Stéphane Gin	NSAC19_test2	1.24	0	57.24	12.42
Stéphane Gin	NSAC19_test3	1.24	0	57.24	12.42

Partie 1 : Création d'une Base de Données sur l'Altération des Verres

Partie 2 : Entraînement d'un réseau de neurones sur la base de données



20 Documents valides →

735 verres

Tableau d'altération des verres

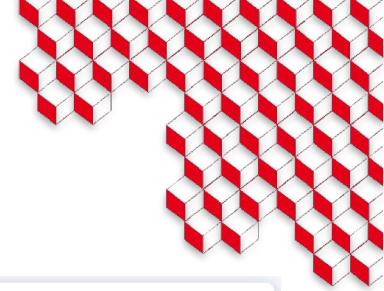
Télécharger le tableau en Excel

Télécharger un fichier PDF scanné
Browse... No file selected.

Télécharger

Premier Auteur	Type de Verre	Composition du verre en mol%			
		Li	B	O	Na
Stéphane Gin	NBS14/18_test1	0	10.71	60.65	8.45
Stéphane Gin	NBS14/18_test2	0	10.71	60.65	8.45
Stéphane Gin	NBS14/18_test3	0	10.71	60.65	8.45
Stéphane Gin	NBSA	0	10.09	60.59	7.99
Stéphane Gin	NBSAC	0	9.74	60.25	7.65
Stéphane Gin	NBSACZ	0	9.58	60.32	7.61
Stéphane Gin	NBSAZ	0	9.94	60.7	7.84
Stéphane Gin	NBSC	0	10.37	60.24	8.17
Stéphane Gin	NBSCZ	0	10.2	60.37	8.0
Stéphane Gin	NBS12/28_test1	0	15.61	60.79	6.54
Stéphane Gin	NBS12/28_test2	0	15.61	60.79	6.54
Stéphane Gin	NBS6/21	0	12.12	54.05	21.19
Stéphane Gin	NBS31/15	0	9.08	55.8	18.71
Stéphane Gin	NSAC19_test1	1.24	0	57.24	12.42
Stéphane Gin	NSAC19_test2	1.24	0	57.24	12.42
Stéphane Gin	NSAC19_test3	1.24	0	57.24	12.42

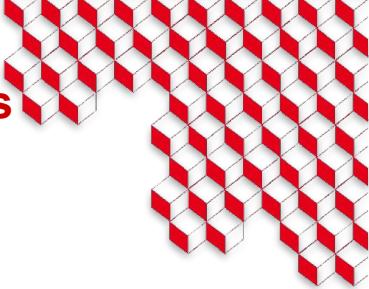
Showing 1 to 735 of 735 entries



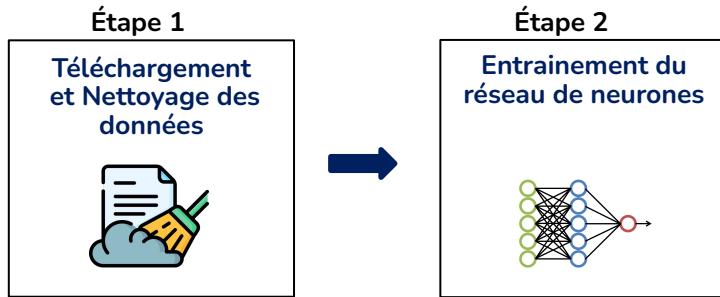
Partie 1 : Création d'une Base de Données sur l'Altération des Verres

Partie 2 : Entraînement d'un réseau de neurones sur la base de données

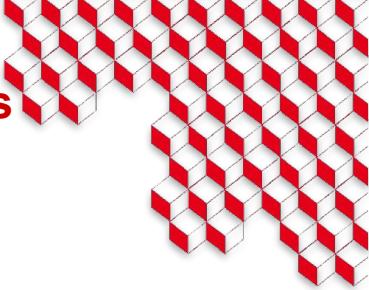
Partie 2 : Entraînement d'un réseau de neurones sur la base de données



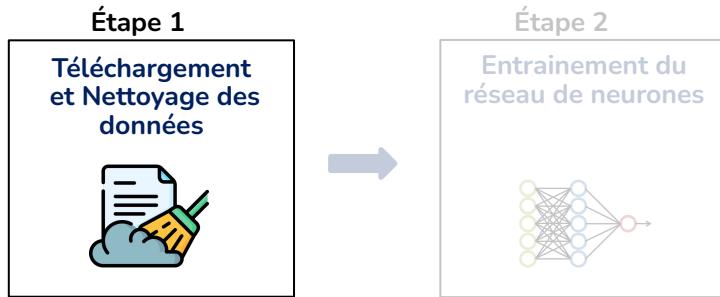
L'entraînement du réseau de neurones sur la base de données s'est déroulé en deux étapes:



Partie 2 : Entraînement d'un réseau de neurones sur la base de données

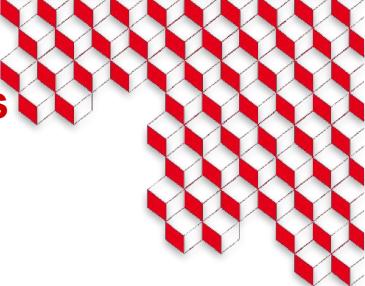


L'entraînement du réseau de neurones sur la base de données s'est déroulé en deux étapes:



Partie 1 : Création d'une Base de Données sur l'Altération des Verres

Partie 2 : Entraînement d'un réseau de neurones sur la base de données



Étape 1



Étape 2



Tableau d'altération des verres

Télécharger un fichier PDF scanné

Browse... No file selected.

Télécharger

Composition du verre en mol%

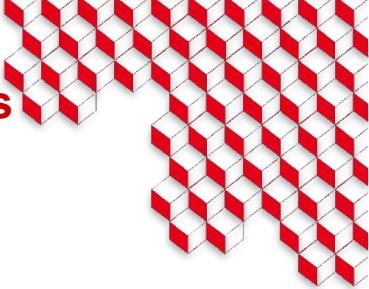
Premier Auteur	Type de Verre	Li	B	O	Na
Stéphane Gin	NBS14/18_test1	0	10.71	60.65	8.45
Stéphane Gin	NBS14/18_test2	0	10.71	60.65	8.45
Stéphane Gin	NBS14/18_test3	0	10.71	60.65	8.45
Stéphane Gin	NBSA	0	10.09	60.59	7.99
Stéphane Gin	NBSAC	0	9.74	60.25	7.65
Stéphane Gin	NBSACZ	0	9.58	60.32	7.61
Stéphane Gin	NBSAZ	0	9.94	60.7	7.84
Stéphane Gin	NBSC	0	10.37	60.24	8.17
Stéphane Gin	NBSCZ	0	10.2	60.37	8.0
Stéphane Gin	NBS12/28_test1	0	15.61	60.79	6.54
Stéphane Gin	NBS12/28_test2	0	15.61	60.79	6.54
Stéphane Gin	NBS36/21	0	12.12	54.05	21.19
Stéphane Gin	NBS31/15	0	9.08	55.8	18.71
Stéphane Gin	NSAC19_test1	1.24	0	57.24	12.42
Stéphane Gin	NSAC19_test2	1.24	0	57.24	12.42
Stéphane Gin	NSAC19_test3	1.24	0	57.24	12.42

Showing 1 to 735 of 735 entries

Télécharger les données



Partie 2 : Entraînement d'un réseau de neurones sur la base de données



Étape 1



Étape 2

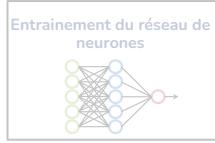


Tableau d'altération des verres

Télécharger un fichier PDF scanné
Browse... No file selected.

Télécharger les données

Premier Auteur	Type de Verre	Li	B	O	Na
Stéphane Gin	NBS14/18_test1	0	10.71	60.65	8.45
Stéphane Gin	NBS14/18_test2	0	10.71	60.65	8.45
Stéphane Gin	NBS14/18_test3	0	10.71	60.65	8.45
Stéphane Gin	NBSA	0	10.09	60.59	7.99
Stéphane Gin	NBSAC	0	9.74	60.25	7.65
Stéphane Gin	NBSACZ	0	9.58	60.32	7.61
Stéphane Gin	NBSAZ	0	9.94	60.7	7.84
Stéphane Gin	NBSC	0	10.37	60.24	8.17
Stéphane Gin	NBSCZ	0	10.2	60.37	8.0
Stéphane Gin	NBS12/28_test1	0	15.61	60.79	6.54
Stéphane Gin	NBS12/28_test2	0	15.61	60.79	6.54
Stéphane Gin	NBS36/21	0	12.12	54.05	21.19
Stéphane Gin	NBS31/15	0	9.08	55.8	18.71
Stéphane Gin	NSAC19_test1	1.24	0	57.24	12.42
Stéphane Gin	NSAC19_test2	1.24	0	57.24	12.42
Stéphane Gin	NSAC19_test3	1.24	0	57.24	12.42

Showing 1 to 735 of 735 entries

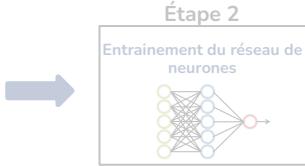
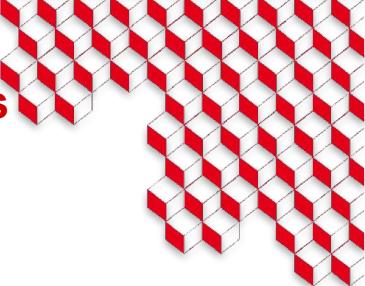
Télécharger les données



735 verres

107 paramètres

Partie 2 : Entraînement d'un réseau de neurones sur la base de données



Nettoyage de la base de données:

Altération vitesse (Spm ou SBET)	$V_a(Si) \text{ (g m}^{-2} \text{ d}^{-1}\text{)}$	r^2	Ordonnée origine	$V_a(B) \text{ (g m}^{-2} \text{ d}^{-1}\text{)}$	Ordonnée origine	$V_a(Na) \text{ (g}$
12.0	0.999	0.2	Not available	Not available	Not available	Not avail
17.7	0.989	0.1	Not available	Not available	Not available	Not avail
12.2	0.999	0.1	Not available	Not available	Not available	Not avail
2.6	0.997	-0.02	Not available	Not available	Not available	Not avail
9.9	0.986	-0.08	Not available	Not available	Not available	Not avail
8.2	0.99	0.2	Not available	Not available	Not available	Not avail
2.2	0.997	0.04	Not available	Not available	Not available	Not avail
52.2	1.0	0.4	Not available	Not available	Not available	Not avail
56.4	0.999	0.8	Not available	Not available	Not available	Not avail
192	1.0	0.4	Not available	Not available	Not available	Not avail
202	0.995	0.6	Not available	Not available	Not available	Not avail
47370	0.986	-31.2	Not available	Not available	Not available	Not avail
2070	0.999	3.8	Not available	Not available	Not available	Not avail
5.6	0.998	0.05	Not available	Not available	Not available	Not avail
7.2	0.999	-0.1	Not available	Not available	Not available	Not avail
5.5	0.984	-0.12	Not available	Not available	Not available	Not avail

735 verres



107 paramètres

370 verres



38 paramètres

230 verres



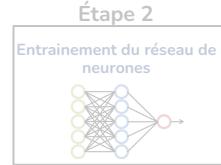
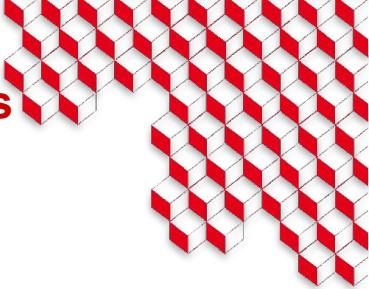
38 paramètres

Variable dépendante $V_0(B)$

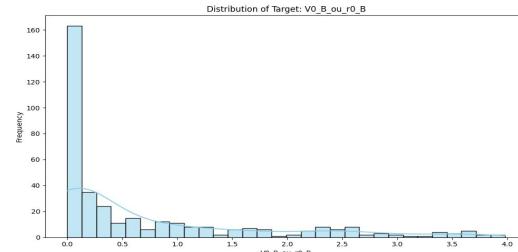
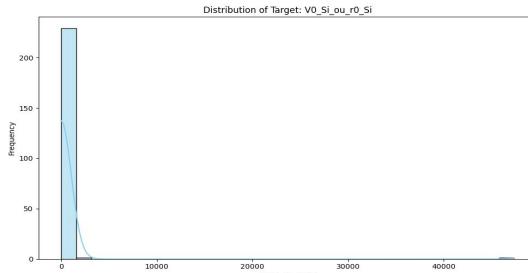
Variable dépendante $V_0(Si)$

Variable dépendante = Le paramètre qu'on veut prédire à la fin pour chaque verre

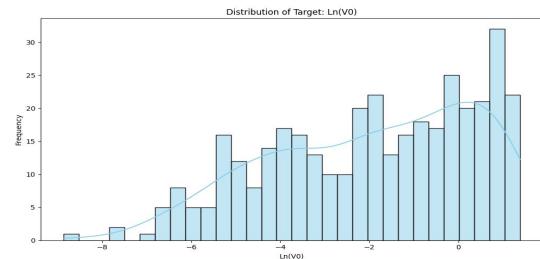
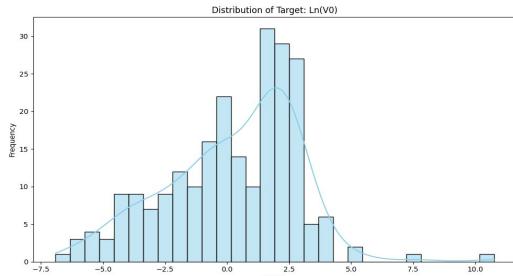
Partie 2 : Entraînement d'un réseau de neurones sur la base de données



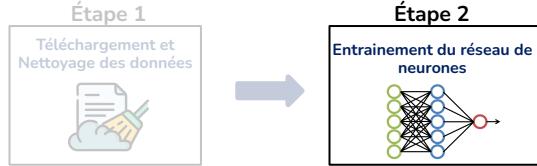
Transformation de la variable dépendante par le logarithme népérienne :



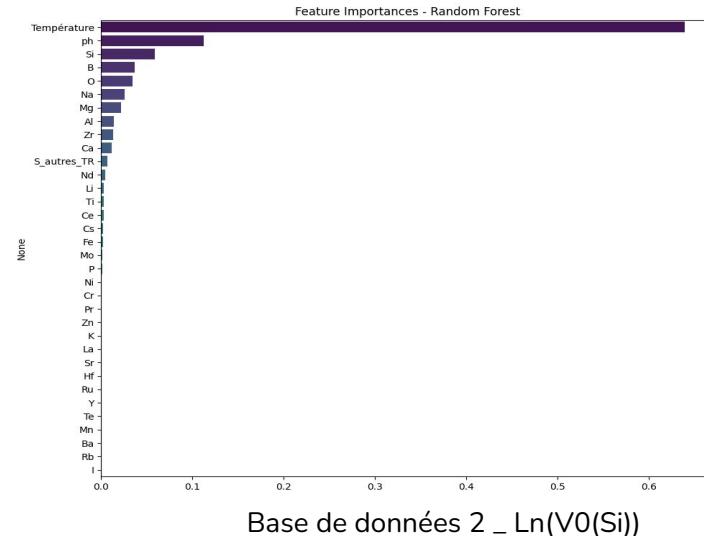
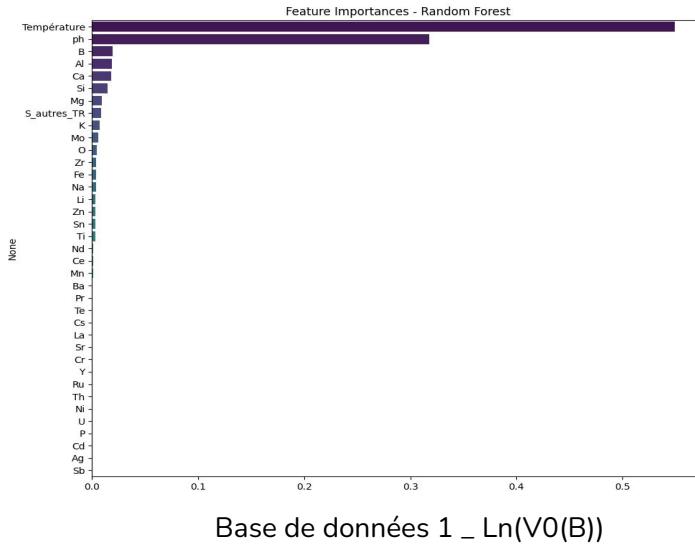
$\ln(V0)$



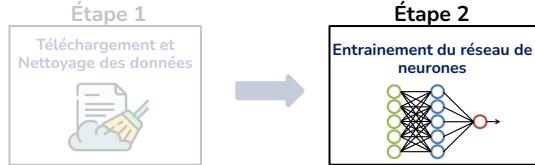
Partie 2 : Entraînement d'un réseau de neurones sur la base de données



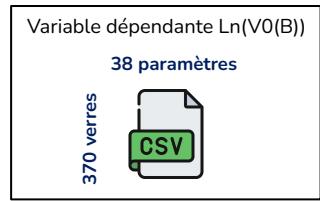
Les paramètres d'entrée influents pour les deux bases de données :



Partie 2 : Entraînement d'un réseau de neurones sur la base de données

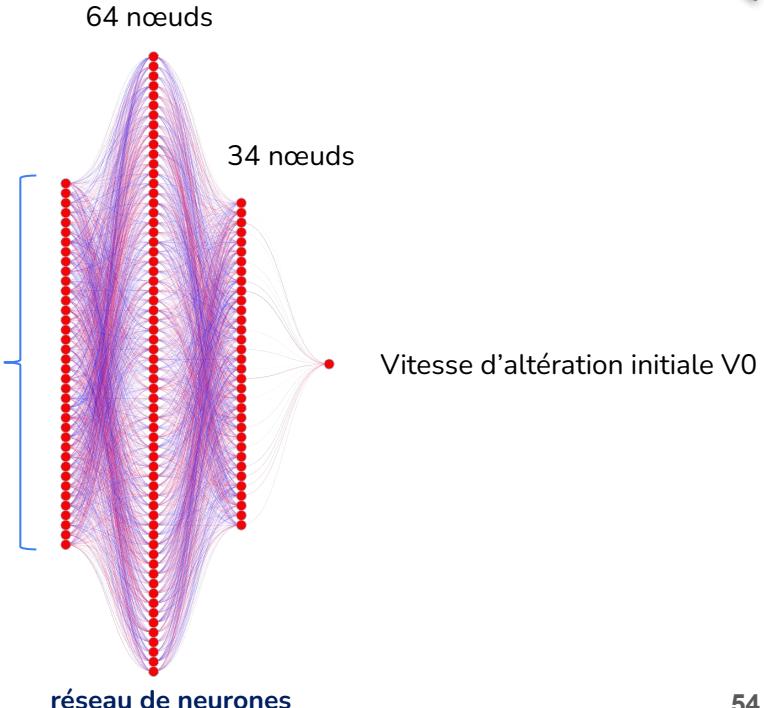


Entrainement du réseau de neurones sur les deux bases :

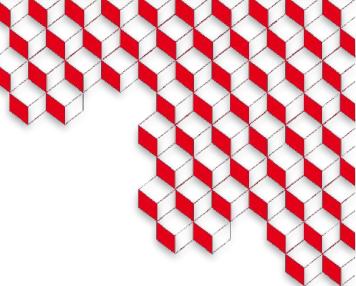


80% entraînement
20% test

38 variables d'entrée



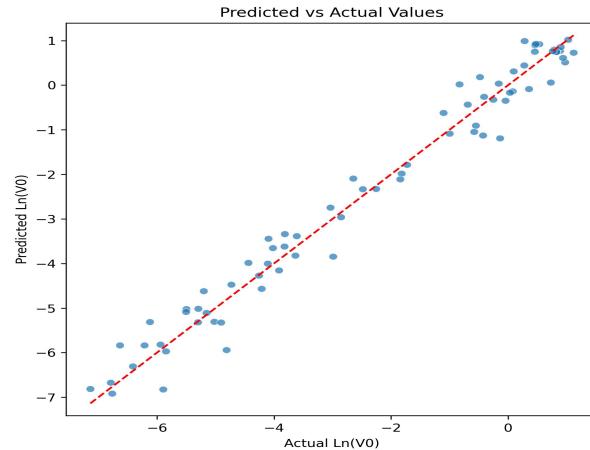
Partie 2 : Entraînement d'un réseau de neurones sur la base de données



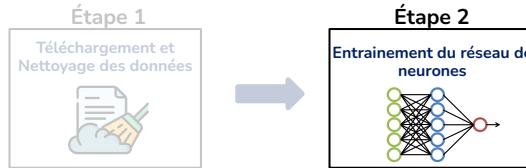
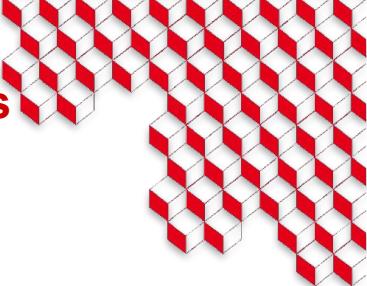
Résultats de prédiction de la vitesse d'altération initiale $\ln(V_0(B))$:

```
Skipping registering GPU devices...
3/3 0s 20ms/step
✓ Neural Network Regression Performance:
R² Score      : 0.9734
MAE          : 0.3366
RMSE         : 0.4269
```

Le modèle arrive à expliquer 97 % des données



Partie 2 : Entraînement d'un réseau de neurones sur la base de données



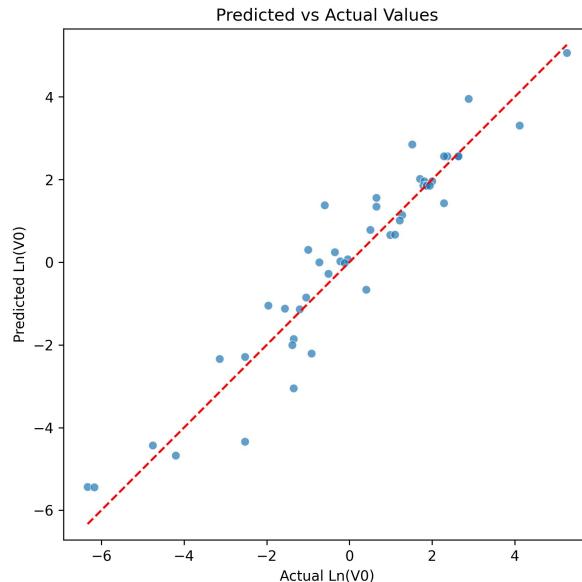
Résultats de prédition de la vitesse d'altération initiale V0(Si):

```
libraries. Please make sure the missing libraries mentioned above are installed properly if you would like to use GPU. Follow the guide at https://www.tensorflow.org/install/gpu for how to download and setup the required platform.
Skipping registering GPU devices...
2/2 ━━━━━━ 0s 39ms/step

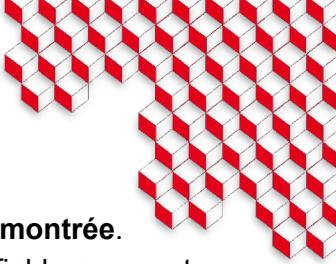
✓ Neural Network Regression Performance:
R² Score      : 0.9050
MAE           : 0.5515
RMSE          : 0.7443

o (3.11,11)  ao280403@VRH050897:~/Bureau/MI_Model$
```

Le modèle arrive à expliquer 90 % des données



Conclusions et perspectives



✓ Conclusion

- Les résultats obtenus sont **très satisfaisants** :
 - la **faisabilité de l'extraction automatique de données** par IA générative est **clairement démontrée**.
 - Malgré le nombre de données limité dans la base de données, le modèle prédictif est assez fiable avec un taux d'explication de plus de 90%.
- Le **prompting** s'est révélé être un **levier essentiel** pour assurer la **pertinence** des extractions.

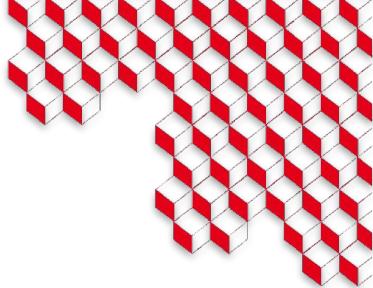
❖ Perspectives

- **Enrichissement de la base de données** avec :
 - Les **notes techniques internes** en diffusion limitée.
 - D'autres **sources documentaires externes**
- **Extension des travaux de modélisation** à d'autres paramètres prédictifs
- **Déploiement de l'infrastructure IA du CEA**, stabilisée et **opérationnelle dès décembre 2025**, à destination de l'ensemble des salariés

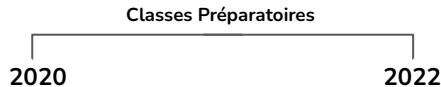
Merci pour votre attention

Annexe

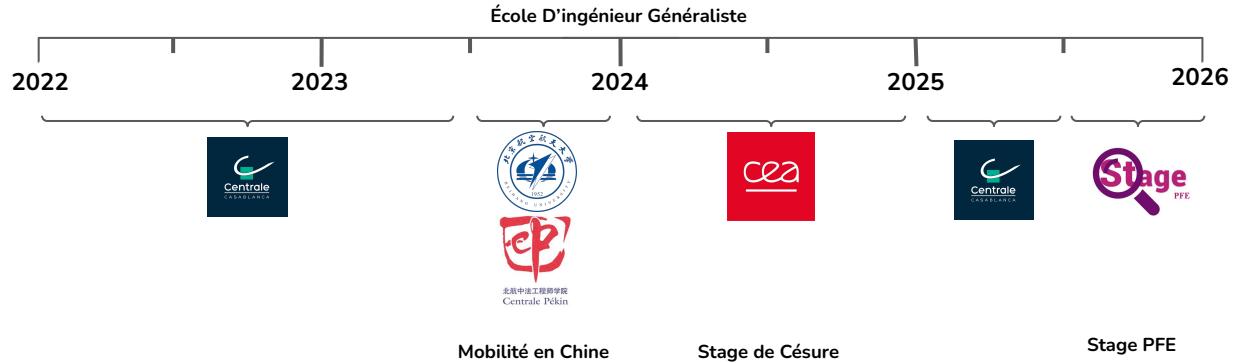
Je me présente :



LYCEE
MOHAMMED VI
D'EXCELLENCE



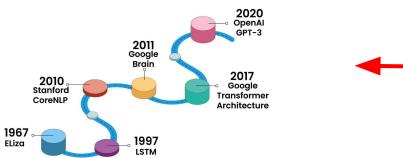
Filière: Technologies et Sciences Industriels



Présentation du stage IA générative au service de la constitution d'une base de données et construction d'un modèle prédictif de la vitesse d'altération - Abdelhak OUANZOUGUI

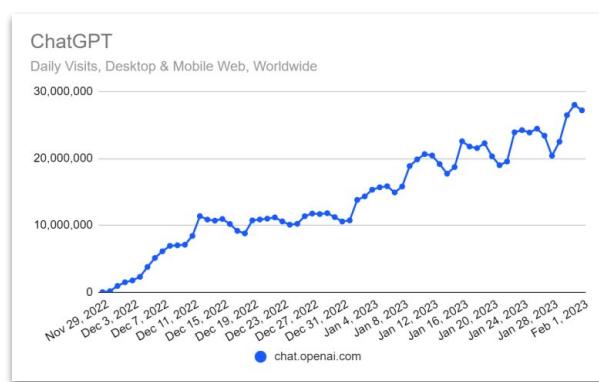
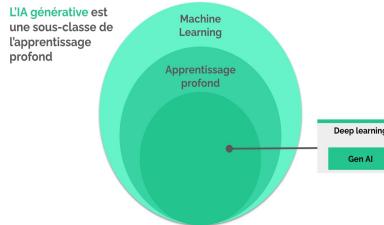
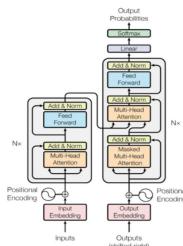
Contexte : Généralités sur l'IA Générative

- **2017** : Introduction des Transformers avec l'article « Attention is All You Need », révolutionnant le traitement du langage.
- **2018-2020** : Émergence de modèles comme GPT et BERT, améliorant la génération de texte et la compréhension du contexte.
- **2020** : Lancement de GPT-3 par OpenAI, avec 175 milliards de paramètres, pour une génération de texte quasi humaine.
- **2022** : ChatGPT démocratise l'IA générative avec des dialogues accessibles.
- **Depuis 2022** : Expansion des usages, croissance de l'open-source, et débats sur l'éthique et la souveraineté des données.



Transformer

Attention Is All You Need

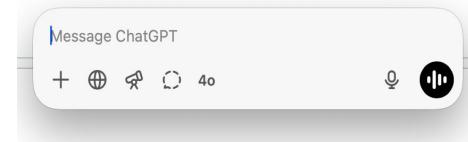


Source : <https://wisernotify.com/blog/chatgpt-users/>, dernière visite du site 03.06.2025

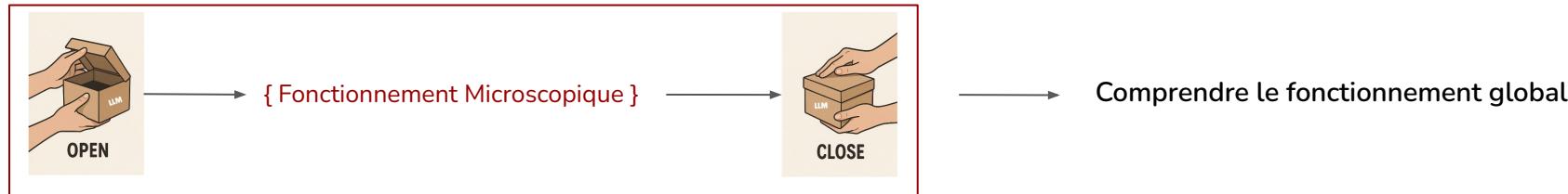
Contexte : Généralités sur l'IA Générative

Questions légitimes :

- Qu'est-ce qu'un LLM vraiment ?
- Que se passe-t-il derrière les interfaces conversationnelles comme ChatGPT ?
- À qui parlez-vous vraiment derrière la barre de recherche ?



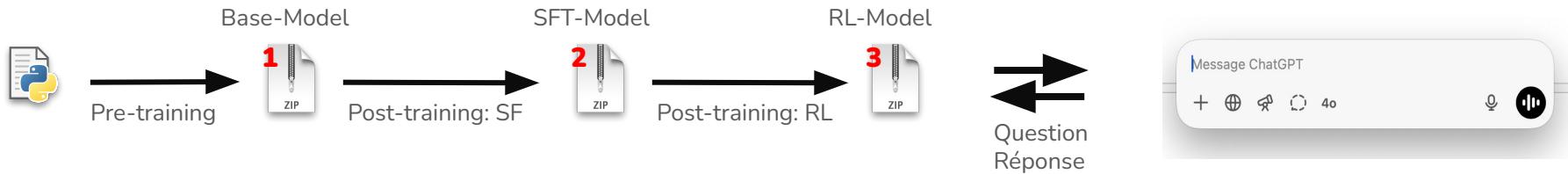
Objectif de cette section : Comprendre le fonctionnement microscopique d'un modèle de langage (LLM), tout en conservant une vision d'ensemble de son mécanisme global.



Pour répondre aux questions précédentes, nous allons voir comment on peut construire un LLM à partir de zéro.

Contexte : Généralités sur l'IA Générative

Les étapes de construction d'un LLM :

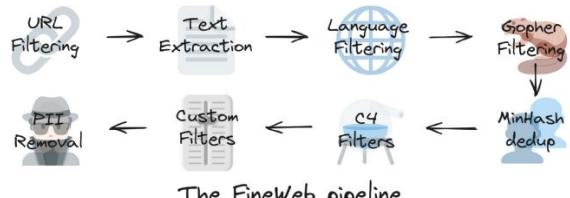


- **Fichier Python (quelques centaines de lignes de code):**
 - Pré-entraînement du modèle sur l'ensemble des données d'Internet, la phase la plus coûteuse, représentant des millions d'euros, et que nous détaillerons dans cette présentation.
- **Dossier ZIP - Modèle de base optimisé (capable de prédire uniquement le mot suivant dans une séquence de texte):**
 - Post-entraînement et optimisation supervisée : conférer une personnalité au modèle pour interagir avec l'utilisateur et répondre à ses questions.
- **Dossier ZIP - Modèle renforcé (capable d'effectuer des raisonnements approfondis et de réaliser des recherches sur Internet)**
 - Post-entraînement et apprentissage par renforcement : étendre les capacités du modèle pour effectuer des recherches sur Internet et améliorer ses réponses.

Contexte : Généralités sur l'IA Générative

Pré-entraînement d'un modèle:

Étape 1 : Télécharger et prétraiter les données d'Internet



<https://huggingface.co/spaces/HuggingFaceFW/blogpost-fineweb-v1>

Exemple de base de données

Base de données FineWeb (44 TB d'espace et 15 milliards de tokens)

[Viewing Single Post From: Spoilers for the Week of February 11th]\n[Lil] | Feb 1 2013, 09:58 AM|I don't care about Chloe/Tanice/Jen-Jen. Don't care about Sami, really, but hoping that we get some good "SAMANTHA GENE!!" Marlena Death-Stares out of it. And "newfound" feelings. Please. If only.\n[STEFANO!! STEFANO, STEFANO!!! :cheers! for the week of February 11th - 1 DAY! New Spoilers & Discussion => Fundamentals of communication, let's do some more! Sorry about the attitude, community! You have set yourself on fire, do not run.\nOKAY? OKAY? Please?!\nLock, D, you had two months to say to Harvard in private emails, "I'm sorry, I shouldn't have been using that animation in my paid presentations. I want use it again. I really do like 'Inner Life', though, and would love to use it in classroom presentations, from the BioVisions site, if that is acceptable.\nI'm sorry here, for two months, for that, I haven't anything to happen, and it did. Two months, and you have done, your own and have written similar post to me, and I understand that you have benefit of the doubt-- maybe you didn't know the credits weren't visible to the audience, and I wouldn't have said a word beyond this, as its Harvards problem, not mine. This would have been a funny joke to those of us involved in dealing with people, but it would have been a PR non-issue for you.\nBut after you set yourself on fire, you didn't just sit it out with a bucket of cold water, and an accountable person ran, and you still running.\nWhy not just state the scenario? Sorry everyone, and move on. What the excuse?!\nIt's the damn fault of the people language? Why the vague words and cryptic capitalizations? Why the writes and rewrites of your 'press release'? We know it wasn't written of your own volition, or it would have been done before Harvard had to take action. And, your behavior before this, regarding this issue, is not indicative of someone who made an innocent mistake. Its weird, and I think with this frantic behavior is the intent to see what happens a patternical feature of Creationists? Or are they doing something? Is it both? Or is it none? Now, how can I get Carl working in on the issue, according to cre8id at AboveTopSecret.com--- PBS/Nova online - Intelligent Design on trial:\n...to my knowledge, Discovery Institute has neither authorized nor received nor is making use of any presentation that used that animation. We have had nothing to do with creating or selling a DVD of that animation, nor do we have had anything to do with placing that animation on the video. I think what he talked about that last part, but the first part sounds similar to DB's claims post-Dover ("WE HAD NOTHING TO DO WITH DOVER!")...\nMaybe Luskin is telling the truth. Maybe this was a magic non-science Creation-friendly narration with convenient edits that AIG or ICR would have killed for... but only Dembski could find it... but he cant tell us where... and he didn't share it with anyone, and its subsequently disappeared from the Internet.\n[...] that simple is that I've been told that there are all kinds of videos of people doing or projecting out of his screen to remain silent. But what if we find more videos of more DI fellows presenting this animation? No shrug, and move on. A novel two-step immunotherapy approach has shown clinically beneficial responses in patients with advanced ovarian cancer. Following Lifestyle Recommendations Reduce Risk of Cancer Death\nPeople who follow the diet and lifestyle recommendations laid out by the WCRF and the AICR have a 20 percent reduced risk of dying from cancer. UCSF launches Social Networking Site for Patients and Families with Hereditary Cancers\nFor Immediate Release May 14.

Remarques :

Recherche des URLs : Identifier l'ensemble des URLs disponibles sur Internet.

Filtrage des contenus : Supprimer les URLs malveillantes ou contenant des contenus inappropriés (racistes, agressifs, etc.).

Extraction du texte : Récupérer tout le texte présent dans les URLs filtrées.

Concaténation : Regrouper les textes extraits pour former une unique séquence textuelle.

Résultat du traitement : Malgré l'immensité d'Internet, le nettoyage et le traitement réduisent le volume à quelques dizaines de téraoctets de texte.

Contexte : Généralités sur l'IA Générative

Pré-entraînement d'un modèle:

Étape 2 : Tokenization

Convertit le texte <----> séquences de symboles (/tokens)

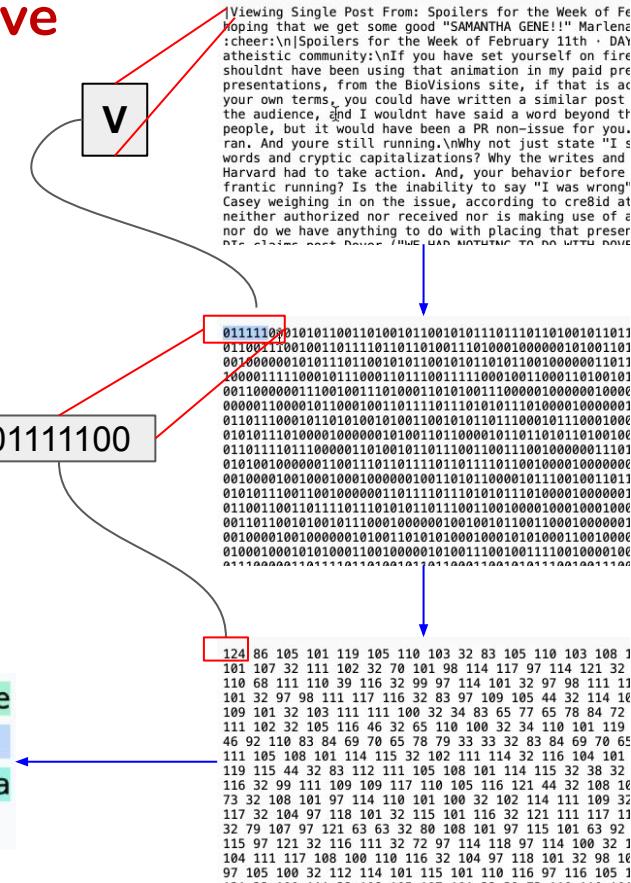
- Commence par un flux d'octets (256 tokens).
- Exécute l'algorithme de codage par paires d'octets (fusionne itérativement la paire de tokens la plus fréquente pour créer un nouveau token).

Exemple : ~5000 caractères de texte

- ~ 40 000 bits (avec un vocabulaire de 2 tokens : bits 0/1)
- ~ 5000 octets (avec un vocabulaire de 256 tokens : octets)
- ~ 1300 tokens GPT-4 (taille du vocabulaire : 100 277)

<https://tiktokenizer.vercel.app/>

|Viewing Single Post From: Spoilers for the Week of February 11th|\\n|Lil| |Feb 1 2013, 09:58 AM|\\nDon't care about Chloe/Taniel/Jen-Jen. Don't care about Sami, really, but hoping that we get



Contexte : Généralités sur l'IA Générative

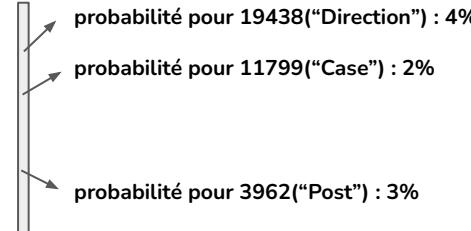
Pré-entraînement d'un modèle:

Étape 3 : Entraînement de réseau de neurons



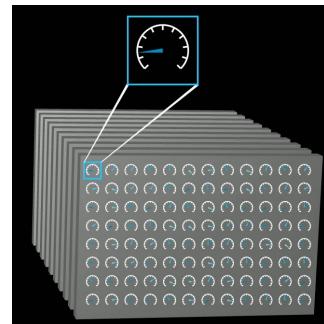
réseau de neurons

100,277 probabilités pour le token suivant



La réponse correcte est « Post » (3962), donc le réseau de neurons doit corriger ces poids pour que sa probabilité soit la plus élevée.

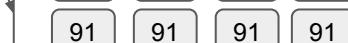
La structure interne d'un modèle :



Jetons de séquence d'entrée x
de 1 à, par exemple, 8 000 jetons



Paramètres (/ « poids ») w
généralement des milliards de ces derniers

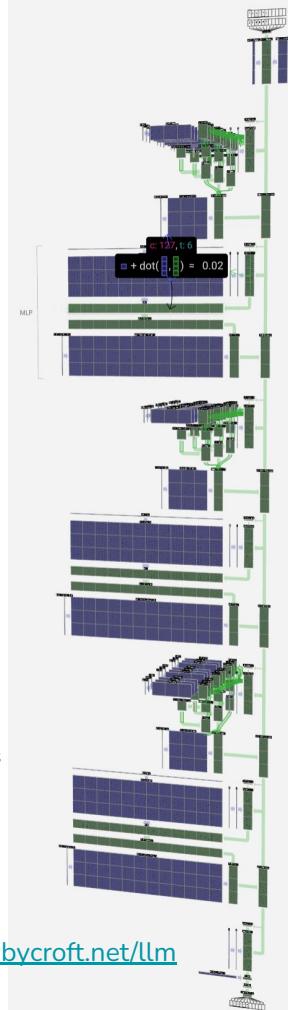


Expression mathématique complexe

$$\begin{aligned} & 1/(1 + \exp(-(w_0 * (1/(1 + \exp(-(w_1 * x_1 + w_2 * x_2 + w_3)))) + w_4 \\ & \quad * (1/(1 + \exp(-(w_5 * x_1 + w_6 * x_2 + w_7)))) \\ & \quad + w_8 * (1/(1 + \exp(-(w_9 * x_1 + w_{10} * x_2 + w_{11})))) + w_{12}))) \end{aligned}$$

100,277 probabilités

<https://bbycroft.net/llm>

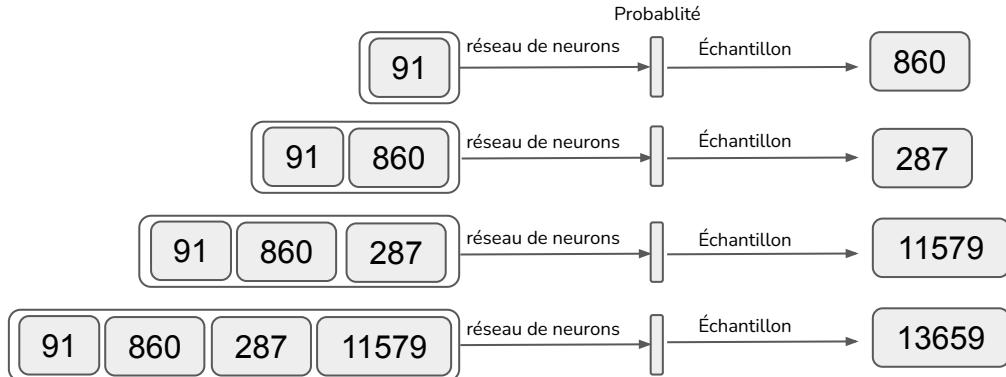


Contexte : Généralités sur l'IA Générative

Pré-entraînement d'un modèle:

Étape 4 : Inférence

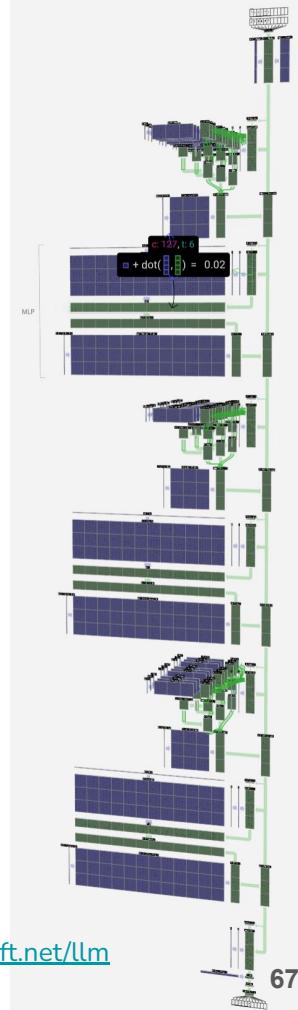
Pour générer des données, il suffit de prédire un token à la fois.



Résultats : Modèle de base capable de prédire le mot suivant, mais incapable de mener des conversations avec des utilisateurs.



<https://bbycroft.net/llm>



Contexte : Généralités sur l'IA Générative

Post-entraînement d'un modèle - SUPERVISED FINE TUNING:

Objectif : Donner une personnalité au modèle et lui apprendre comment répondre aux questions et comment respecter des instructions, et résoudre le problème des hallucinations.

Conversations

Humain : « Combien font $2+2$? »

Assistant : « $2+2 = 4$ »

Humain : « Et si c'était * au lieu de + ? »

Assistant : « $2*2 = 4$, pareil que $2+2$! »

- Premiers travaux : 2022 - Article Instruct GPT.
- Les étiqueteurs humains rédigent des conversations basées sur des instructions d'étiquetage.
- Aujourd'hui, une grande partie de l'étiquetage est assistée par des LLM (par exemple, les humains modifient plus qu'ils n'écrivent), ou est entièrement synthétique.

Excerpt of labeling instructions on the API prompt distribution

You are given a text-based description of a task, submitted by a user. This task description may be in the form of an explicit instruction (e.g. "Write a story about a wise frog."). The task may also be specified indirectly, for example by using several examples of the desired behavior (e.g. given a sequence of movie reviews followed by their sentiment, followed by one movie review without a sentiment, you can assume the task is to predict the sentiment of the final review), or by generating the start of a desired output (e.g. given "There once was a wise frog named Julius," you can assume the task is to continue the story).

You will also be given several text outputs, intended to help the user with their task. Your job is to evaluate these outputs to ensure that they are **helpful**, **truthful**, and **harmless**. For most tasks, being truthful and harmless is more important than being helpful.

Résultats : Modèle fine tuned capable de répondre aux questions des utilisateurs de manière interactive mais il est limité sur sa base de données et souffre souvent du problème d'hallucinations .

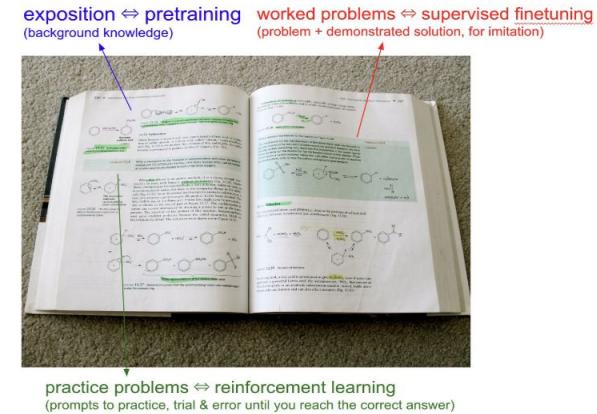


Contexte : Généralités sur l'IA Générative

Post-entraînement d'un modèle - REINFORCEMENT LEARNING:

Objectif : Enseigner au modèle à apprendre, à effectuer des recherches sur Internet pour rester à jour, à réaliser des raisonnements approfondis et à accomplir davantage de tâches qu'une simple conversation.

- On donne un énoncé de problème (prompt) et la réponse finale.
- Nous voulons pratiquer des solutions qui nous mènent de l'énoncé du problème à la réponse, et les « internaliser » dans le modèle.
- On répète autant que nous le pouvons.



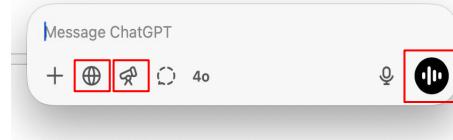
Résultats : Modèle optimisé pour un raisonnement plus approfondi et pour plus de capacités.



Contexte : Généralités sur l'IA Générative

Exemples d'utilisation des modèles d'inférence modernes :

- Conversation simple
- Recherche sur le web
- Recherche approfondie
- Raisonnement approfondi
- Assistant vocale



Les limites et les inconvénients des modèles d'inférence :

- Base de connaissances limitée : si vous demandez des informations sur des documents qui n'existent pas sur Internet, le modèle ne peut pas vous répondre.
- Les modèles d'inférence modernes ne peuvent pas interagir avec d'autres outils, par exemple, remplir un tableau Excel.
- Les modèles d'inférence modernes sont limités dans la taille de leurs réponses.
- L'enjeu de la confidentialité des documents internes.

Solutions Alternatives : Outils Open source !