Meet **Magento** FRANCE

Inférence Machine Learning en PHP par l'exemple

Guillaume MOIGNEU

#MM25FR

Guillaume Moigneu

VP, Advocacy | **Upsun & Platform.sh**

nls.io on bluesky

Qui est familier avec les **Transformers**?

Qu’est-ce qu’un Transformer?

Une architecture de réseau neuronal qui a révolutionné le NLP et au-delà.

Introduite dans l'article de 2017 "Attention is All You Need"
Alimente des modèles comme GPT, BERT et T5

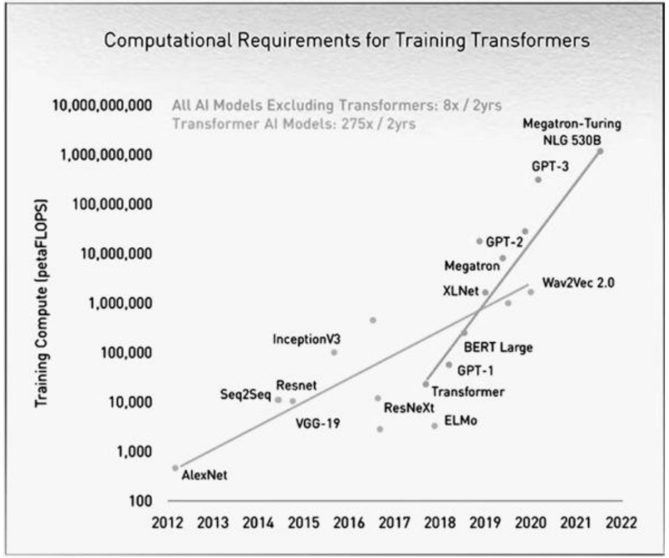
Une façon de gérer une tâche spécifique au-dessus d'un modèle. Plus efficacement.

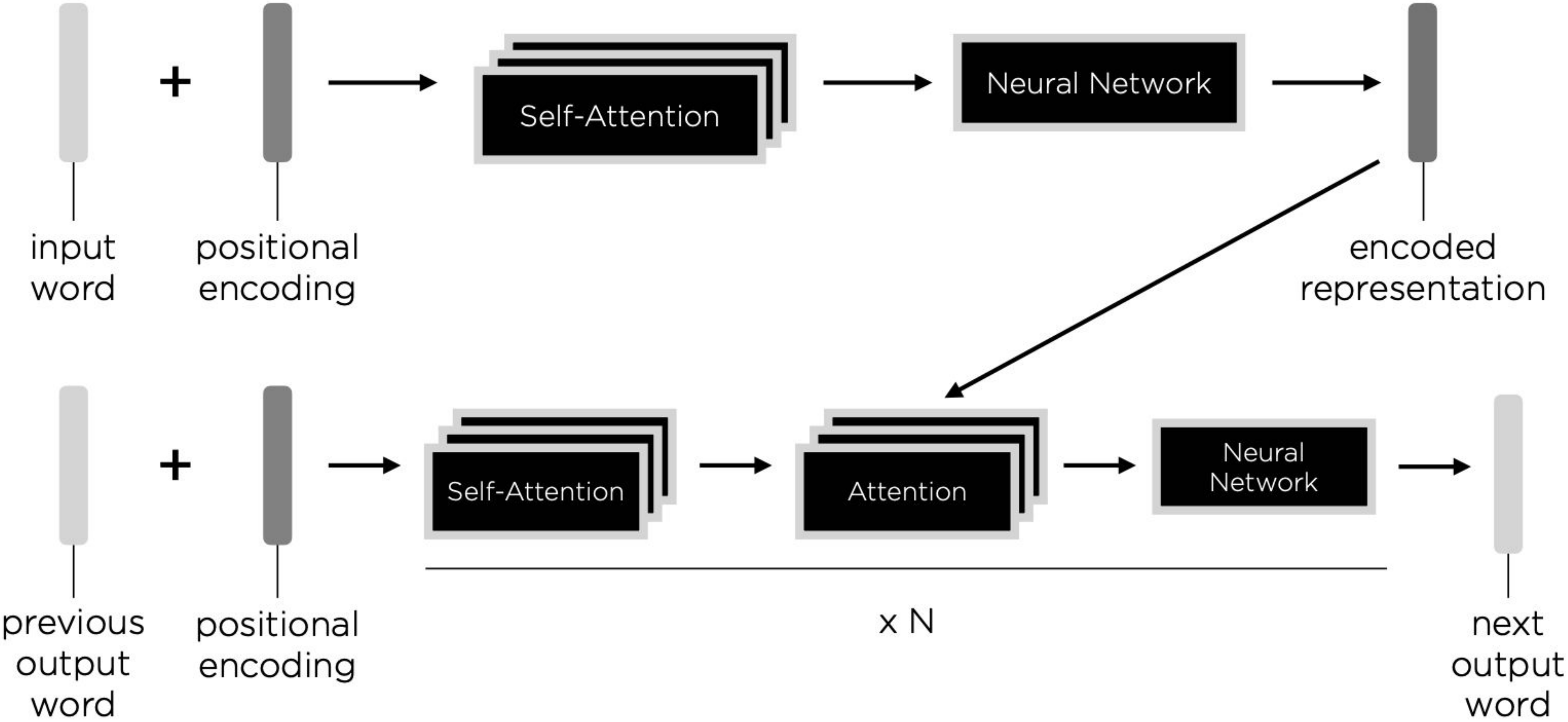
Innovation clé : Traite toutes les entrées simultanément au lieu de séquentiellement

Structure encodeur-décodeur
Couches d'auto-attention
Réseaux de neurones feed-forward
Encodages de position

Avantages:

- Traitement parallèle et réduction du coût d'entraînement et d'inférence
- Compréhension étendue du contexte
- Entraînement et inférence plus rapides
- Meilleur captage des relations
- Utilisation réduite des ressources




input word
+
positional encoding
Self-Attention
Neural Network
encoded representation
previous output word
+
positional encoding
Self-Attention
Attention
Neural Network
x N
next output word

Démo et liens

SCAN ME

Demo:

https://huggingface.co/spaces/webml-community/attention-visualization

Learn more:

https://cdn.cs50.net/ai/2023/x/lectures/6/lecture6.pdf

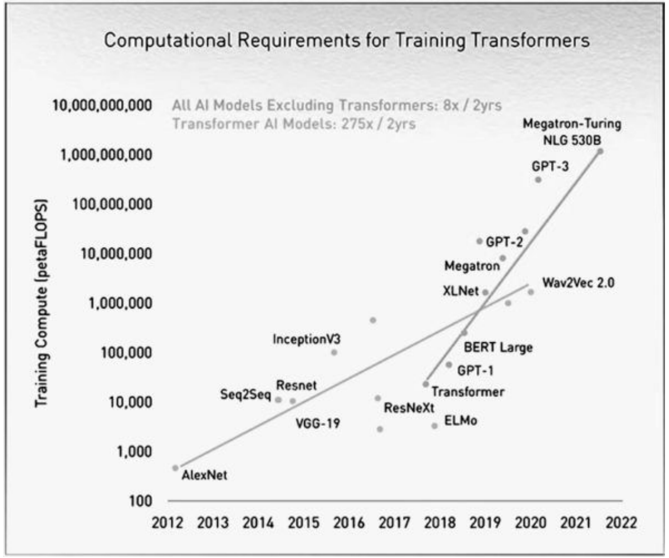
Inférence Machine Learning en PHP par l'exemple

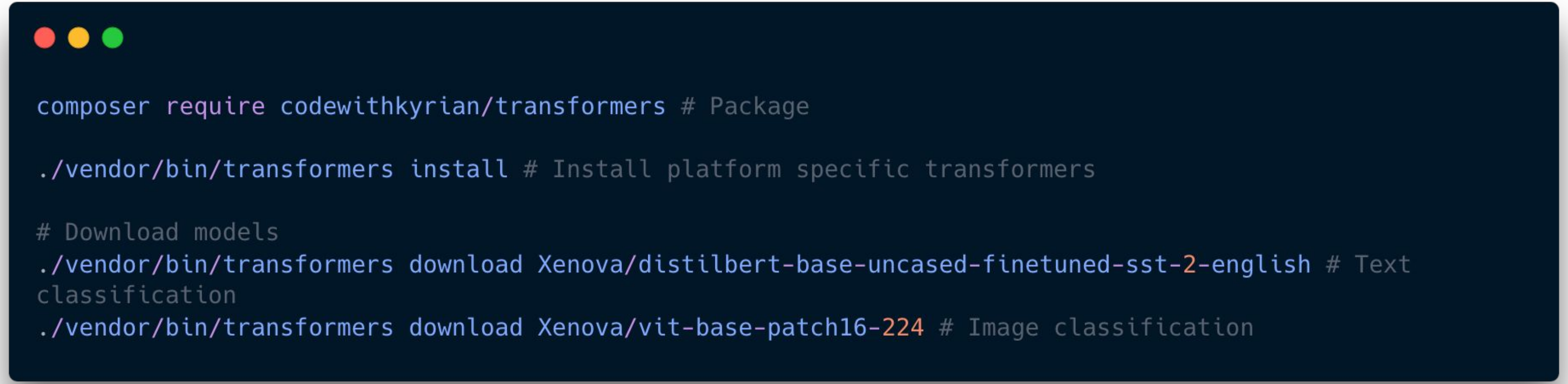
Implémentation de 3 cas d'usages en PHP

Pourquoi faire de l'inférence locale en PHP?

- Aucune souscription tierce
- Pas de "fuite" de données à l'extérieur
- Pas de changement d'architecture/setup

- Millions de modèles disponibles + vos propres modèles
- Fine-tuning possible
- Meilleur contrôle (test A/B, performance, uptime)

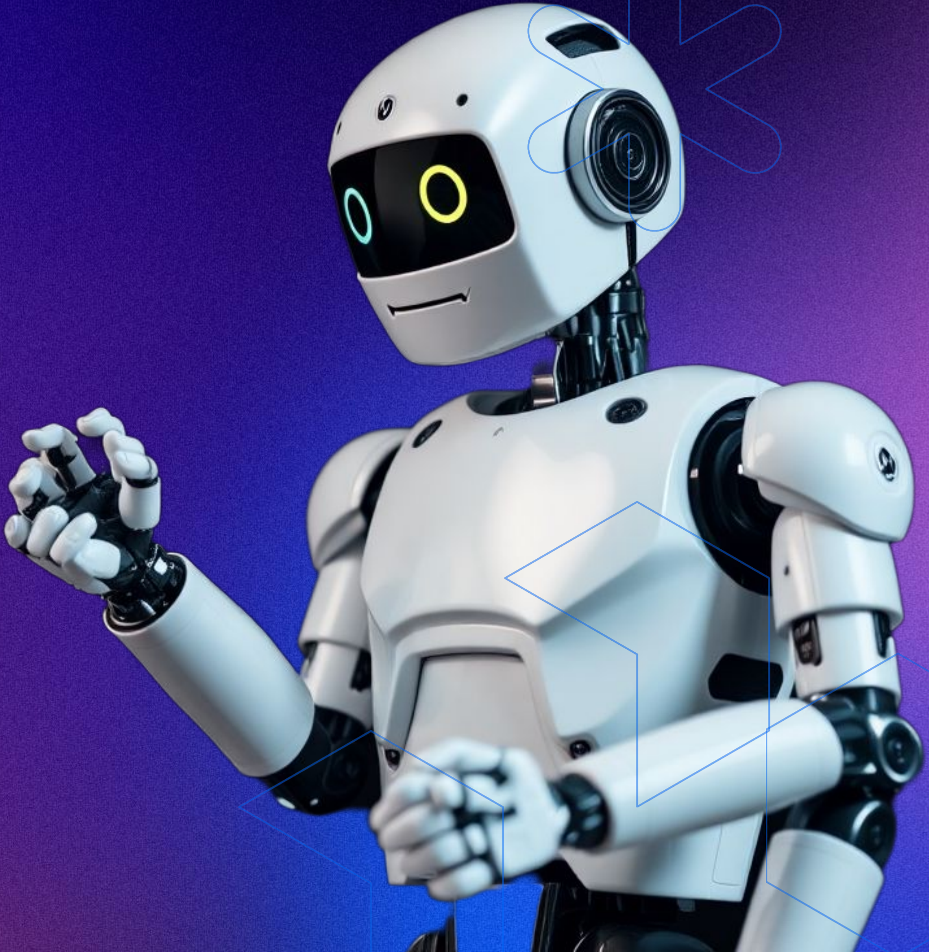


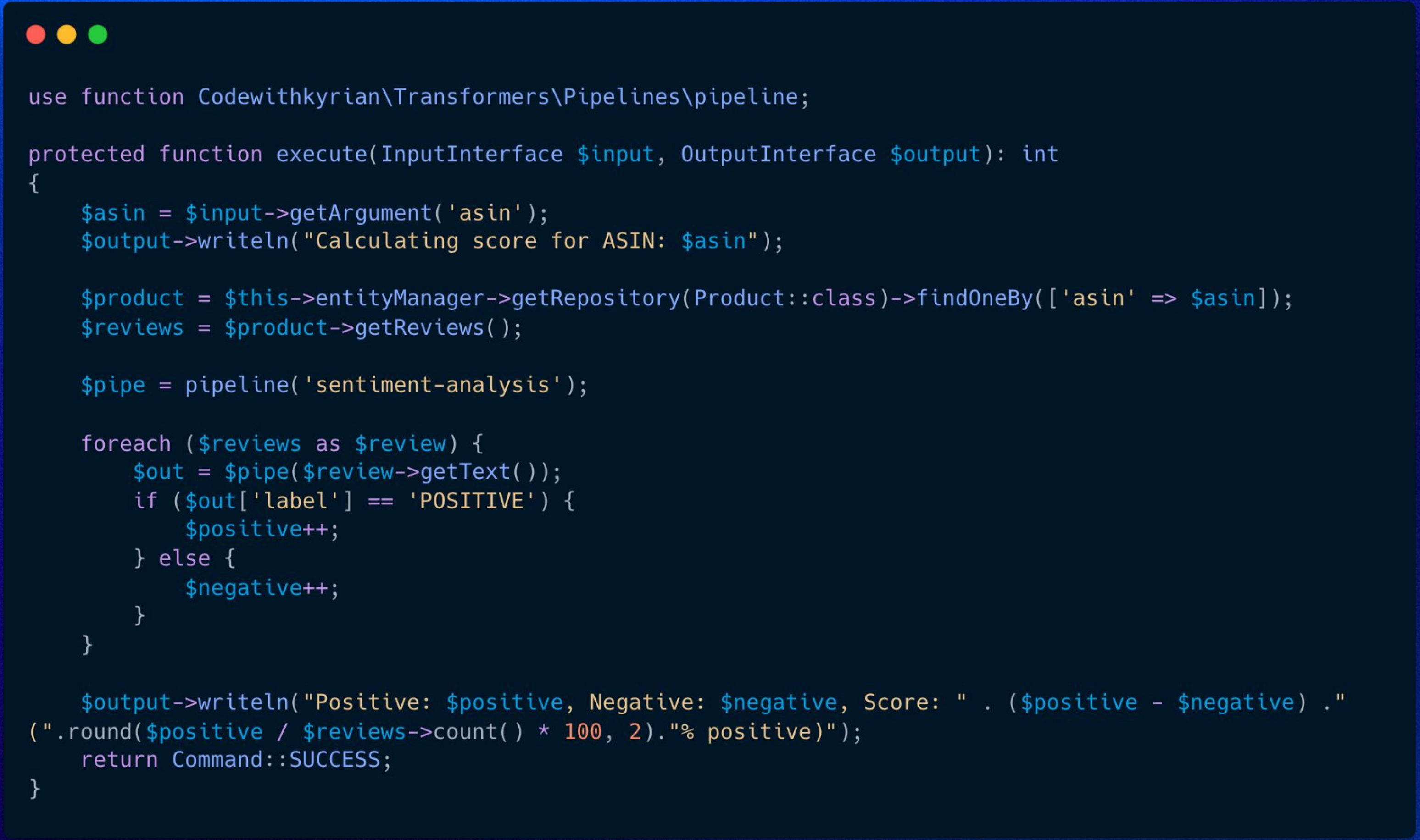


```
composer require codewithkyrian/transformers # Package

./vendor/bin/transformers install # Install platform specific transformers

# Download models
./vendor/bin/transformers download Xenova/distilbert-base-uncased-finetuned-sst-2-english # Text classification
./vendor/bin/transformers download Xenova/vit-base-patch16-224 # Image classification
```

Cas d'usage #1

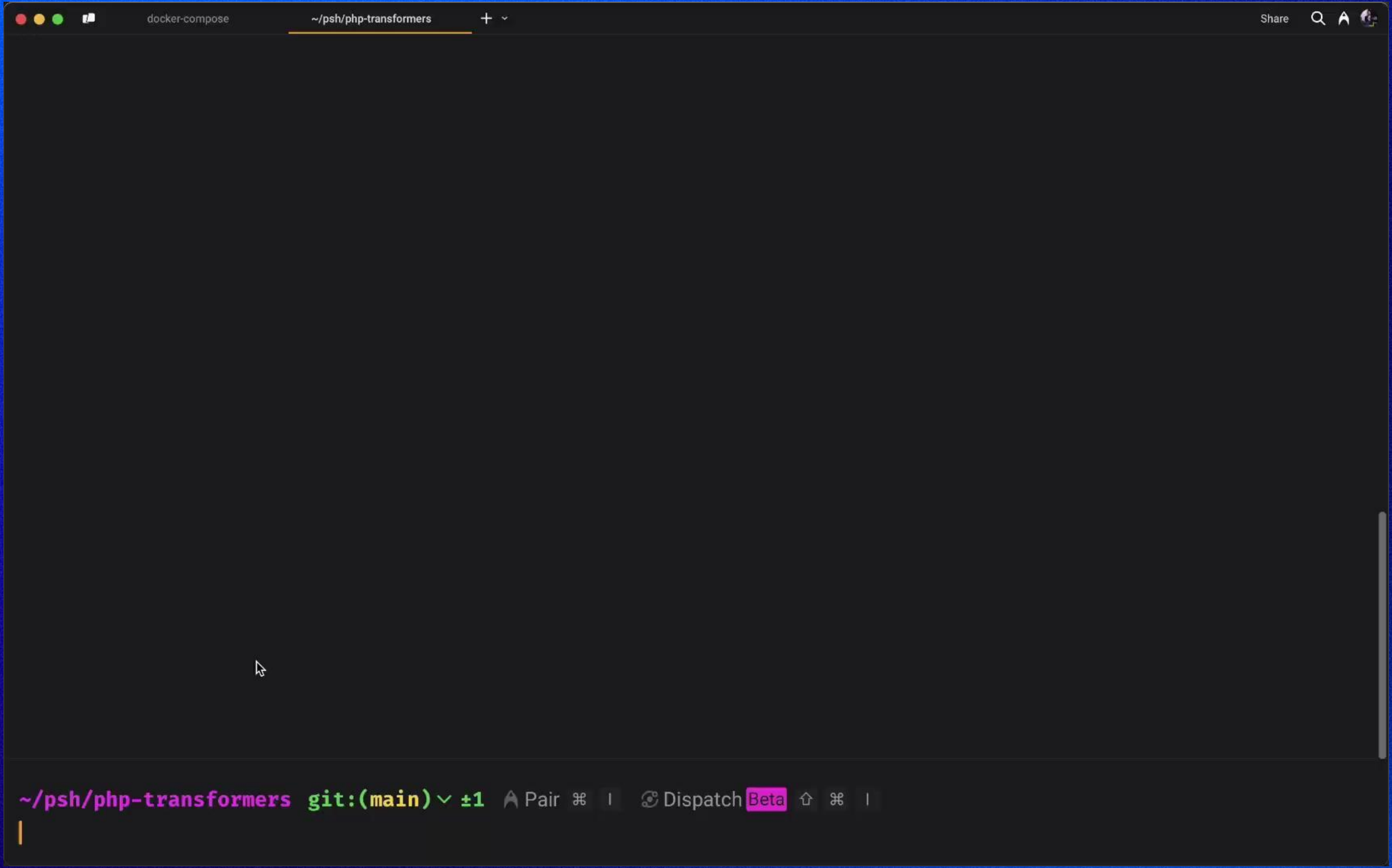
Classification de texte





```
use function Codewithkyrian\Transformers\Pipelines\pipeline;

protected function execute(InputInterface $input, OutputInterface $output): int
{
    $asin = $input->getArgument('asin');
    $output->writeln("Calculating score for ASIN: $asin");

    $product = $this->entityManager->getRepository(Product::class)->findOneBy(['asin' => $asin]);
    $reviews = $product->getReviews();

    $pipe = pipeline('sentiment-analysis');

    foreach ($reviews as $review) {
        $out = $pipe($review->getText());
        if ($out['label'] == 'POSITIVE') {
            $positive++;
        } else {
            $negative++;
        }
    }

    $output->writeln("Positive: $positive, Negative: $negative, Score: " . ($positive - $negative) ."
(".round($positive / $reviews->count() * 100, 2)."% positive)");
    return Command::SUCCESS;
}
```

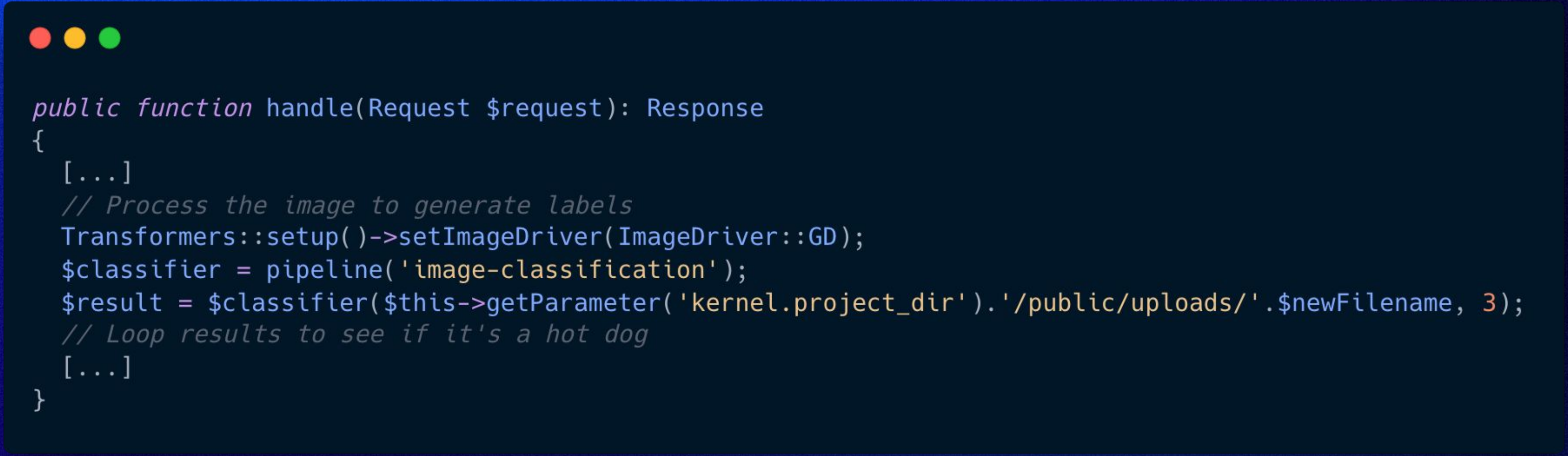
```
php bin/console app:score B07VGRJDFY
```



```
~/psh/php-transformers git:(main) ✓ ±1
```

Cas d'usage #2

Classification d'images

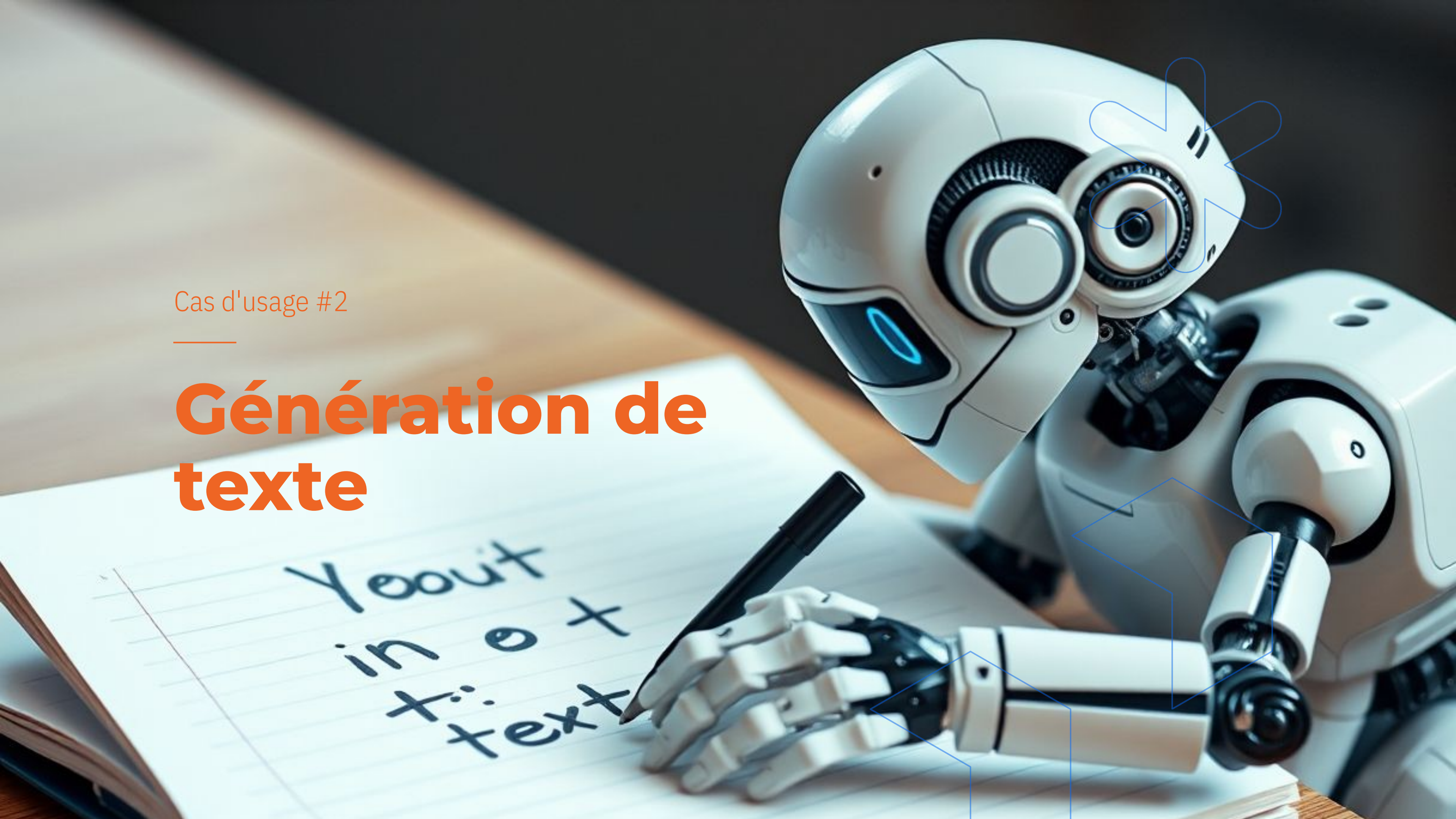


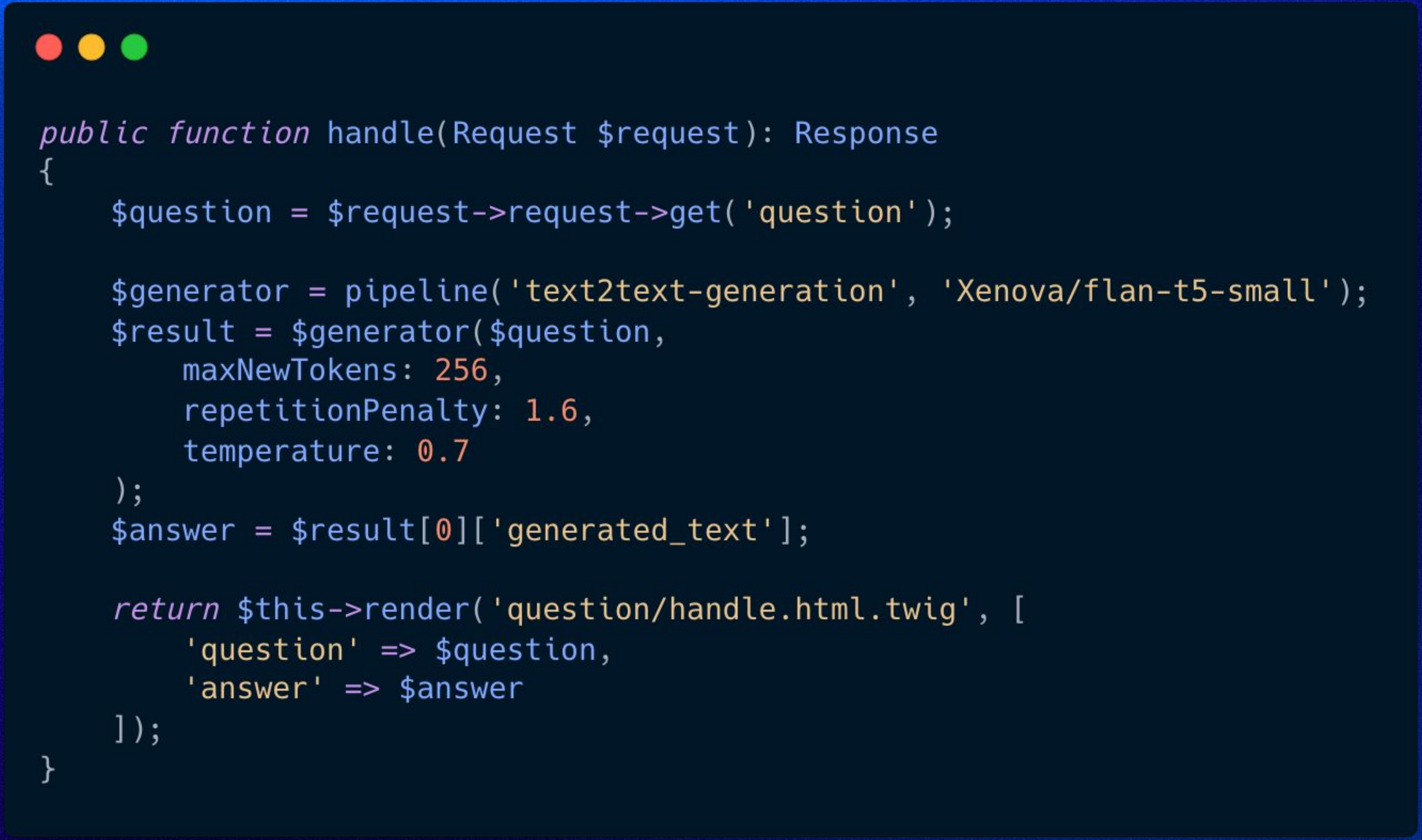
```
public function handle(Request $request): Response
{
  [...]
  // Process the image to generate labels
  Transformers::setup()->setImageDriver(ImageDriver::GD);
  $classifier = pipeline('image-classification');
  $result = $classifier($this->getParameter('kernel.project_dir').'/public/uploads/'.$newFilename, 3);
  // Loop results to see if it's a hot dog
  [...]
}
```

https://mm25fr.moigneu.net/hotdog

Cas d'usage #2

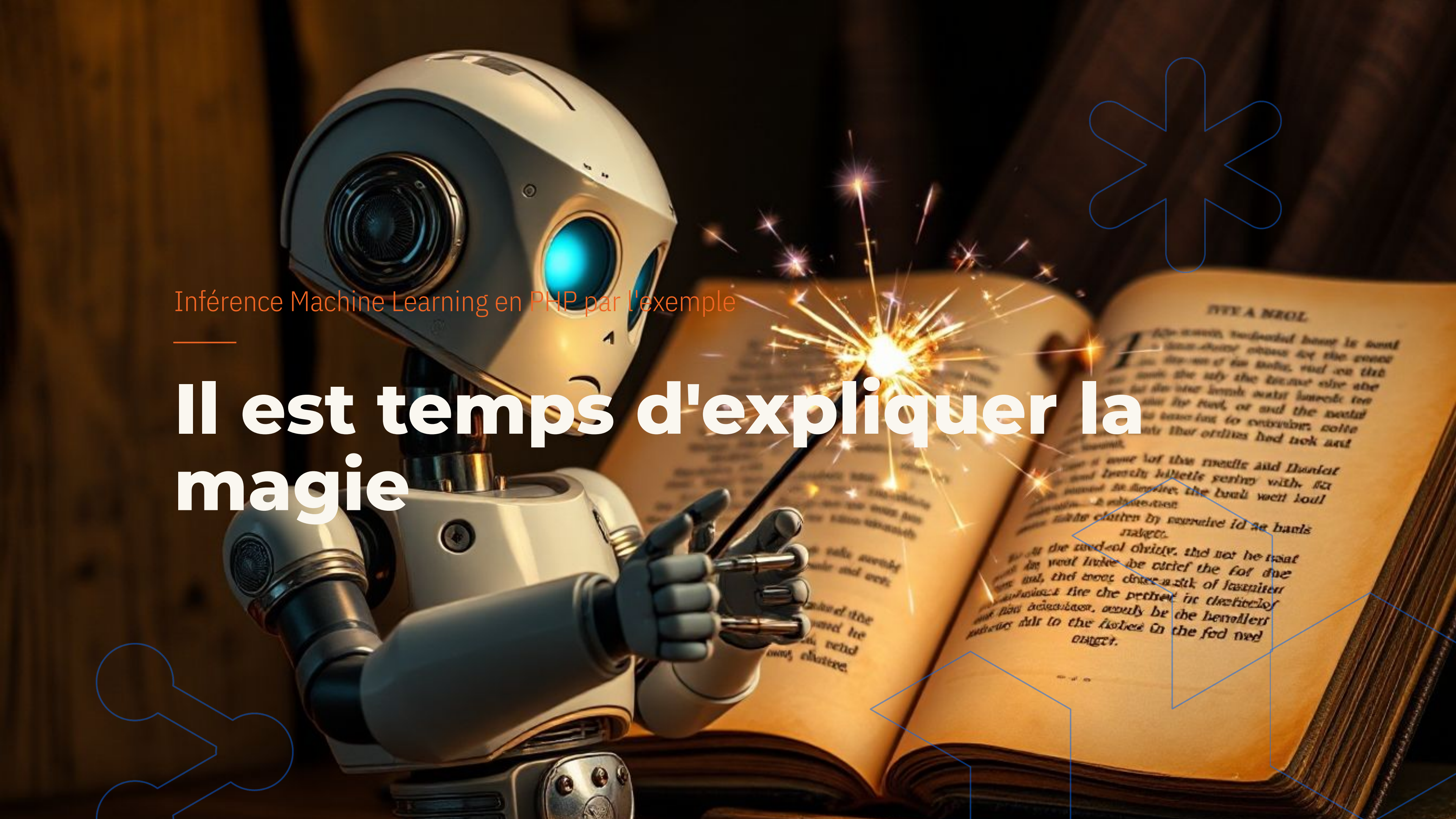
Génération de texte





```
public function handle(Request $request): Response
{
    $question = $request->request->get('question');

    $generator = pipeline('text2text-generation', 'Xenova/flan-t5-small');
    $result = $generator($question,
        maxNewTokens: 256,
        repetitionPenalty: 1.6,
        temperature: 0.7
    );
    $answer = $result[0]['generated_text'];

    return $this->render('question/handle.html.twig', [
        'question' => $question,
        'answer' => $answer
    ]);
}
```

https://mm25fr.moigneu.net/question



Inférence Machine Learning en PHP par l'exemple

Il est temps d'expliquer la magie

PHP FFI - Foreign Function Interface

FFI extension allows PHP code to directly **call functions and manipulate data from C libraries** without writing additional C code or PHP extensions.

Open Neural Network Exchange

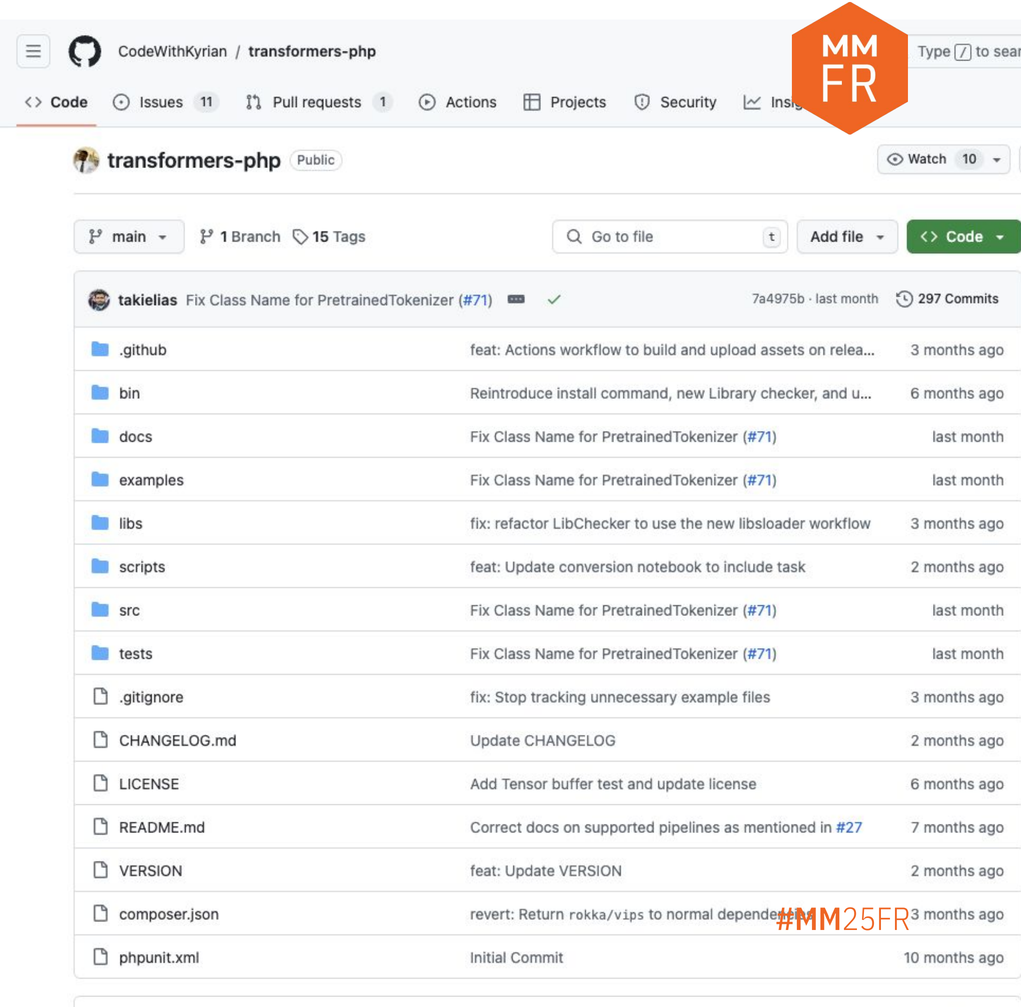
ONNX (Open Neural Network Exchange) is an **open standard format** that allows AI models to be shared between different machine learning frameworks like PyTorch, TensorFlow, and many others.

Transformers & pipelines

Fournis par le package TransformersPHP.

Managé par **Kyrian Obikwelu**
Et 7 contributeurs principaux

Pipelines supportées



takielias Fix Class Name for PretrainedTokenizer (#71) ✓ 7a4975b · last month 297 Commits

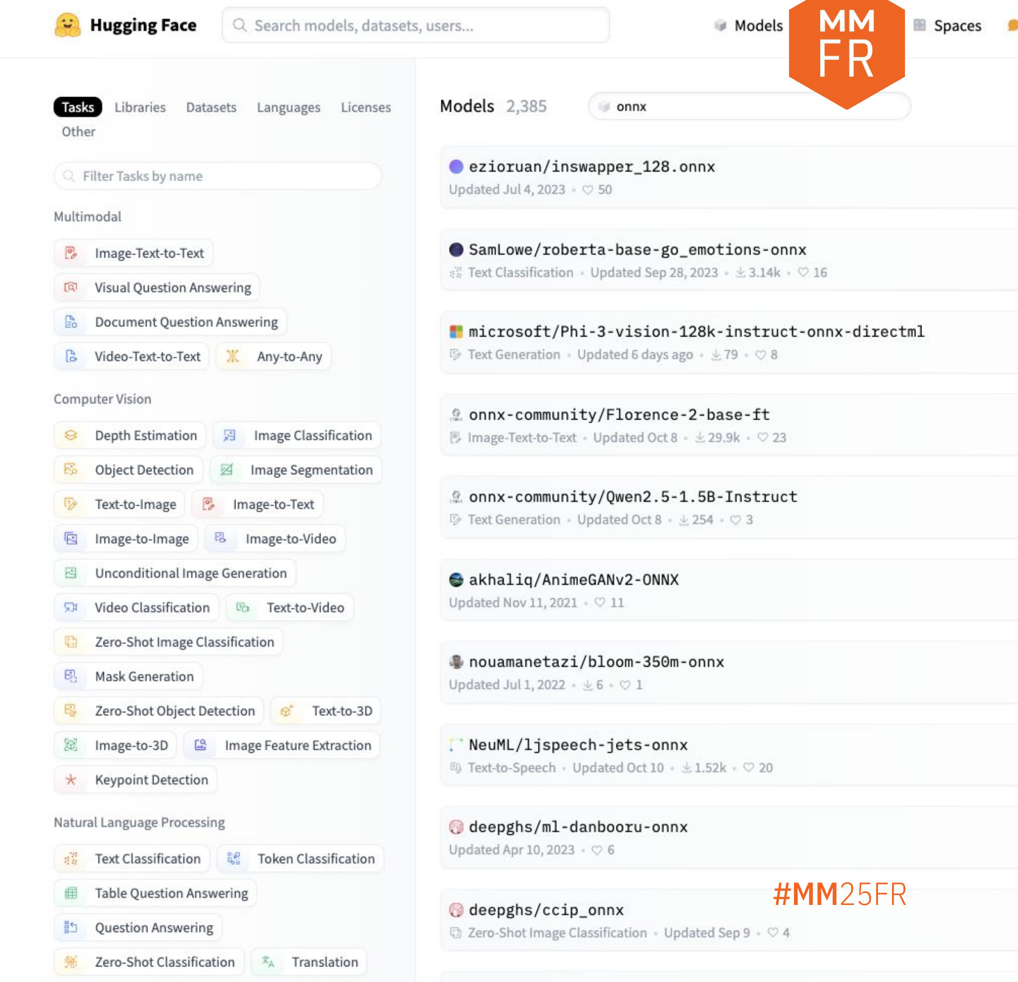
.github	feat: Actions workflow to build and upload assets on relea...	3 months ago
bin	Reintroduce install command, new Library checker, and u...	6 months ago
docs	Fix Class Name for PretrainedTokenizer (#71)	last month
examples	Fix Class Name for PretrainedTokenizer (#71)	last month
libs	fix: refactor LibChecker to use the new libsloader workflow	3 months ago
scripts	feat: Update conversion notebook to include task	2 months ago
src	Fix Class Name for PretrainedTokenizer (#71)	last month
tests	Fix Class Name for PretrainedTokenizer (#71)	last month
.gitignore	fix: Stop tracking unnecessary example files	3 months ago
CHANGELOG.md	Update CHANGELOG	2 months ago
LICENSE	Add Tensor buffer test and update license	6 months ago
README.md	Correct docs on supported pipelines as mentioned in #27	7 months ago
VERSION	feat: Update VERSION	2 months ago
composer.json	revert: Return `rokka/vips` to normal dependencies	3 months ago
phpunit.xml	Initial Commit	10 months ago

Models

Centaines de modèles dispos sur 🤗.

Vous pouvez convertir n'importe quel modèle en format ONNX via

- Python directement
- ou un Notebook



Hugging Face

Search models, datasets, users...

Models

Spaces

Tasks Libraries Datasets Languages Licenses Other

Filter Tasks by name

Multimodal

- Image-Text-to-Text
- Visual Question Answering
- Document Question Answering
- Video-Text-to-Text
- Any-to-Any

Computer Vision

- Depth Estimation
- Image Classification
- Object Detection
- Image Segmentation
- Text-to-Image
- Image-to-Text
- Image-to-Image
- Image-to-Video
- Unconditional Image Generation
- Video Classification
- Text-to-Video
- Zero-Shot Image Classification
- Mask Generation
- Zero-Shot Object Detection
- Text-to-3D
- Image-to-3D
- Image Feature Extraction
- Keypoint Detection

Natural Language Processing

- Text Classification
- Token Classification
- Table Question Answering
- Question Answering
- Zero-Shot Classification
- Translation

Models 2,385 onnx

eziorUan/inswapper_128.onnx
Updated Jul 4, 2023 • ♡ 50

SamLowe/roberta-base-go_emotions-onnx
Text Classification • Updated Sep 28, 2023 • ⤓3.14k • ♡ 16

microsoft/Phi-3-vision-128k-instruct-onnx-directml
Text Generation • Updated 6 days ago • ⤓79 • ♡ 8

onnx-community/Florence-2-base-ft
Image-Text-to-Text • Updated Oct 8 • ⤓29.9k • ♡ 23

onnx-community/Qwen2.5-1.5B-Instruct
Text Generation • Updated Oct 8 • ⤓254 • ♡ 3

akhaliq/AnimeGANv2-ONNX
Updated Nov 11, 2021 • ♡ 11

nouamanetazi/bloom-350m-onnx
Updated Jul 1, 2022 • ⤓6 • ♡ 1

NeuML/ljspeech-jets-onnx
Text-to-Speech • Updated Oct 10 • ⤓1.52k • ♡ 20

deepghs/ml-danbooru-onnx
Updated Apr 10, 2023 • ♡ 6

deepghs/ccip_onnx
Zero-Shot Image Classification • Updated Sep 9 • ♡ 4

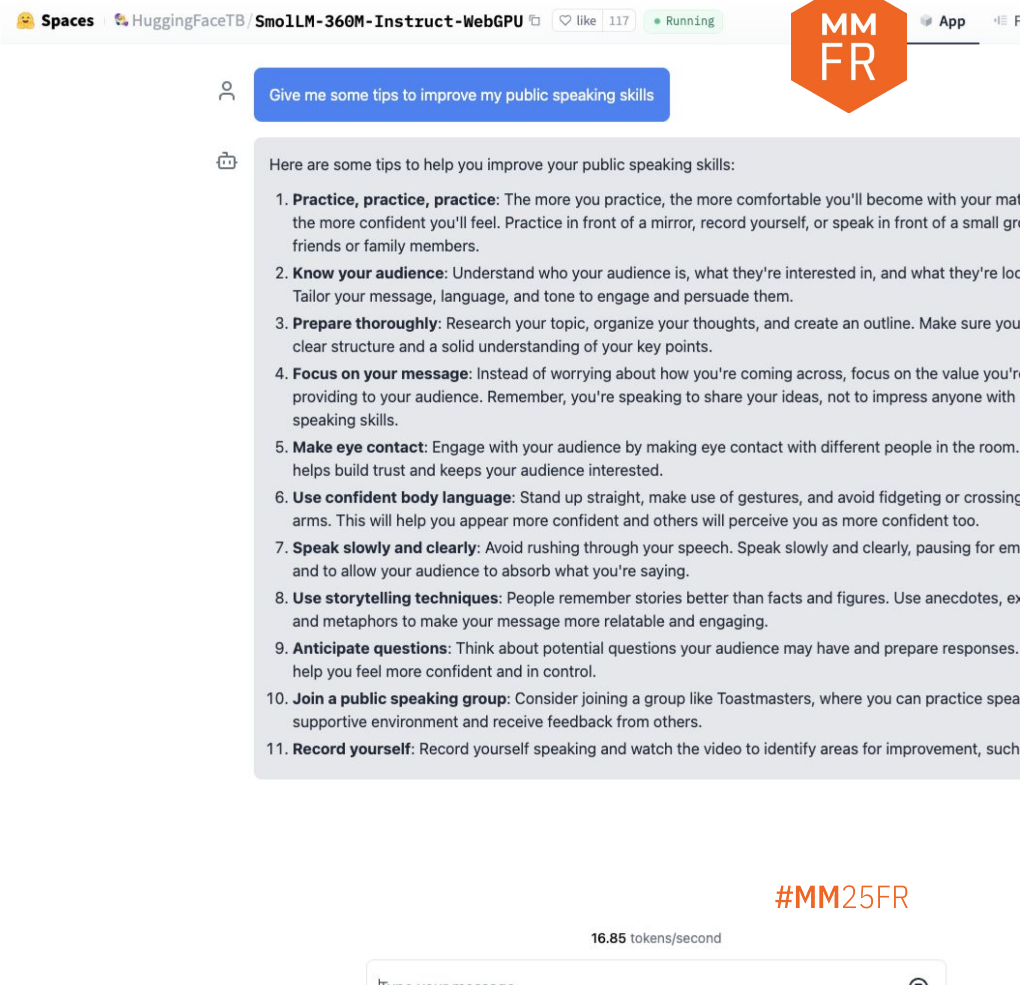
MM FR

#MM25FR

SmolLM

Modèle LLM prometteur et efficace

Fourni par l'équipe HF
Seulement 538 Mo
Excellente alternative aux modèles complets
Peut même fonctionner sur le navigateur



Give me some tips to improve my public speaking skills

Here are some tips to help you improve your public speaking skills:

1. **Practice, practice, practice**: The more you practice, the more comfortable you'll become with your mat the more confident you'll feel. Practice in front of a mirror, record yourself, or speak in front of a small gr friends or family members.
2. **Know your audience**: Understand who your audience is, what they're interested in, and what they're loo Tailor your message, language, and tone to engage and persuade them.
3. **Prepare thoroughly**: Research your topic, organize your thoughts, and create an outline. Make sure you clear structure and a solid understanding of your key points.
4. **Focus on your message**: Instead of worrying about how you're coming across, focus on the value you'r providing to your audience. Remember, you're speaking to share your ideas, not to impress anyone with speaking skills.
5. **Make eye contact**: Engage with your audience by making eye contact with different people in the room. helps build trust and keeps your audience interested.
6. **Use confident body language**: Stand up straight, make use of gestures, and avoid fidgeting or crossing arms. This will help you appear more confident and others will perceive you as more confident too.
7. **Speak slowly and clearly**: Avoid rushing through your speech. Speak slowly and clearly, pausing for em and to allow your audience to absorb what you're saying.
8. **Use storytelling techniques**: People remember stories better than facts and figures. Use anecdotes, ex and metaphors to make your message more relatable and engaging.
9. **Anticipate questions**: Think about potential questions your audience may have and prepare responses. help you feel more confident and in control.
10. **Join a public speaking group**: Consider joining a group like Toastmasters, where you can practice spea supportive environment and receive feedback from others.
11. **Record yourself**: Record yourself speaking and watch the video to identify areas for improvement, such

16.85 tokens/second

Plus grande limitation actuellement :
Pas de support GPU. Pas encore.

Ma recommandation

Bien que cela soit **excellent pour les petites tâches** (classification, étiquetage d'images, etc.), tout travail sur les LLM devrait être **délégué à une infrastructure basée sur du GPU**.

Utilisez un service SaaS : OpenAI, API Claude, etc.

Ou utilisez les points d'accès d'inférence HF pour déployer le modèle de votre choix
Et interrogez ce point d'accès API depuis votre application.

Ou "Construisez le vôtre", à vos risques et périls !

Meet **Magento** FRANCE

MERCI !

Guillaume Moigneu
nls.io on bluesky

plz follow!

#MM25FR