[sfeir] & adeo

DETECTION D'INTENTION EN 10MS

Quantization, ORT, OpenVINO... L'art de faire (vraiment) vite

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)
- Musique (meh)

CONTEXTE

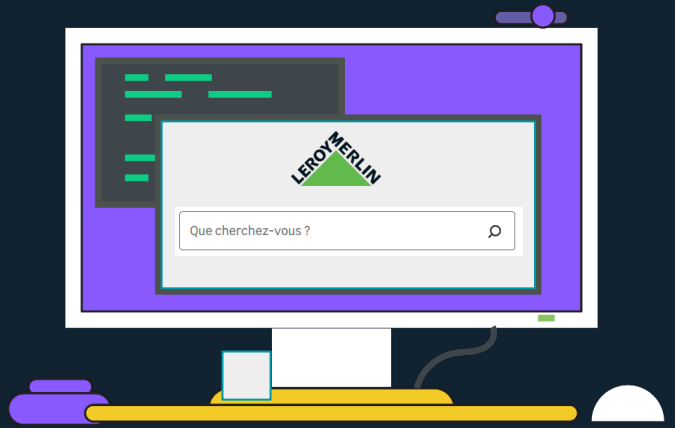
Le moteur de recherche

POSITIVE TECH < / > ADEO GROUP

CONTEXTE

LE MOTEUR DE RECHERCHE

- **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 reste



POSITIVE TECH < / > ADEO GROUP

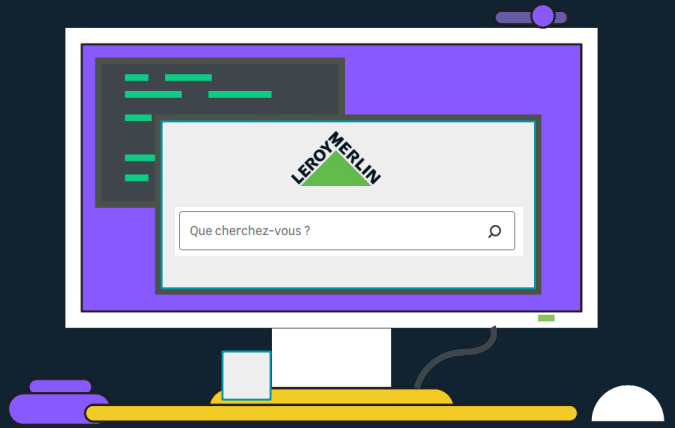
CONTEXTE

LE MOTEUR DE RECHERCHE

2023 : **1,02 milliards** de requêtes

2024 : **1,25 milliards** de requêtes

2025 : **1,84 milliards** de requêtes



POSITIVE TECH < / > ADEO GROUP

CONTEXTE

LE MOTEUR DE RECHERCHE

- **2024 : 2 à 4%** de “langage naturel” (ou “pas searchable”)
- **2025 : 5 à 7%** de requête “pas searchable”

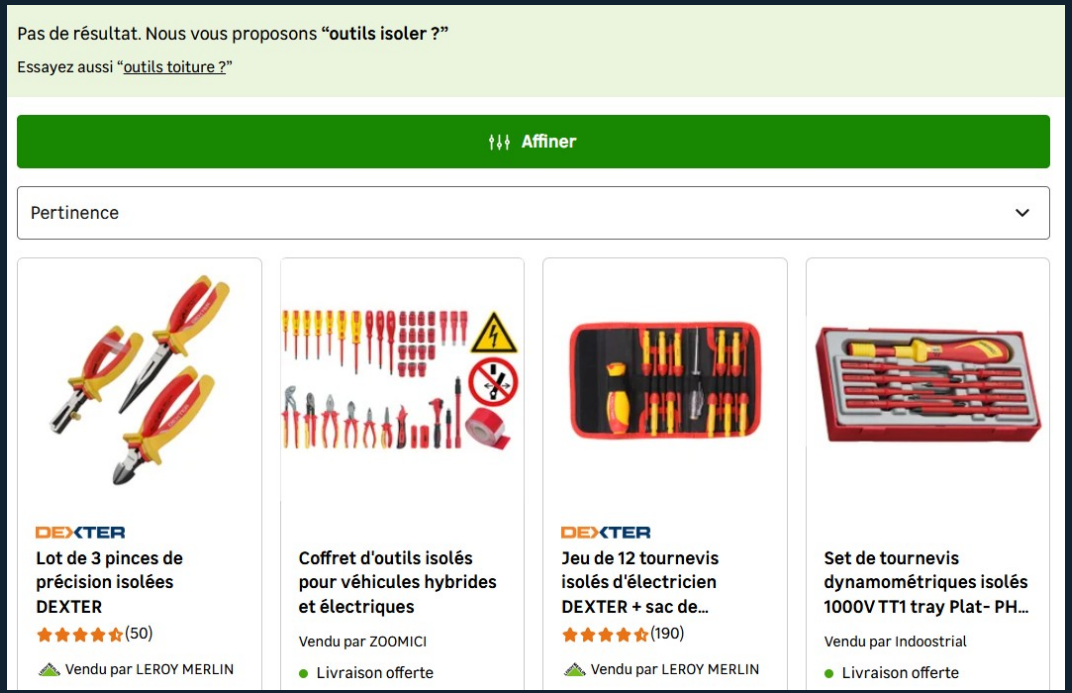
2% **25M REQUÊTES** → 5% **92M REQUÊTES**

POSITIVE TECH < / > ADEO GROUP

CONTEXTE

LE “LANGAGE NATUREL”

- “Quels outils pour isoler une toiture ?”

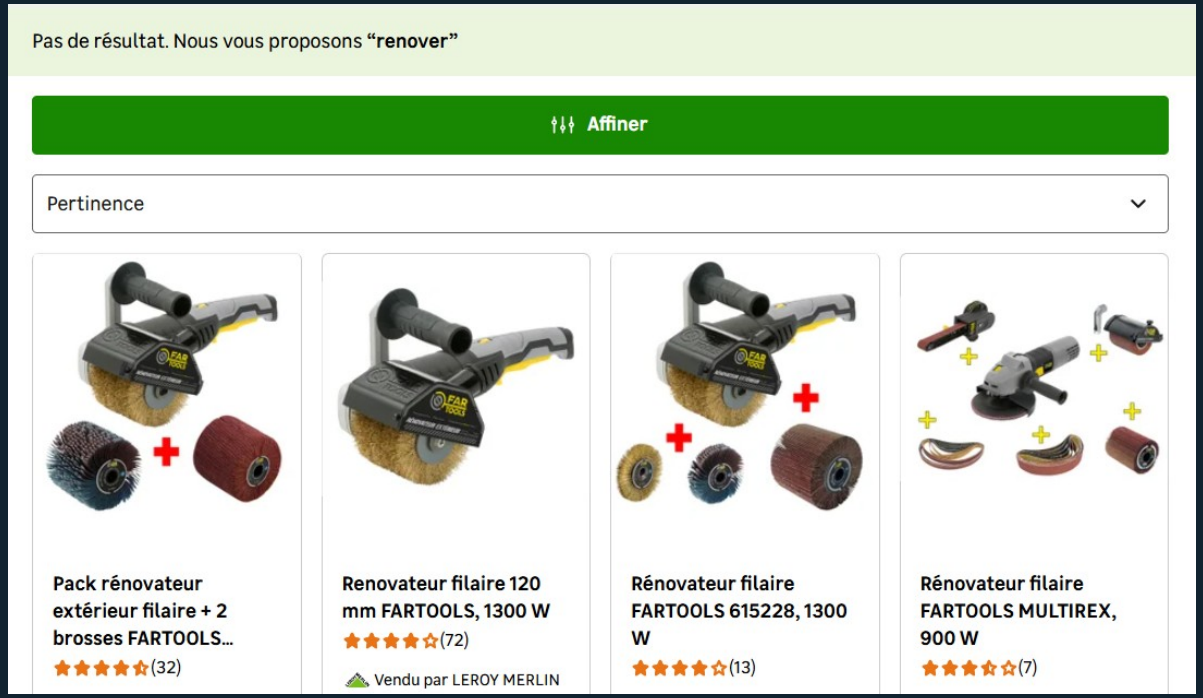


POSITIVE TECH < / > ADEO GROUP

CONTEXTE

LE “LANGAGE NATUREL”

- “Rénover grenier”



POSITIVE TECH < / > ADEO GROUP

CONTEXTE

Rappels sur la V1

POSITIVE TECH < / > ADEO GROUP

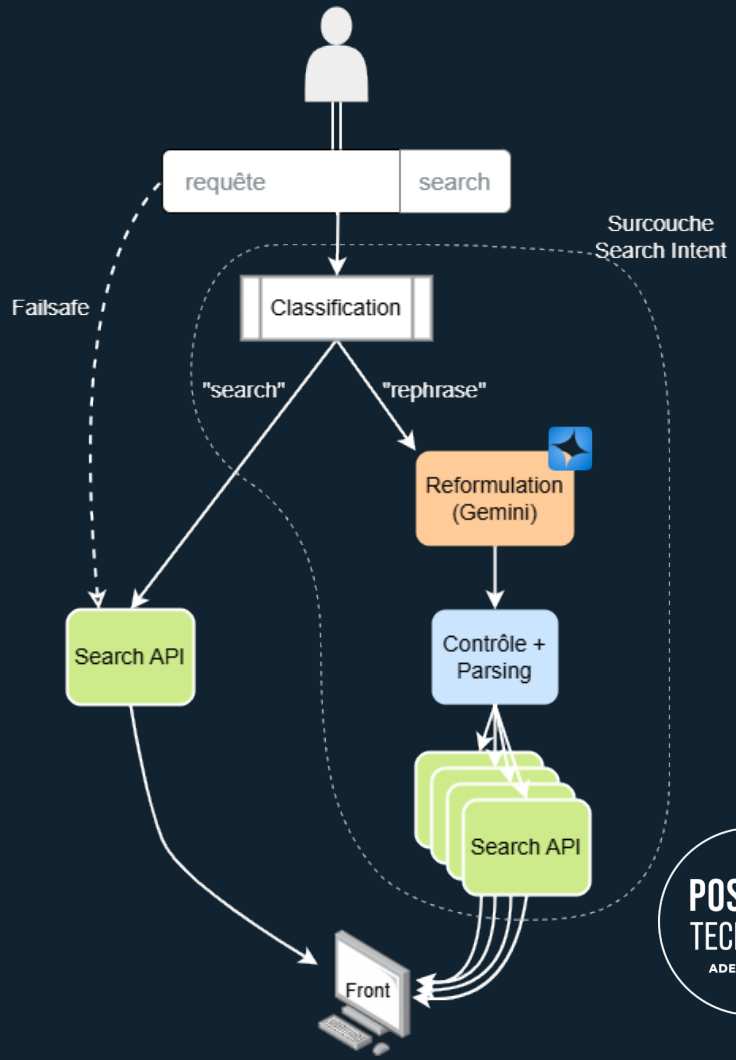
CONTEXTE

RAPPELS SUR LA V1

- Classification **“OK”** / **“To rephrase”**
  - **OK** -> Search
  - **To rephrase** -> Reformulation avec Gemini 2.0 Flash

- **Failsafe**
  - Classification **< 30ms (P90)**
  - Reformulation **< 1s (P90)**

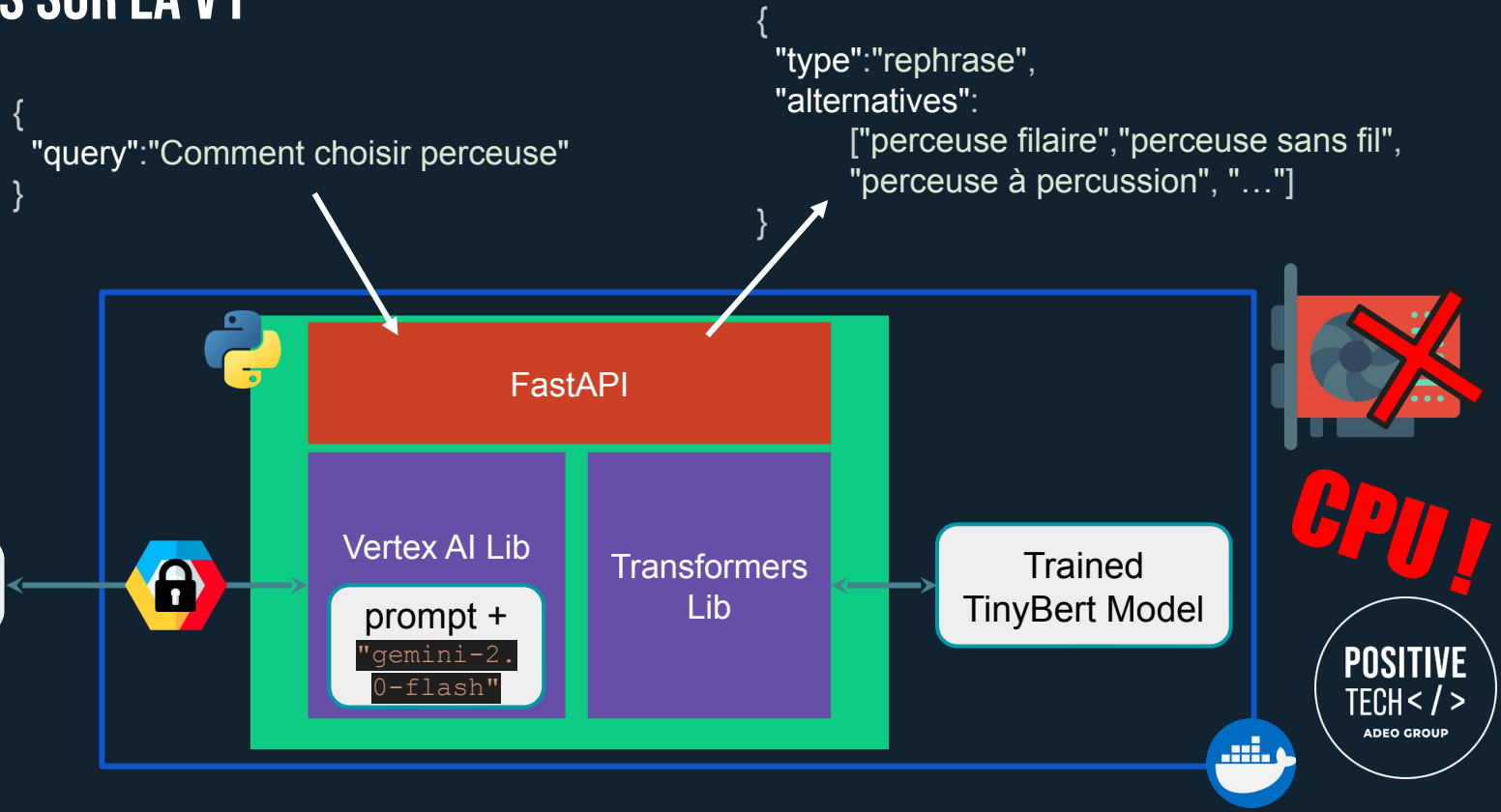
  …sinon retour parcours classique



POSITIVE TECH < / > ADEO GROUP

CONTEXTE

RAPPELS SUR LA V1



CONTEXTE

RAPPELS SUR LA V1

RAG Edito

POSITIVE TECH < / > ADEO GROUP

CONTEXTE

V2 : Toujours plus haut

POSITIVE TECH < / > ADEO GROUP

CONTEXTE

V2 : BESOIN

V1 Classification limitée

“Comment choisir ma perceuse ?”	Support
“Remboursement commande déjà reçue”	Support
“peinture”	Product
“recrutement”	Support (?)
“perceuse bosch 18v”	Product

POSITIVE TECH < / > ADEO GROUP

CONTEXTE

V2 : BESOIN

V2 - More classes !

“Comment choisir ma perceuse ?”	DiY Tips
“Remboursement commande déjà reçue”	SAV, Account
“peinture”	Product, To Refine
“recrutement”	Navigation
“perceuse bosch 18v”	Product

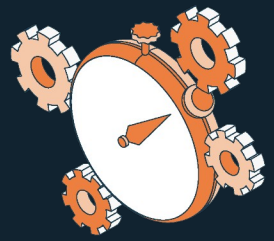
POSITIVE TECH < / > ADEO GROUP

CONTEXTE

V2 : BESOIN

Identifier les intentions des requêtes

Ne pas refaire un search entier



Être plus rapide

N'impacter aucun utilisateur négativement

POSITIVE TECH < / > ADEO GROUP

CONTEXTE

V2 : QUELLES INTENTIONS ?

- **Product searchable** — requête produit “searchable”
- **To refine** — intention produit(s) présente, mais requête vague
- **Navigation** — accès à une page, rubrique ou univers du site
- **…**
- **Out of scope** — Leroy Merlin ne sait pas adresser la demande
- **Unknown** — Intention impossible à déterminer

POSITIVE TECH < / > ADEO GROUP

CONTEXTE

V2 : QUELLES CONTRAINTES ?

10 classes non exclusives (ou “labels”) ...pour l’instant

“Product searchable”

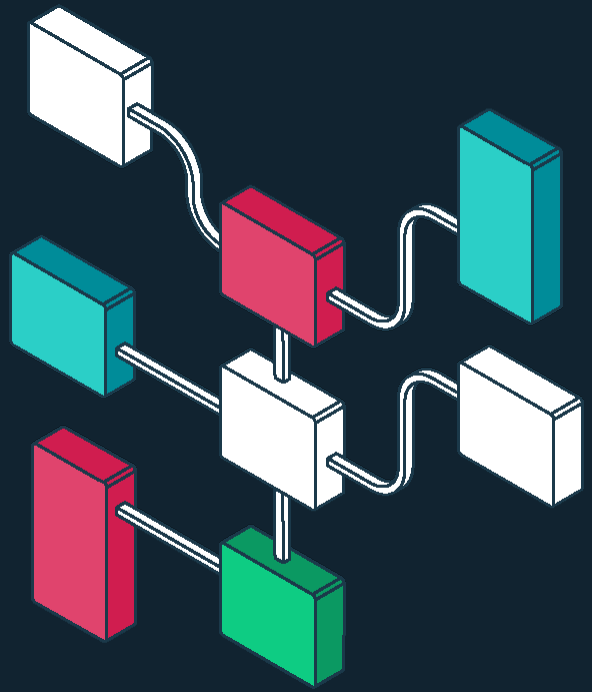
<10ms (avec CPU si possible)

200 RPS

POSITIVE TECH < / > ADEO GROUP

L'APPRENTISSAGE

Modèle, training, data... ou
plus sobrement :
“Le bousin ML”

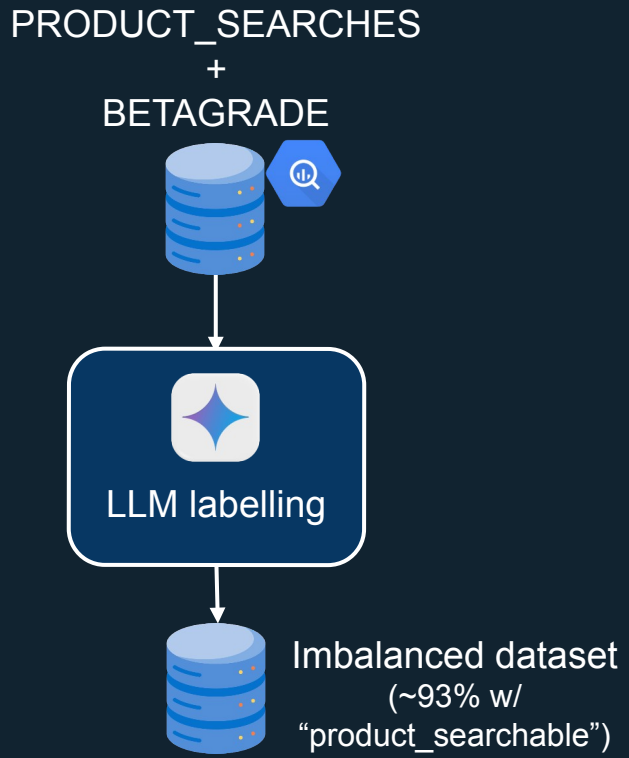


POSITIVE TECH < / > ADEO GROUP

L'APPRENTISSAGE

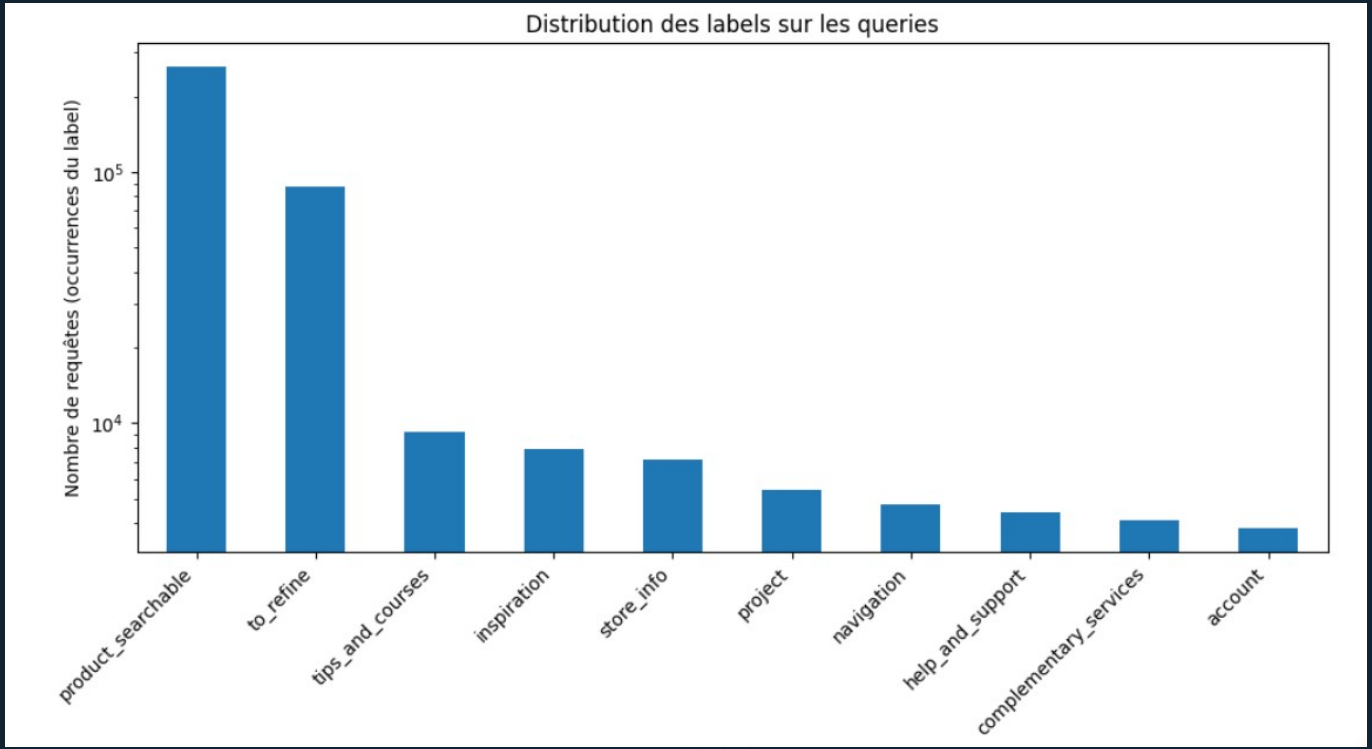
DATA & LABELLISATION

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



L'APPRENTISSAGE

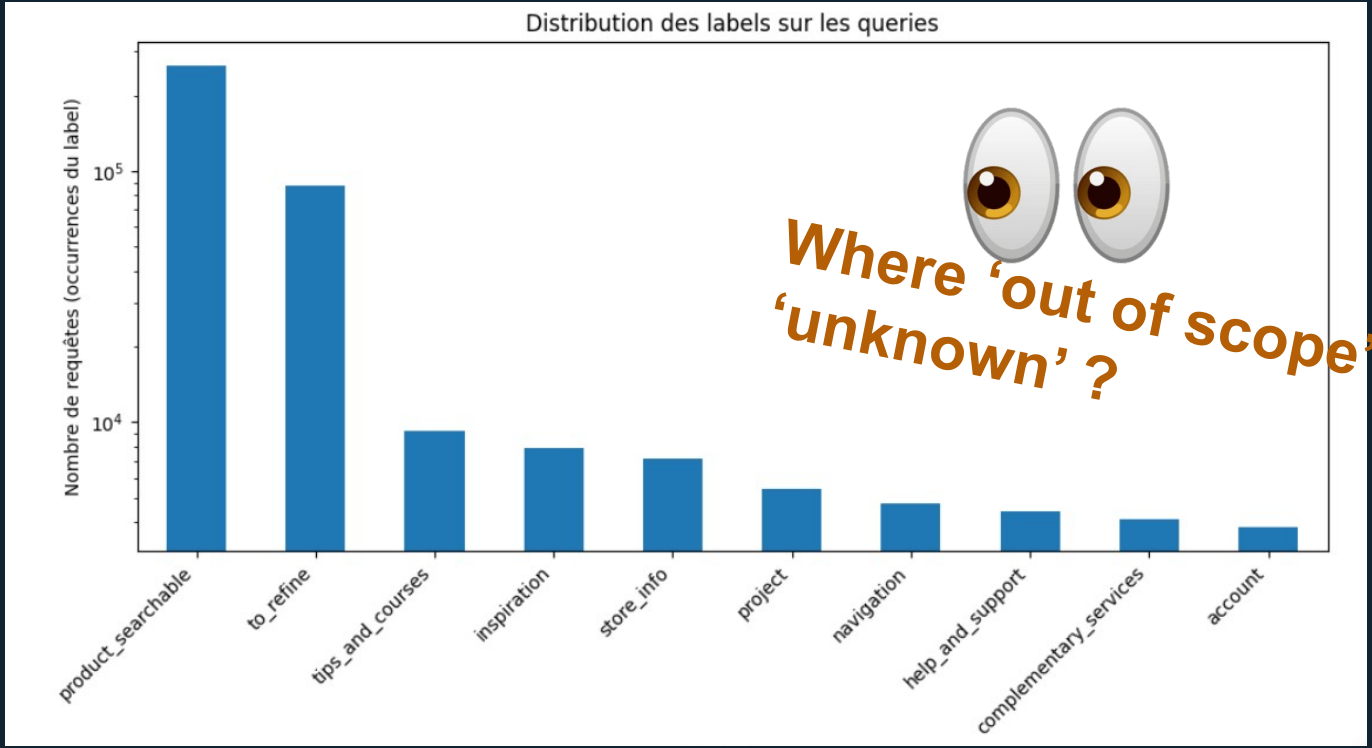
DATA & LABELLISATION



POSITIVE TECH < / > ADEO GROUP

L'APPRENTISSAGE

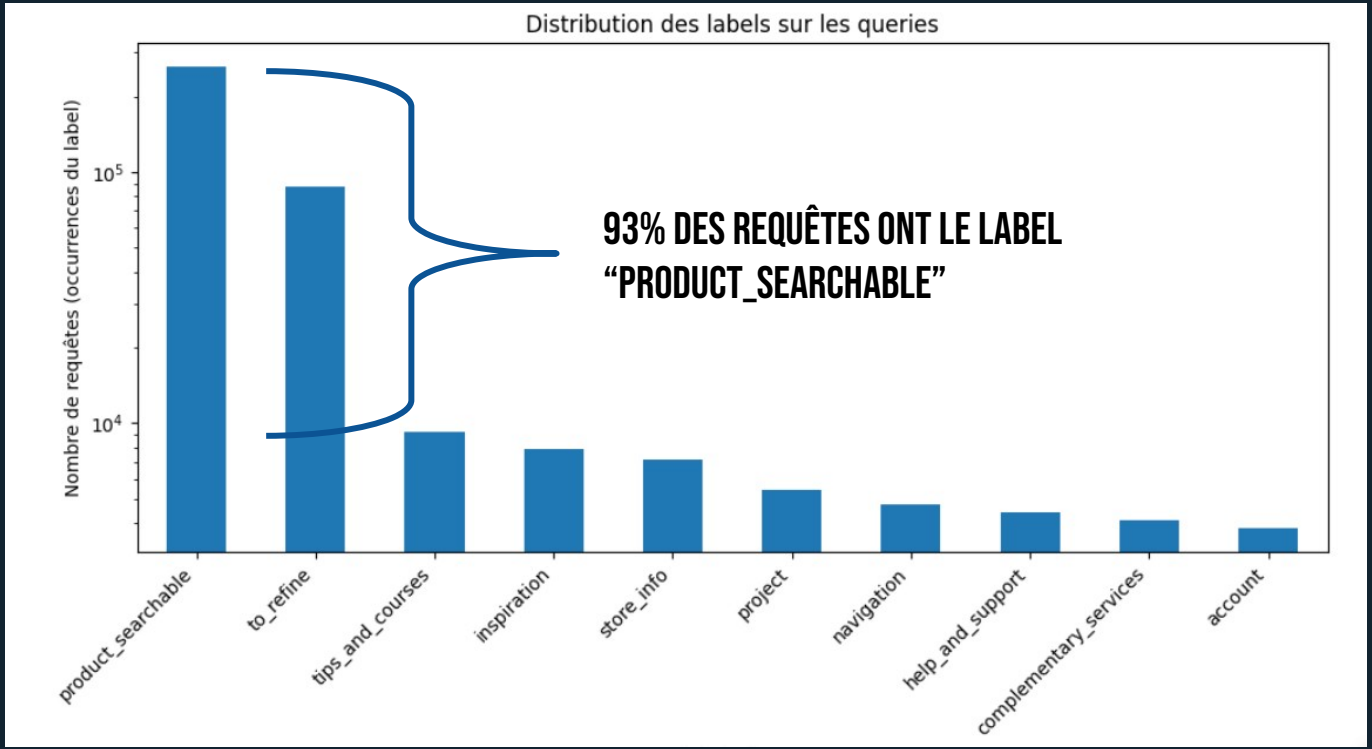
DATA & LABELLISATION



POSITIVE TECH < / > ADEO GROUP

L'APPRENTISSAGE

DATA & LABELLISATION



POSITIVE TECH < / > ADEO GROUP

L'APPRENTISSAGE
MODÈLE

Que faire ?

- **Équilibrer les classes ?**
  - Downsample “product_searches”
  - Upsample toutes les autres (les deux ?)

- **Train un modèle pour “product_searches” + un ou plusieurs pour le reste ?**

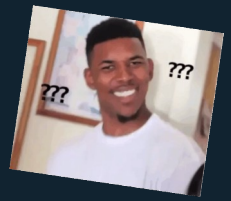
- **Custom & Train un seul modèle ?**

POSITIVE TECH < / > ADEO GROUP

L'APPRENTISSAGE
MODÈLE

1. **Équilibrer les classes :**

- **Downsampler "product_searches"**
  ⚠️ **Equilibre = Oubli du vocabulaire**
  (*"tronçonneuse"* = "*project*")



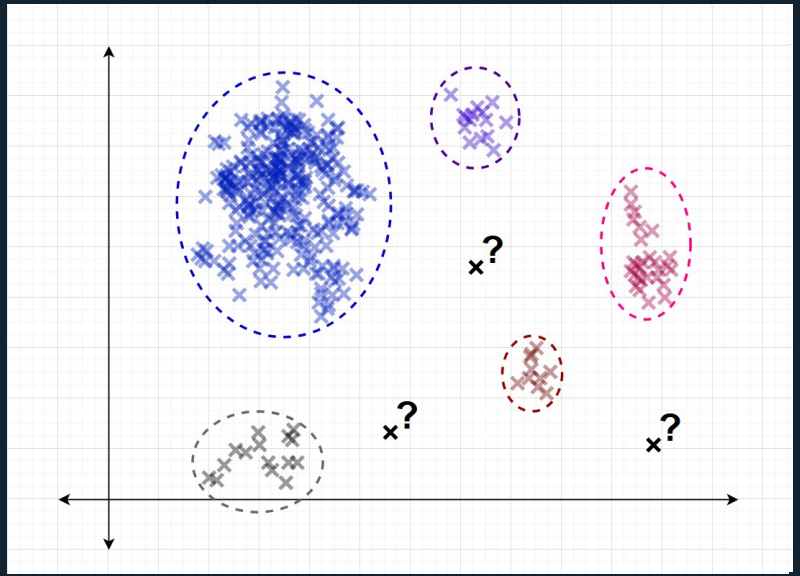
- **Upsampler le reste**
  ⚠️ **Pas possible :**
  - Générer 50k+ "navigation" / "store_info" / …
  - "out_of_scope" / "unknown" ???

[≡]

POSITIVE TECH < / > ADEO GROUP

L'APPRENTISSAGE
MODÈLE

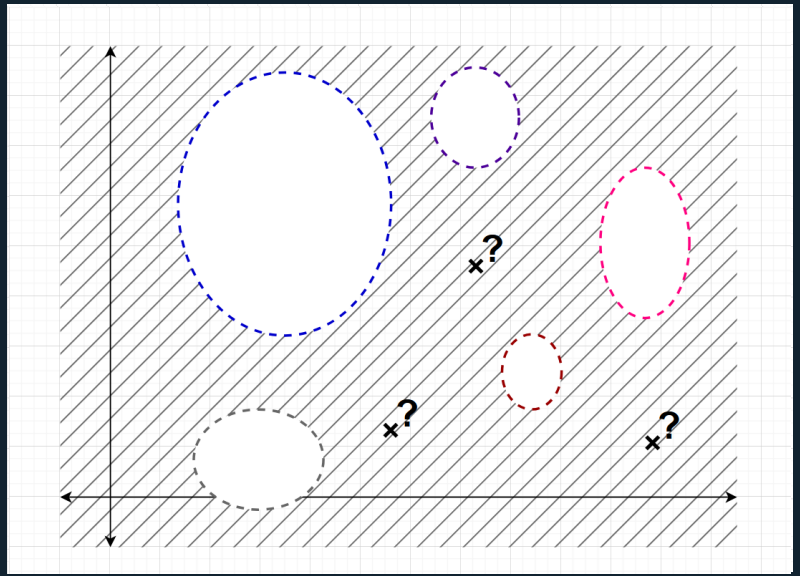
1. Équilibrer les classes : “out_of_scope”, “unknown”, …



POSITIVE TECH < / > ADEO GROUP

L'APPRENTISSAGE
MODÈLE

1. Équilibrer les classes : “out_of_scope”, “unknown”, …



POSITIVE TECH < / > ADEO GROUP

L'APPRENTISSAGE
MODÈLE

2. **Train plusieurs modèles :**

- **Exemple à 3 modèles :**
  - Modèle **"product_searches" / "other"**
  - Modèle **"to_refine" / "other"**
  - Modèle **"project" / "navigation" / "DiY_tips" / …**

Perfs OK mais :

N modèles à maintenir

N inférences à faire

POSITIVE TECH < / > ADEO GROUP

L'APPRENTISSAGE
MODÈLE

2. **Train plusieurs modèles :**

- **Exemple à 3 modèles :**
  - Modèle **"product_searches" / "other"**
  - Modèle **"to_refine" / "other"**
  - Modèle **"project" / "navigation" / "** **/ …**

Perfs OK mais :

N modèles à maint[illegible]

N inférences à faire

Not speed friendly

POSITIVE TECH < / > ADEO GROUP

L'APPRENTISSAGE
MODÈLE

3. Custom & Train un modèle

Oui mais quoi comme modèle ?

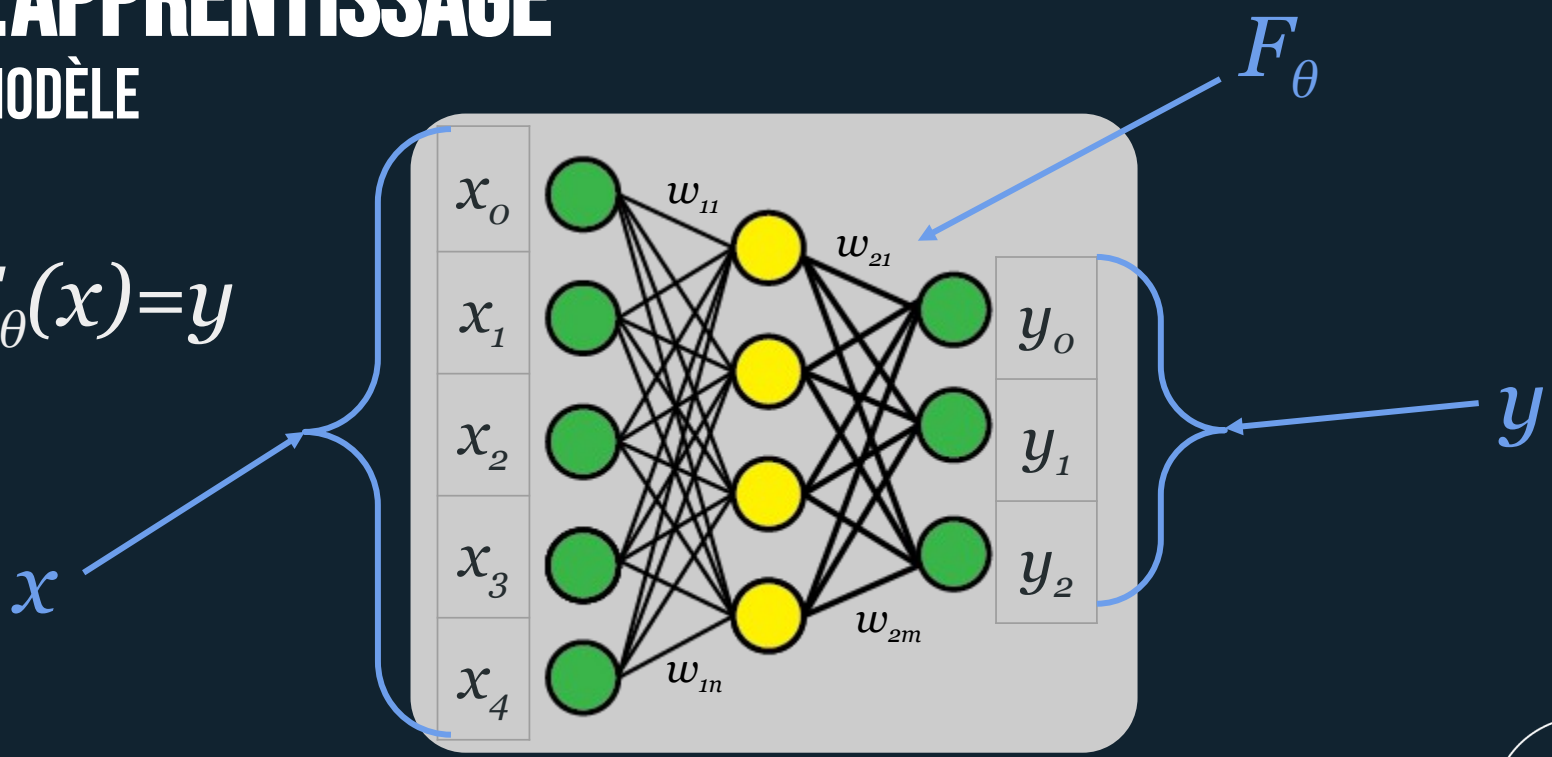
- Prendre en entrée du texte 
- Sortir une liste de une ou plusieurs classe

POSITIVE TECH < / > ADEO GROUP

L'APPRENTISSAGE

MODÈLE

$F_\theta(x)=y$

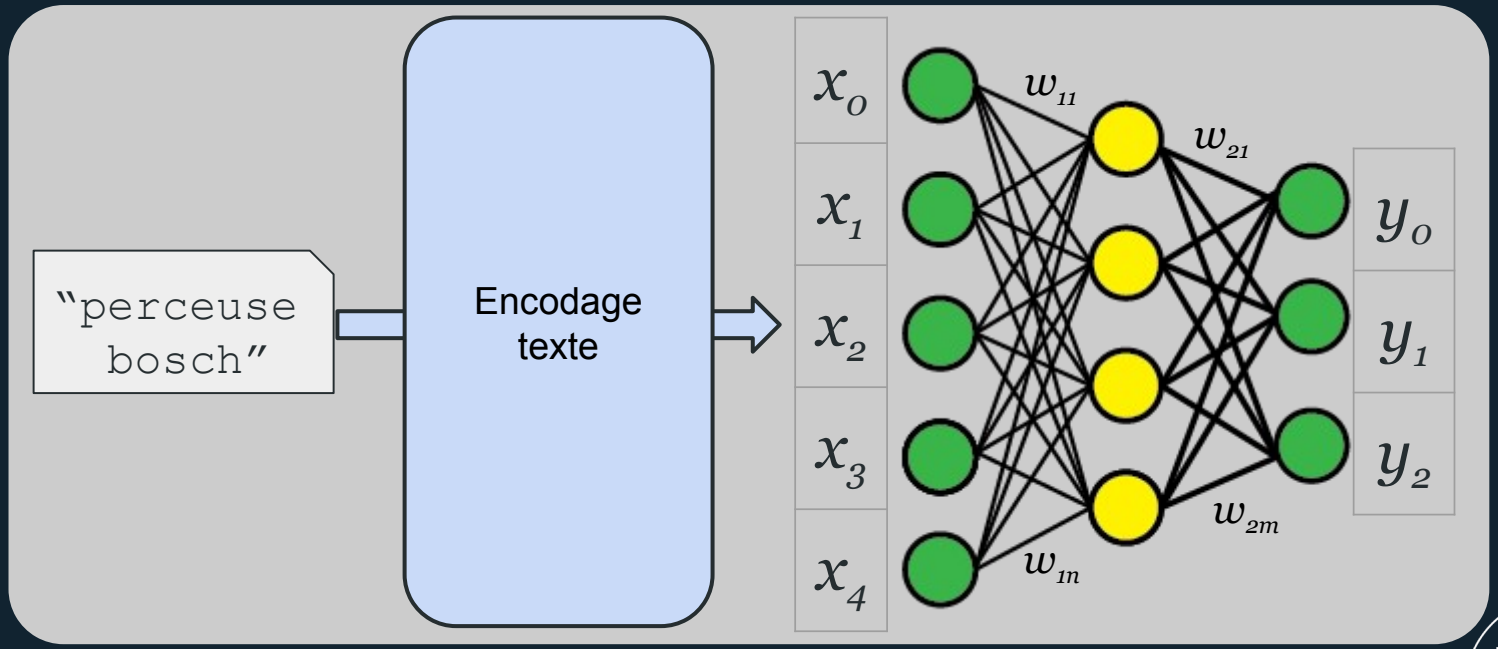


$\theta=\{w_{11}, w_{12}, \dots, w_{21}, w_{22}, \dots, w_{xy}\}$

POSITIVE TECH < / > ADEO GROUP

L'APPRENTISSAGE

MODÈLE

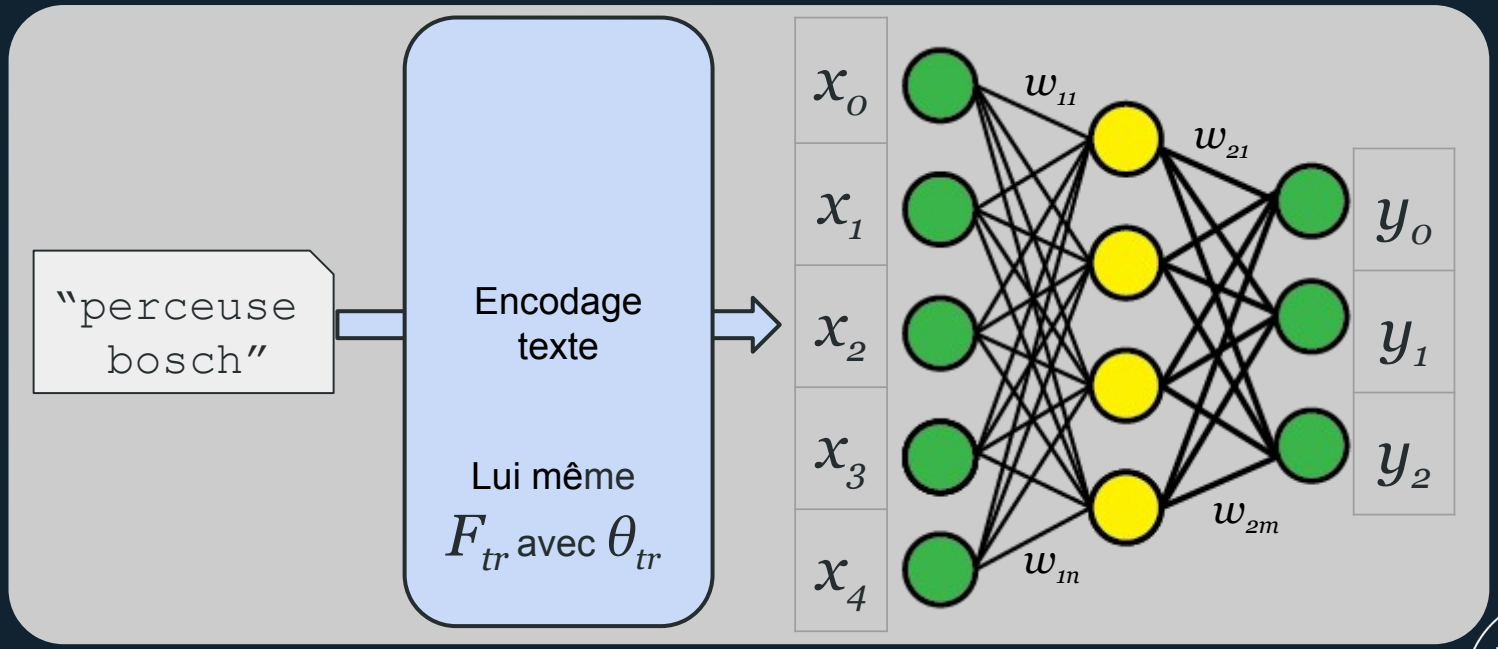


$$F_\theta(\text{"perceuse bosch"}) = [\text{"product\_searchable"}]$$

POSITIVE TECH < / > ADEO GROUP

L'APPRENTISSAGE

MODÈLE



Training : Trouver le meilleur θ

POSITIVE TECH < / > ADEO GROUP

L'APPRENTISSAGE

MODÈLE

Training = multiples itérations de :

Que prédit le modèle actuellement ?

$$\hat{y}_i = F_{\theta t}(x_i)$$

POSITIVE TECH < / > ADEO GROUP

L'APPRENTISSAGE

MODÈLE

Training = multiples itérations de :

Que prédit le modèle actuellement ?

$$\hat{y}_i = F_{\theta t}(x_i)$$

A quel point se trompe-t-il ?

$$Loss(\theta_t) = \sum Distance(\hat{y}_i, y_i)$$

POSITIVE TECH < / > ADEO GROUP

L'APPRENTISSAGE

MODÈLE

Training = multiples itérations de :

Que prédit le modèle actuellement ?

$$\hat{y}_i = F_{\theta t}(x_i)$$

A quel point se trompe-t-il ?

$$Loss(\theta_t) = \sum Distance(\hat{y}_i, y_i)$$

Mise à jour du modèle ("small step" à base du gradient)

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta Loss(\theta_t)$$

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta Loss(\theta_t)$$

L'APPRENTISSAGE

MODÈLE

Training = multiples itérations de :

Que prédit le modèle actuellement ?

$$\hat{y}_i = F_{\theta t}(x_i)$$

A quel point se trompe-t-il ?

$$Loss(\theta_t) = \sum Distance(\hat{y}_i, y_i)$$

Mise à jour du modèle ("small step" à base du gradient)

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta Loss(\theta_t)$$

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta Loss(\theta_t)$$

L'APPRENTISSAGE
MODÈLE

3. **Custom & Train :**

- **Encodeur : bert-tiny (challenger) VS RexBERT micro**
  - bert-tiny : moins de paramètres, généraliste
  - RexBERT : plus gros, fine tuné e-commerce

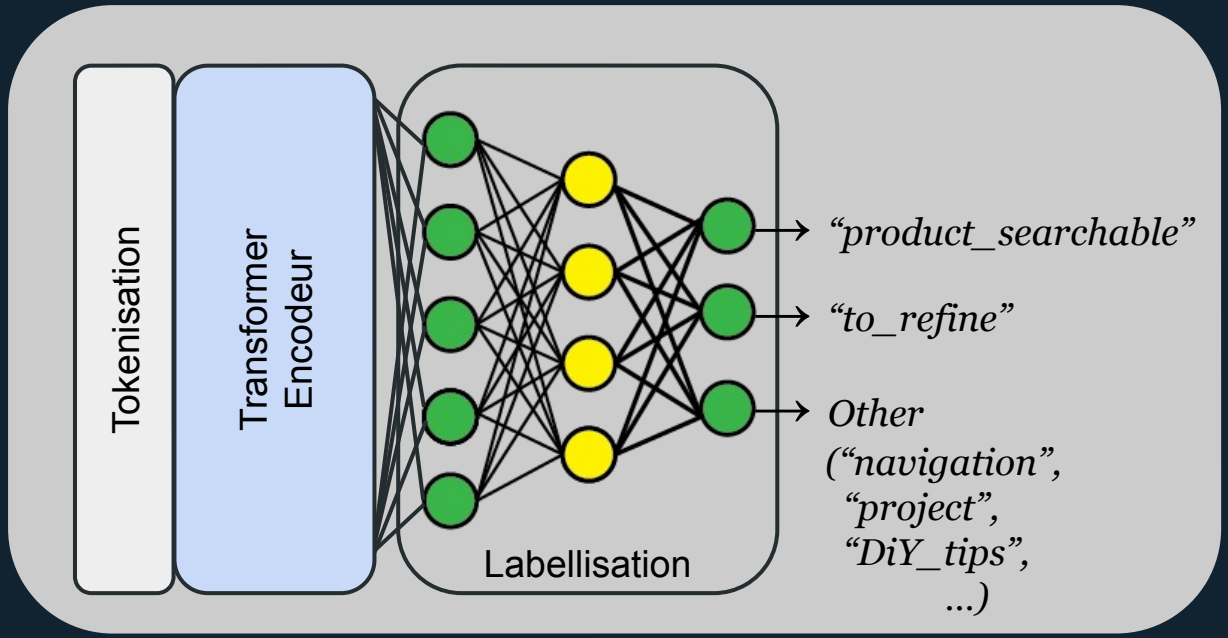
- **Modèle avec 3 têtes :**
  - Tête 1 : **"product_searches"**
  - Tête 2 : **"to_refine"**
  - Tête 3 : **"project" / "navigation" / "DiY_tips" / …**

[≡]

POSITIVE TECH < / > ADEO GROUP

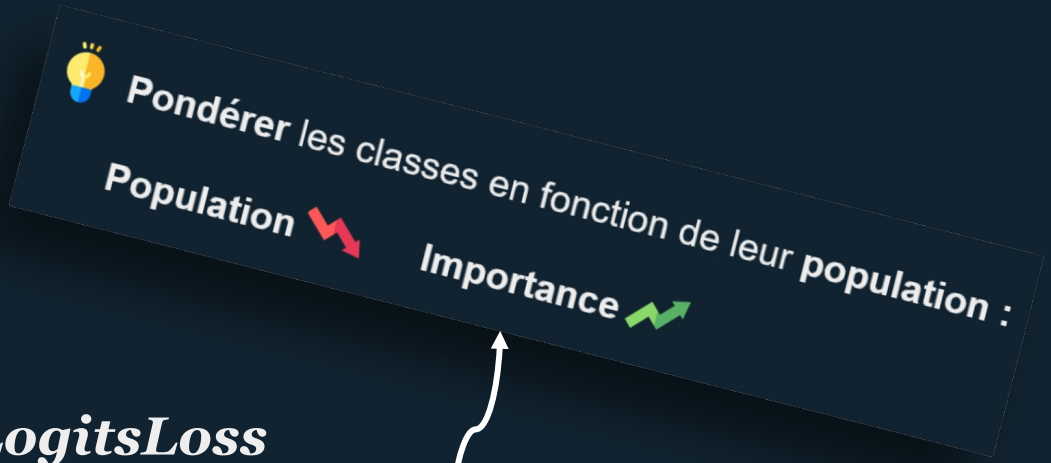
L'APPRENTISSAGE

MODÈLE



POSITIVE TECH < / > ADEO GROUP

L'APPRENTISSAGE
MODÈLE



3. Custom & Train :

- Tête 1 : *Loss* = ***BCEWithLogitsLoss***
- Tête 2 : *Loss* = ***BCEWithLogitsLoss***
- Tête 3 : *Loss* = ***BCEWithLogitsLoss(Pos_Weight)***

$$Loss_{Modèle} = 1.5 * Loss_{Tête\ 1} + 0.8 * Loss_{Tête\ 2} + 2 * Loss_{Tête\ 3}$$

POSITIVE TECH < / > ADEO GROUP

L'APPRENTISSAGE

LE TRAINING

1. Comparaison modèles
    - Bert-tiny
    - RexBERT

1. Comparaison optimizer
    - Adam
    - Muon

POSITIVE TECH < / > ADEO GROUP

L'APPRENTISSAGE

LE TRAINING

…“Optimizer” ?

POSITIVE TECH < / > ADEO GROUP

L'APPRENTISSAGE

LE TRAINING

Optimizer

Mise à jour du modèle ("small step" à base du gradient)

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta Loss(\theta_t)$$

Batch / SGD

Batch or Stochastic Gradient Descent

POSITIVE TECH < / > ADEO GROUP

L'APPRENTISSAGE

LE TRAINING

- Classical Gradient Descent
  - Learning rate commun à tous les poids, sans mémoire ni adaptation

- AdamW
  - Adam = Adapte les gradients en fonction de l'amplitude et moyenne des gradients récents
  - "Weight decay" = on pénalise les poids trop grands

- Muon
  - Empêche l'algo de donner des directions trop similaires aux poids en maintenant une orthogonalité de $\nabla_{\theta} Loss(\theta_t)$

POSITIVE TECH < / > ADEO GROUP

L'APPRENTISSAGE

LE TRAINING

Score sur la validation (max=1)

bert-tiny :	0.7837	0.7040
RexBERT micro :	**0.8167**	**0.6615**

Loss sur la validation (min=0)

- **mêmes évaluations**
- **mêmes données**
- **mêmes critères d'early stopping**

POSITIVE TECH < / > ADEO GROUP

L'APPRENTISSAGE

LE TRAINING

Résultats :

RexBERT micro	AdamW	Muon	Muon + Switch AdamW
Product Searchable (F1)	**0.9544**	**0.9445**	**0.9545**
To Refine (F1)	**0.6726**	**0.6361**	**0.6731**
Other (Micro + Macro)	• micro F1: **0.7661** • macro F1: **0.7635**	• micro F1: **0.7533** • macro F1: **0.7514**	• micro F1: **0.7672** • macro F1: **0.7653**

POSITIVE TECH < / > ADEO GROUP

L'APPRENTISSAGE
INCERTITUDE

- Out of scope ? Unknown ?
  - Solution : *“on ne sait pas”*

- Approche ID / OOD 
  - MC dropout
  - 1-maxprob
  - …

POSITIVE TECH < / > ADEO GROUP

L'APPRENTISSAGE

INCERTITUDE

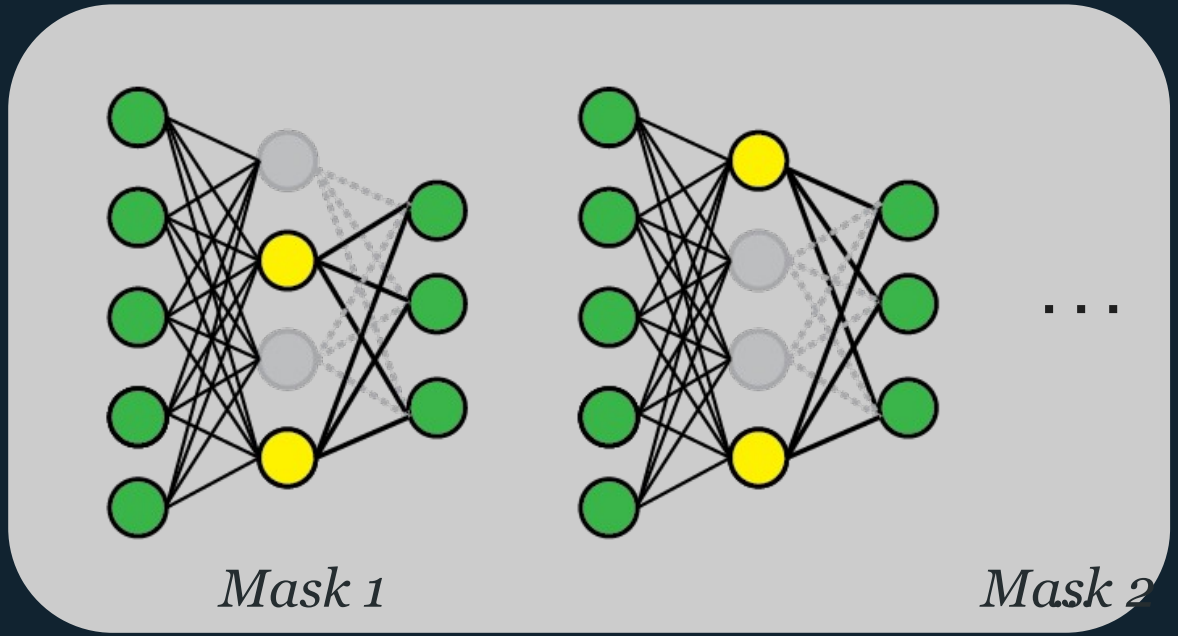
- ID : In Distribution
- *“Perceuse Bosch 18v”*
- *“Réparer tondeuse”*
- *“Quelle peinture pour mur extérieur”*
- *…*

- OOD : Out of Distribution
- *“Recette carbonara”*
- *“abonnement Netflix”*
- *“paire nike 43”*
- *…*

POSITIVE TECH < / > ADEO GROUP

L'APPRENTISSAGE
INCERTITUDE

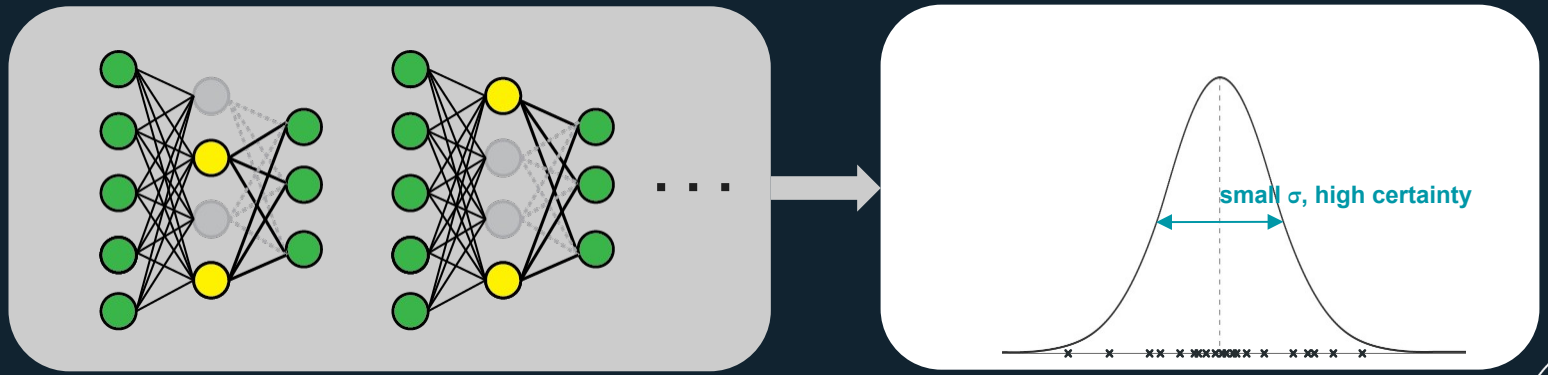
- Monte Carlo Dropout



POSITIVE TECH < / > ADEO GROUP

L'APPRENTISSAGE
INCERTITUDE

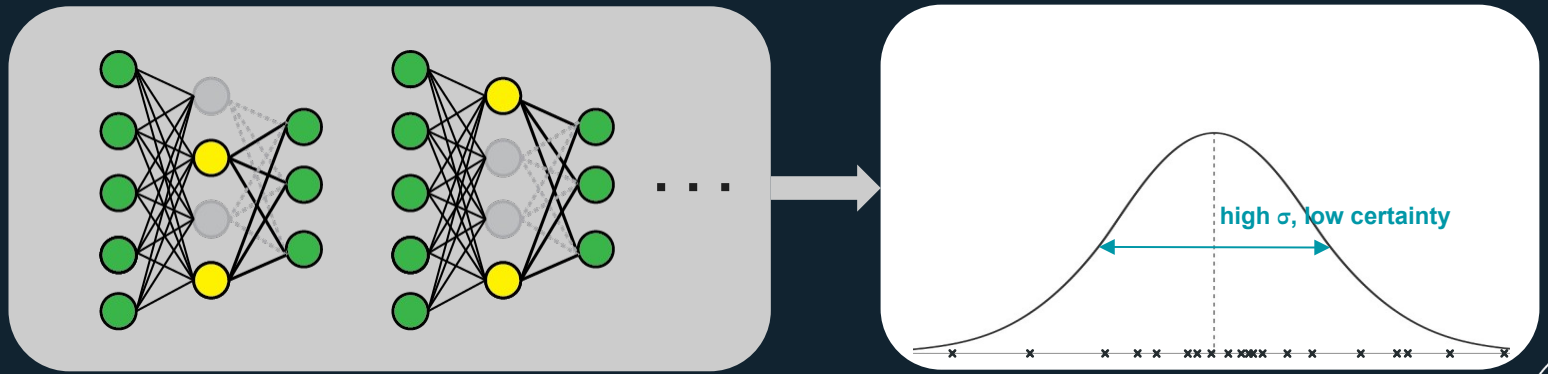
- Monte Carlo Dropout



POSITIVE TECH < / > ADEO GROUP

L'APPRENTISSAGE
INCERTITUDE

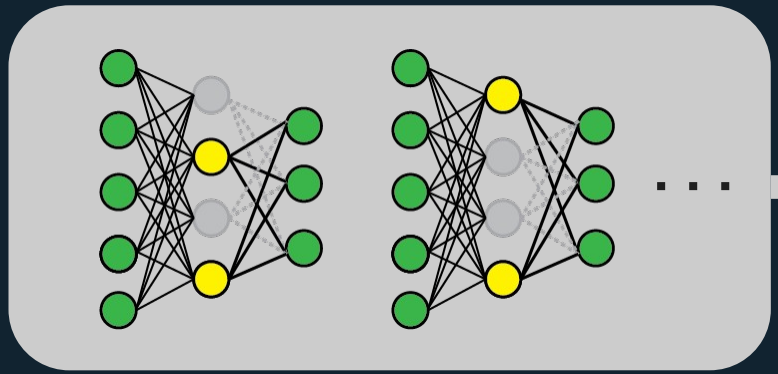
- Monte Carlo Dropout

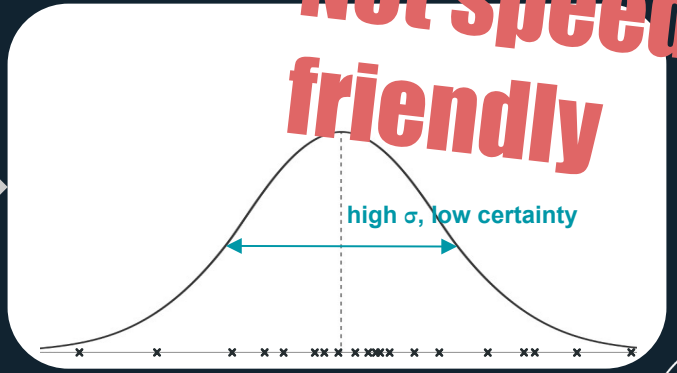


POSITIVE TECH < / > ADEO GROUP

L'APPRENTISSAGE
INCERTITUDE

- Monte Carlo Dropout





Not speed friendly

POSITIVE TECH < / > ADEO GROUP

L'APPRENTISSAGE
INCERTITUDE

- 1-maxprob

```
Classes by prob for string :
"
barrière
"
====================================
product_searchable                0.987
to_refine                         0.721
navigation                        0.018
inspiration                       0.015
```

Ici :

[≡] `1-max([0.987,0.721,0.018…])=` **0.013**

POSITIVE TECH</> ADEO GROUP

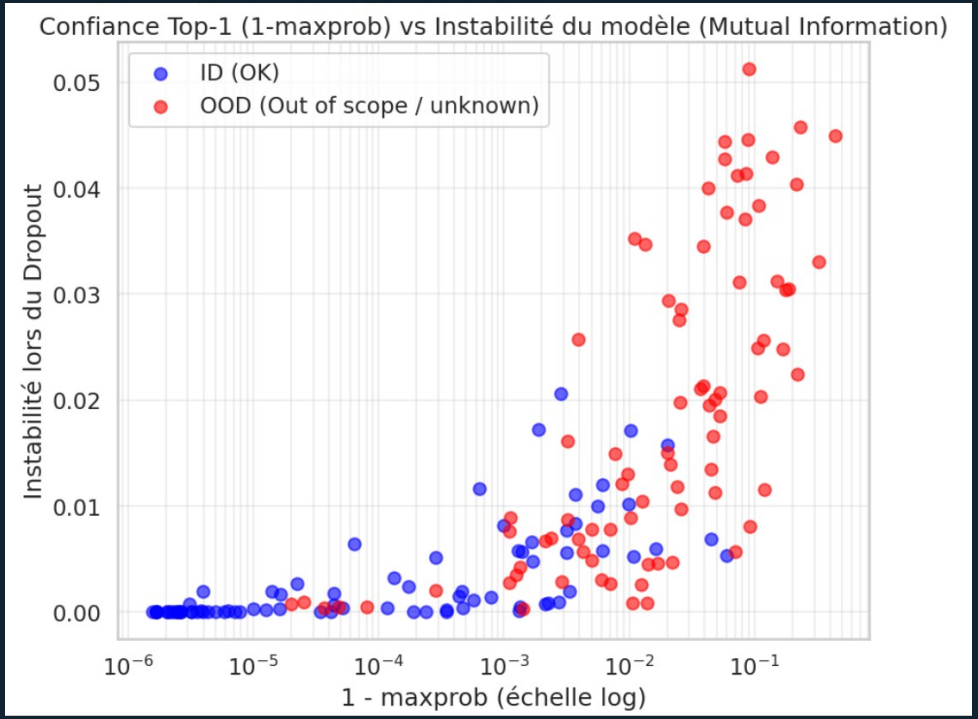
L'APPRENTISSAGE

INCERTITUDE

Corrélation (Spearman)
= 0.887

score(1-maxprob)
= 0.913

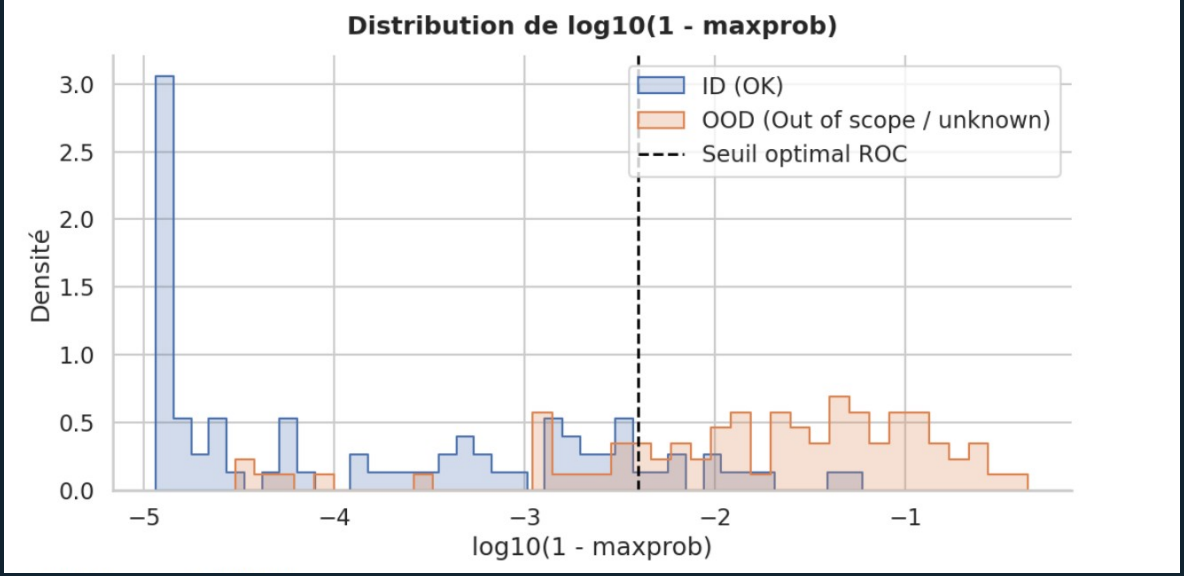
score(dropout)
= 0.876



POSITIVE TECH < / > ADEO GROUP

L'APPRENTISSAGE

INCERTITUDE



Pour le seuil

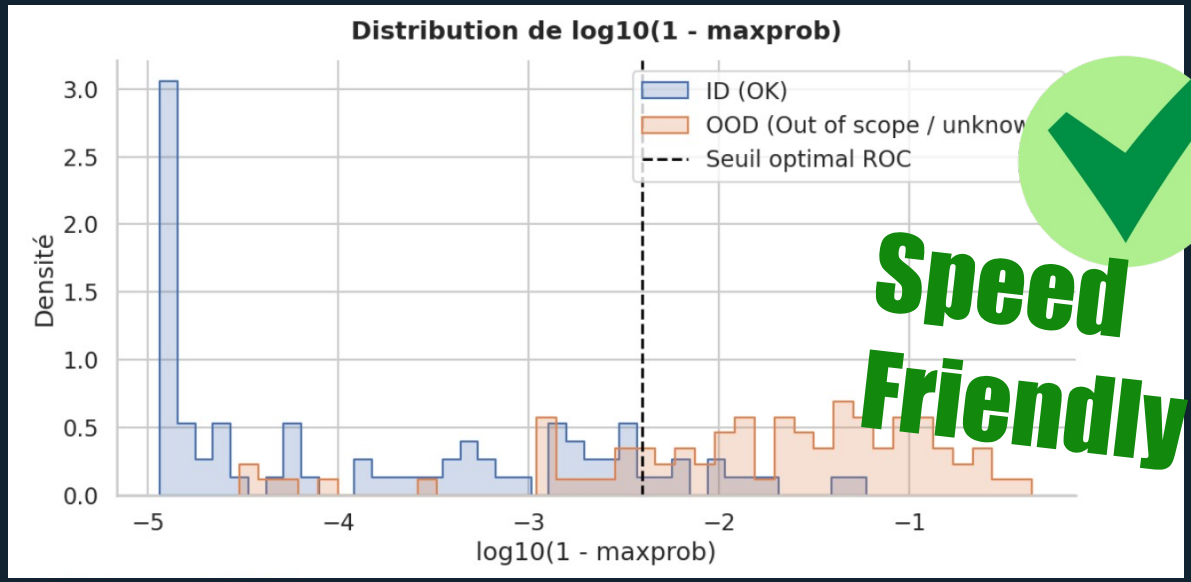
0.00394

FP (ID → OOD)	FN (OOD → ID)	TN (ID → ID)	TP (OOD → OOD)
12.35%	**20.48%**	**87.65%**	**79.52%**

POSITIVE TECH < / > ADEO GROUP

L'APPRENTISSAGE

INCERTITUDE



Pour le seuil
0.00394

FP (ID → OOD)	FN (OOD → ID)	TN (ID → ID)	TP (OOD → OOD)
12.35%	**20.48%**	**87.65%**	**79.52%**

[≡]

POSITIVE TECH< / > ADEO GROUP

LE SERVING

POSITIVE TECH < / > ADEO GROUP

LE SERVING
TEST INITIAL

- Perf ?

```
model_cpu = copy.deepcopy(model).to(device_cpu)
```

Boucle For (~4000 itérations) “Juste pour voir” :

mean=6.76ms p90=8.98ms

POSITIVE TECH < / > ADEO GROUP

LE SERVING

TEST INITIAL

- Perf ?

```
model_cpu = copy.deepcopy(model).to(device_cpu)
```

Boucle For (~4000 itérations) “Juste pour voir”

mean=6.76ms p90=8.98ms

POSITIVE TECH< / > ADEO GROUP

LE SERVING
CHOIX DE LA STACK

- ONNX Runtime
- OpenVINO
- Quantization

POSITIVE TECH < / > ADEO GROUP

LE SERVING

CHOIX DE LA STACK

- ONNX : Format ouvert
  - Représente le modèle en un graphe d'opérateurs
  - Framework agnostique

- ORT : Moteur d'inférence
  - ***OPTIMISE***
  - Multiples *"ExecutionProvider"*

POSITIVE TECH < / > ADEO GROUP

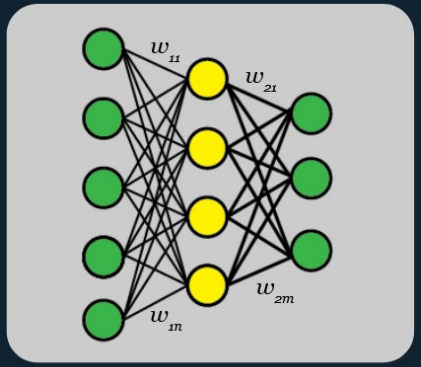
LE SERVING
CHOIX DE LA STACK

- Toolkit… dont moteur d'inférence
- Spécialisé Intel (CPU / GPU / NPU)
  - Supporte des mode HETERO / MULTI devices
- Optimise et compile pour le hardware
  - Quantization 16 bits supportée en CPU

[≡]

POSITIVE TECH< / > ADEO GROUP

LE SERVING
CHOIX DE LA STACK



Quantization

$\theta=\{w_{11}, w_{12}, \dots, w_{21}, w_{22}, \dots, w_{xy}\}$

- Diminuer la précision pour gagner
  - en place
  - en vitesse
  - … parfois au détriment de la précision
- Plusieurs “catégories” de quantization

[≡]

POSITIVE TECH < / > ADEO GROUP

LE SERVING
CHOIX DE LA STACK

Quantization 32 → 16 bits

- “Simple arrondi” …parfois 
- **FP16** : Floating Point standard 16 bits
- **BF16** : Brain Float - même ordre de grandeur que FP32, mais moins précis

POSITIVE TECH < / > ADEO GROUP

LE SERVING
CHOIX DE LA STACK

Quantization 32 → 8 bits ou 4 bits, ou…

- Continue → discret : besoin de mapping
- Statique : on calibre avec un jeu de données pour ne pas sortir de la plage des 8bits

OU

- Quantization Dynamique : à chaud

LE SERVING
CHOIX DE LA STACK

- Qu’est ce qu’on benchmark ?

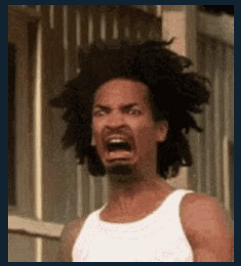
	ORT	OpenVINO Core
FP32	✓	✓
FP16	Since the CPU version of ONNX Runtime doesn’t support float16 ops and the [illegible] needs to measure the accuracy [illegible]	✓
BF16		✓
Int8	✗	✗

POSITIVE TECH < / > ADEO GROUP

LE SERVING
CHOIX DE LA STACK

- Quantization int8 - Aperçu des performances :

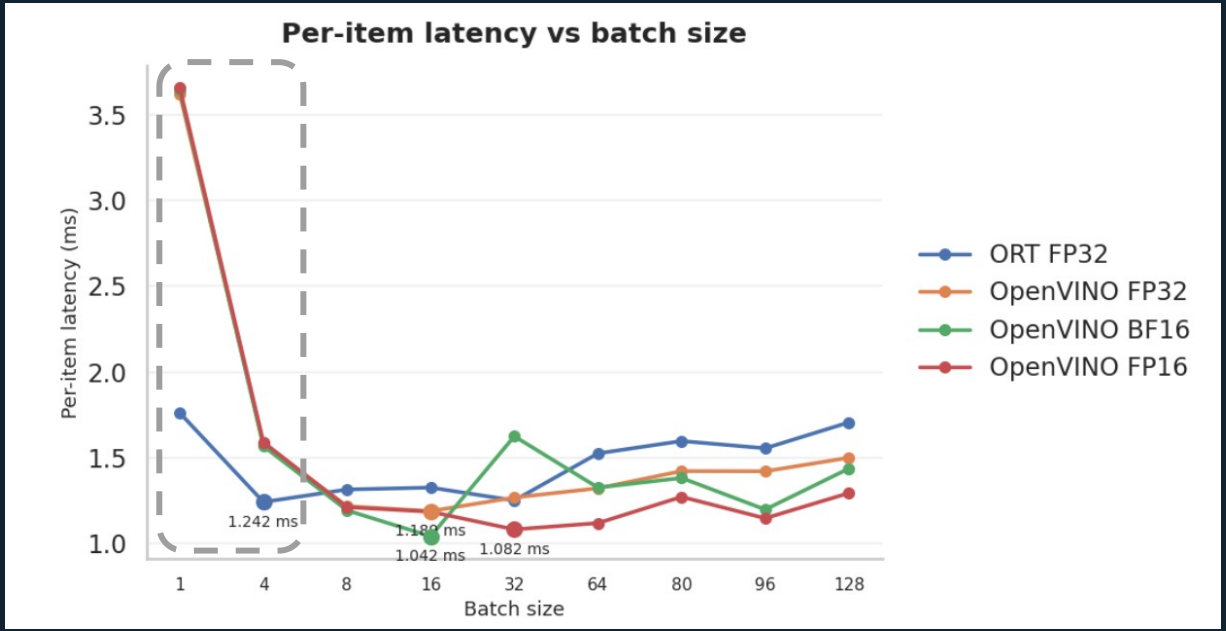
```
=== Base vs ORT Quant INT8 ===
   PS agreement    : 91%
   TR agreement    : 80%
   OTHER agreement: 37%
```



POSITIVE TECH < / > ADEO GROUP

LE SERVING
CHOIX DE LA STACK

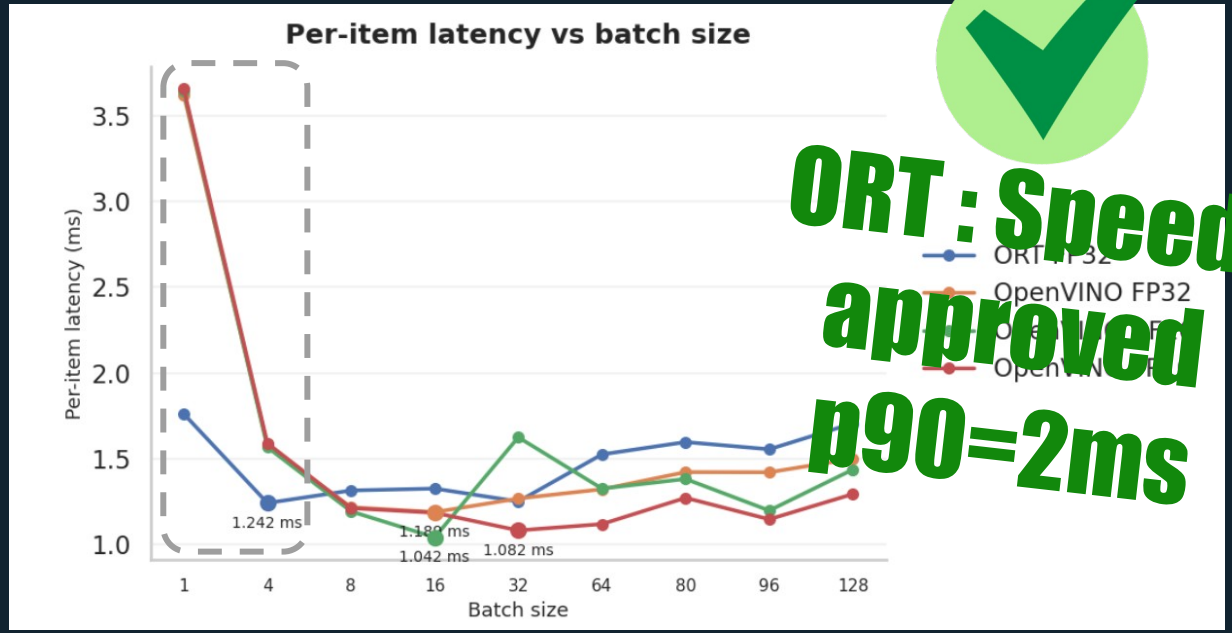
- Benchmarks



POSITIVE TECH < / > ADEO GROUP

LE SERVING
CHOIX DE LA STACK

- Benchmarks



ORT : Speed approved p90=2ms

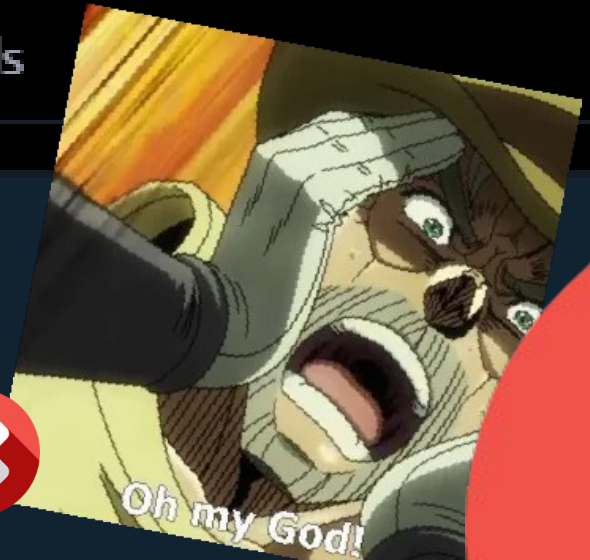
POSITIVE TECH < / > ADEO GROUP

LE SERVING

API & STRESS TEST

p90 Response time percentile 90

0,025 seconds

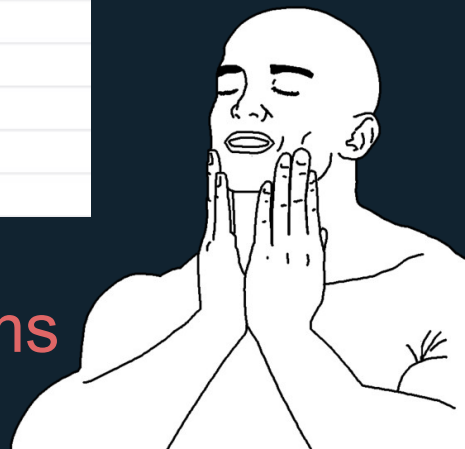


LE SERVING
API & STRESS TEST

- API / Modèle : Stress test colocalisé dans kube

DATE	SERVICE	RESOURCE	DURATION	METHOD	STATUS CODE
Jan 12 11:54:36.322	intent-detection	predict	4.68ms		
Jan 12 11:54:35.613	intent-detection	predict	4.91ms		
Jan 12 11:54:34.740	intent-detection	predict	4.21ms		
Jan 12 11:54:34.740	opus-intent-detection	GET /predict	5.44ms	GET	200
Jan 12 11:54:33.745	opus-intent-detection	GET /predict	5.54ms	GET	200
Jan 12 11:54:29.478	opus-intent-detection	GET /predict	6.22ms	GET	200

Octoperf Injecteur Belgique : +20ms



LE SERVING
API & STRESS TEST

- API / Modèle : Stress test colocal ns kube

DATE	SERVICE	RESOURCE	DURATION	METHOD	STATUS CODE
Jan 12 11:54:36.322	intent-detection	predict	4.68ms		
Jan 12 11:54:35.613	intent-detection	predict	4.91m		
Jan 12 11:54:34.740	intent-detection	predict	4.21m		
Jan 12 11:54:34.740	opus-intent-detection	GET /predict	5.44ms	GET	200
Jan 12 11:54:33.745	opus-intent-detection	GET /predict	5.54		200
Jan 12 11:54:29.478	opus-intent-detection	GET /predict	6.22		

Speed approved p90=6ms

Octoperf Injecteur Belgique : +20ms

POSITIVE TECH < / > ADEO GROUP

CONCLUSION

POSITIVE TECH < / > ADEO GROUP

CONCLUSION

- Test de bert-tiny / rexBERT micro avec tête custom
- Test d'optimizer AdamW / Muon / AdamW + Muon
- Benchmark local de OpenVINO / ONNX Runtime / PyTorch
- Benchmark local de Quant 16bits / Non Quant / …
- Benchmark local de Batch 1 / 4 / 8 / 16 / 32 / 64 / …
- Benchmark prod w/ octoperf
- Benchmark prod intra cluster w/ Datadog
- Comparaison MC Dropout / 1-maxprob / …

POSITIVE TECH < / > ADEO GROUP

CONCLUSION

1ms / item !

- Test de bert-tiny / **rexBERT micro** avec tête custom
- Test d'optimizer AdamW / Muon / **AdamW + Muon**
- Benchmark local de OpenVINO / **ONNX Runtime** / PyTorch
- Benchmark local de Quant 16bits / **Non Quant** / …
- Benchmark local de **Batch 1 / 4** / 8 / 16 / 32 / 64 / …
- Benchmark prod w/ octoperf
- Benchmark **prod intra cluster w/ Datadog**
- Comparaison MC Dropout / **1-maxprob** / …

"Pour un classifieur rapide et fiable"

POSITIVE TECH < / > ADEO GROUP

CONCLUSION

- + de quantization = pas toujours mieux
- Tester le Quant. Aware Training ?
- Attention à ce que vous mesurez
- Testez pleins de chose !
- La police de “*f(x)*” c’est Georgia en Italique

POSITIVE TECH < / > ADEO GROUP

[sfeir] & adeo

MERCI !

POSITIVE TECH < / > ADEO GROUP

[≡]

L'APPRENTISSAGE

MODÈLE

Training = multiples itérations de :

- Pour chaque exemple connu (x_i, y_i) du Dataset d'entraînement, le réseau, dans son état actuel θ_t, produit une prédiction :

$$\hat{y}_i = F_{\theta t}(x_i)$$

- Calcul de l'écart entre la prédiction et la vérité avec une fonction de coût ("Loss")

$$Loss(\theta_t) = \sum Distance(\hat{y}_i, y_i)$$

- Calcul de la variation de la Loss en fonction de chaque poids du réseau $\nabla_\theta Loss(\theta_t)$ par rétropropagation et mise à jour des poids :

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta Loss(\theta_t)$$

POSITIVE TECH < / > ADEO GROUP