Évaluation comparative d'AlphaAudio-v0 pour la reconnaissance automatique de la parole en français

Théo Hubert, *AlphaEdge*

Mars 2026 (v1.0)

Résumé

Ce document présente une évaluation comparative d'AlphaAudio-v0 sur deux corpus publics de parole en français, Common Voice v24 et MLS. Les modèles sont comparés selon deux axes : le temps d'inférence et le taux d'erreur par mot (Word Error Rate, WER), après application d'un protocole commun de normalisation textuelle. Dans le cadre expérimental retenu, AlphaAudio-v0 obtient la latence la plus faible sur les deux corpus évalués, avec un avantage d'environ 5× par rapport a Whisper Large v3. En précision, il atteint le meilleur WER sur Common Voice et le deuxième meilleur WER sur MLS, avec un écart limité de 0,36 point par rapport au meilleur score observé. Ces résultats mettent en évidence un compromis précision–latence très favorable et situent AlphaAudio-v0 parmi les systèmes les plus performants de ce benchmark pour l'ASR en français.

1 Introduction

Le présent document rapporte un benchmark du modèle AlphaAudio-v0 en reconnaissance automatique de la parole (ASR) pour le français. L'objectif est de situer ce modèle par rapport à un ensemble de systèmes de référence selon deux critères centraux en recherche appliquée comme en contexte industriel : la qualité de transcription et le temps d'inférence.

AlphaAudio-v0 est un modèle développé par AlphaEdge. Dans notre terminologie interne, il appartient à une famille de modèles dits *Efficient Language Models* (ELM). Contrairement aux approches où la réduction de taille implique une dégradation des performances, les ELM exploitent cette contrainte comme un levier, de sorte qu'une architecture plus compacte puisse conduire à de meilleures performances. Le présent rapport se concentre toutefois sur l'évaluation empirique, il n'a pas pour objet de détailler l'architecture interne du modèle.

L'ambition de ce benchmark n'est pas de couvrir exhaustivement l'ensemble des systèmes ASR existants, mais de fournir une comparaison documentée sur deux jeux de données publics couramment utilisés pour l'évaluation du français parlé. Les résultats rapportés ici montrent qu'un modèle conçu pour l'efficience peut néanmoins atteindre un niveau de performance élevé sur des corpus de référence en français.

2 Protocole expérimental

2.1 Jeux de données

L'évaluation repose sur deux jeux de données publics :

— **Common Voice v24 (fr)** : l'ensemble de validation complet, soit 16 200 échantillons. Ce corpus présente une forte variabilité inter-locuteurs, d'accents et de conditions d'enregistrement.

— **MLS (Multilingual LibriSpeech, fr)** : l'ensemble de validation complet, soit 2 200 échantillons. Ce corpus est issu de livres audio et correspond à un cadre acoustique plus contrôlé.

L'usage conjoint de ces deux corpus permet de comparer les modèles dans un environnement hétérogène et bruité (Common Voice) et un environnement plus homogène (MLS).

2.2 Modèles comparés

Les modèles comparés dans ce benchmark sont les suivants : AlphaAudio-v0, Whisper Small, Whisper Large v3, Whisper Large v3 Turbo, NVIDIA Canary, Qwen3-ASR 0.6B, Qwen3-ASR 1.7B, Voxtral Mini v2 (API) et Gladia async (API, voir section 4).

2.3 Normalisation des transcriptions

Afin d'assurer une comparaison homogène des sorties, les transcriptions de référence et les hypothèses de tous les modèles ont été soumises au même protocole de normalisation textuelle. Plus précisément :

— toutes les chaînes sont converties en bas de casse (ex. : « Bonjour » → « bonjour »),
— les échantillons contenant des parenthèses ou des crochets sont exclus,
— les formules de civilité ainsi que certaines liaisons par tiret ou élision sont régularisées (ex. : « M. » → « monsieur », « y-a » → « il y a »),
— la ponctuation et les marqueurs de disfluence du type « euh » sont retirés,
— les nombres, unités et pourcentages sont ramenés à des formes canoniques (ex. : « vingt pour cent » → « 20 % »)

Ce protocole vise à limiter les écarts imputables à des conventions d'écriture plutôt qu'à la reconnaissance elle-même.

2.4 Environnement d'exécution

Tous les modèles exécutés localement ont été évalués sur le même matériel, à savoir un GPU NVIDIA RTX 6000 Pro. Aucune optimisation spécifique n'a été appliquée à un modèle particulier, afin de conserver des conditions de comparaison aussi homogènes que possible. Toutes les mesures locales ont été réalisées avec un batch effectif de 1 (*batch size* = 1).

Les précisions numériques utilisées pour les modèles exécutés localement sont résumées dans le tableau 1.

Modèle	**Précision numérique**
AlphaAudio-v0	float32
Whisper Small	float32
NVIDIA Canary	float32
Whisper Large v3 Turbo	float32
Qwen3-ASR 0.6B	bfloat16
Qwen3-ASR 1.7B	bfloat16
Whisper Large v3	float32

TABLE 1 – Précision numérique utilisée pour les modèles exécutés localement.

Les modèles accessibles uniquement via API (notamment Voxtral Mini v2 et Gladia) ne donnent pas accès au matériel sous-jacent. Pour ces systèmes, les temps reportés correspondent au temps minimal renvoyé par le service, après exclusion de la latence réseau côté client.

2.5 Métriques

Deux métriques sont reportées :

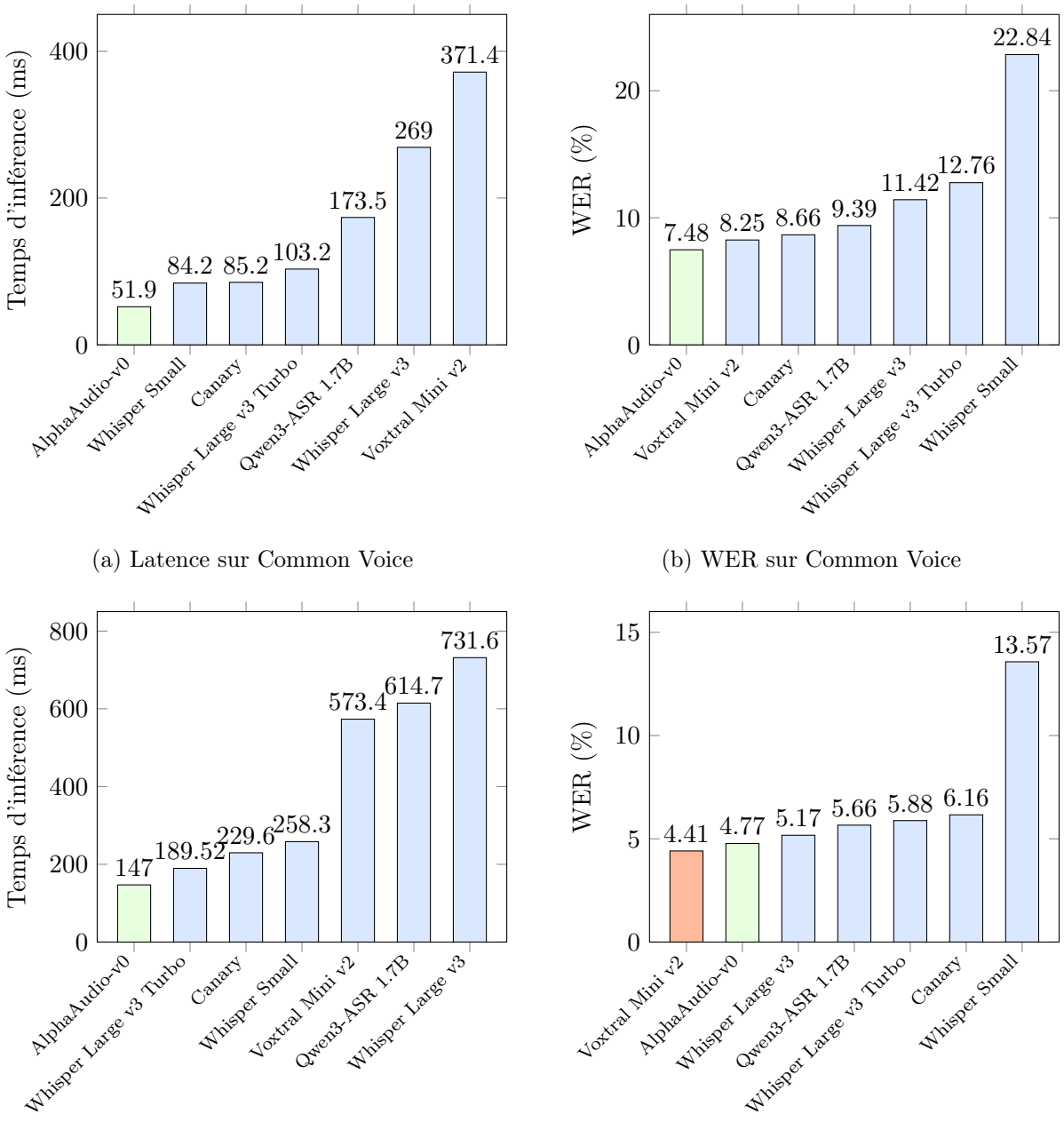
— **Temps d'inférence** : mesuré en millisecondes,
— **WER** (*Word Error Rate*) : taux d'erreur par mot, en pourcentage,
— **CER** (*Character Error Rate*) : taux d'erreur par mot, en pourcentage.

3 Résultats

3.1 Synthèse des observations

Deux résultats ressortent de cette évaluation. Premièrement, **AlphaAudio-v0 présente la latence la plus faible sur les deux corpus testés**. Sur Common Voice, le temps mesuré est de 51,9 ms, contre 84,2 ms pour le deuxième modèle le plus rapide évalué dans cette comparaison. Sur MLS, AlphaAudio-v0 atteint 147,0 ms, contre 229,6 ms pour le deuxième modèle le plus rapide.

Deuxièmement, **les performances de transcription restent de premier plan**. Sur Common Voice, AlphaAudio-v0 obtient le meilleur WER observé (7,48 %). Sur MLS, il obtient le deuxième meilleur WER (4,77 %), derrière Voxtral Mini v2 (4,41 %), avec un écart absolu limité à 0,36 point.



(a) Latence sur Common Voice

(b) WER sur Common Voice

(c) Latence sur MLS

(d) WER sur MLS

FIGURE 1 – Comparaison du temps d'inférence et du WER sur Common Voice v24 et MLS

3.2 Mesures détaillées

Temps d'inférence

Modèle	Common Voice (ms)	Écart absolu vs Alpha	MLS (ms)	Écart absolu vs Alpha
AlphaAudio-v0	**51,9**	—	**147,0**	—
Whisper Small	84,2	+32,3	258,3	+111,3
NVIDIA Canary	85,2	+33,3	229,6	+82,6
Whisper Large v3 Turbo	103,2	+51,3	189,52	+42,52
Qwen3-ASR 0.6B	164,1	+112,2	573,3	+426,3
Qwen3-ASR 1.7B	173,5	+121,6	614,7	+467,7
Whisper Large v3	269,0	+217,1	731,6	+584,6
Voxtral Mini v2 (API)	371,4	+319,5	573,4	+426,4

TABLE 2 – Temps d'inférence mesurés dans le protocole expérimental retenu.

Performances de transcription (WER)

Modèle	Common Voice (%)	Écart absolu vs Alpha	MLS (%)	Écart absolu vs Alpha
AlphaAudio-v0	**7,48**	—	4,77	—
Voxtral Mini v2 (API)	8,25	+0,77	**4,41**	-0,36
NVIDIA Canary	8,66	+1,18	6,16	+1,39
Qwen3-ASR 1.7B	9,39	+1,91	5,66	+0,89
Whisper Large v3	11,42	+3,94	5,17	+0,40
Whisper Large v3 Turbo	12,76	+5,28	5,88	+1,11
Qwen3-ASR 0.6B	13,42	+5,94	9,55	+4,78
Whisper Small	22,84	+15,36	13,57	+8,80

TABLE 3 – Taux d'erreur par mot (WER).

Performances de transcription (CER)

Modèle	Common Voice (%)	Écart absolu vs Alpha	MLS (%)	Écart absolu vs Alpha
AlphaAudio-v0	**2,95**	—	1,84	—
Mistral Voxtral (API)	3,05	+0,10	**1,79**	-0,05
NVIDIA Canary	2,98	+0,03	2,32	+0,48
Qwen3-ASR 1.7B	3,43	+0,48	2,30	+0,46
Whisper Large v3	4,20	+1,25	2,31	+0,47
Whisper Large v3 Turbo	4,76	+1,81	2,75	+0,91
Qwen3-ASR 0.6B	5,12	+2,17	3,74	+1,90
Whisper Small	9,29	+6,34	5,25	+3,40

TABLE 4 – Taux d'erreur par caractère (CER).

Analyse ciblée : Whisper Small. Whisper Small constitue un point de comparaison utile pour analyser le compromis précision–latence sur un modèle de taille plus réduite. Dans ce benchmark, il reste plus lent qu'AlphaAudio-v0 sur les deux corpus (84,2 ms contre 51,9 ms sur Common Voice ; 258,3 ms contre 147,0 ms sur MLS) tout en présentant un WER nettement plus élevé (22,84 % contre 7,48 % sur Common Voice, 13,57 % contre 4,77 % sur MLS). Dans cette configuration expérimentale, la réduction de taille du modèle ne se traduit donc pas par un meilleur compromis global.

4 Étude de cas : comparaison avec Gladia

La comparaison avec Gladia est fournie à titre indicatif sur un sous-échantillon de 500 échantillons par corpus. Le service étant accessible via API, le matériel sous-jacent n'est pas observable. Afin d'isoler au mieux le temps d'exécution du modèle, la latence réseau côté client a été retirée et le temps retenu correspond au temps minimal reporté par le service.

Dataset	**AlphaAudio-v0 (ms)**	**Gladia (ms)**	**Écart mesuré**
Common Voice 24.0	**74,5**	4195,4	+4120,9 ms
MLS	**164,3**	4928,9	+4764,6 ms

Table 5 – Comparaison de latence avec Gladia. Les temps reportés pour Gladia correspondent au temps minimal renvoyé par l'API après exclusion de la latence réseau côté client.

Dataset	**AlphaAudio-v0 (%)**	**Gladia (%)**	**Écart absolu**
Common Voice 24.0	**7,59**	12,20	+4,61 pt
MLS	**4,19**	4,74	+0,55 pt

Table 6 – Comparaison de précision (WER) avec Gladia.

Dataset	**AlphaAudio-v0 (%)**	**Gladia (%)**	**Écart absolu**
Common Voice 24.0	**2,95**	4,51	+1,56 pt
MLS	**1,83**	2,02	+0,19 pt

Table 7 – Comparaison de précision (CER) avec Gladia.

Sur ce sous-échantillon, AlphaAudio-v0 obtient un WER et un CER inférieur à celui de Gladia sur les deux corpus. L'écart est marqué sur Common Voice et plus contenu sur MLS. Malgré le caractère indirect de la comparaison matérielle, le résultat observé dans cette section reste nettement favorable à AlphaAudio-v0 sur les deux dimensions reportées. Il convient de préciser que Gladia n'est pas nativement conçu pour être évalué sur des échantillons audio courts issus de datasets académiques. En conséquence, les scores de temps d'inférence observés nous semblent peu représentatifs de ses performances réelles en conditions opérationnelles.

5 Conclusion

Dans le protocole expérimental retenu, AlphaAudio-v0 se distingue par un profil de performance particulièrement solide. Le modèle obtient la meilleure latence sur les deux jeux de données évalués et le meilleur WER sur Common Voice, tout en restant très proche du meilleur score sur MLS. Ce positionnement est notable car il associe une avance nette en vitesse à un niveau de précision de premier plan parmi les modèles comparés.

Pris ensemble, ces résultats montrent qu'un modèle conçu pour l'efficience peut atteindre un niveau élevé de performance en ASR français, sans se limiter à un compromis défavorable entre rapidité et exactitude. Dans ce benchmark, AlphaAudio-v0 apparaît ainsi comme l'un des systèmes les plus aboutis du point de vue du compromis précision–latence.

Une version multilingue du modèle, incluant notamment l'anglais, est actuellement en cours de préparation.