Actes de CORIA-TALN 2026
Nantes, France, 29 juin au 3 juillet 2026

Évaluation de la reconnaissance automatique de la parole par les grands modèles de langage génératifs

Thibault Bañeras-Roux[1], Shashi Kumar[1,2], Driss Khalil[1], Sergio Burdisso[1], Petr Motlicek[1,3], Shiran Liu[1], Mickael Rouvier[4], Jane Wottawa[5], Richard Dufour[6]
(1) Idiap Research Institute, Martigny, Switzerland
(2) EPFL, Lausanne, Switzerland
(3) Brno University of Technology, Czech Republic
(4) Avignon University, Avignon, France
(5) Le Mans University, Le Mans, France
(6) Nantes University, Nantes, France
thibault.roux@idiap.ch

RÉSUMÉ

La reconnaissance automatique de la parole (RAP) est traditionnellement évaluée par le taux d'erreur mot (WER), une métrique insensible au sens. Les métriques sémantiques basées sur les plongements sont mieux corrélées à la perception humaine, mais les grands modèles de langage (LLM) décodeurs restent peu explorés pour cette tâche. Cet article évalue leur pertinence selon trois approches : (1) sélection de la meilleure hypothèse parmi deux candidats, (2) calcul de distance sémantique via embeddings génératifs, et (3) classification qualitative des erreurs. Sur le jeu de données HATS, les meilleurs LLM atteignent 92–94 % d'accord avec les annotateurs humains pour la sélection d'hypothèses, contre 63 % pour le WER, surpassant aussi les métriques sémantiques. Les embeddings issus de LLM décodeurs montrent des performances comparables aux modèles encodeurs. Enfin, les LLM offrent une perspective prometteuse pour une évaluation interprétable et sémantique de la RAP.

ABSTRACT

Evaluation of Automatic Speech Recognition Using Generative Large Language Models

Automatic Speech Recognition (ASR) is traditionally evaluated using Word Error Rate (WER), a metric that is insensitive to meaning. Embedding-based semantic metrics are better correlated with human perception, but decoder-based Large Language Models (LLMs) remain underexplored for this task. This paper evaluates their relevance through three approaches : (1) selecting the best hypothesis between two candidates, (2) computing semantic distance using generative embeddings, and (3) qualitative classification of errors. On the HATS dataset, the best LLMs achieve 92–94% agreement with human annotators for hypothesis selection, compared to 63% for WER, also outperforming semantic metrics. Embeddings from decoder-based LLMs show performance comparable to encoder models. Finally, LLMs offer a promising direction for interpretable and semantic ASR evaluation.

MOTS-CLÉS : Reconnaissance automatique de la parole, grands modèles de langage, similarité sémantique, modèles génératifs, perception humaine.

KEYWORDS: Automatic Speech Recognition, Large Language Models, semantic similarity, generative models, human perception.

Actes de l'atelier sur l'évaluation des modèles génératifs (LLM) et challenges (EvalLLM 2026), pages 75–86.

1 Introduction

La reconnaissance automatique de la parole (RAP) est une technologie fondamentale pour les applications de contrôle vocal, de transcription et d'accessibilité. L'évaluation fiable des systèmes de RAP est essentielle pour mesurer les progrès et orienter l'amélioration des modèles. Historiquement, le taux d'erreur mot (WER pour *Word Error Rate*) a dominé ce paysage, mais cette métrique présente des limitations bien connues : elle est rigide et sensible à la casse, ne capturant pas nécessairement les informations importantes pour la tâche ou pour la perception humaine.

Face à ces enjeux, plusieurs alternatives ont été proposées. Les métriques sémantiques (Kim *et al.*, 2021; Zhang *et al.*, 2019) basées sur les plongements de mots (*embeddings*) ont montré des performances prometteuses, en particulier celles exploitant les modèles d'encodeurs contextuels comme BERT (Devlin *et al.*, 2019). Ces approches capturent mieux les nuances sémantiques du langage, ce qui explique leur meilleure corrélation avec les jugements humains. Cependant, alors que les encodeurs ont été souvent explorés pour cette tâche (Bañeras-Roux *et al.*, 2022; Kim *et al.*, 2021; Le *et al.*, 2016; Gordeeva *et al.*, 2021), les grands modèles de langage (LLM) de type décodeur – tels que GPT (Radford *et al.*, 2018), Llama (Touvron *et al.*, 2023) ou Gemma (Team *et al.*, 2024) – n'ont reçu que peu d'attention. Ces modèles ont pourtant démontré des capacités remarquables et polyvalentes dans une vaste gamme de tâches (Ahuja *et al.*, 2023; Labrak *et al.*, 2024) suggérant qu'ils pourraient offrir des perspectives novatrices pour l'évaluation de la RAP. Au-delà de l'évaluation des transcriptions, la sélection efficace de données d'entraînement demeure un défi crucial pour l'adaptation de modèles ASR à de nouveaux domaines, particulièrement avec des ressources limitées (Rangappa *et al.*, 2025b,a).

Cette étude explore précisément les capacités des LLM décodeurs dans le cadre spécifique de l'évaluation des systèmes de reconnaissance automatique de la parole, selon deux axes complémentaires. Le premier axe s'intéresse au paradigme « LLM comme juge » : peut-on tirer parti de la compréhension contextuelle et sémantique profonde des LLM pour évaluer directement la qualité des hypothèses de transcription ? Le second axe examine la qualité des représentations vectorielles extraites de ces modèles : comment les embeddings issus de décodeurs LLM se comparent-ils à ceux des encodeurs pour capturer les similarités sémantiques pertinentes ?

Nos expériences reposent sur le jeu de données HATS (Bañeras-Roux *et al.*, 2023), qui offre des annotations humaines sur les erreurs de reconnaissance de la parole et a montré que les métriques sémantiques corrèlent significativement mieux avec la perception humaine que le WER. Nous évaluons (1) la capacité des LLM à sélectionner la meilleure hypothèse parmi deux candidates, (2) l'efficacité de différentes stratégies d'agrégation (pooling) appliquées aux embeddings de LLM pour construire une métrique de similarité sémantique, et (3) la capacité des LLM à assigner des étiquettes qualitatives aux erreurs de transcription. Nos résultats révèlent que les meilleurs LLM surpassent non seulement le WER et le CER, mais aussi les métriques sémantiques sophistiquées basées sur des encodeurs, tout en offrant une meilleure interprétabilité des erreurs.

2 État de l'art

2.1 Limitations du taux d'erreur mot

Comme dit précédemment, le taux d'erreur mot a été la métrique de référence pour évaluer les systèmes de reconnaissance automatique de la parole. Pourtant, le WER présente des limitations importantes bien documentées dans la littérature (Wang *et al.*, 2003; Morris *et al.*, 2004; He *et al.*, 2011). Premièrement, cette métrique opère au niveau lexical sans aucune considération du contexte sémantique, traitant de manière identique les erreurs graves affectant la compréhension et les erreurs mineures sans impact sur le sens. Deuxièmement, le WER ne corrèle pas nécessairement avec les tâches en aval (Favre *et al.*, 2013) qui dépendent des transcriptions produites, suggérant qu'une amélioration du WER ne garantit pas une amélioration perceptible pour l'utilisateur final. Enfin, plusieurs études ont démontré une faible corrélation entre le WER et la perception humaine de la qualité des transcriptions (Kim *et al.*, 2022; Thennal *et al.*, 2025), remettant en question son utilité pour évaluer la performance réelle des systèmes de RAP destinée à la lecture humaine.

Ces limitations ont des implications pratiques importantes. En effet, des études ont montré que le classement des systèmes de RAP peut différer de manière significative selon la métrique utilisée pour l'évaluation (Bañeras-Roux *et al.*, 2024a). Pour un objectif spécifique – par exemple, générer des sous-titres à lire par des humains – il peut être plus pertinent d'utiliser une métrique qui corrèle davantage avec la perception humaine plutôt que d'optimiser un WER qui ne capture pas les nuances perceptives pertinentes (Kafle & Huenerfauth, 2017; Nam & Fels, 2019).

2.2 Métriques sémantiques basées sur les plongements

Face aux limitations du WER, plusieurs alternatives ont été proposées. Parmi celles-ci, les métriques sémantiques fondées sur les plongements de mots se sont avérées particulièrement prometteuses. Ces approches exploitent les représentations vectorielles du langage obtenues par des modèles de plongement de mots ou de contexte, permettant de capturer les similarités sémantiques entre la référence et l'hypothèse de transcription.

Une première génération de ces métriques a exploité les plongements de mots statiques pour améliorer l'évaluation de la RAP en contexte de la traduction automatique (Le *et al.*, 2016). Plus récemment, l'émergence de modèles contextuels puissants, notamment les encodeurs bidirectionnels comme BERT et ses variantes multilingues, a permis le développement de métriques sémantiques plus sophistiquées. Ces modèles contextuels offrent des représentations plus riches, sensibles à la position des mots et au contexte entourant chaque unité linguistique. Parmi les approches sémantiques notables, la métrique *SemDist* (Kim *et al.*, 2021) propose une distance dérivée de la similarité cosinus entre les embeddings d'une référence et d'une hypothèse. Cette métrique a démontré une corrélation significativement meilleure avec la perception humaine comparée au WER, validant l'hypothèse que la sémantique est un facteur clé pour l'évaluation perceptive de la qualité des transcriptions.

McCowan *et al.* définit la métrique idéale pour la reconnaissance automatique de la parole sur plusieurs critères dont l'interprétabilité. Le problème étant que les métriques sémantiques sont principalement construites à partir de similarités cosinus entre plongements, ce qui produit des scores comme par exemple : 0,62, 0,75 ou ou 0,81. Contrairement à un pourcentage d'erreur (comme le font le WER et le CER), ces scores sont peu interprétables.

Pour répondre à cette problématique, des travaux (Gordeeva *et al.*, 2021; Roy, 2021) ont proposé une métrique sémantique que l'on peut adapter à une tâche en aval, et qui pouvait être interprété comme un taux d'erreur mot pondérée en fonction de la gravité sémantique et lexicale d'une erreur. D'autres travaux (Bañeras-Roux *et al.*, 2024b) ont également proposé des paradigmes pour interpréter les erreurs de RAP en fonction de leur sévérité et de leur impact sur les métriques sémantiques à l'aide d'embeddings contextuels. Ces approches offraient une compréhension plus granulaire des erreurs, permettant d'identifier quels types d'erreurs affectent le plus la qualité perçue des transcriptions.

2.3 Perception humaine et évaluation interprétable

Un axe de recherche complémentaire s'est concentré sur l'alignement explicite entre les métriques automatiques et la perception humaine. Plusieurs travaux ont proposé des modèles d'évaluation fondés sur la perception humaine (Itoh *et al.*, 2015; Sasindran *et al.*, 2024).

Des études empiriques ont collecté des annotations humaines pour évaluer comment les locuteurs perçoivent la qualité des transcriptions automatiques. Ces annotations ont permis de créer des ensembles de données comme HATS qui intègre des jugements humains explicites sur la qualité des erreurs de reconnaissance, permettant de valider et comparer différentes métriques d'évaluation. Cette étude a montré une corrélation plus importante avec la perception humaine que pour le WER, confirmant une autre étude (Kim *et al.*, 2022) effectuée sur des jeux de données privées intégrant la perception humaine. Une autre étude récente (Thennal *et al.*, 2025) a également démontré à l'aide d'annotations humaines que le taux d'erreur caractère (CER) est plus pertinent que le WER dans un cadre multilingue.

3 Expériences

Nous nous appuyons sur les résultats rapportés dans le jeu de données HATS afin d'établir un point de référence pour l'évaluation des métriques. Ce jeu de données français contient 1 000 triplets, chacun étant composée d'une transcription de référence annotée manuellement et deux hypothèses (A et B) générées par différents système de RAP, chacune associée au nombre de fois où l'hypothèse a été considéré comme meilleure par un annotateur (au moins 7 par triplet). Le tableau 1 présente le pourcentage d'accord entre différentes métriques automatiques et les jugements humains.

Afin de mieux interpréter ces résultats, trois sous-ensembles du corpus sont considérés, construits en fonction du niveau d'accord entre annotateurs humains : (i) les exemples avec un accord strict de 100 %, (ii) ceux avec un accord d'au moins 70 %, et (iii) l'ensemble complet des données. Ce découpage permet d'analyser le comportement des métriques en fonction de la fiabilité des annotations humaines : un fort accord est supposé refléter des cas plus "évidents", tandis qu'un accord plus faible correspond à des exemples plus ambigus ou où les choix pourraient avoir été mis au hasard par les annotateurs.

Les colonnes Accord indiquent la proportion de cas où la métrique est cohérente avec la préférence humaine, tandis que les colonnes Égal correspondent aux situations où la métrique ne permet pas de départager les sorties, ce qui n'arrive pas pour les métriques basées sur la similarité cosinus puisque

Métrique	= 100 %		≥ 70 %		Complet	
	Accord	Égal	Accord	Égal	Accord	Égal
Taux d'erreur mot	63	23	53	28	49	28
Taux d'erreur caractère	77	17	64	21	60	22
BERTScore CamemBERT-large	80	0	68	0	65	0
SemDist CamemBERT-large	80	0	71	0	67	0
SemDist Phrases CamemBERT-large	90	0	78	0	73	0

TABLE 1 – Pourcentage d'accord entre des métriques et la perception humaine, selon le niveau d'accord des annotateurs. Les mentions **100 %** et **70 %** désignent des sous-ensembles du jeu de données HATS en fonction du pourcentage d'accord entre les annotateurs.

ce sont des valeurs continues. L'accord se calcule comme suivant :

$$\text{Accord} = \frac{\max(n_A, n_B)}{n_A + n_B}$$

où n_A et n_B correspondent respectivement au nombre d'annotateurs ayant préféré les hypothèses A et B. Étant donnée les plus faibles performances de BERTScore, nous nous concentrerons dans cette étude sur SemDist. Pour nos expériences, nous utilisons la boîte à outils SDialog (Burdisso *et al.*, 2026).

3.1 Sélection de la meilleure hypothèse à l'aide d'un LLM

Nous utilisons différents LLM pour déterminer, parmi deux hypothèses de transcription, laquelle est la plus proche d'une référence donnée. Cette expérience vise à évaluer la capacité de différents LLM à comparer deux hypothèses incorrectes et à identifier celle qui correspond le mieux à la transcription de référence, en tenant compte à la fois de la fidélité lexicale et de la cohérence sémantique.

Le modèle est interrogé via un prompt avec un exemple (*one-shot*). Il est explicitement demandé au modèle de justifier sa réponse, puis de conclure par une décision sous la forme `A` ou `B`. Cette formulation permet d'exploiter le raisonnement du modèle (*chain-of-thougth*) tout en facilitant l'extraction automatique de la prédiction finale.

Le prompt utilisée est le suivant :

```
User: Une référence est une transcription exacte d'un audio.
Deux hypothèses fausses sont proposées. Explique ta réflexion et
finis ta phrase en écrivant 'A', 'B' ou 'a' ou 'b' si indécis.
Référence : c' est à lui même
Hypothèse A : êtes à lui même
Hypothèse B : c' est euh à lui-même

Assistant: Même si l'hypothèse B contient une disfluence ('euh'),
elle correspond beaucoup mieux à la référence en termes de
mots et de sens. La disfluence peut être tolérée si elle fait
partie de l'original, tandis que l'erreur grammaticale de
l'hypothèse A est plus problématique. Donc, la transcription
la plus acceptable est l'hypothèse B.
```

```
User: Référence : <référence>
Hypothèse A : <hypothèse A>
Hypothèse B : <hypothèse B>
```

Les résultats présentés dans le tableau 2 révèlent que les LLM génératifs sont capables de sélectionner efficacement la meilleure hypothèse de transcription parmi deux candidates, en accord avec la perception humaine. On observe une progression générale des performances avec la taille et la génération des modèles, bien que la taille ne constitue pas le seul critère déterminant : ainsi, `qwen3-8B`, pourtant nettement plus petit, surpasse `gemma3-27b`, et `qwen3.5-27b` dépasse `qwen3-30b` malgré un nombre de paramètres comparable. Il est possible que les données d'entraînement et l'architecture qui peut intégrer ou non un mélange d'experts (MoE) ainsi que l'utilisation de raisonnement caché explique ces différences. Ce sont les modèles les plus récents, `gpt-4.1` et `qwen3.5-35b`, qui atteignent les meilleures performances avec respectivement 94 % et 92 % d'accord avec les annotateurs sur le sous-ensemble à accord total. Il est à noter qu'un modèle en libre accès comme `qwen3.5-35b` présente des performances comparables à celles de `gpt-4.1`, suggérant que des alternatives accessibles peuvent rivaliser avec des modèles propriétaires de pointe pour cette tâche.

LLM	$=$ **100 %**	$\geq$ **70 %**	**Complet**
GPT-4o	92	83	78
GPT-4.1	**94**	**85**	**79**
gemma3-27b	72	63	61
gemma4-31b	87	78	73
Qwen3-0.6B	50	47	47
Qwen3-1.7B	59	58	56
Qwen3-8B	80	74	72
Qwen3-30B	84	75	71
Qwen3.5-27B	91	83	77
Qwen3.5-35B	**92**	**83**	**78**

TABLE 2 – Pourcentage d'accord entre les choix du LLM (sélection de la meilleure hypothèse) et la perception humaine, selon le niveau d'accord des annotateurs.

Plus remarquable encore, les performances des meilleurs LLM surpassent celles des métriques automatiques d'évaluation ASR classiques telles que le WER et le CER, qui opèrent au niveau des caractères ou des mots sans considération sémantique, mais aussi celles de métriques plus sophistiquées comme SemDist, y compris dans sa meilleure configuration utilisant les plongements de phrases de CamemBERT-large. Ce résultat suggère que les LLM génératifs, grâce à leur compréhension contextuelle et sémantique profonde du langage, parviennent à capturer des nuances perceptives que les métriques traditionnelles ne modélisent pas, comme la tolérance aux disfluences ou la préférence pour la cohérence grammaticale et le sens global de l'énoncé. Ces capacités ouvrent également des perspectives prometteuses pour l'annotation automatique de données perceptives, où les LLM pourraient se substituer partiellement ou compléter l'annotation humaine à moindre coût. Par ailleurs, cette approche permet également d'effectuer des comparaisons directes entre les performances de deux systèmes de RAP, en évaluant leurs sorties de manière relative plutôt qu'absolue, ce qui est particulièrement pertinent dans des scénarios où l'on cherche à mettre en production le meilleur système vis-à-vis de la perception humaine.

3.2 Métrique sémantique basée sur les plongements de LLM décodeurs

Nous évaluons la corrélation de la métrique *SemDist*, définie comme une distance dérivée de la similarité cosinus entre les embeddings de la référence et de l'hypothèse, avec les annotations humaines du jeu de données HATS.

Pour obtenir les représentations des transcriptions, nous utilisons différentes familles de LLM, couvrant plusieurs tailles et méthodes d'entraînement. Le texte est d'abord tokenisé en une séquence de tokens, puis transformé par le LLM en une séquence d'embeddings, correspondant aux représentations vectorielles de chaque token.

La métrique *SemDist* nécessitant une représentation de dimension fixe, cette séquence est agrégée via une opération de *pooling*. Nous comparons plusieurs stratégies de pooling, allant de méthodes simples comme la moyenne des embeddings à des approches pondérées.

L'objectif est d'évaluer l'impact de ces choix sur la qualité des représentations, en particulier leur capacité à capturer les similarités sémantiques reflétées par les annotations humaines. L'ensemble des combinaisons entre familles de LLM et méthodes de pooling est ainsi évalué en termes de corrélation entre *SemDist* et les annotations du jeu de données HATS.

Stratégies de pooling. On considère les différentes stratégies de pooling suivantes, où t_i désigne l'embedding du i-ème token et n la longueur de la séquence :

- **Dernier token** : utilisation de l'embedding du dernier token, t_n.
- **Avant-dernier** : utilisation de l'embedding du token précédant le dernier, t_{n-1}.
- **Moyenne** : moyenne des embeddings de tous les tokens, $\frac{1}{n}\sum_{i=1}^{n} t_i$.
- **Moyenne sans dernier token** : moyenne des embeddings en excluant le dernier token, $\frac{1}{n-1}\sum_{i=1}^{n-1} t_i$.
- **Moyenne pondérée** : moyenne pondérée des embeddings de tous les tokens, où les poids augmentent linéairement avec la position (i.e., poids i pour le token t_i), $\frac{\sum_{i=1}^{n} i\, t_i}{\sum_{i=1}^{n} i}$.
- **Moyenne pondérée sans dernier token** : moyenne pondérée des embeddings en excluant le dernier token, avec les mêmes poids positionnels, $\frac{\sum_{i=1}^{n-1} i\, t_i}{\sum_{i=1}^{n-1} i}$.
- **Moyenne des 4 derniers tokens** : moyenne des embeddings des quatre derniers tokens, $\frac{1}{4}\sum_{i=n-3}^{n} t_i$.

Contrairement aux intuitions initiales, l'analyse révèle un faible lien entre la taille du modèle et la corrélation des plongements avec la perception humaine. Alors qu'il est courant d'observer de meilleures performances pour les modèles de taille conséquente, nous observons peu ce phénomène voir parfois l'inverse : Qwen3.5-35B ayant des performances plus faibles que pour Qwen3.5-27B. Chose intéressante : les LLM ayant les meilleures performances pour classifier une hypothèse comme étant meilleure (voir Table 2) n'ont pas nécessairement les embeddings les plus pertinents pour la métrique SemDist (Qwen3.5-27B a un score de 91 % pour la sélection de la meilleure hypothèse contre 78 % avec la moyenne des 4 derniers tokens).

Un résultat notable est que les représentations basées uniquement sur le dernier token sont, dans la majorité des cas, surpassées par les alternatives. Par exemple, Qwen3.5-27B atteint un score de seulement 59 % avec un pooling du dernier token contre au moins 74 % pour le pooling pondérée, la seconde pire méthode pour ce LLM et qui a un écart de 15 points. Nous attribuons ceci à l'objectif

Modèle	Dern.	Moy.	Moy.*	Pond.	Pond.*	4 dern.	Av.dern.
gemma-2-2b	73	79	79	79	78	82	83
gemma-2b	76	80	80	79	80	81	80
gemma-3-1b-pt	75	76	74	77	76	76	75
gemma-3-27b-it	77	83	83	81	82	82	82
gemma-4-31B	72	75	73	77	76	76	75
gemma-7b	65	69	71	71	72	73	80
OLMo-2-1124-7B	73	82	80	83	80	83	83
Mistral-7B-v0.3	79	83	84	84	85	85	85
Qwen3-0.6B	70	80	80	78	78	81	81
Qwen3-0.6B-Base	74	79	79	81	80	80	80
Qwen3-1.7B	72	83	83	80	81	87	87
Qwen3-1.7B-Base	74	82	81	81	82	86	86
Qwen3-30B	72	84	84	83	82	84	82
Qwen3-4B	77	85	83	82	81	79	80
Qwen3-4B-Base	74	82	83	82	82	82	83
Qwen3-8B	75	85	83	83	83	85	85
Qwen3-8B-Base	71	78	78	80	77	82	82
Qwen3-Embedding-0.6B	**88**	84	85	85	84	83	82
Qwen3-Embedding-4B	**88**	86	85	86	84	85	86
Qwen3-Embedding-8B	**88**	**89**	**87**	**89**	**87**	**88**	**89**
Qwen3.5-27B	59	77	76	74	76	78	77
Qwen3.5-35B	68	68	68	67	68	67	67
Moyenne	*74*	*80*	*80*	*80*	*80*	*81*	*81*

TABLE 3 – Performance SemDist par modèle et technique de pooling sur le sous-ensemble du jeu de données HATS faisant consensus entre annotateurs.
* = sans dernier token | Dern. = Dernier token | Av.dern. = Avant-dernier

d'entraînement des LLM : la tâche de prédiction du prochain token optimise les représentations pour prédire les tokens futurs plutôt que de capturer le contenu sémantique global. Le dernier token, positionné en fin de séquence où il doit prédire une continuation peu informative, porte peu d'information sémantique sur la phrase complète.

Malgré sa potentielle capacité à effacer et perdre des informations, le pooling par moyenne se montre particulièrement efficace, égalant ou dépassant souvent des stratégies plus sophistiquées. Nous expliquons ceci par une propriété géométrique de l'espace d'embeddings : lorsque l'hypothèse et la référence ont des longueurs comparables (ce qui est attendue dans une tâche de reconnaissance automatique de la parole), les paires sémantiquement similaires exhibent des trajectoires d'embeddings convergentes, causant leurs moyennes à se regrouper dans l'espace latent.

Sans surprise, les modèles explicitement affinés pour les tâches d'embeddings surpassent considérablement les LLM génériques de taille comparable, même pour l'utilisation de l'embedding du dernier token, ce dernier étant celui ayant été utilisé pour l'affinage du LLM (Zhang *et al.*, 2025). Qwen3-Embedding-8B atteint avec le pooling de moyenne un score de 89 % contre 85 % pour Qwen3-8B, une amélioration de 4 points. Remarquablement, cet affinement réduit l'écart entre le dernier token et les autres méthodes, suggérant que l'entraînement spécifique à la tâche réoriente le dernier token pour servir de représentation sémantique cohérente plutôt que simplement de prédicteur du prochain token.

3.3 LLM génératif pour classifier les hypothèses

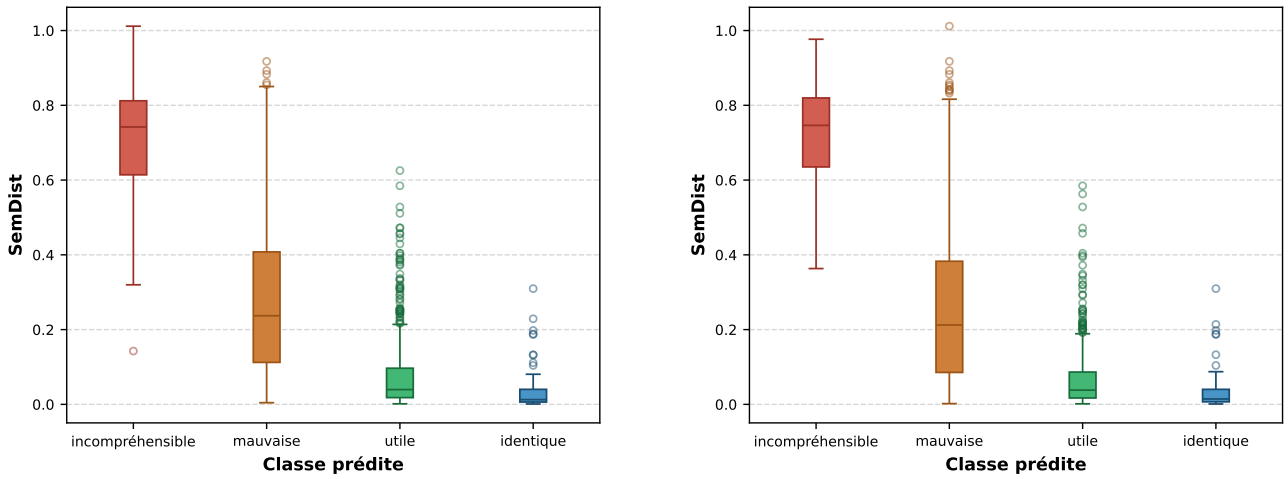
Dans cette section, nous évaluons la capacité des modèles de langage à classer une hypothèse étant donné une référence. Contrairement à la Section 3.1, où la tâche consistait à sélectionner la meilleure hypothèse parmi deux candidates pour une même référence, nous considérons ici chaque paire (référence, hypothèse) de manière indépendante, et le modèle doit attribuer une classe qualitative parmi quatre catégories. L'objectif est de faire une évaluation intrinsèque et objective du modèle.

Étant donnée une référence et une hypothèse, le modèle doit assigner l'une des étiquettes suivantes :

- **identique** : l'hypothèse est identique à la référence (ou différence de casse ou de tiret).
- **utile** : le sens est préservé malgré des erreurs mineures (normalisation, ponctuation, orthographe, majuscules, abréviations, légères variations syntaxiques sans perte de compréhension).
- **mauvaise** : le sens est partiellement altéré (erreurs significatives sur des mots-clés, substitutions ou omissions importantes, mais une partie du contenu reste compréhensible).
- **incompréhensible** : le sens est totalement perdu (la phrase ne peut pas être comprise ou interprétée correctement en lien avec la référence).

Chaque paire est annotée automatiquement par le modèle considéré, sans comparaison explicite avec d'autres hypothèses. Cette formulation permet d'évaluer la capacité intrinsèque du modèle à juger la qualité d'une hypothèse de manière absolue, plutôt que relative.

En parallèle de cette annotation catégorielle, nous calculons un score SemDist entre la référence et l'hypothèse en utilisant le LLM encodeur le plus performant à nos connaissance : `sentence-camembert-large`. Cette mesure fournit un signal quantitatif indépendant, permettant d'analyser la cohérence entre les jugements qualitatifs du modèle et une métrique continue permettant de simuler la perception humaine.



(a) GPT-4.1

(b) Qwen3.5-35B

FIGURE 1 – Distribution en boîte à moustaches des scores SemDist selon la classe annotée par un LLM.

La Figure 1 présente la distribution des scores de similarité pour chacune des classes prédites. L'axe des abscisses correspond aux quatre catégories (identique, utile, mauvaise, incompréhensible), tandis que l'axe des ordonnées représente les scores de similarité. Cette visualisation permet d'observer dans quelle mesure les classes produites par le modèle s'alignent avec des niveaux de similarité attendus,

par exemple des scores élevés pour la classe identique et plus faibles pour les classes mauvaise ou incompréhensible. Les modèles GPT-4.1 et Qwen3.5-35B obtiennent une corrélation significative de Spearman de -0.66 et -0.60 avec les scores SemDist. Ces résultats suggèrent que la classification des erreurs de reconnaissance de la parole par des LLM constitue un indicateur imparfait mais pertinent des performances du système, tout en apportant un gain d'interprétabilité par rapport aux métriques fondées uniquement sur des similarités cosinus.

4 Conclusion

Cet article a évalué systématiquement les capacités des grands modèles de langage pour l'évaluation de la reconnaissance automatique de la parole selon trois approches complémentaires. Premièrement, les LLM génératifs surpassent les métriques classiques et sémantiques existantes dans la tâche de sélection de la meilleure hypothèse, avec des performances atteignant 92-94 % d'accord avec les annotateurs humains, démontrant des capacités efficaces pour de l'annotation automatique et pour trouver le meilleur système de RAP. Deuxièmement, les embeddings extraits de LLM décodeurs fournissent des représentations de qualité, même avec des stratégies d'agrégation simples. Pour cette tâche, la taille du modèle semble ne pas être un facteur déterminant et les modèles affinés explicitement obtiennent les meilleures performances pour toutes les techniques de pooling. Troisièmement, les LLM semblent capables de classifier qualitativement les erreurs de transcription de manière cohérente, offrant une interprétabilité supérieure aux métriques numériques opaques.

Ces résultats suggèrent que les LLM offrent une voie prometteuse et interprétable pour l'évaluation perceptive des systèmes de RAP, ouvrant la voie à des métriques d'évaluation de nouvelle génération qui alignent mieux la performance mesurée avec la perception humaine réelle.

5 Considérations éthiques

L'utilisation de LLM pour l'évaluation de la RAP présente des enjeux éthiques notables. D'abord, l'inférence à grande échelle engendre des coûts énergétiques et une empreinte carbone significativement supérieurs aux métriques légères comme le WER ou le CER. Cependant, nos résultats montrent que des modèles compacts, affinés pour une tâche sémantique, offrent des performances comparables aux modèles les plus massifs, ouvrant la voie à des déploiements plus raisonnable énergétiquement.

Les LLM sont entraînés sur des corpus potentiellement biaisés (surreprésentation de certaines variantes linguistiques, normes de langue standard, etc.). Ces biais pourraient influencer subtilement l'évaluation des transcriptions.

Remerciements

Ce travail a été soutenu par l'Idiap Research Institute et le projet ELOQUENCE du programme Horizon 2020 de l'Union européenne (numéro de subvention 101070558).

Références

AHUJA K., DIDDEE H., HADA R., OCHIENG M., RAMESH K., JAIN P., NAMBI A., GANU T., SEGAL S., AHMED M. *et al.* (2023). Mega : Multilingual evaluation of generative ai. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, p. 4232–4267.

BAÑERAS-ROUX T., ROUVIER M., WOTTAWA J. & DUFOUR R. (2022). Qualitative evaluation of language model rescoring in automatic speech recognition. In *Interspeech*.

BAÑERAS-ROUX T., ROUVIER M., WOTTAWA J. & DUFOUR R. (2024a). A comprehensive analysis of tokenization and self-supervised learning in end-to-end automatic speech recognition applied on french language. In *2024 32nd European Signal Processing Conference (EUSIPCO)*, p. 141–145 : IEEE.

BAÑERAS-ROUX T., ROUVIER M., WOTTAWA J. & DUFOUR R. (2024b). A paradigm for interpreting metrics and measuring error severity in automatic speech recognition. In *International Conference on Text, Speech, and Dialogue*, p. 174–183 : Springer.

BAÑERAS-ROUX T., WOTTAWA J., ROUVIER M., MERLIN T. & DUFOUR R. (2023). Hats : An open data set integrating human perception applied to the evaluation of automatic speech recognition metrics. In *International Conference on Text, Speech, and Dialogue*, p. 164–175.

BURDISSO S., BAROUDI S., LABRAK Y., GRÜNERT D., CYRTA P., CHEN Y., MADIKERI S., VILLATORO-TELLO E., MARXER R. & MOTLICEK P. (2026). Sdialog : A python toolkit for end-to-end agent building, user simulation, dialog generation, and evaluation. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 3 : System Demonstrations)*, p. 320–340.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). Bert : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics : human language technologies, volume 1 (long and short papers)*, p. 4171–4186.

FAVRE B., CHEUNG K., KAZEMIAN S., LEE A., LIU Y., MUNTEANU C., NENKOVA A., OCHEI D., PENN G., TRATZ S. *et al.* (2013). Automatic human utility evaluation of asr systems : Does wer really predict performance ? In *Proc. Interspeech 2013*, p. 3463–3467.

GORDEEVA L., ERSHOV V., GULYAEV O. & KURALENOK I. (2021). Meaning error rate : Asr domain-specific metric framework. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, p. 458–466.

HE X., DENG L. & ACERO A. (2011). Why word error rate is not a good metric for speech recognizer training for the speech translation task ? In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 5632–5635 : IEEE.

ITOH N., KURATA G., TACHIBANA R. & NISHIMURA M. (2015). A metric for evaluating speech recognizer output based on human-perception model. In *Proc. Interspeech 2015*, p. 1285–1288.

KAFLE S. & HUENERFAUTH M. (2017). Evaluating the usability of automatically generated captions for people who are deaf or hard of hearing. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*, p. 165–174.

KIM S., ARORA A., LE D., YEH C.-F., FUEGEN C., KALINLI O. & SELTZER M. L. (2021). Semantic distance : A new metric for asr performance analysis towards spoken language understanding. In *Proc. Interspeech 2021*, p. 1977–1981.

KIM S., LE D., ZHENG W., SINGH T., ARORA A., ZHAI X., FUEGEN C., KALINLI O. & SELTZER M. (2022). Evaluating user perception of speech recognition system quality with semantic distance metric. In *Proc. Interspeech 2022*, p. 3978–3982.

LABRAK Y., BAZOGE A., MORIN E., GOURRAUD P.-A., ROUVIER M. & DUFOUR R. (2024). Biomistral : A collection of open-source pretrained large language models for medical domains. In *Findings of the association for computational linguistics : acl 2024*, p. 5848–5864.

LE N.-T., SERVAN C., LECOUTEUX B. & BESACIER L. (2016). Better evaluation of asr in speech translation context using word embeddings. In *Interspeech 2016*.

MCCOWAN I. A., MOORE D., DINES J., GATICA-PEREZ D., FLYNN M., WELLNER P. & BOURLARD H. (2004). On the use of information retrieval measures for speech recognition evaluation.

MORRIS A. C., MAIER V. & GREEN P. D. (2004). From wer and ril to mer and wil : improved evaluation measures for connected speech recognition. In *Interspeech*, volume 4-8, p. 2004.

NAM S. & FELS D. (2019). Simulation of subjective closed captioning quality assessment using prediction models. *International Journal of Semantic Computing*, **13**(01), 45–65.

RADFORD A., NARASIMHAN K., SALIMANS T., SUTSKEVER I. *et al.* (2018). Improving language understanding by generative pre-training. *OpenAI Blog*.

RANGAPPA P., CAROFILIS A., PRAKASH J., KUMAR S., BURDISSO S., MADIKERI S., VILLATORO-TELLO E., SHARMA B., MOTLICEK P., HACIOGLU K. *et al.* (2025a). Efficient data selection for domain adaptation of asr using pseudo-labels and multi-stage filtering. In *Proc. Interspeech 2025*, p. 4928–4932.

RANGAPPA P., ZULUAGA-GOMEZ J., MADIKERI S., CAROFILIS A., PRAKASH J., BURDISSO S., KUMAR S., VILLATORO-TELLO E., NIGMATULINA I., MOTLICEK P. *et al.* (2025b). Speech data selection for efficient asr fine-tuning using domain classifier and pseudo-label filtering. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 1–5 : IEEE.

ROY S. (2021). Semantic-wer : A unified metric for the evaluation of asr transcript for end usability. *arXiv preprint arXiv :2106.02016*.

SASINDRAN Z., YELCHURI H. & PRABHAKAR T. (2024). Semascore : A new evaluation metric for automatic speech recognition tasks. In *Proc. Interspeech 2024*, p. 4558–4562.

TEAM G., MESNARD T., HARDIN C., DADASHI R., BHUPATIRAJU S., PATHAK S., SIFRE L., RIVIÈRE M., KALE M. S., LOVE J. *et al.* (2024). Gemma : Open models based on gemini research and technology. *arXiv preprint arXiv :2403.08295*.

THENNAL D., JAMES J., GOPINATH D. P. *et al.* (2025). Advocating character error rate for multilingual asr evaluation. In *Findings of the Association for Computational Linguistics : NAACL 2025*, p. 4926–4935.

TOUVRON H., LAVRIL T., IZACARD G., MARTINET X., LACHAUX M.-A., LACROIX T., ROZIÈRE B., GOYAL N., HAMBRO E., AZHAR F. *et al.* (2023). Llama : Open and efficient foundation language models. *arXiv preprint arXiv :2302.13971*.

WANG Y.-Y., ACERO A. & CHELBA C. (2003). Is word error rate a good indicator for spoken language understanding accuracy. In *2003 IEEE workshop on automatic speech recognition and understanding (IEEE Cat. No. 03EX721)*, p. 577–582 : IEEE.

ZHANG T., KISHORE V., WU F., WEINBERGER K. Q. & ARTZI Y. (2019). Bertscore : Evaluating text generation with bert. In *International Conference on Learning Representations*.

ZHANG Y., LI M., LONG D., ZHANG X., LIN H., YANG B., XIE P., YANG A., LIU D., LIN J. *et al.* (2025). Qwen3 embedding : Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv :2506.05176*.