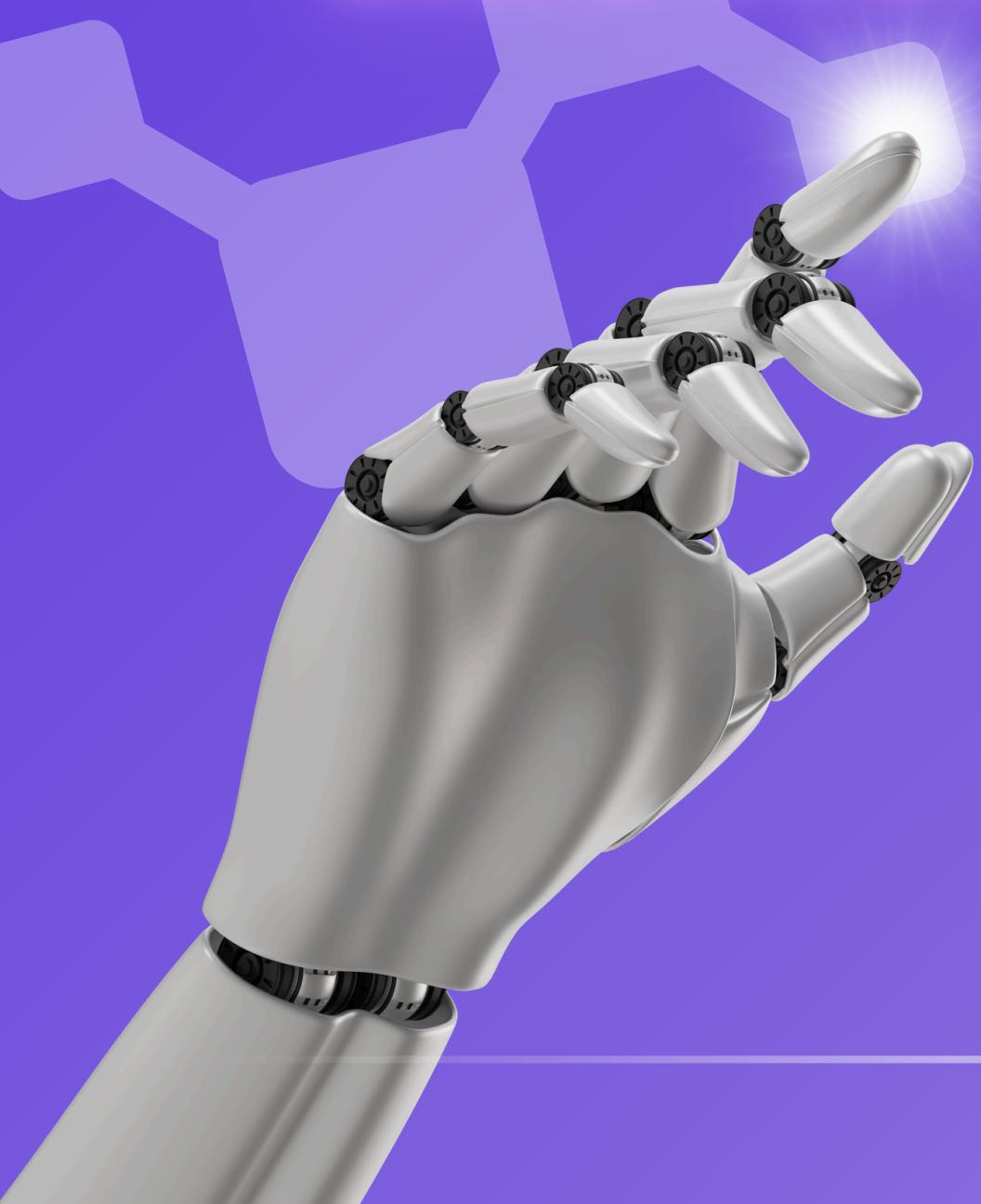


ANALYSE MULTI-MODALE DES ETATS EMOTIONNELS GRACE AUX LLM



Aldorith CHOUNA
Gilles TAGNE
Raphael NOUBIENGANG
Josue MIENGUE



Introduction

La détection automatique du mensonge est un domaine émergent à l'intersection :

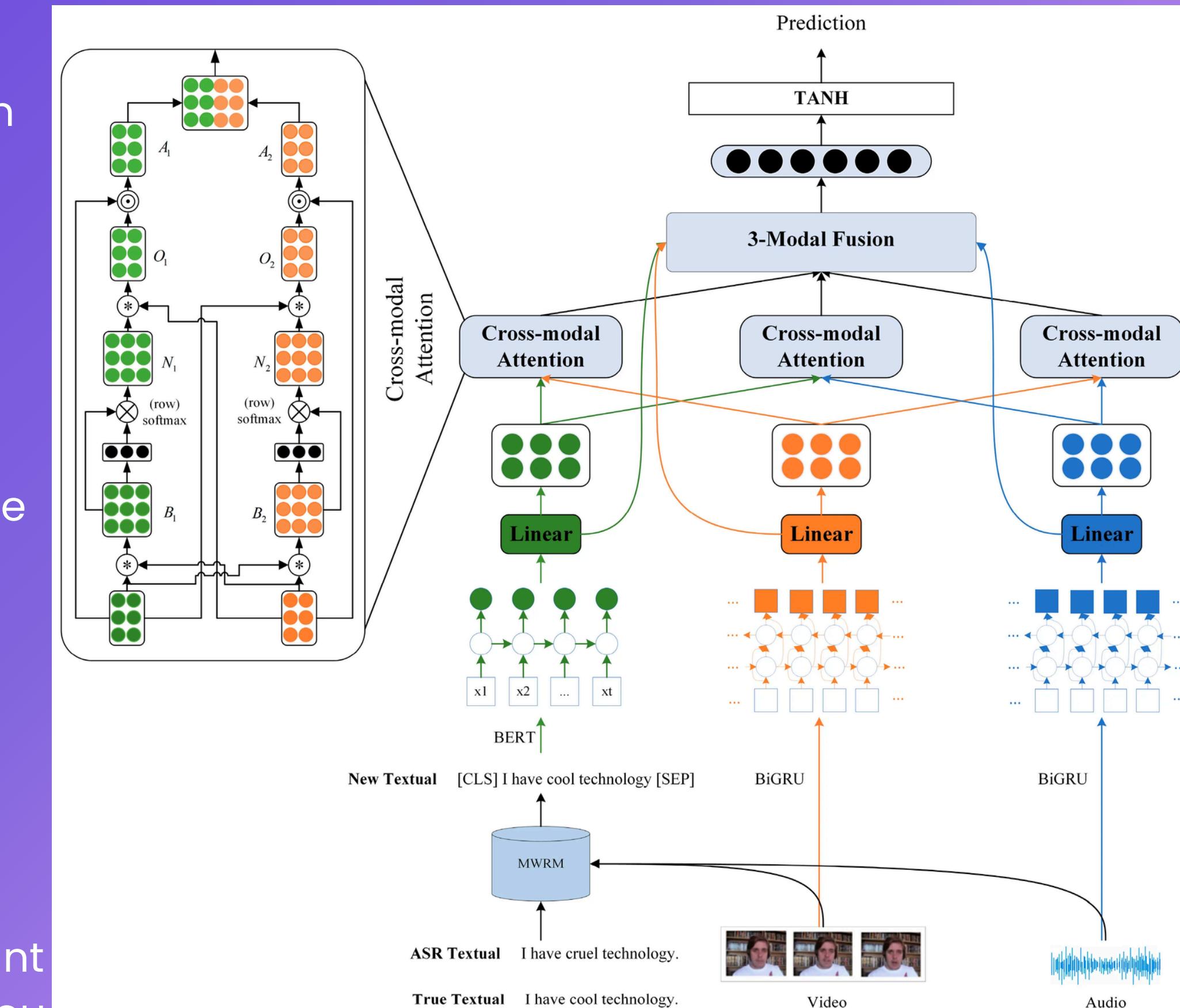
- de la psychologie comportementale,
- de la linguistique,
- et de l'intelligence artificielle multimodale.

L'objectif du projet est de développer un modèle capable d'analyser simultanément :

- le texte prononcé,
- la voix et ses variations acoustiques,
- les expressions faciales et micro-mouvements,

afin d'estimer si une déclaration est vraie ou mensongère.

Ce travail repose sur le **dataset MU3D**, fournissant des vidéos d'interviews annotées comme Truth ou Lie avec leurs composantes multimodales alignées.



CONTEXTE

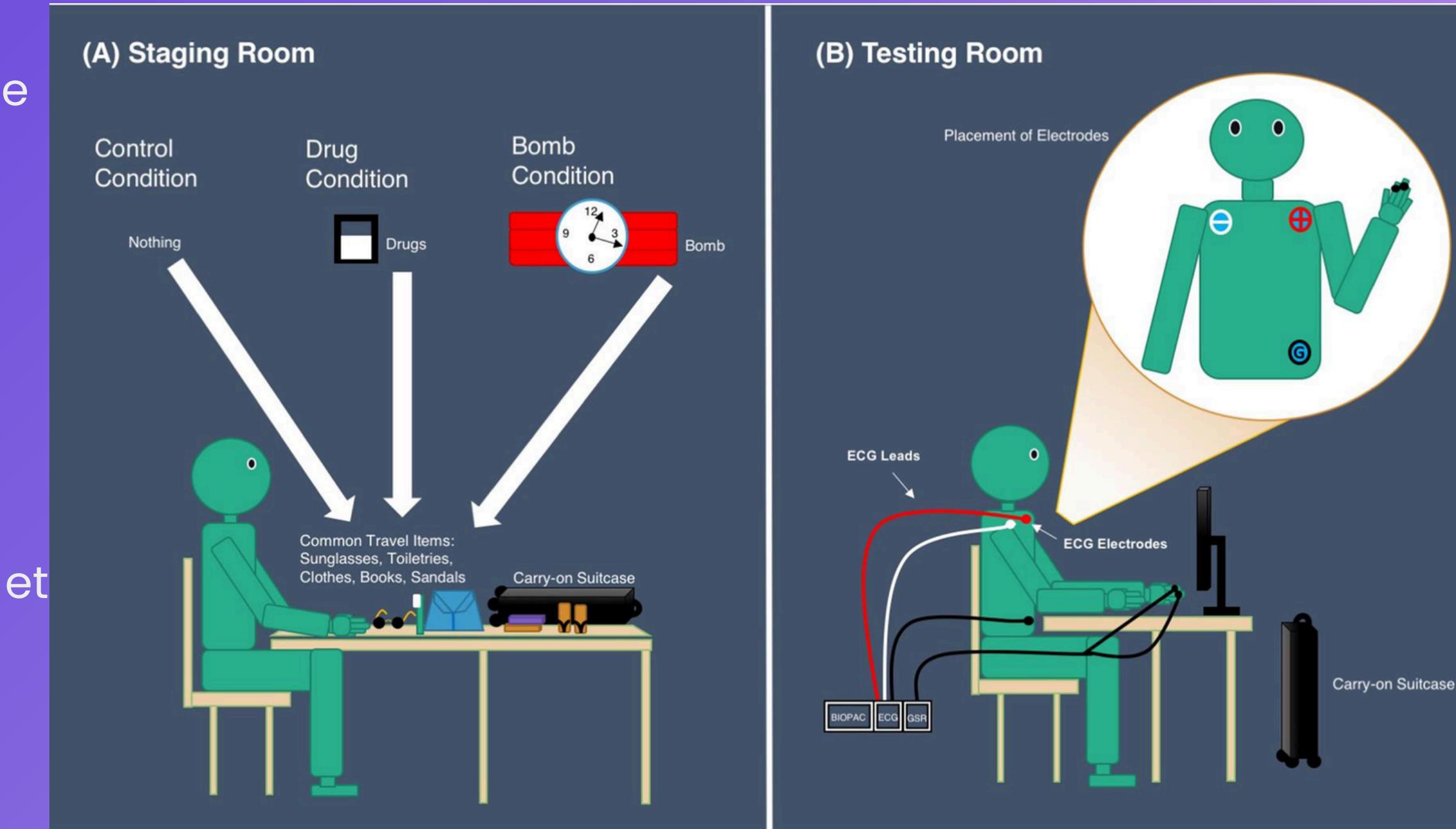
Depuis plus de 50 ans, la recherche en psychologie montre que le mensonge laisse souvent apparaître des indices comportementaux, par exemple :

- micro-expressions involontaires,
- variation de la voix (pitch, intensité, rythme),
- hésitations,
- incohérences verbales.

Cependant, ces signaux sont faibles, subtils et variés selon les individus, ce qui rend la détection humaine peu fiable (~54 %).

L'essor du Deep Learning et des modèles multimodaux permet désormais :

- d'extraire automatiquement des caractéristiques riches,
- d'apprendre des corrélations entre texte, audio, et vidéo,
- d'améliorer la capacité d'analyse au-delà du jugement humain.



Problématique

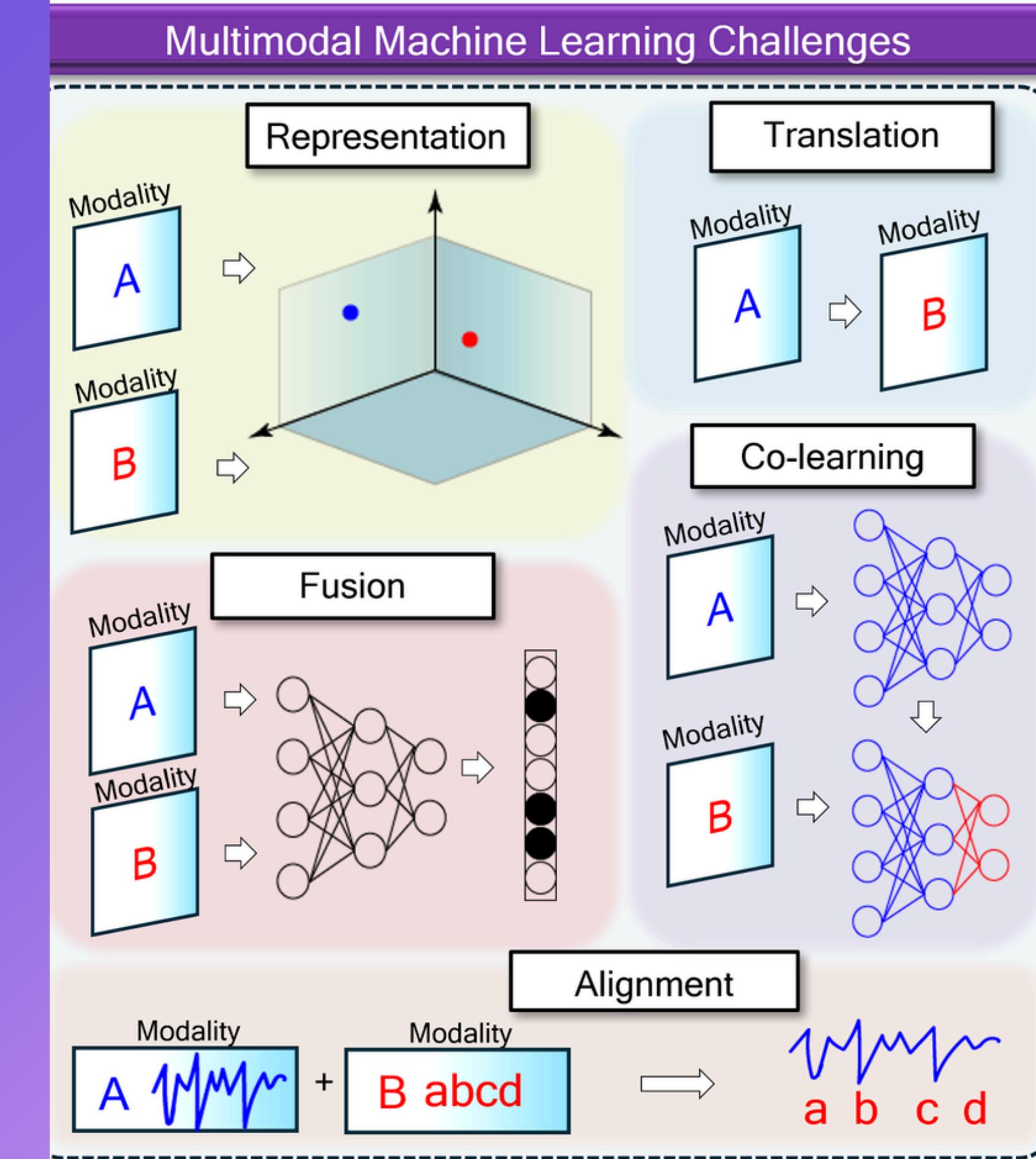
? PROBLEMATIQUE

Comment concevoir un modèle multimodal capable d'intégrer simultanément des informations textuelles, audio et vidéo pour prédire la véracité d'une déclaration, malgré la subtilité des indices du mensonge et la faible quantité de données disponibles ?

OBJECTIF FINAL

Construire un modèle capable de :
comprendre le contenu verbal,
analyser les indices vocaux,
lire les indices faciaux,

-
-
-
- **pondérer dynamiquement les modalités grâce à un mécanisme de gating / attention,**
- → **puis classifier Vérité / Mensonge avec la meilleure robustesse possible sur un petit dataset**



Présentation complète et illustrée du dataset MU3D

MU3D (Multimodal Multi-trait Deception Dataset) est un jeu de données conçu pour étudier la détection de mensonge en utilisant simultanément :

- Texte (transcriptions verbales de réponses)
- Audio (ton, rythme, prosodie)
- Vidéo (expressions faciales, micro-expressions)

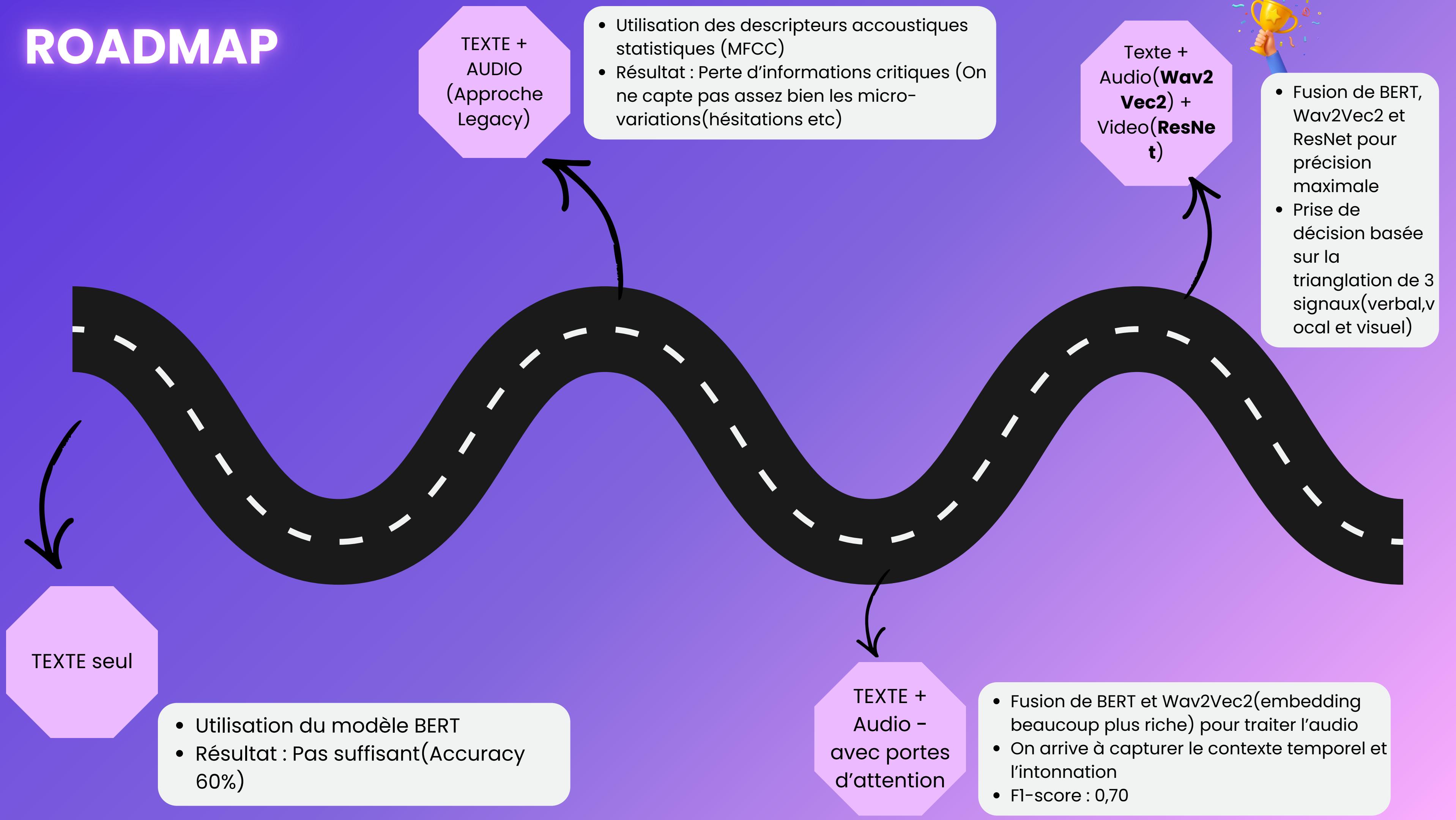
Il contient des interviews filmées dans lesquelles des participants répondent à des questions en disant parfois la vérité et parfois en mentant.

Structure du dataset

- VideoID → ID unique de la vidéo
- ParticipantID
- Transcription → texte de la réponse
- Veracity → 1 = vérité, 0 = mensonge
- AudioFile → fichier .wav correspondant
- VideoFile → fichier .mp4 correspondant

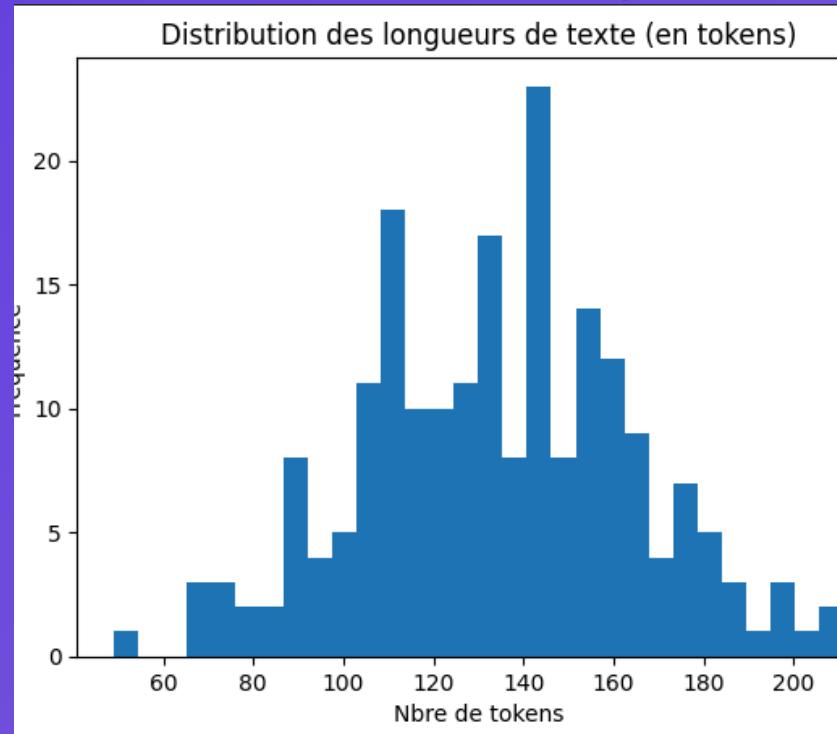


ROADMAP



Méthodologie BERT pour la Détection de Mensonge (Texte)

Prétraitement et Tokenisation
Tokenisation subword WordPiece (BERT-base-uncased)



Entraînement BERT

- Modèle : AutoModelForSequenceClassification (2 classes)
- Fine-tuning complet du modèle

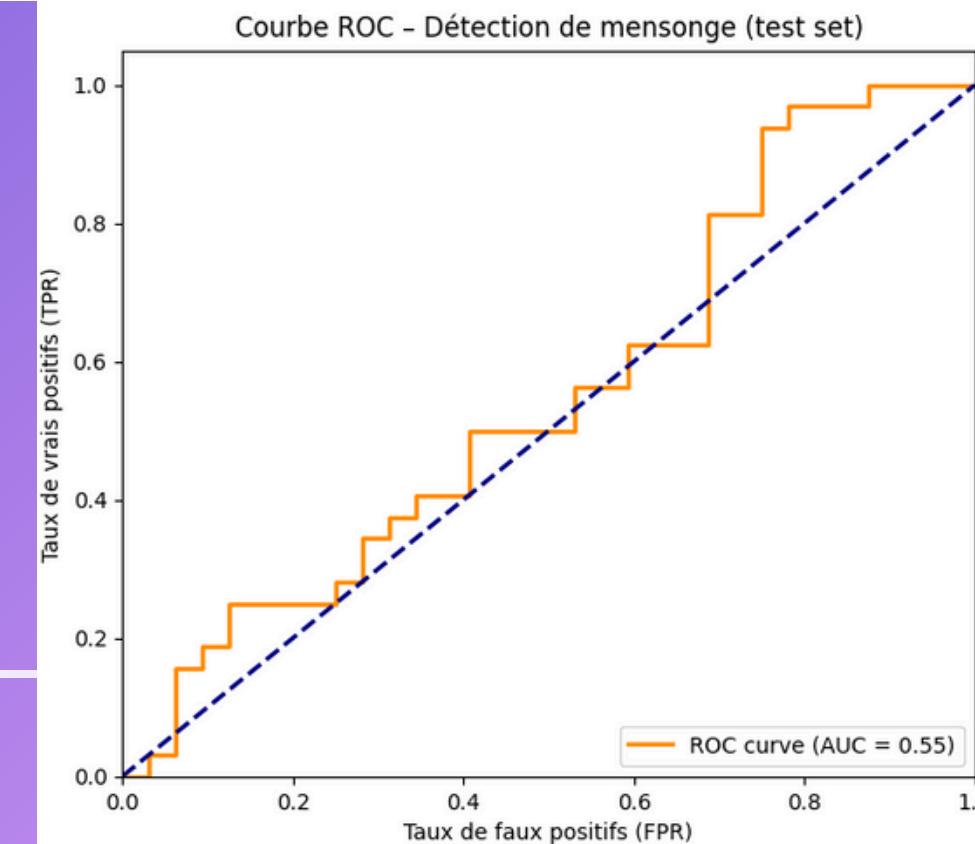
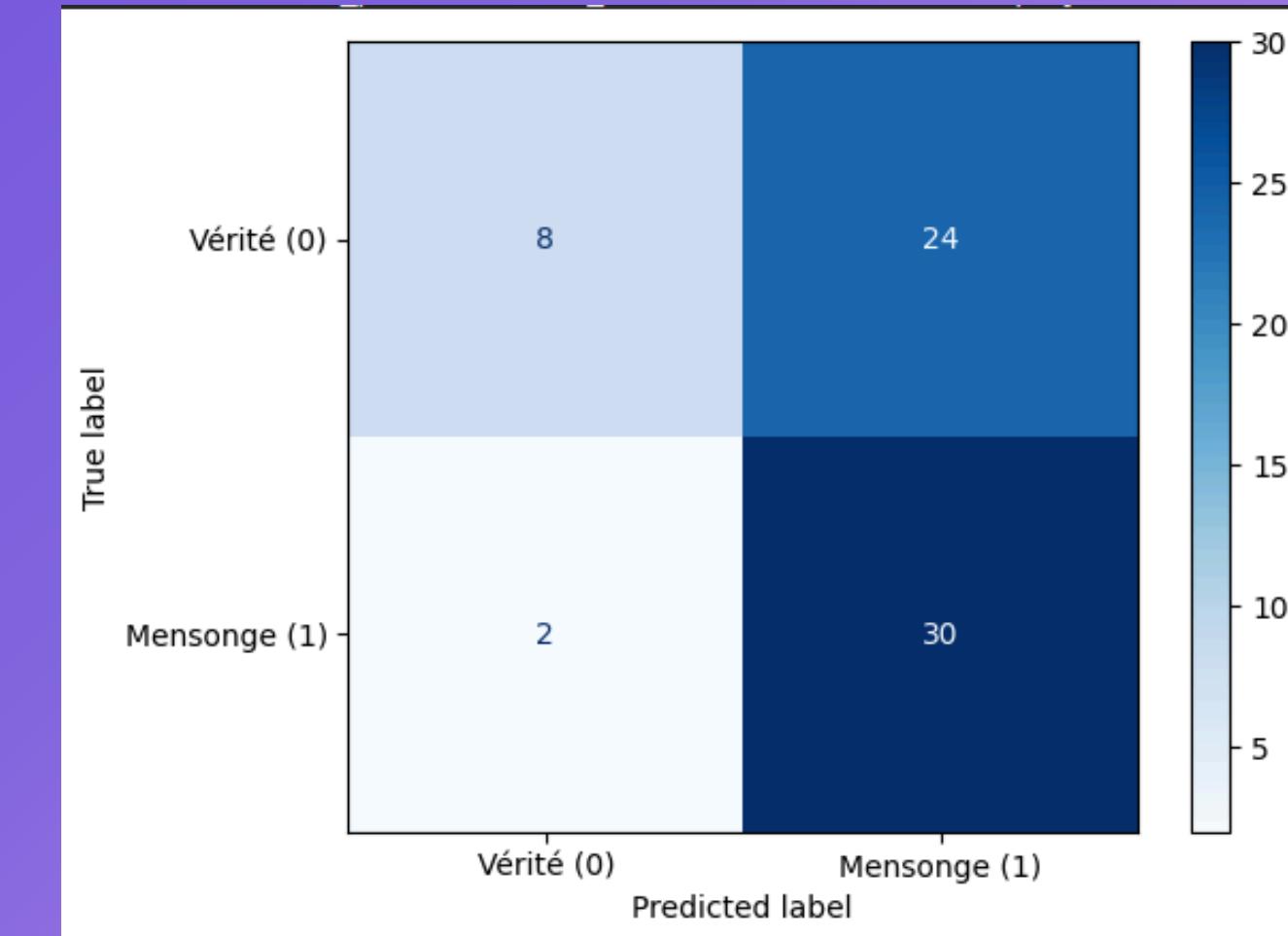
Optimisation avancée (Hyperparamètres)

Learning rates : 2e-5, 3e-5, 5e-5

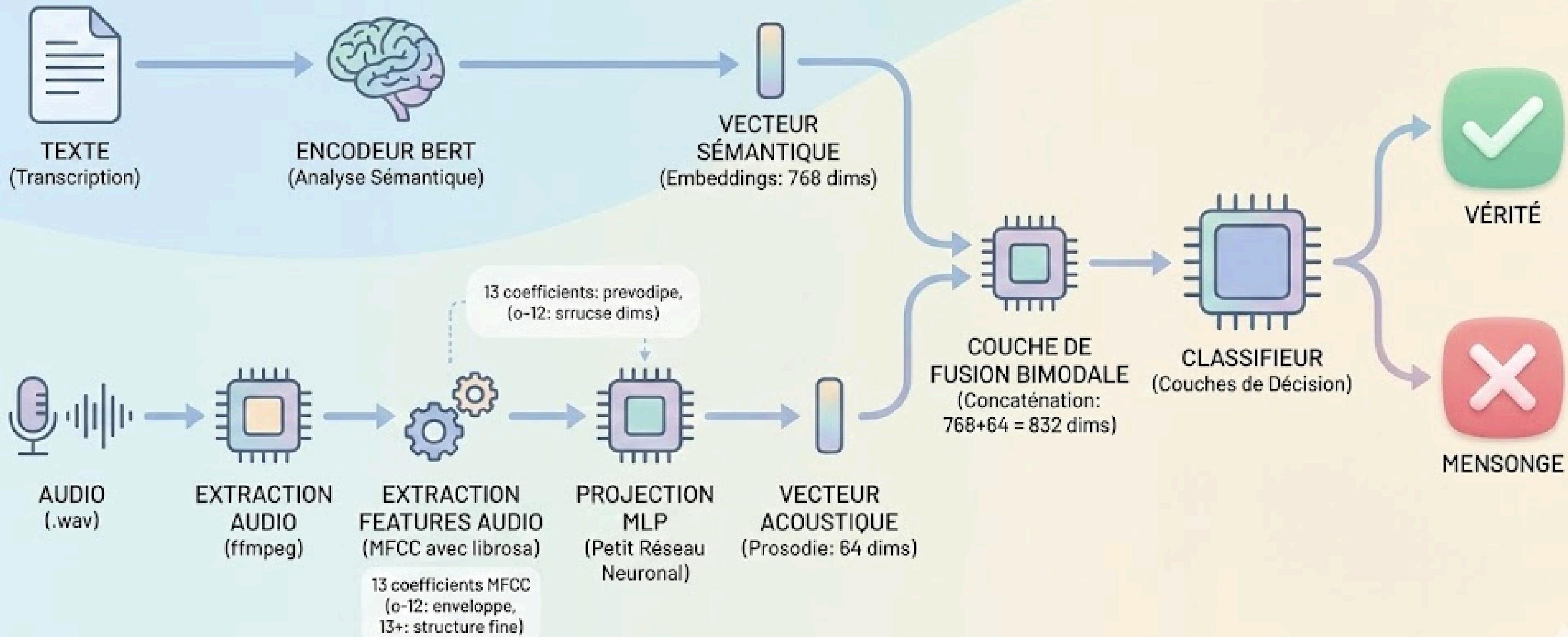
Batch sizes : 8, 16

Epochs max : 3, 5, 8

Évaluation finale



MODÈLE BIMODAL DE DÉTECTION DE MENSONGES (TEXTE + AUDIO)



Modèle multimodal DistilBERT + Wav2Vec2 (embeddings)

Objectif :

DistilBert pour la transcription

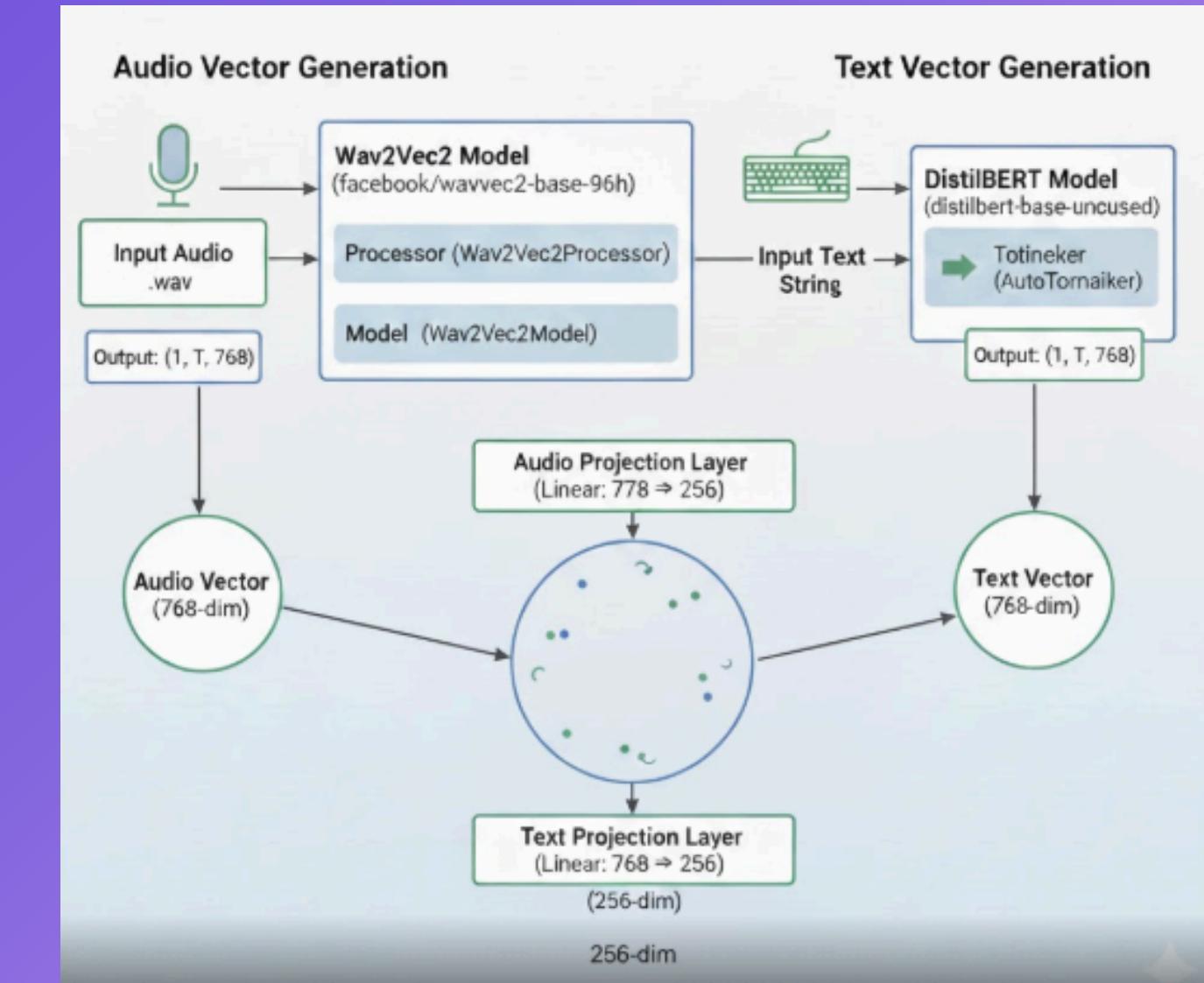
- Wav2Vec2 pour l'audio_emb (768 D apres resampling en 16khz)

DistilBert pour le text_emb (768D)

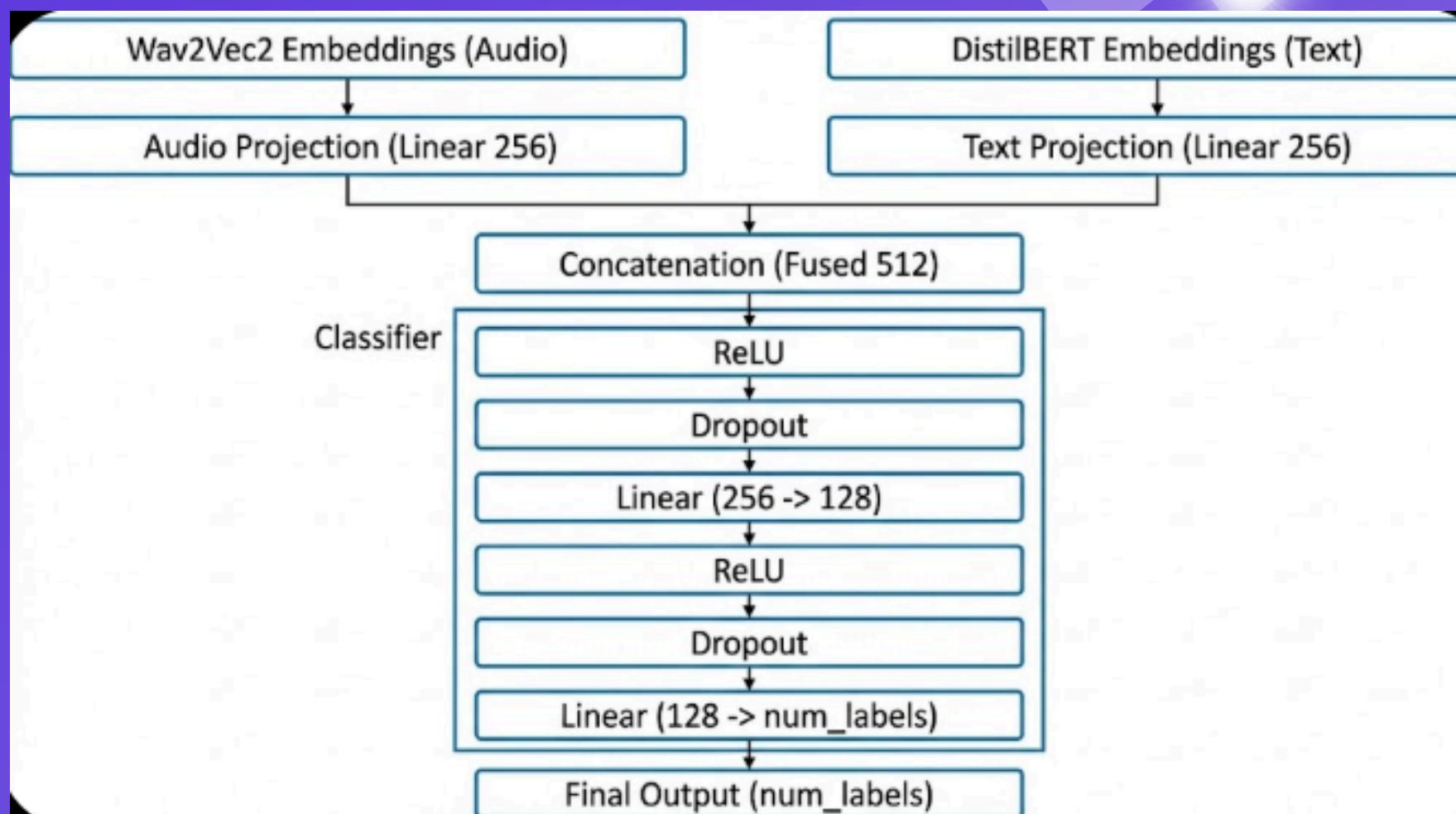
text_emb (768)	audio_emb (768)	label
[0.12, -0.44, 0.88, ...]	[-0.02, 0.33, 0.51, ...]	0
[0.07, 0.56, -0.21, ...]	[0.12, -0.03, -0.44, ...]	1

Entraînement (70%)

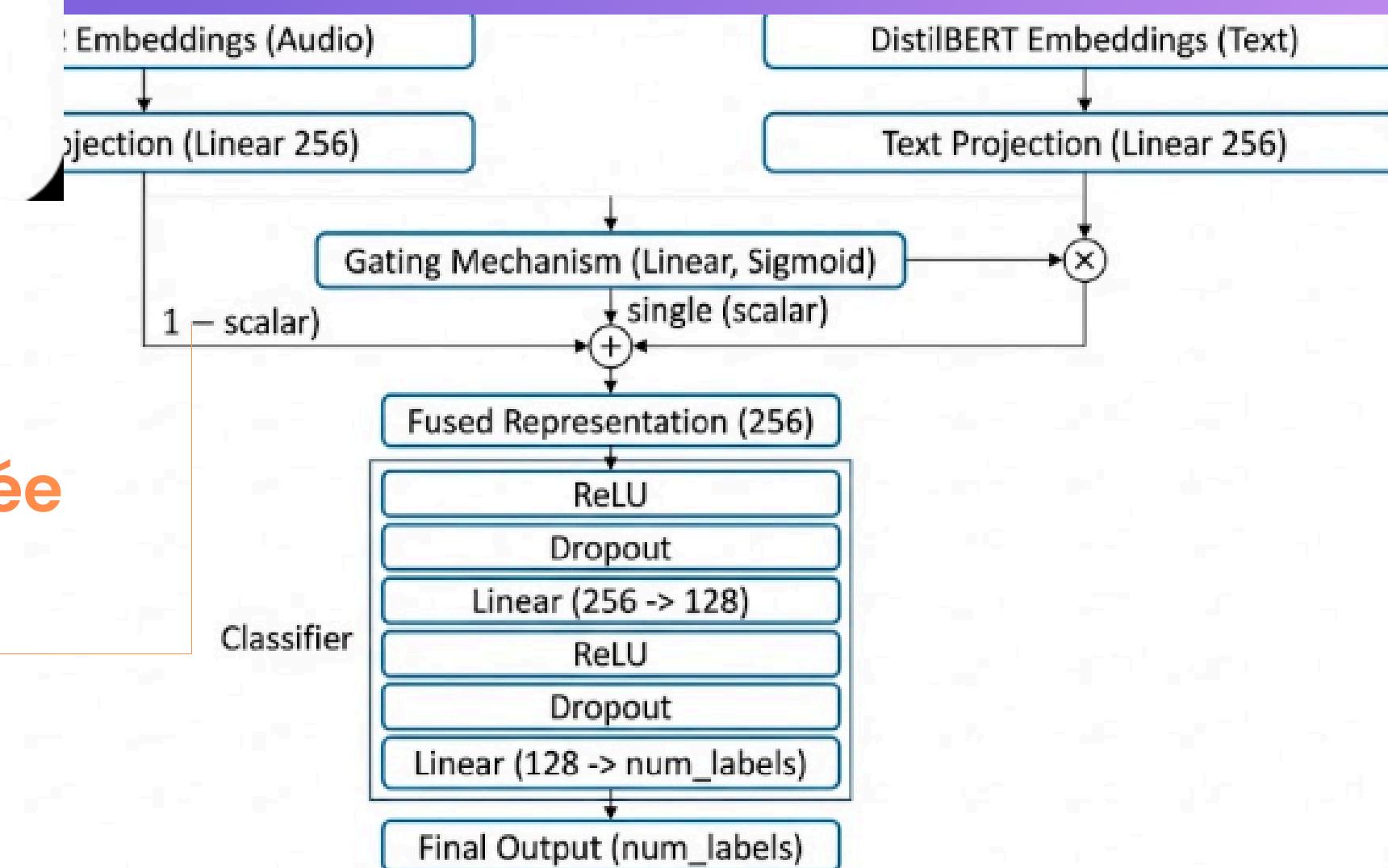
- Modèle audio : facebook/wav2vec2-base-960h



Modèle multimodal DistilBERT + Wav2Vec2 (embeddings)

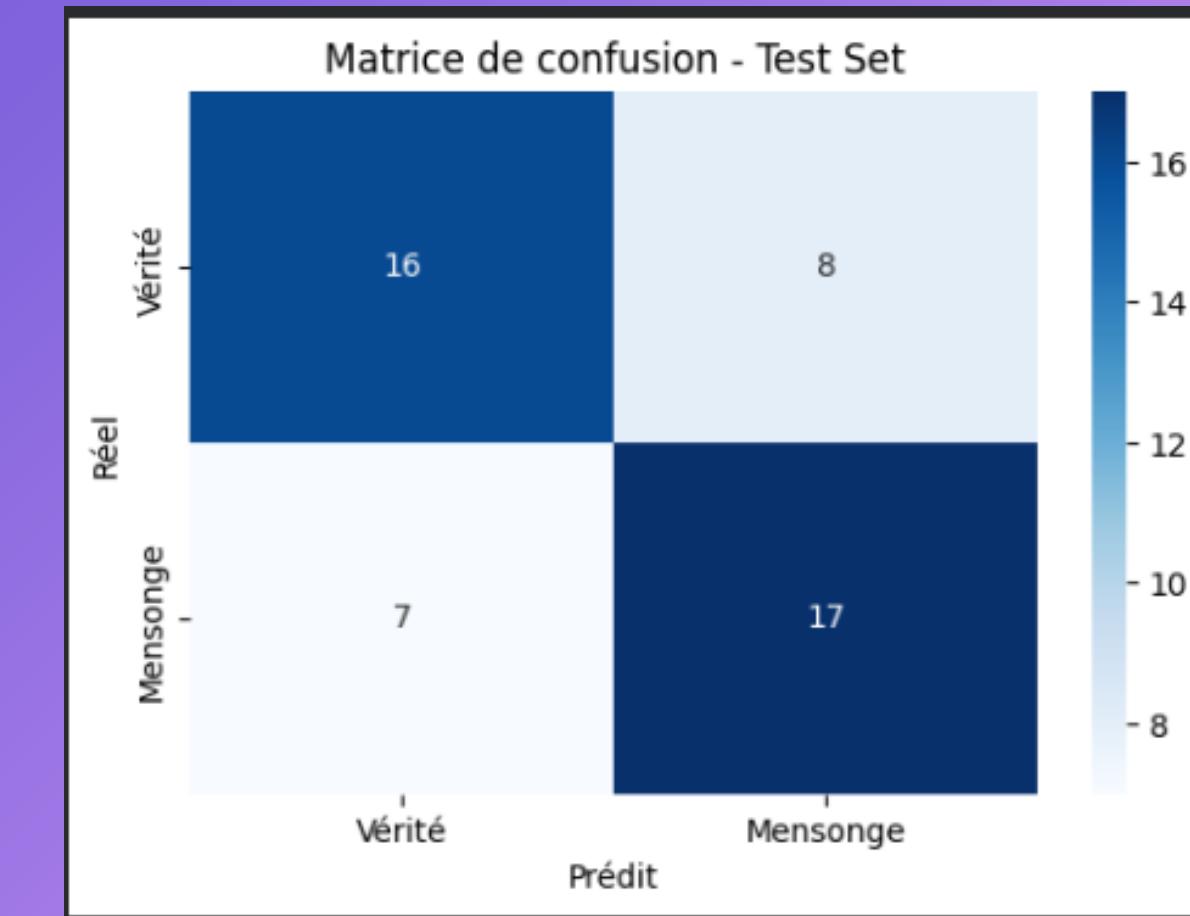
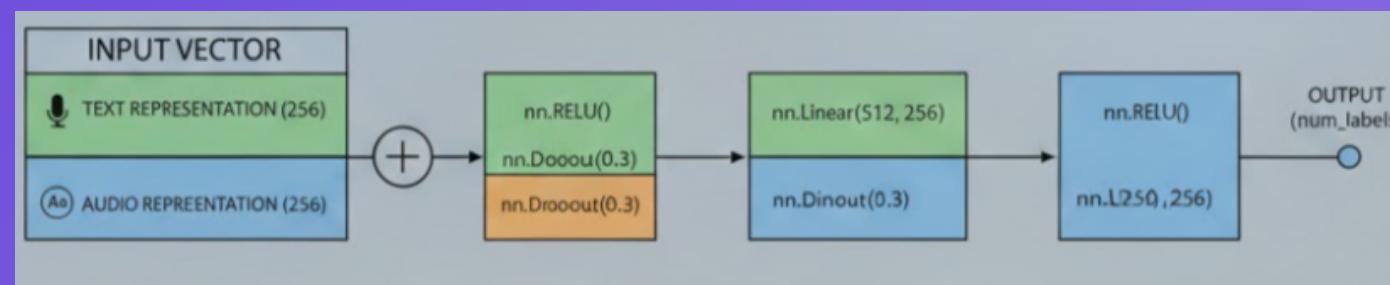
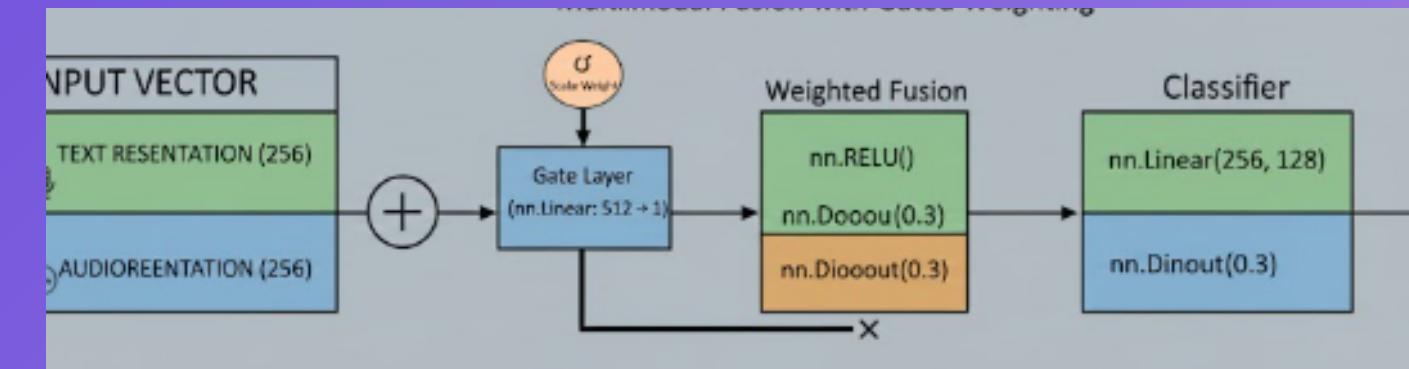
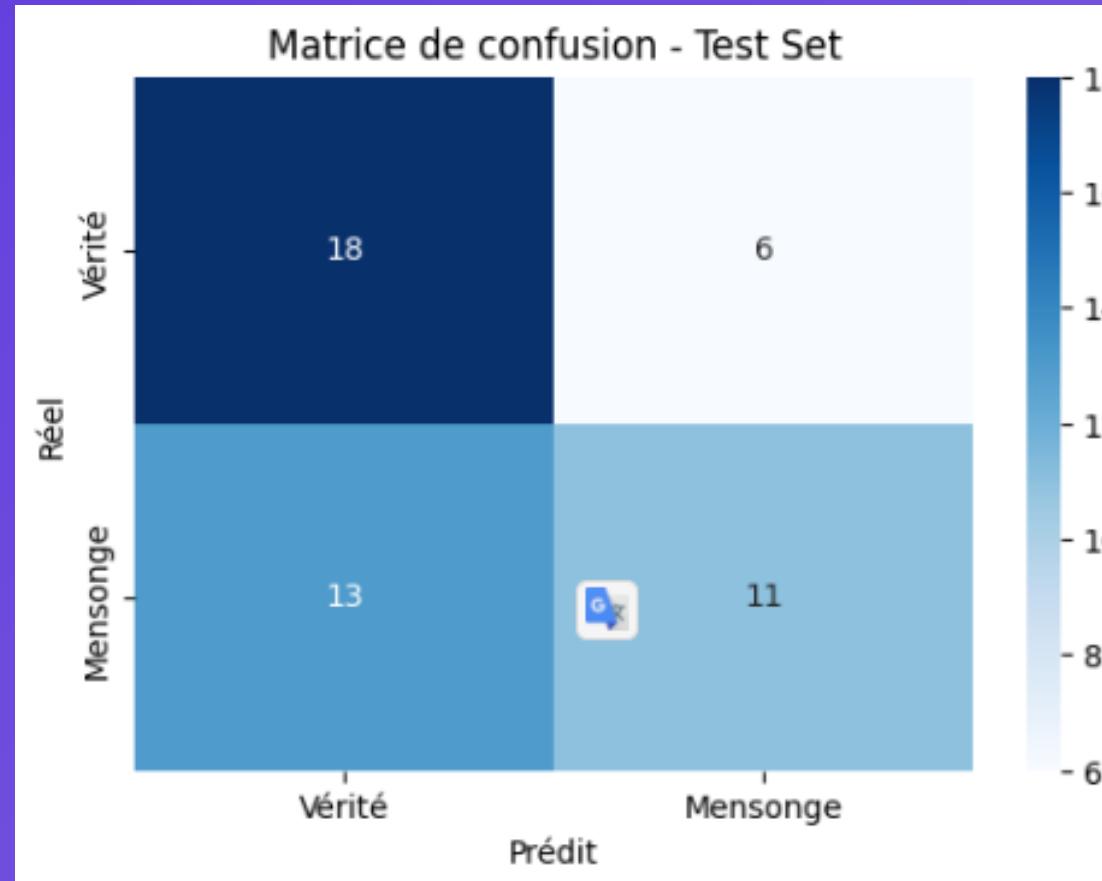


Concatenation simple



Fusion pondérée

Modèle multimodal DistilBERT + Wav2Vec2 (embeddings)



Détection multimodale de mensonge (Texte + Audio + Vidéo)

Objectif

Concevoir un modèle plus robuste qu'un modèle purement texte

→ en exploitant 3 modalités complémentaires :

- **Texte (DistilBERT)** → contenu linguistique
- **Audio (Wav2Vec2-base-960h)** → rythme, intonation, hésitations
- **Vidéo (ViT / CNN facial features)** → micro-expressions, regard, mouvement

Pipeline global – Étapes suivies

1. Préparation des données

Nettoyage du codebook MU3D

Alignment Videoid ↔ texte ↔ audio ↔ vidéo

Extraction :

Features audio : Wav2Vec2 → vecteurs 768D

Features vidéo : ViT-B/16 → features 512D

• Embeddings texte : DistilBERT → embeddings CLS 768D

Gating layer (attention scalaire)

$$\text{gate} = \text{softmax}(W \cdot [t, a, v])$$

Où :

- t = embedding texte projeté
- a = embedding audio projeté
- v = embedding vidéo projeté
- $\text{gate} = [w_{\text{text}}, w_{\text{audio}}, w_{\text{video}}]$

Fusion pondérée

$$h = w_t \cdot t + w_a \cdot a + w_v \cdot v$$

Puis un MLP fait la classification vérité/mensonge.

RESULTATS



Ce qu'on observe :

- Le modèle détecte très bien la Vérité :
→ 20 sur 24 bien classées
- Le modèle a du mal à détecter le Mensonge :
→ 12 mensonges détectés,
→ 12 mensonges confondus avec Vérité même avec 3 modalités.

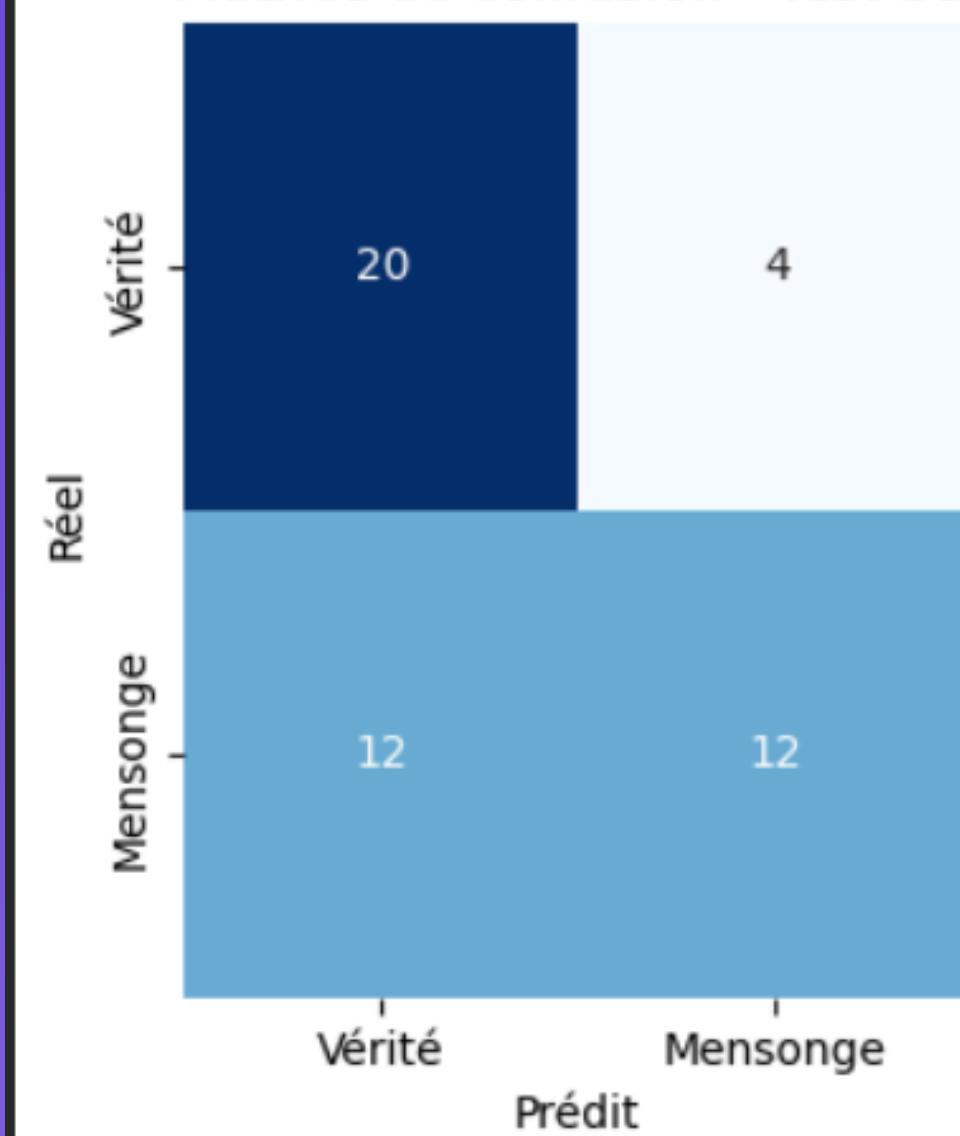
== Résultats test ==

Loss : 1.6725337704022725

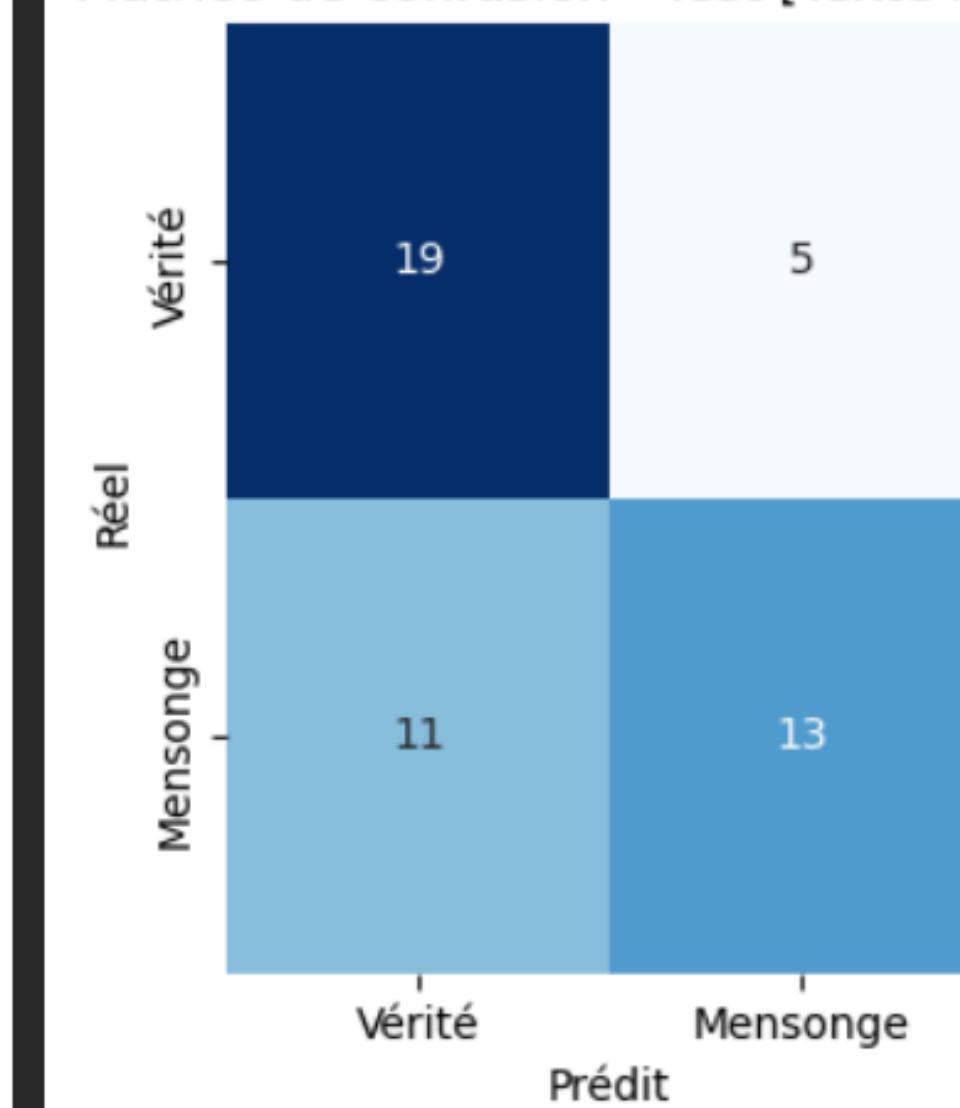
Accuracy : 0.6666666666666666

	precision	recall	f1-score	support
Vérité	0.62	0.83	0.71	24
Mensonge	0.75	0.50	0.60	24
accuracy			0.67	48
macro avg	0.69	0.67	0.66	48
weighted avg	0.69	0.67	0.66	48

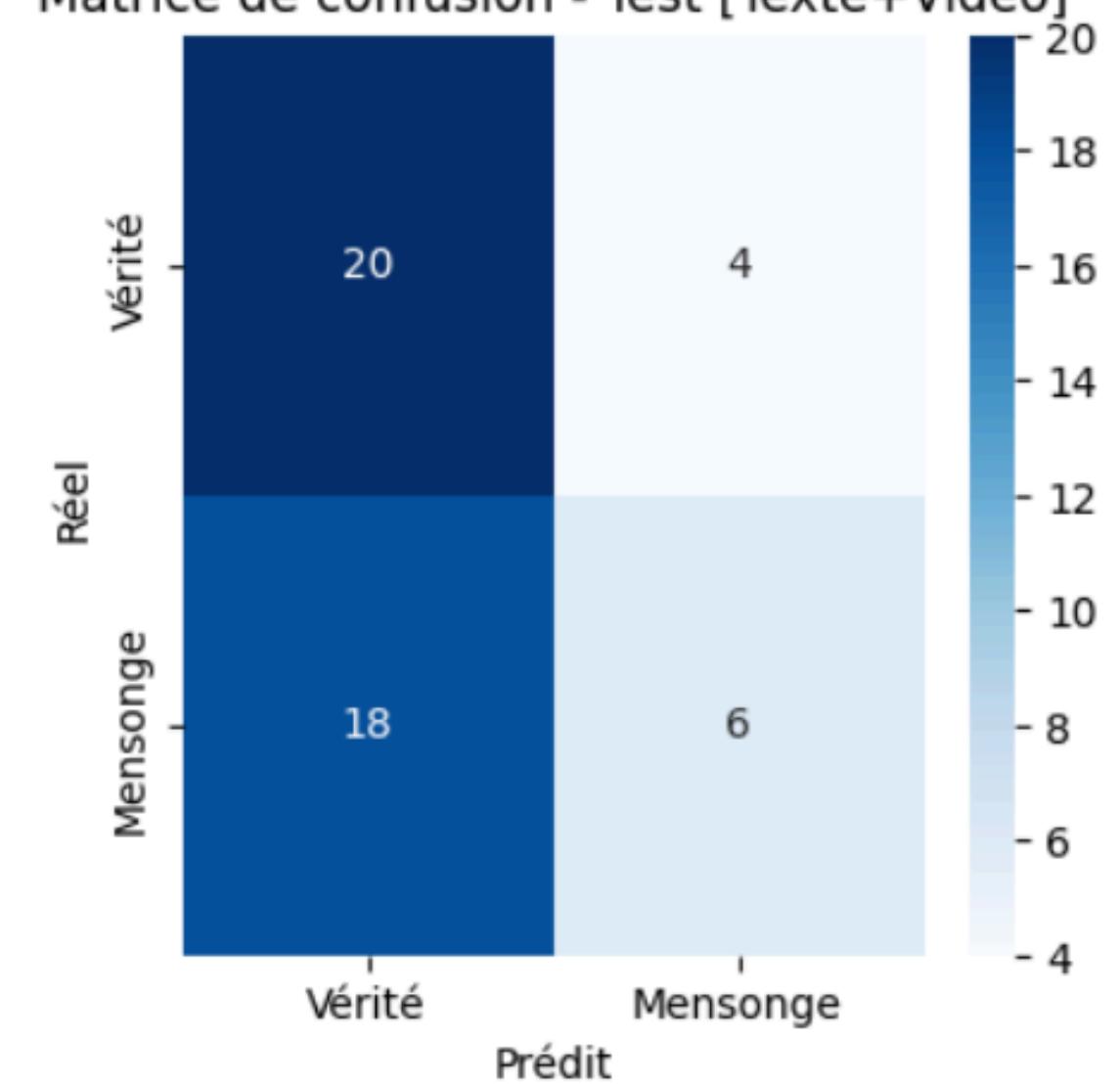
Matrice de confusion - Test Set



Matrice de confusion - Test [Texte+Audio]



Matrice de confusion - Test [Texte+Vidéo]



Synthèse interprétable

« La matrice de confusion montre que le modèle reconnaît très bien la classe Vérité, mais reste hésitant sur la classe Mensonge, avec de nombreux faux négatifs.

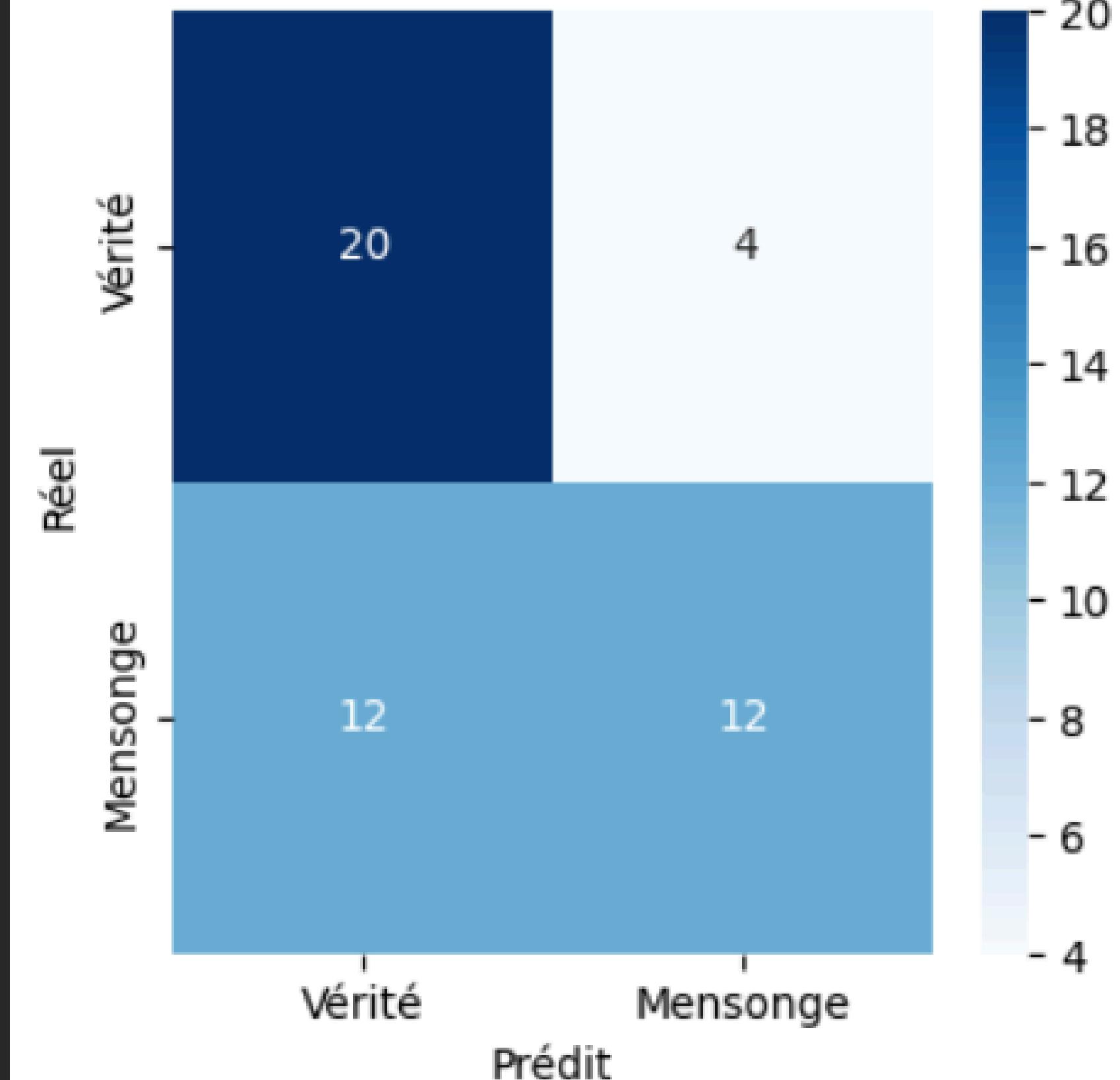
Cependant, l'analyse ACC@k révèle que le modèle détecte malgré tout des signaux de mensonge :

même lorsqu'il prédit Vérité, la classe Mensonge apparaît dans les classes les plus probables dans plus de 85% des cas (ACC@7 = 0.857).

Cela signifie que le modèle perçoit les indices du mensonge, mais ne les juge pas assez forts pour en faire la classe finale.

C'est typique d'un modèle qui manque encore de données ou d'une meilleure calibration. »

Matrice de confusion - Test Set



ACC@2 = 0.500

ACC@5 = 0.800

ACC@7 = 0.857



THANK YOU For Attention

(Don't forget that it's ALL YOU NEED.)

