Fiche explicative – AudioLDM

🔹 **Nom du modèle :**
AudioLDM

🔹 **Type :**
Modèle de diffusion latent pour audio (Latent Diffusion Model appliqué au son)

🔹 **Développeur :**
University of Surrey (équipe de recherche sur l'audio IA)

🔹 **Date de sortie :**
Février 2023

🎯 Objectif

AudioLDM est un modèle de génération audio capable de créer des sons et des ambiances **à partir de descriptions textuelles** simples ("text-to-audio").
Il utilise une approche basée sur les **modèles de diffusion latente**, qui ont montré d'excellents résultats dans la génération d'images (ex : Stable Diffusion).

✅ Résultat : AudioLDM permet de générer des sons réalistes (musique, bruitages, ambiances) en réponse à un simple **prompt texte**.

🔍 Fonctionnement simplifié

Étape	Description
Entrée	Texte descriptif (ex : "sons d'oiseaux dans une forêt tropicale")
Encodage	Texte transformé en représentation latente
Génération	Diffusion latente pour produire un spectrogramme
Reconstruction	Vocodeur reconvertit le spectrogramme en audio .wav

Techniques utilisées :

- **Latent Diffusion Models (LDM)** adaptés à l'audio
- **CLAP** (Contrastive Language-Audio Pretraining) pour aligner le texte et le son
- **HiFi-GAN** comme vocodeur pour convertir les spectrogrammes en signaux audio

---

🎼 Applications concrètes

- **Génération de bruitages** réalistes pour films, jeux vidéo, expériences VR
- **Création de courtes ambiances sonores** thématiques
- **Design sonore assisté par IA** pour artistes et sound designers
- **Exploration musicale** pour expérimenter de nouvelles textures sonores

---

🧩 Exemples d’usage

Domaine	Exemple
Films / Documentaires	Générer une ambiance réaliste de jungle ou de ville futuriste
Jeux vidéo	Produire des bruitages dynamiques adaptés aux actions du joueur
Musique expérimentale	Créer des textures sonores à partir de concepts textuels ("sons liquides", "vent numérique")

⚙️ Détails techniques

Caractéristique	Valeur
Architecture	Latent Diffusion Model (inspiré de Stable Diffusion)
Framework	PyTorch
Vocodeur utilisé	HiFi-GAN
Encodage texte/audio	CLAP model (pré-entraîné)
Dataset d'entraînement	AudioCaps, Freesound, UrbanSound8K
Résolution audio	16kHz (standard pour sons environnementaux)

📚 Ressources officielles et utiles

- 📄 Publication scientifique officielle sur arXiv (AudioLDM)
- 🛠️ Code source AudioLDM (GitHub officiel)
- 📚 CLAP : Modèle texte-audio utilisé

🚀 Démonstrations & alternatives pratiques

Google Colab utilisables aujourd'hui

- 📚 Colab AudioLDM Demo (officiel)

✍️ Exemple de code simple pour générer du son avec AudioLDM

https://colab.research.google.com/drive/1G-G2CXCJD6yGFHAxcmcSu8aioU18TMiZ

⚖️ Tableau des avantages / inconvénients

✅ Avantages	❌ Inconvénients
Génère des sons complexes à partir d'un simple texte	Limité à de courtes durées (quelques secondes)
Grande variété sonore possible (musique, bruitages, ambiances)	Sons parfois flous ou bruités si le prompt est trop vague
Démo Colab facile à utiliser	Pas encore optimisé pour générer de longues musiques structurées
Basé sur des techniques modernes de diffusion	Nécessite GPU pour des temps de génération rapides