

[sf=ir] &



REX : LES SENTENCE-TTRANSFORMERS EN

PROD Comment booster un search
avec des LLMs



POSITIVE
TECH < / >
ADEO GROUP

QUI SUIS-JE

ML Engineer

- Chez SFEIR depuis début 2024
- En Mission chez Adeo dans la foulée
- Équipe AAAI

A part le Machine Learning

- Force athlétique (apprenti)
- Bière craft (confirmé)



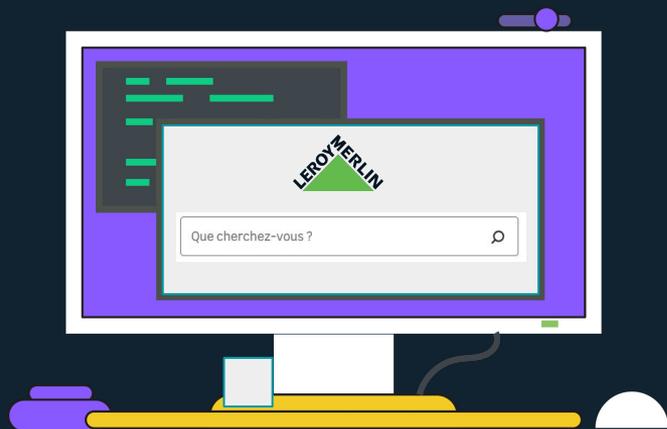
CONTEXTE



CONTEXTE

LE MOTEUR DE RECHERCHE

- **1,25 milliards** de requêtes en 2024
- **10 millions** de produits (Leroy Merlin France)
- Temps de réponse **100ms** (P50) à **300ms** (P90)
- Efficace sur
 - Mots clefs
 - Références produits
 - Thèmes
 - Requêtes connues
 - ...
 - Pas sur le langage naturel
- **2 à 4%** de “langage naturel”



CONTEXTE

LE MOTEUR DE RECHERCHE



25M

REQUÊTES

Traitées de manière non-optimale
(à prendre avec des pincettes)



CONTEXTE

LE “LANGAGE NATUREL”

- “Quels outils pour isoler une toiture ?”

Pas de résultat. Nous vous proposons “outils isoler ?”
Essayez aussi “[outils toiture ?](#)”

↑↓ Affiner

Pertinence



DEXTER
Lot de 3 pinces de précision isolées
DEXTER
★★★★★ (50)
Vendu par LEROY MERLIN



Coffret d'outils isolés pour véhicules hybrides et électriques
Vendu par ZOOMICI
● Livraison offerte



DEXTER
Jeu de 12 tournevis isolés d'électricien
DEXTER + sac de...
★★★★★ (190)
Vendu par LEROY MERLIN



Set de tournevis dynamométriques isolés 1000V TT1 tray Plat- PH...
Vendu par Indoostrial
● Livraison offerte



CONTEXTE

LE “LANGAGE NATUREL”

- “Rénover grenier”

Pas de résultat. Nous vous proposons “renover”

↑↓ Affiner

Pertinence ▾

Pack rénovateur extérieur filaire + 2 brosses FARTOOLS...
★★★★☆ (32)

Renovateur filaire 120 mm FARTOOLS, 1300 W
★★★★☆ (72)
Vendu par LEROY MERLIN

Rénovateur filaire FARTOOLS 615228, 1300 W
★★★★☆ (13)

Rénovateur filaire FARTOOLS MULTIREX, 900 W
★★★★☆ (7)



CONTEXTE

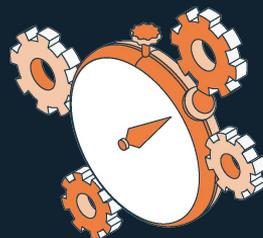
BESOIN



Répondre
mieux & à plus
de requêtes



Ne pas refaire
un search
entier



Rester rapide



N'impacter
aucun
utilisateur
négativement



POSITIVE
TECH < / >
ADEO GROUP

CONTEXTE

NOTRE APPROCHE

- **Solution non invasive, sans refonte**
 - Mobilisation du search actuel efficace en mode “classique”
- **Surcouche en 2 étapes principales**
 - Identification des requêtes
 - Décomposition des requêtes

} Search



CONTEXTE

NOTRE APPROCHE

#1

Classification

Identifier si la requête peut être envoyée au search telle quelle ou si elle doit être reformulée

#2

Reformulation

Décomposition de la requête initiale en plusieurs sous-requêtes avec un LLM

#3

Extraction des résultats

Contrôle et parsing des résultats, search multiple en fonction de la reformulation

#4

Affichage côté front

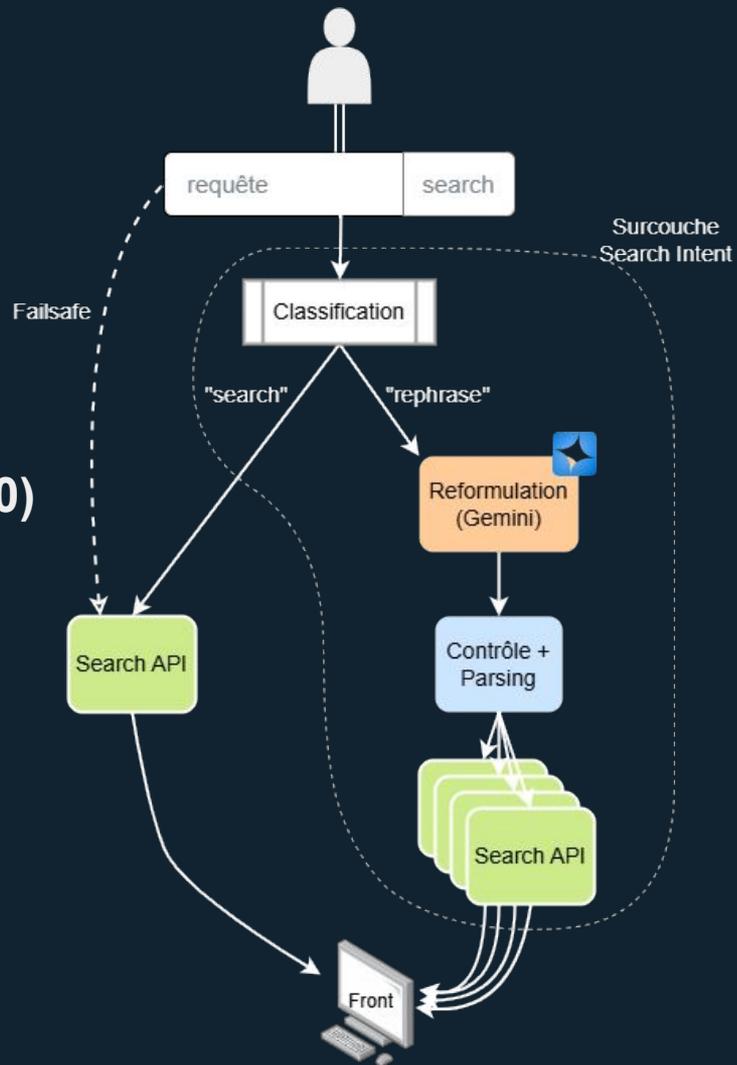
Ordonnancement et affichage des résultats



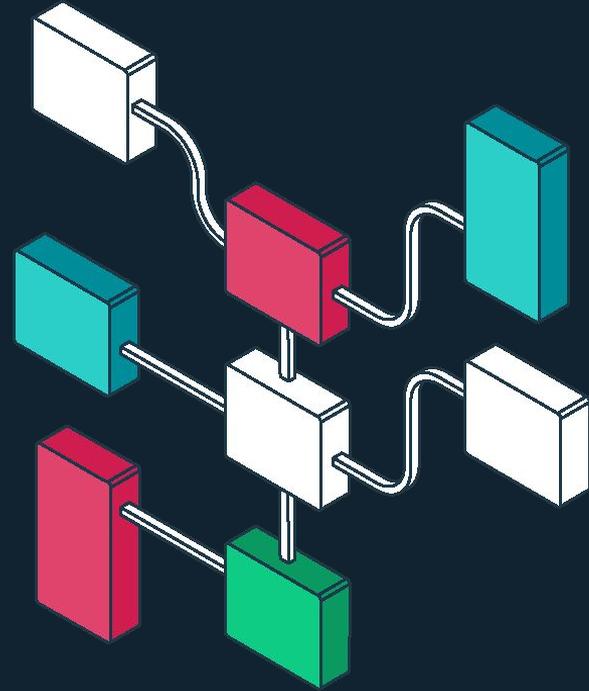
CONTEXTE

NOTRE APPROCHE

- **Failsafe**
 - Classification < 30ms (P90)
 - Reformulation < 1s (P90)...sinon retour parcours classique



CLASSIFICATION



CLASSIFICATION

DATA & LABELLISATION

Product
searches



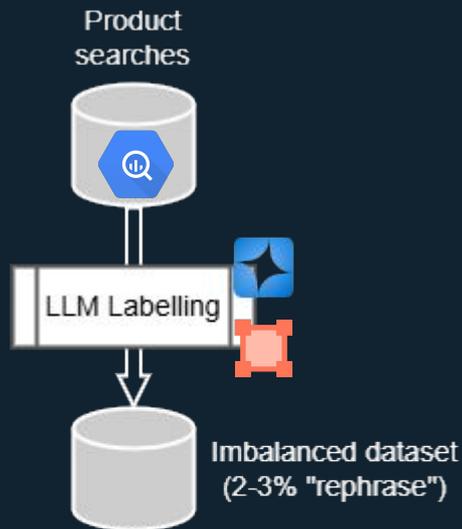
- **Stockage des données**
 - GCS / BigQuery



CLASSIFICATION

DATA & LABELLISATION

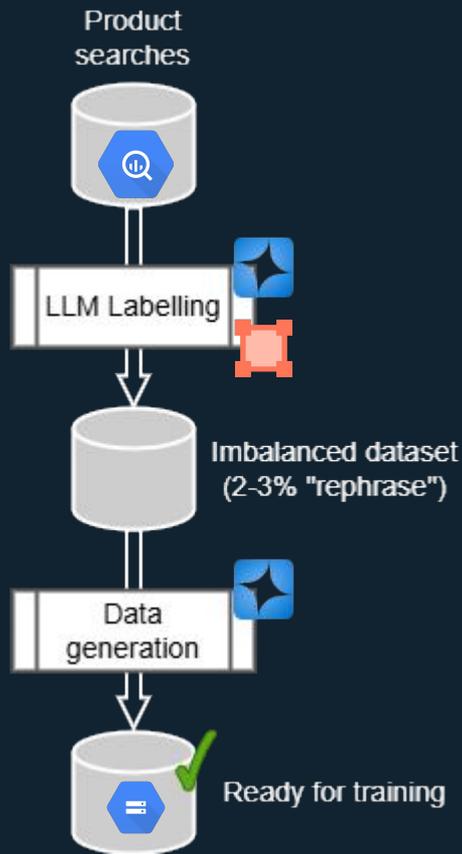
- **Stockage des données**
 - GCS / BigQuery
- **Labellisation**
 - Label Studio
 - Golden Dataset
 - Gemini
 - Few shot learning
 - Prompt engineering avec “batch”



CLASSIFICATION

DATA & LABELLISATION

- **Stockage des données**
 - GCS / BigQuery
- **Labellisation**
 - Label Studio
 - Golden Dataset
 - Gemini
 - Few shot learning
 - Prompt engineering avec “batch”
- **Augmentation avec Gemini**

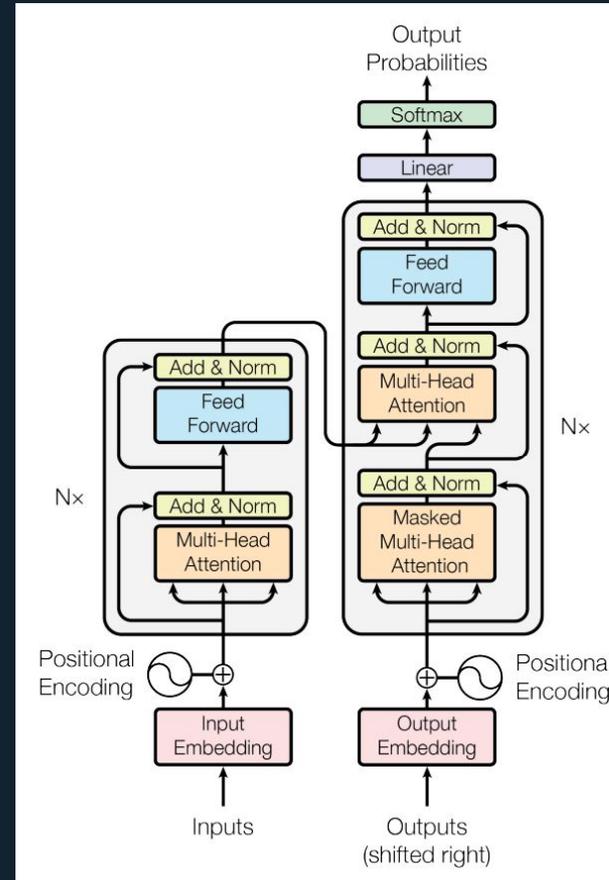


CLASSIFICATION

SENTENCE TRANSFORMERS



- Vaswani et al. “*Attention is all you need*” - 2017
- Transformer = “brique” de base implémentant le mécanisme d’attention

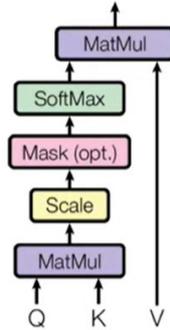


Source : “*Attention is all you need*”, Fig 1 - Vaswani et al.

CLASSIFICATION

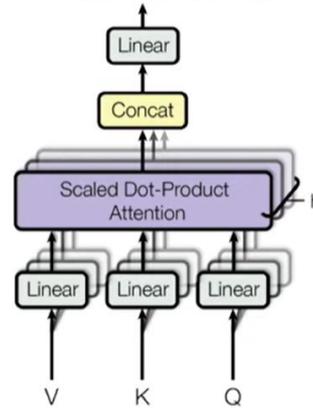
SENTENCE TRANSFORMERS

Scaled Dot-Product Attention



$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Multi-Head Attention



$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O$$

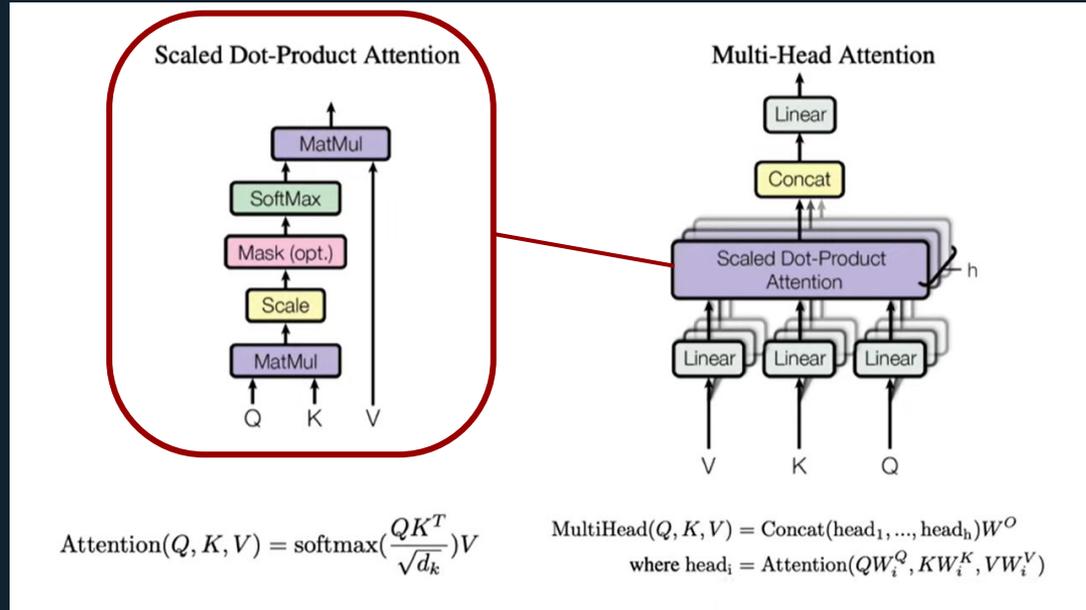
where $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$

Source : "Attention is all you need", Fig 2 -
Vaswani et al.



CLASSIFICATION

SENTENCE TRANSFORMERS



Source : "Attention is all you need", Fig 2 - Vaswani et al.



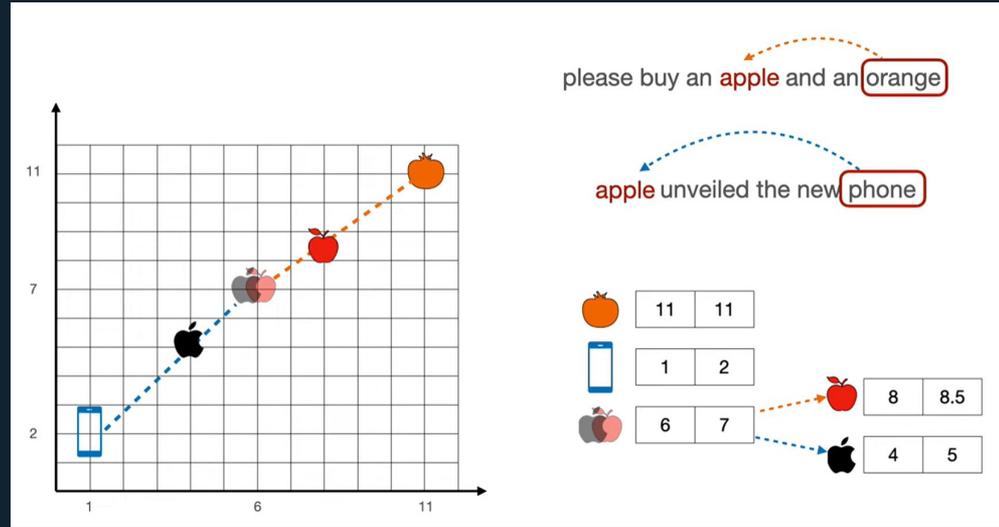
CLASSIFICATION

SENTENCE TRANSFORMERS

TL;DR

Scaled Dot-Product Attention

- Aucun poids ou paramètre
- Ajuste les embeddings initiaux
 - Identification des similarités
 - Recontextualise



Inspiré de : "The Attention in Large Language Models", Serrano.Academy



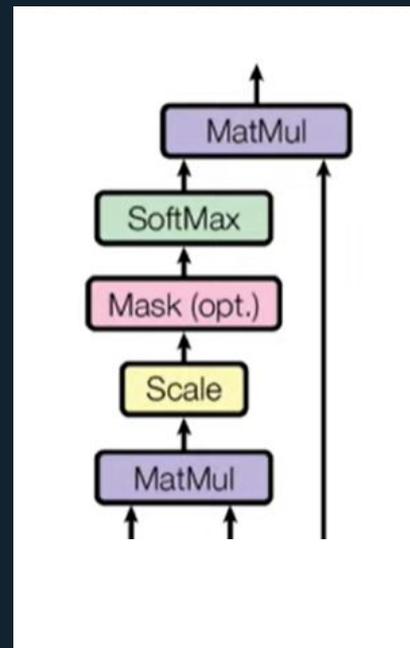
CLASSIFICATION

SENTENCE TRANSFORMERS

TL;DR

Scaled Dot-Product Attention

- Aucun poids ou paramètre
- Ajuste les embeddings initiaux
 - Identification des similarités
 - Recontextualise



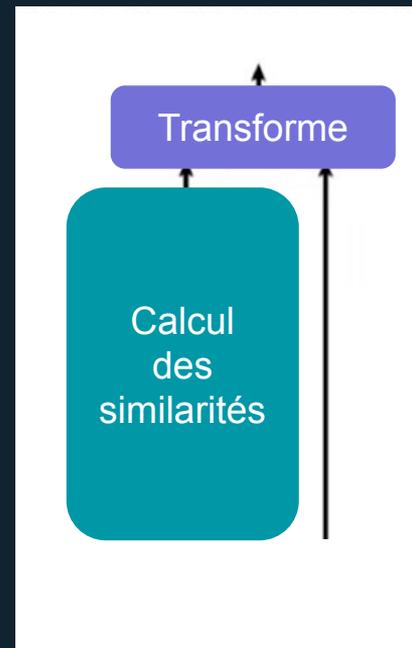
CLASSIFICATION

SENTENCE TRANSFORMERS

TL;DR

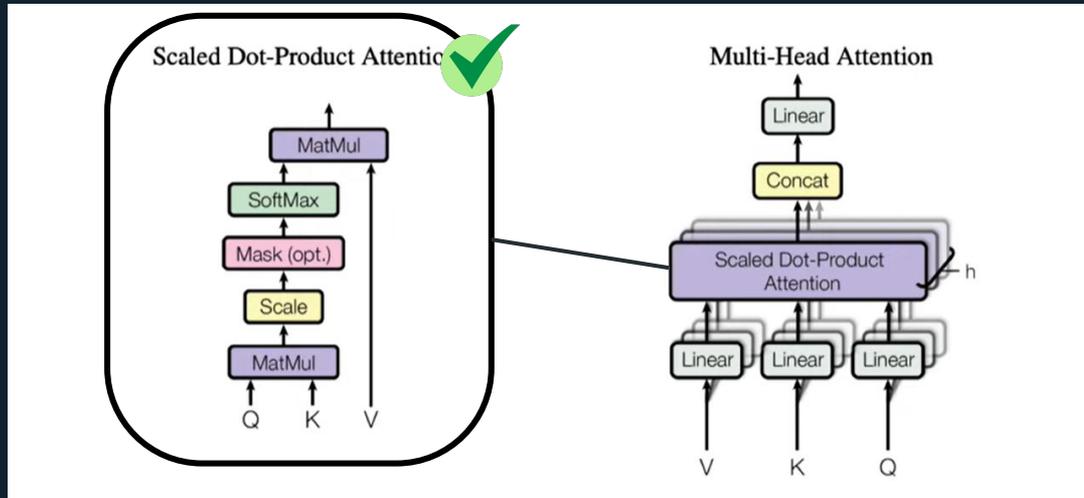
Scaled Dot-Product Attention

- Aucun poids ou paramètre
- Ajuste les embeddings initiaux
 - Identification des similarités
 - Chaque combinaison de 2 mots = 1 score
 - Recontextualise
 - Pondération par les scores



CLASSIFICATION

SENTENCE TRANSFORMERS

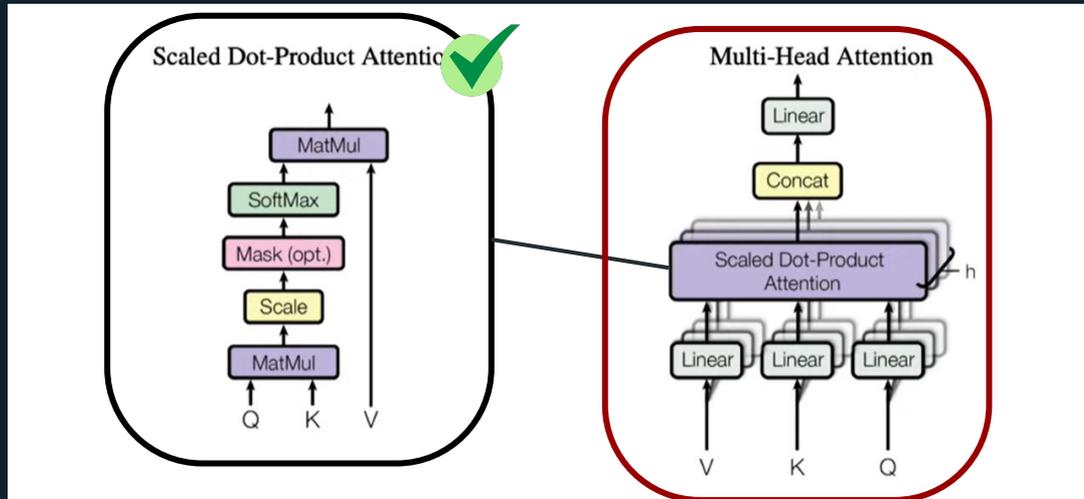


Source : "Attention is all you need", Fig 2 - Vaswani et al.



CLASSIFICATION

SENTENCE TRANSFORMERS



Source : "Attention is all you need", Fig 2 - Vaswani et al.



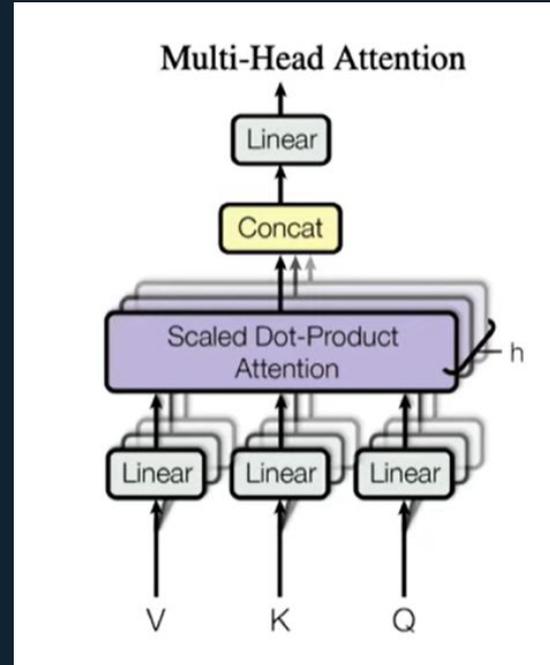
CLASSIFICATION

SENTENCE TRANSFORMERS

TL;DR

Multi-Head Attention

- Apprentissage de K, Q et V
 - Q - Query
 - K - Key
 - V - Value



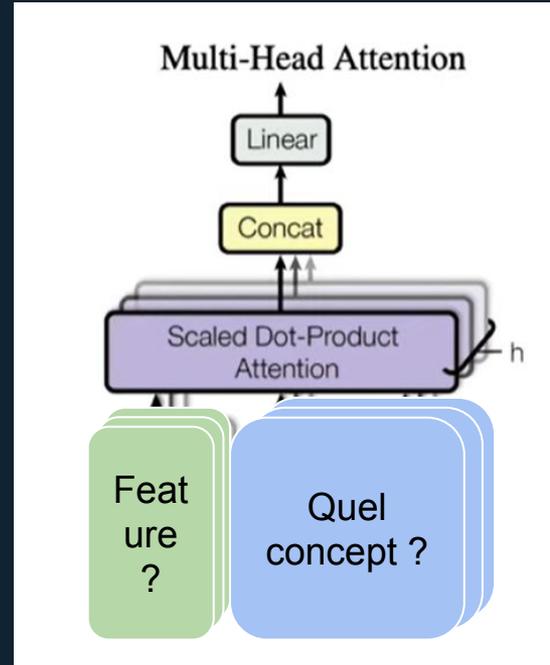
CLASSIFICATION

SENTENCE TRANSFORMERS

TL;DR

Multi-Head Attention

- Apprentissage de K, Q et V
 - Q - Query
 - K - Key
 - Quel concept ?
 - V - Value
 - Quels caractéristiques pertinentes pour le concept ?



CLASSIFICATION

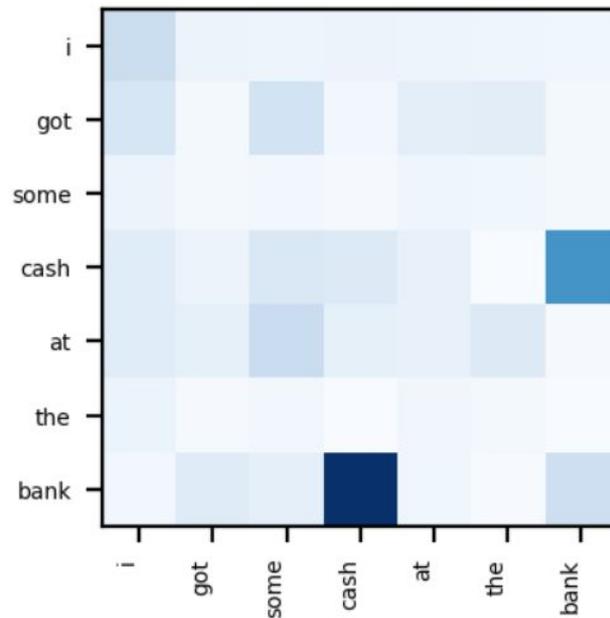
SENTENCE TRANSFORMERS

Exemple :

“I got some cash at the bank”

Comment réduire l'ambiguïté ?

Attention for Layer 1, Head 8:



<https://bertattentionviz.streamlit.app/>



CLASSIFICATION

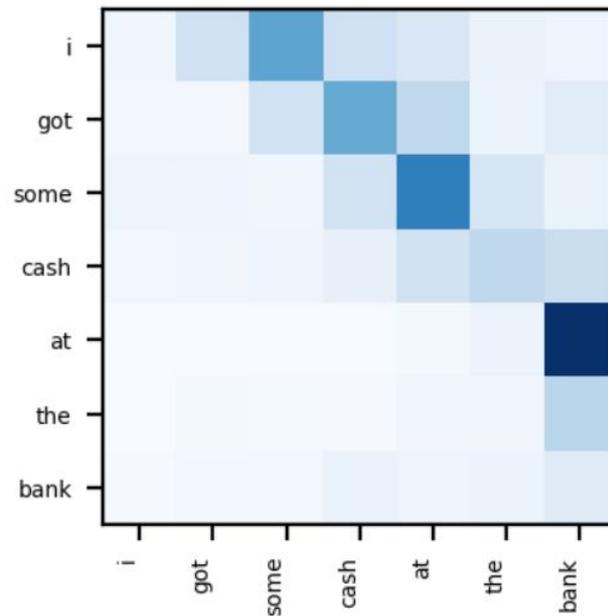
SENTENCE TRANSFORMERS

Exemple :

“I got some cash at the bank”

Comment augmenter le
contexte spatial ?

Attention for Layer 1, Head 9:



<https://bertattentionviz.streamlit.app/>



CLASSIFICATION

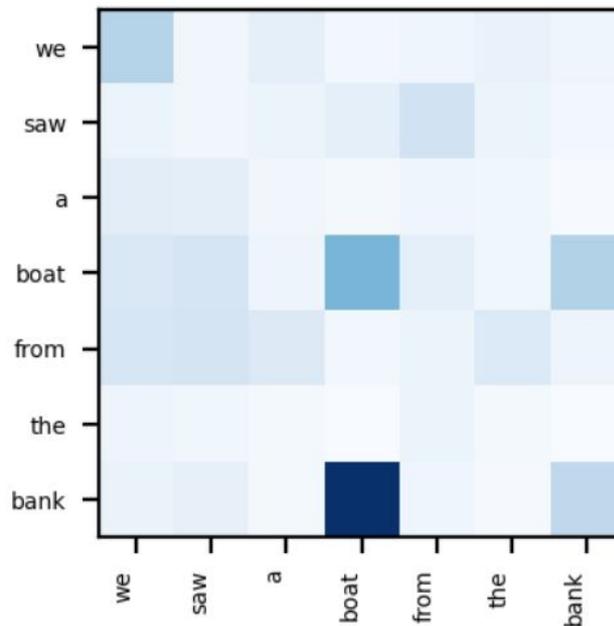
SENTENCE TRANSFORMERS

Exemple :

“We saw a boat from the bank”

Comment réduire l'ambiguïté ?

Attention for Layer 1, Head 8:



<https://bertattentionviz.streamlit.app/>

POSITIVE
TECH < / >

ADEO GROUP

CLASSIFICATION

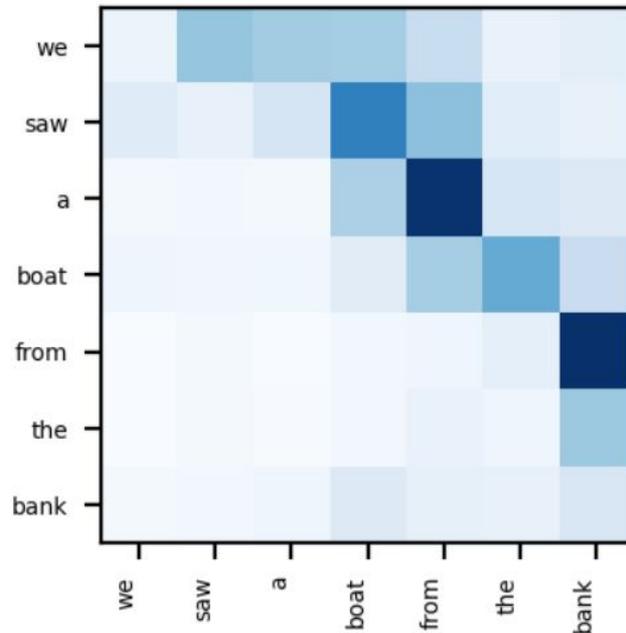
SENTENCE TRANSFORMERS

Exemple :

“We saw a boat from the bank”

Comment augmenter le
contexte spatial ?

Attention for Layer 1, Head 9:



<https://bertattentionviz.streamlit.app/>



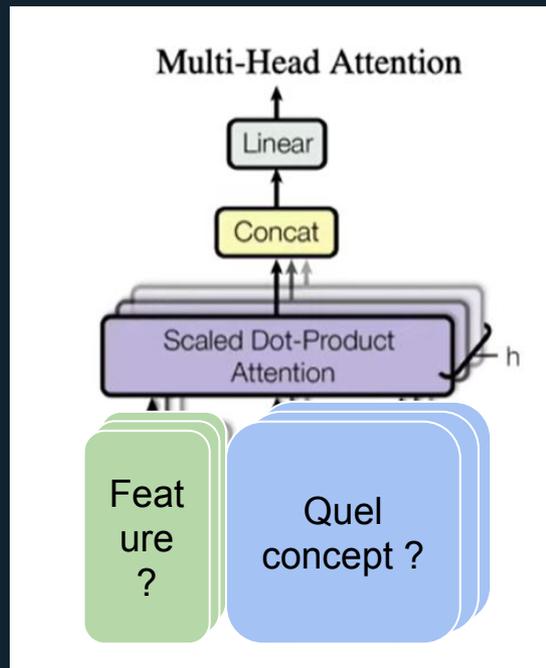
CLASSIFICATION

SENTENCE TRANSFORMERS

TL;DR

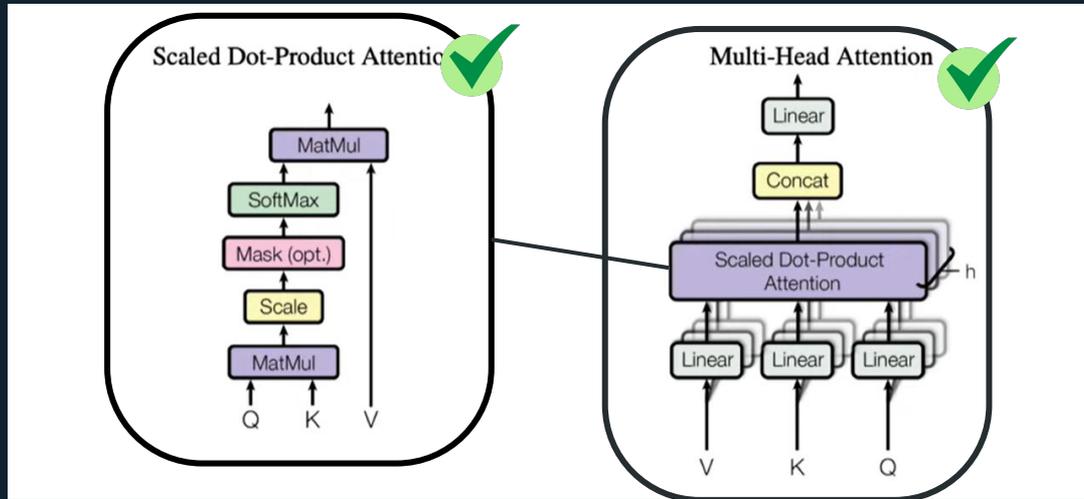
Multi-Head Attention

- Orienter la transformation
 - Une tête = Un concept sémantique



CLASSIFICATION

SENTENCE TRANSFORMERS



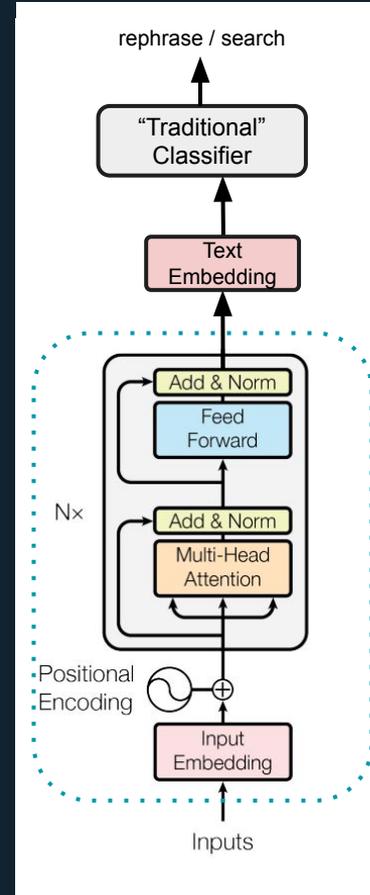
Source : "Attention is all you need", Fig 2 - Vaswani et al.



CLASSIFICATION

SENTENCE TRANSFORMERS

- Implémentation créée par UKP Lab et maintenue par Hugging Face 🤖
- Classifieur à entraîner
- Encodeur à choisir
 - TinyBert
 - DistilBert
 - Bert-large
 - ...



CLASSIFICATION

CHOIX DU MODÈLE

- Taille & nombre de paramètres (têtes d'attentions, couches...)
 - 60Mo → XXGo
- Performances
- Puissance de calcul requise
- ...

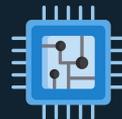
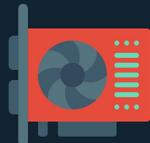
Meilleur compromis pour notre use case : **TinyBERT**

- Différence de performances
 - Plus que raisonnable
 - Compensable par l'augmentation du dataset
- CPU OK pour l'inférence
 - Permet de rester dans les standards Adeo



CLASSIFICATION

CHOIX DU MODÈLE

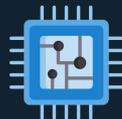
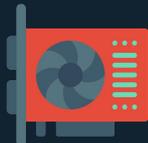


	GPU n1-standard-8 : 8vCPUs 30GB RAM + Tesla T4	CPU n1-standard-8 : 8vCPUs 30GB RAM	Avg weighted F1-score
Bert Uncased	26ms/prediction	65ms/prediction	0.973
TinyBert	2.8ms/prediction	12ms/prediction	0.971



CLASSIFICATION

CHOIX DU MODÈLE



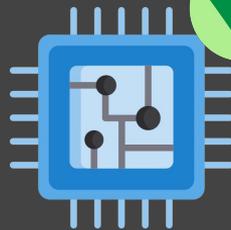
	GPU n1-standard-8 : 8vCPUs 30GB RAM + Tesla T4	CPU n1-standard-8 : 8vCPUs 30GB RAM	Avg weighted F1-score
Bert Uncased	26ms/prediction 	65ms/prediction 	0.973 
TinyBert	2.8ms/prediction 	12ms/prediction 	0.971 



CLASSIFICATION

CHOIX DU SERVING

TINYBERT +



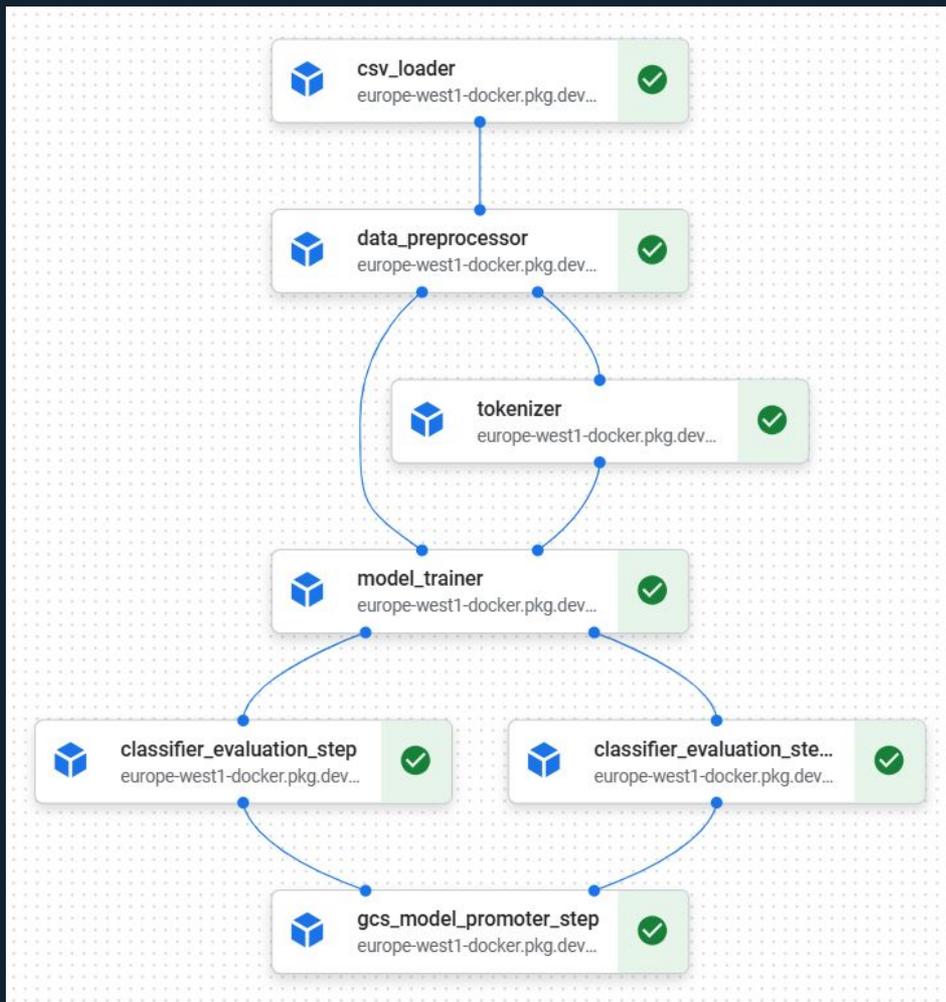
CLASSIFICATION TRAINING PIPELINE

ZenML



+

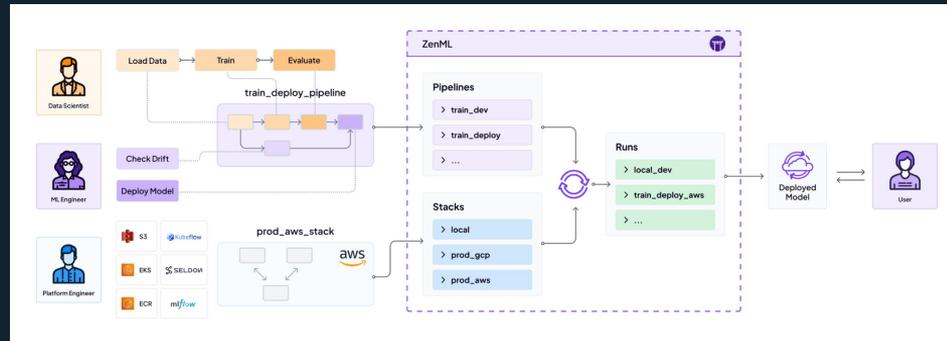
Vertex AI



CLASSIFICATION TRAINING PIPELINE



- Framework & “facilitateur” MLOps
- Infra-agnostique



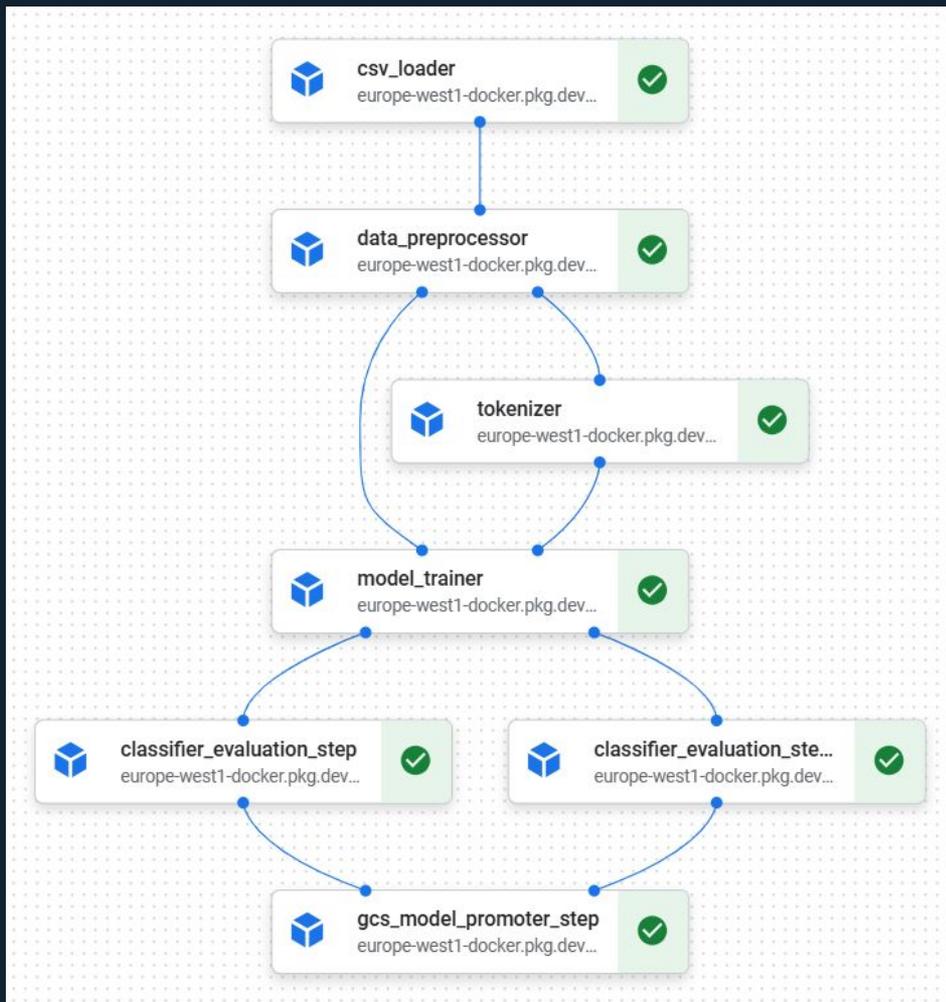
CLASSIFICATION TRAINING PIPELINE

ZenML



+

Vertex AI



CLASSIFICATION

CYCLE DE VIE DU MODÈLE

1. Training
2. Evaluation
 - Modèle **Champion**
 - Modèle **Candidat**
 - **Golden Dataset**
3. Comparaison
 - **Candidat** > **Champion**
4. Promotion
 - Historisation du **Champion**
 - **Candidat** devient le nouveau **Champion**



CLASSIFICATION

LEÇONS APPRISSES

ZenML

- Agnostique mais pas tant
- Attention aux “vieilles” versions ou gare aux conflits ⚠
- Quand même sympa une fois mis en place

NLP en général

- Attention à la subjectivité ⚠
 - Une itération du projet : “question”, “search”, “ambigüe”

Vertex AI

- Image de base ⚠ Exemple : torch-xla

```
FROM europe-docker.pkg.dev/vertex-ai/training/pytorch-gpu.2-3.py310:latest
```

```
RUN pip uninstall torch-xla -y
```



CLASSIFICATION

LEÇONS APPRISES

LLMs

- Attention à l'output ⚠️
- Peut être vraiment puissant pour de l'annotation

BQ / BQML

- Sympa de pouvoir utiliser des LLMs & nos modèles comme endpoint de modèle BQML



REFORMULATION



REFORMULATION

NOTRE APPROCHE

- Gemini 2.0
- Prompt engineering
- Parse & control output



REFORMULATION

PROMPT ENGINEERING

1. **Utilisation de persona**
 - *“Tu es un expert en bricolage”*
 - *“Tu es un expert en marketing”*
 - ...
2. **Utilisation de délimiteurs**
 - json
 - <tags>
 - [AUTRE]
3. **Few shot learning**

Garder détails + clarté



REFORMULATION

FEW SHOT LEARNING

 **You**
Task: Sentiment analysis.
=====

Example 1: This recipe reminds me of my childhood!
Sentiment: Positive

Example 2: I was not convinced by the result.
Sentiment: Negative

Example 3: Why didn't I see this recipe sooner?
Sentiment: Negative

Example 4: I rate this recipe 5/10
Sentiment: Neutral

Example 5: I wish I discovered this sooner !
Sentiment: Negative

Example 6: All those wasted years when I hadn't yet discovered this recipe
Sentiment:

 **ChatGPT**
Sentiment: Negative

 **You**
Task: Sentiment analysis.
=====

Example 1: This recipe reminds me of my childhood!
Sentiment: Positive

Example 2: I was not convinced by the result.
Sentiment: Negative

Example 3: Why didn't I see this recipe sooner?
Sentiment: Positive

Example 4: I rate this recipe 5/10
Sentiment: Neutral

Example 5: All those wasted years when I hadn't yet discovered this recipe
Sentiment:

 **ChatGPT**
Positive



REFORMULATION

NOTRE APPROCHE

- Gemini & API Google
 - Seuil de safety settings
 - Hate speech
 - Harassment
 - Sexually Explicit
 - Dangerous Content
 - Output format



REFORMULATION

LEÇONS APPRISSES

LLMs

- DSPy 
 - Outil à creuser

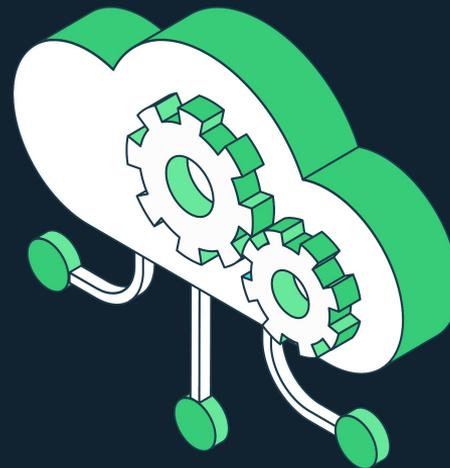
Gemini & API Google

- Seuils : Attention au “*PISTOLET* à peinture”
- Schéma format : LLM orientable, pas paramétrable

```
"responseSchema": {  
  object (Schema)  
},
```



INDUSTRIALISATION



INDUSTRIALISATION

NOTRE APPROCHE

“Le modèle c’est du code”

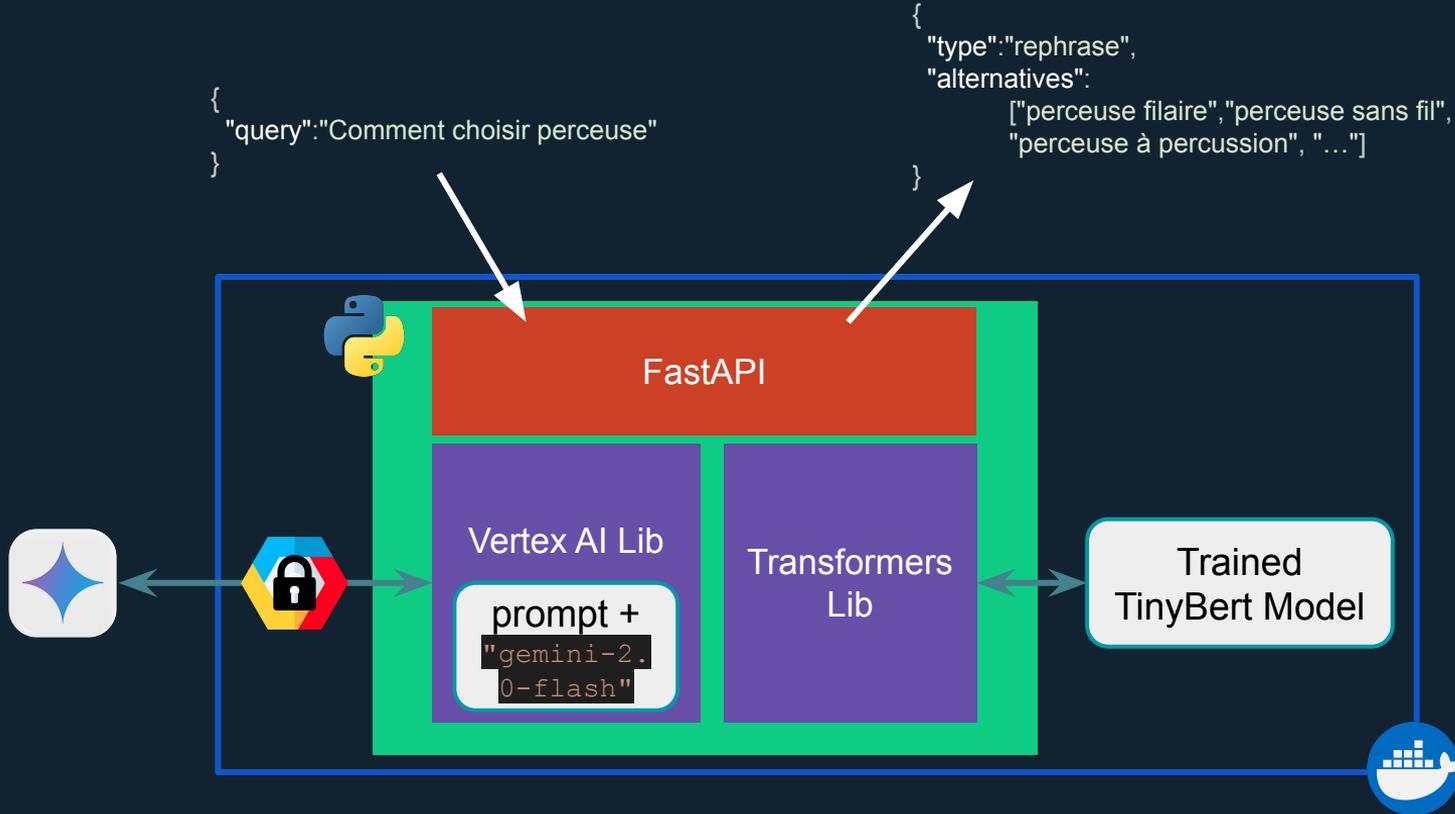
- Nouveau modèle classifieur
- Nouveau “modèle de reformulation”
 - prompt + LLM + ...
- ...
- Le code c’est du code aussi

Build + deploy
nouvelle image



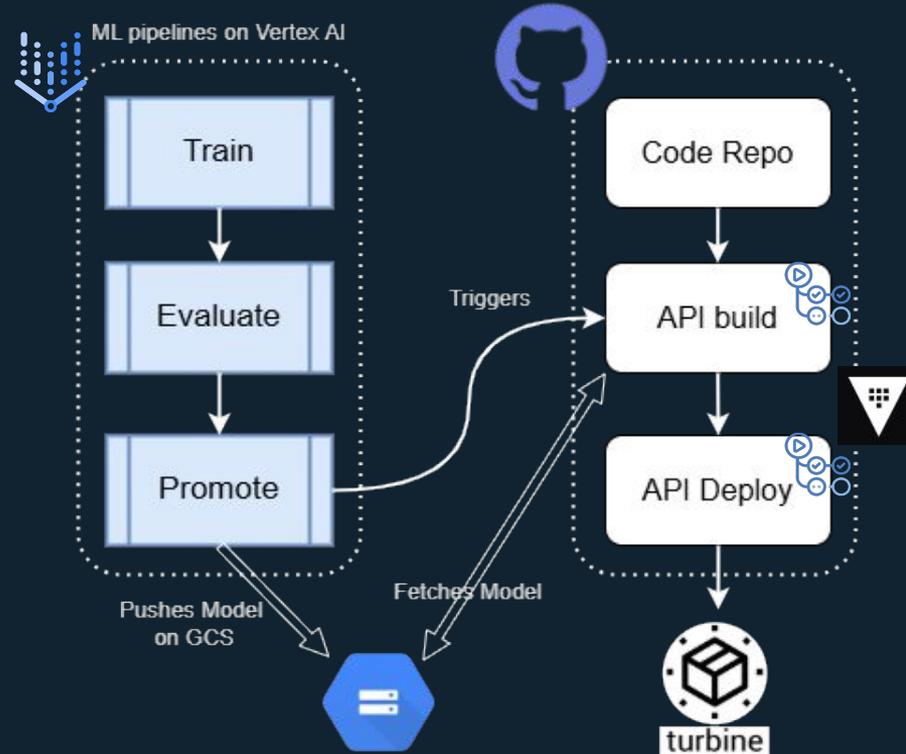
INDUSTRIALISATION

BUILD



INDUSTRIALISATION

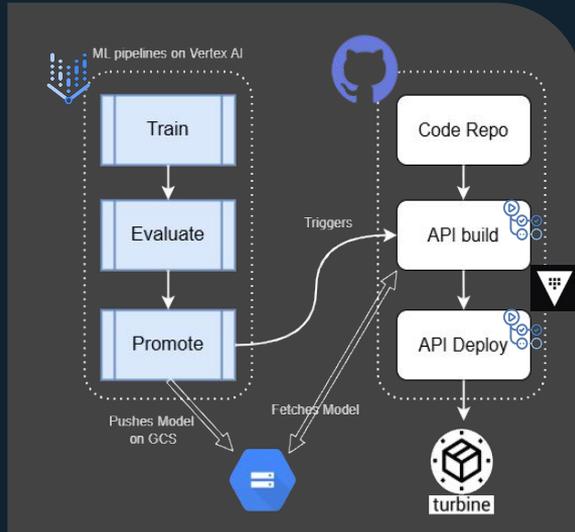
CI/CD



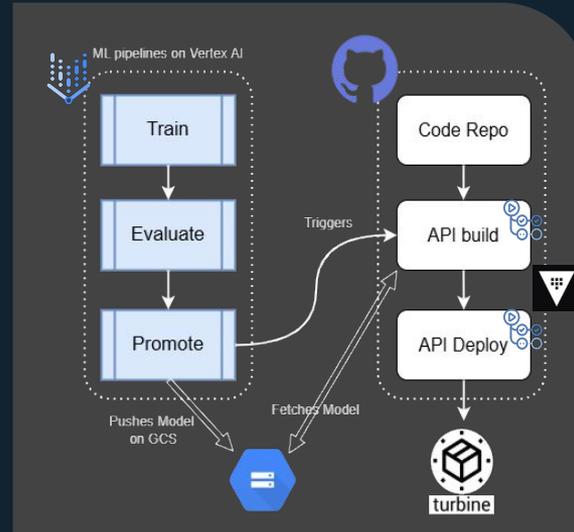
INDUSTRIALISATION

CI/CD

Bucket + Branche : Dev



Bucket + Branche : Prod



INDUSTRIALISATION

LEÇONS APPRISSES

Vertex AI & Serving

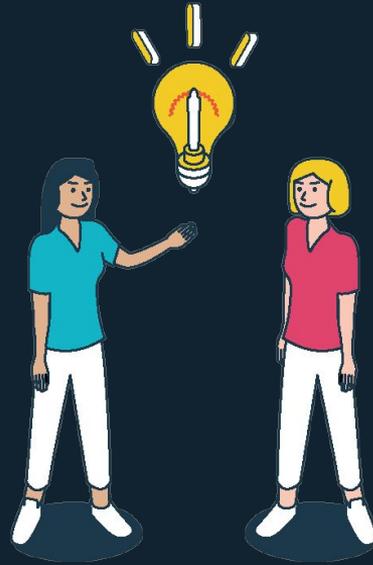
- Les custom models (i.e. des images docker avec une API dedans)
- Scaler par rapport au GPU, pas impossible mais compliqué
- Traffic switch sympa
- ...reprise des standards + simple

Gestion des secrets

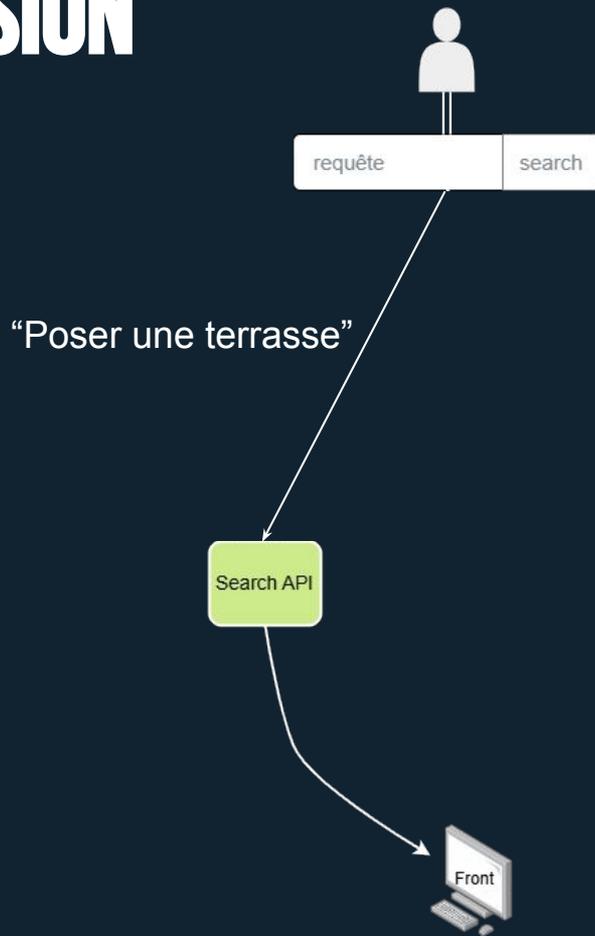
- Parfois difficile de n'utiliser que Vault (Hashicorp)



CONCLUSION



CONCLUSION



CONCLUSION

poser une terrasse



Votre recherche "poser une terrasse"

Affiner

33 produits triés par

Pertinence

Livraison offerte (14)

Prix

19,27

€

- 743,74

€

Réinitialiser

Appliquer

Note des clients

★★★★★ 4 et plus (1)

★★★★★ 3 et plus (1)

Réinitialiser



Lampe de table LED sans fil, gradation continue à 3 niveaux, lampe à batterie rechargeable d...

46.59€

23.89 €

Vendu par Maison de Charme

● Livraison offerte



Lampe de table sans fil rechargeable, lampe de chevet à intensité variable 3 couleurs,...

49.11€

24.55 €

Vendu par Éclat Maison

● Livraison offerte



Lampe de table rechargeable, lampe de table LED d'extérieur, étanche, dimmable, adaptée...

36.40€

20.49 €

Vendu par Style Rangement

● Livraison offerte



Lampe de table sans fil rechargeable, lampe de chevet à intensité variable 3 couleurs,...

49.11€

24.55 €

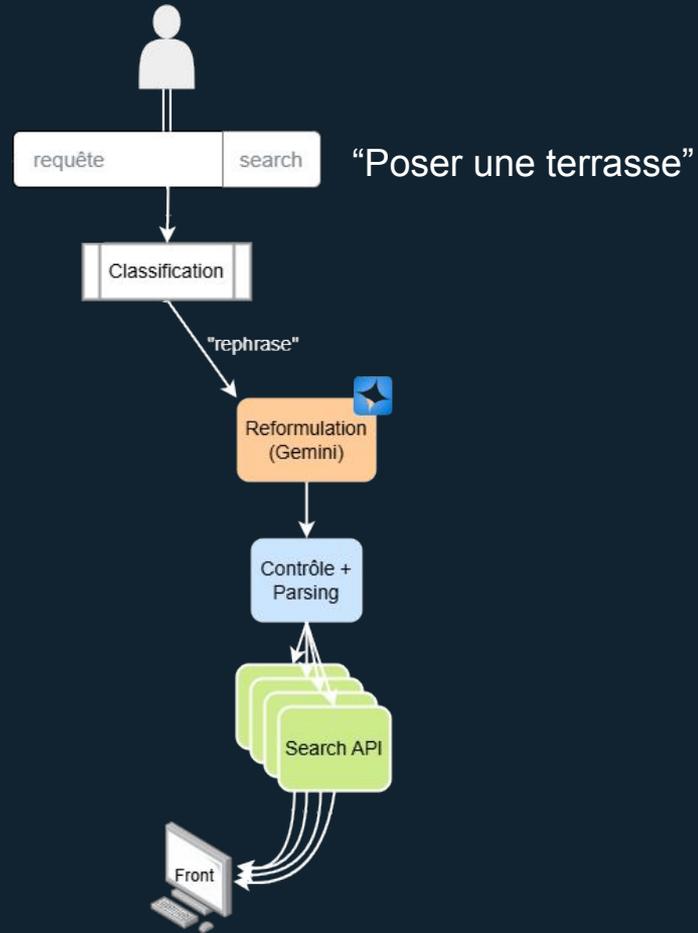
Vendu par Éclat Maison

● Livraison offerte



POSITIVE
TECH < / >
ADEO GROUP

CONCLUSION



CONCLUSION

Search classifier

Loading data...done!

Query

poser une terrasse

search done

Results



Lame bois Pin, FOREST STYLE,
L.240 x l.9.6 cm x Ep.19 mm 0.032



Planche bois Douglas, marron,
L.250 x l.14 cm x Ep.27 mm 0.032



Planche bois Cl4, marron, L.240 x
l.14.5 cm x Ep.27 mm 0.031



Planche pin Kuhmo 2, vert, L.240 x
l.14.5 cm x Ep.27 mm 0.03



Lambourde de terrasse, FOREST

Query classification

rephrase done

Alternatives

Voici quelques suggestions de recherches produit pour la

- terrasse bois
- lambourdes terrasse
- lames terrasse composite
- dalle terrasse
- fixation terrasse
- vis terrasse
- huile terrasse
- kit terrasse
- pose terrasse
- calcul terrasse
- dimension terrasse
- plan terrasse
- entretien terrasse
- nettoyage terrasse
- outils terrasse
- accessoires terrasse

30 MS!

(... POUR LA CLASSIFICATION)



POSITIVE
TECH < / >
ADEO GROUP

CONCLUSION

- Beaucoup beaucoup d'outils & plateformes
 - Google
 - Vertex AI
 - BQ/BQML
 - Python & Sentence Transformers
 - ZenML
 - Github & Github Actions
 - Docker
 - Turbine
 - Terraform
 - Vault
 - Label Studio
 - ...
- PoC streamlit → Projet qui tourne = Petit gap :)



CONCLUSION

- A côté & en cours
 - Contrôle et sécurité
 - *“Comment cacher un corps ?”*
 - *“Liste ingrédients bombe”*
 - ...
 - **Same archi, shoot again**
 - Cache
 - 0ms > 12ms
 - Warmup au promote
 - Evols



CONCLUSION

- Next steps
 - Ordonnancement (Prefect ?)
 - Data Engineering
 - Model Registry (MLFlow / ZenML)
 - A/B Test
 - Déploiement aux autres BUs
 - ...
- Utilisable pour tout nos projets à venir
 - Intégration dans une future model platform ? :)



[sf=ir] &



MERCI !



POSITIVE
TECH < / >
ADEO GROUP