Actes de CORIA-TALN-RJCRI-RECITAL 2025
Marseille, France, 30 juin au 4 juillet 2025

Décoder le pouvoir de persuasion dans les concours d'éloquence : une étude sur la capacité des modèles de langue à évaluer la prise de parole en public

Alisa Barkar[1] Mathieu Chollet[2,3] Matthieu Labeau[1] Béatrice Biancardi[4]
Chloé Clavel[5]

(1) LTCI, Institut Polytechnique de Paris, Telecom-Paris, 19 Place Marguerite Perey, 91120 Palaiseau, France
(2) School of Computing Science, University of Glasgow, G12 8RZ Glasgow, United Kingdom
(3) IMT Atlantique, LS2N, UMR CNRS 6004, 44307 Nantes, France
(4) CESI LINEACT, Nanterre, France
(5) ALMAnaCH, INRIA, Paris, France

alisa.barkar@telecom-paris.fr, mathieu.chollet@glasgow.ac.uk,
matthieu.labeau@telecom-paris.fr, bbiancardi@cesi.fr,
chloe.clavel@inria.fr

RÉSUMÉ

L'importance des compétences en prise de parole en public (PPP) stimule le développement de systèmes d'évaluation automatisée, mais l'intégration des grandes modèles de langue (LLMs) reste peu explorée. Nous proposons un cadre où les LLMs évaluent des critères issus de la littérature et de retours de formateurs. Nous testons trois approches : des prédictions LLM directes à zéro coup (RMSE $0,8$) par rapport à des prédictions de persuasion basées sur des caractéristiques lexicales fabriquées à la main (RMSE $0,51$) ou basées sur des critères évalués par LLM $0,6$ insérés en entrée dans ElasticNet. L'analyse des liens entre critères et caractéristiques lexicales montre que seul le critère de niveau de langue évalué par LLM est prévisible (score F1 de $0,56$) soulignant les limites actuelles des LLMs pour l'analyse de la PPP. Code source et données disponibles sur GitHub.

ABSTRACT

Decoding Persuasion in Public Speaking with LLMs : A Critical Assessment

The growing importance of public speaking (PS) skills fuels the development of automated evaluation systems, yet the use of large language models (LLMs) remains underexplored. This study examines the application of LLMs to assessing persuasiveness. We propose a framework where LLMs evaluate criteria derived from educational literature and coach feedback. We test three approaches : direct zero-shot LLM predictions (RMSE 0.8) against persuasiveness predictions based on hand-crafted lexical features (RMSE 0.51) or based on LLM-evaluated criteria 0.6 inserted as an input to ElasticNet. By analyzing the relationship between new criteria and lexical features, we show that only the LLM-evaluated language level criterion is predictable (F1-score of 0.56), challenging assumed links. Our findings highlight the current limitations of LLMs in accurately analyzing PS. All code and materials are publicly available on GitHub.

MOTS-CLÉS : Évaluation de la prise de parole en public, Grandes modèles de langue(LLMs), Prédiction du pouvoir de persuasion, Caractéristiques interprétables, Modalité textuelle, Évaluation automatique du discours, Modèles open-source.

Actes de l'atelier Évaluation des modèles génératifs (LLM) et challenge 2025 (EvalLLM), pages 77–90.

KEYWORDS: Public speaking assessment, Large Language Models (LLMs), Persuasiveness prediction, Interpretable features, Textual modality, Automatic speech evaluation, Open-source models.

ARTICLE : **Accepté à** ICAART 2025 : 17th International Conference on Agents and Artificial Intelligence.

1 Introduction

La prise de parole en public (PPP) constitue une compétence essentielle tant pour la réussite professionnelle que pour le développement de la confiance en soi (Schreiber & Hartranft, 2017). Ainsi, des systèmes d'évaluation automatisée ont émergé afin de proposer des solutions de formation accessibles aux étudiants et aux jeunes professionnels (Rodero & Larrea, 2022; Schneider *et al.*, 2015; Kurihara *et al.*, 2007). Récemment, les grandes modèles de langue (LLMs) ont également investi le domaine de l'évaluation automatique de la PPP. Des systèmes commerciaux, tels que Poised, Yoodli et des services basés sur GPT [1], illustrent cette tendance. Parallèlement, la recherche académique s'est principalement concentrée sur des modèles de classification/régression (Chen *et al.*, 2015) et des réseaux neuronaux (Ashwin & Rajendran, 2022; Tun *et al.*, 2023), créant ainsi un écart important entre la recherche et les pratiques commerciales. Notre étude vise à combler cet écart en exploitant les LLMs pour la prédiction de dimensions subjectives (*par exemple*, le pouvoir de persuasion, l'engagement, etc.), une approche fréquemment utilisée dans la littérature sur l'évaluation automatique de la PPP. Plus précisément, nous nous concentrons sur la dimension du pouvoir de persuasion, dont la corrélation avec les caractéristiques lexicales a été largement démontrée. Par exemple, (Larrimore *et al.*, 2011) a analysé le lien entre le nombre de mots, les types de langage et leur efficacité persuasive dans le cadre de la recherche de financement, tandis que (Park *et al.*, 2014) a montré que les caractéristiques lexicales surpassent les caractéristiques visuelles dans la prédiction du pouvoir de persuasion.

Notre objectif est d'évaluer la capacité des LLMs à analyser la PPP en se concentrant sur la modalité textuelle, qui a reçu moins d'attention en comparaison avec les modalités audio et visuelle, largement étudiées (Eyben *et al.*, 2016; Nguyen *et al.*, 2012; Chen *et al.*, 2015). Les approches existantes de l'analyse textuelle manquent souvent d'interprétabilité et de pertinence pour obtenir un retour utile. Par exemple, des travaux antérieurs se sont appuyés sur des représentations non interprétables (Das *et al.*, 2021) ou sur des caractéristiques lexicales définies à partir de connaissances expertes, telles que les uni-/bi-grammes (Park *et al.*, 2014), le débit de parole, la durée des pauses, les mots de remplissage (Dinkar *et al.*, 2020), les lexiques de sentiment (Chen *et al.*, 2015), ou encore les indicateurs émotionnels basés sur LIWC (Pennebaker *et al.*, 2022), pour évaluer le charisme des orateurs (Yang *et al.*, 2020). Afin de répondre à ce problème d'interprétabilité et d'utilité, et en nous appuyant sur des entretiens avec des coachs en prise de parole ainsi que sur des grilles d'évaluation établies (Chollet & Lefebvre, 2022), nous avons élaboré une liste de critères textuels essentiels à une performance en public réussie. Nous avons ensuite développé une méthode d'évaluation basée sur un LLM pour évaluer ces critères, et analysé leur relation avec les caractéristiques manuelles existantes ainsi que leur efficacité dans la prédiction du pouvoir de persuasion.

En nous appuyant sur le succès des LLMs dans des tâches telles que l'identification de discours

1. Poised : https://www.poised.com/about-us ; Yoodli : https://app.yoodli.ai/ ; Services basés sur GPT : https://bit.ly/48eqllR, https://bit.ly/3YeKnYy

politiques (Gilardi *et al.*, 2023) et la classification des sentiments (Zhu *et al.*, 2023; Latif *et al.*, 2023), ainsi que sur leurs limites à justifier l'évaluation automatique de récits (Chhun *et al.*, 2024; Binz & Schulz, 2022; Lamprinidis, 2023), nous avons formulé l'hypothèse que ces modèles pouvaient analyser de manière fiable des textes et en extraire des caractéristiques pertinentes, tout en peinant à évaluer avec précision leur pouvoir de persuasion. Dans une optique d'accessibilité, de reproductibilité et de réduction de l'impact environnemental, nous avons privilégié des LLMs plus petits et open-source, dont l'empreinte carbone est plus modeste (*e.g.* 539 tonnes pour Llama 2 (AI, 2023) contre 552 tonnes pour les modèles GPT (Patterson *et al.*, 2021)). Conscients des progrès rapides des modèles, nous mettons en avant nos contributions méthodologiques et proposons une implémentation open-source accompagnée de résultats adaptables à des modèles plus récents, ainsi que des analyses supplémentaires non incluses dans cet article, disponibles sur GitHub [2]. À partir des lacunes identifiées, nous formulons les questions de recherche (QR) suivantes :

QR-1 : Dans quelle mesure les LLM peuvent-ils prédire avec précision le pouvoir de persuasion d'un discours à partir de sa transcription ?

QR-2 : Les critères évalués par les LLM peuvent-ils constituer une représentation de haut niveau des caractéristiques lexicales définies à partir de connaissances expertes ?

2 Jeu de données 3MT_French

Motivations. Notre objectif à long terme est de créer un système open-source destiné aux étudiants francophones, c'est pourquoi nous concentrons notre travail sur le contexte éducatif. Cependant, la majorité des jeux de données disponibles en PPP ne sont pas adaptés à des contextes éducatifs. Par exemple, MIT Interview (Courgeon *et al.*, 2014) concerne des entretiens d'embauche, POM (Park *et al.*, 2014) se concentre sur des critiques de films, AVSpeech (Ephrat *et al.*, 2018) et YouTube-8M (Abu-El-Haija *et al.*, 2016) ciblent des vidéos issues de YouTube, SAC-LAD (Song *et al.*, 2023) et CREMA-D (Cao *et al.*, 2014) se focalisent sur l'anxiété et les émotions, tandis que NUSMSP (Gan *et al.*, 2017) ne contient pas d'annotations sur le pouvoir de persuasion. Nous utilisons donc le jeu de données 3MT_French (Biancardi *et al.*, 2024), qui fournit des évaluations participatives du pouvoir de persuasion dans les présentations d'étudiants en doctorat francophones :

— Présentations de 3 minutes issues du concours « Ma Thèse en 180 secondes » (*135 femmes, 113 hommes*) portant sur des sujets de recherche variés.
— Chaque présentation a été évaluée à l'aide d'une échelle de Likert en 5 points pour le pouvoir de persuasion, adaptée de la grille d'évaluation dans « Public Speaking Competence Rubric » (Schreiber *et al.*, 2012). Trois annotateurs par échantillon ont jugé dans quelle mesure l'orateur construisait un message crédible et convaincant. Pour plus de détails sur le protocole d'annotation, se référer à (Biancardi *et al.*, 2024).
— Le jeu de données ne contenait pas de transcriptions. Nous les avons générées à l'aide de Whisper (Radford *et al.*, 2022), un modèle de reconnaissance automatique de la parole à la pointe de la technologie. Après une vérification manuelle et la suppression des échantillons corrompus, 227 transcriptions ont été conservées.

2. https://github.com/anonympapers/PublicSpeakingAnalysisWithOpenSourceLLMs

3 Méthodologies pour l'évaluation de la prise de parole en public

Afin de répondre aux **QRs** formulées, notre méthodologie s'est articulée autour de deux étapes principales. Dans un premier temps, nous avons mis en œuvre la méthode d'incitation décrite dans la Section 3.1, en présentant sa structure ainsi que les motivations sous-jacentes à cette approche. Cette incitation nous a permis de traiter la **QR-1**, en cherchant à prédire directement le degré de persuasion à partir de la transcription du discours à l'aide de LLM.

Dans un second temps, afin d'aborder la **QR-2** et d'examiner les relations entre les caractéristiques lexicales et les critères évalués par le LLM à partir de la transcription du discours (tels que la pertinence du langage utilisé, la qualité de la structure, etc.), nous détaillons dans la Section 3.2 le choix des caractéristiques lexicales retenues ainsi que les critères d'évaluation mobilisés dans l'analyse basée sur le LLM.

3.1 Prédiction du pouvoir de persuasion en zéro-shot (QR-1)

Description de l'analyse :
TÂCHE = [évaluer la transcription de la prestation orale fournie dans la section TRANSCRIPTION. Pour comprendre comment l'évaluer, se référer à la description de la dimension donnée dans la section DIMENSION. Donner UNIQUEMENT un nombre en réponse — ce nombre est une note qui doit être attribuée selon l'échelle décrite dans la section éCHELLE.]
ÉCHELLE = [échelle de 1 = pas du tout à 5 = tout à fait.]
DIMENSION = [Selon vous, dans quelle mesure ce discours est-il persuasif, *c'est-à-dire*, la personne parvient-elle à formuler un message convaincant ? Son raisonnement est-il rigoureux ?]
TRANSCRIPTION = [*contenu de la transcription*]

TABLEAU 1 – Exemple de structure de prompt pour l'évaluation du pouvoir de persuasion.

Techniques de sollicitation. Cet article ne se concentre pas sur du prompt enginering, nous évitons donc les techniques avancées telles que Chain of Thought (Wei *et al.*, 2023) ou les Explications Contrastives (Paranjape *et al.*, 2021). À la place, en suivant les exemples de prompts en langage naturel présentés dans (Korini & Bizer, 2023), (Huang *et al.*, 2023) et (Cheng *et al.*, 2023), nous utilisons une approche simple d'instruction en mode Zero-shot, en fournissant les transcriptions et les questions dans des sections séparées. Notre objectif est d'éviter les longues transcriptions nécessaires au few-shot prompting et de maintenir la tâche simple et concise pour les petits modèles. Étant donné la forte influence de la formulation des prompts sur les performances des modèles (Zhao *et al.*, 2021), (Webson & Pavlick, 2022), nous testons plusieurs formulations en anglais, en français et en contexte multilingue [3]. La Tableau 1 présente un exemple utilisé dans la Section 5 pour l'exposition des résultats. Nous structurons les prompts en sections : TÂCHE – contient les instructions destinées au modèle, DIMENSION – fournit la question liée à la dimension, ÉCHELLE – définit l'échelle d'évaluation, et TRANSCRIPTION – contient la transcription. Les prompts sont générées automatiquement pour chaque transcription. Nous désignons un prompt individuelle par p et l'ensemble des prompts testées par P.

Définition du pouvoir de persuasion pour les LLMs. Afin de nous aligner sur les scores humains du jeu de données 3MT_French, nous adaptons la question issue de (Biancardi *et al.*, 2024), initialement formulée en français (voir Tableau 1).

3. Des exemples de prompts sont disponibles dans notre dépôt GitHub open-source

Post-traitement des résultats et notations. Nous extrayons automatiquement les scores à l'aide de la bibliothèque spaCy (Honnibal & Montani, 2017) et filtrons les sorties incohérentes des LLMs (par exemple, plusieurs scores prédits, changements de l'échelle de notation ou génération de texte au lieu d'une notation). Chaque résultat est ensuite revu manuellement afin de confirmer les incohérences. Pour tout modèle noté m appartenant à l'ensemble M, et pour toute prompt $p \in P$, les scores du pouvoir de persuasion prédits par le LLM sont notés $\{\hat{y}_{i_{p,m}}\}_{i \in N}$, où N représente les échantillons du jeu de données (ensemble complet ou ensemble de test).

3.2 Nouvelles Caractéristiques Interprétables vs Caractéristiques Lexicales (QR-2)

Caractéristiques lexicales. Afin de valider l'utilité des nouvelles caractéristiques interprétables, nous avons extrait plusieurs types de caractéristiques lexicales définies à partir de connaissances expertes élaborées à partir de (Barkar *et al.*, 2023), regroupées en trois grandes catégories : Niveau linguistique (propriétés lexicales et syntaxiques du texte), Richesse du vocabulaire et transitions (variété et transitions au sein du vocabulaire) et Processus affectifs, cognitifs et perceptifs (quantification de la présence d'éléments affectifs (émotions), cognitifs (processus de pensée) et perceptifs (éléments sensoriels et expérientiels) du dictionnaire LIWC (Pennebaker *et al.*, 2022)) [4]. Pour plus de détails sur les caractéristiques lexicales, se référer à (Barkar *et al.*, 2023). Nous utilisons spaCy (Honnibal & Montani, 2017) pour l'étiquetage morpho-syntaxique, le tagueur français (Labrak & Dufour, 2022) pour les étiquettes fines, et LIWC (Pennebaker *et al.*, 2022) avec le dictionnaire français (Piolat *et al.*, 2011) pour les caractéristiques LIWC.

Collecte des critères et composition des questions à choix multiple. Afin de répondre à la **QR-2**, nous analysons les discours en nous appuyant sur des grilles d'évaluation françaises reconnues (*par exemple* (EPF, 2013), grille d'évaluation de *"Ma thèse en 180 secondes"*). Ces critères ont été validés par des entretiens avec 11 formateurs et formatrices français en prise de parole (Chollet & Lefebvre, 2022), et complétés par des critères issus de la littérature internationale sur le pouvoir de persuasion et la clarté. Bien que les critères puissent varier selon les cultures, nous visons une applicabilité large, tout en reconnaissant les limites culturelles. Nous avons utilisé des questions à choix multiple (avec les options A, B ou C) pour créer un prompt d'évaluation [5] pour chaque critère et chaque transcription. Les scores ainsi obtenus sont désignés comme critères évalués par LLM. Les critères identifiés sont listés ci-dessous et leur lien avec les caractéristiques lexicales est discuté.

Présentation du sujet : Ce critère évalue la présentation du sujet en termes de clarté et d'originalité. Nous nous sommes appuyés sur (Chollet & Lefebvre, 2022).

Structure : Ce critère évalue la structure du discours en fonction de la clarté, de l'organisation et de l'efficacité des transitions. Nous nous sommes appuyés sur (Chollet & Lefebvre, 2022). Nous émettons l'hypothèse que ce critère pourrait être lié à la catégorie des *caractéristiques lexicales de richesse du vocabulaire et des transitions*.

Niveau de langue : Ce critère évalue l'adéquation du niveau de langue du discours en termes de clarté, d'absence de jargon, d'argot ou de termes offensants, ainsi que la richesse du vocabulaire et des expressions. Nous nous sommes appuyés sur (Chollet & Lefebvre, 2022). Nous émettons l'hypothèse que ce critère pourrait être lié à la catégorie *niveau de langue*.

4. La liste complète des caractéristiques lexicales ainsi que les formules sont disponibles sur notre GitHub
5. Questions sont disponibles sur notre GitHub

Voix passive : Ce critère évalue l'utilisation de la voix passive dans le discours en fonction de sa clarté et de sa pertinence, en tenant compte de son impact potentiel sur l'objectivité et la compréhension. Ce critère est motivé par (Ruelas Inzunza, 2020), qui a mis en évidence la corrélation entre la perception d'un texte (objectivité, ambiguïté, clarté) et l'utilisation de la voix passive.

Concision : Ce critère évalue la longueur des phrases du discours en fonction de la clarté et de la lisibilité, en considérant si elles sont trop courtes, suffisamment variées, ou excessivement longues et confuses. Nous nous sommes appuyés sur (Melloni *et al.*, 2017), qui mesure la concision en fonction de la longueur du texte complet. Cependant, puisque tous les exemples de notre ensemble de données présentaient une longueur globale similaire, nous avons choisi d'utiliser la longueur des phrases comme critère. Ainsi, nous évaluons la concision de chaque phrase. Nous émettons l'hypothèse que ce critère pourrait être lié à la catégorie *niveau de langue*.

Redondance : Ce critère évalue la redondance du contenu du discours, en considérant si elle améliore la clarté, introduit des répétitions utiles à des fins d'insistance, ou présente une redondance excessive qui nuit à la transmission du message. Selon (Cao & Zhuge, 2019), la concision est définie par l'absence de redondance entre les phrases. Nous émettons l'hypothèse que ce critère pourrait être relié à la catégorie *niveau de langue*.

Langage négatif : Ce critère évalue l'utilisation de mots ou d'expressions à connotation négative dans le discours, en considérant s'ils sont utilisés de manière efficace, perceptibles sans impact notable, ou excessivement négatifs au point de nuire à l'impression générale. Nous nous appuyons sur les travaux de (Martin, 2017) qui illustrent les effets différenciés du bouche-à-oreille positif et négatif. Nous formulons l'hypothèse que des liens seront observables avec la catégorie *langage en relation avec les processus affectifs, cognitifs et de perception*.

Métaphore : Ce critère évalue l'utilisation des métaphores dans le discours, en considérant si elles améliorent l'explication, si elles sont présentes mais perfectibles, ou absentes. Nous nous appuyons sur des résultats issus de la littérature, notamment (Goatly, 1997), qui souligne l'importance du langage métaphorique, ainsi que sur des études telles que (Ortony, 1993), qui mettent en évidence son lien avec des explications mémorables et efficaces. Nous avons ainsi intégré un critère spécifique portant sur l'usage du langage métaphorique.

Storytelling : Ce critère évalue la dimension storytelling du discours, en considérant si elle est absente, présente mais hors sujet, ou utilisée de manière pertinente et efficace par rapport au contenu. La littérature propose plusieurs études sur la relation entre narration et maîtrise linguistique (Zuhriyah, 2017; Natasia & Angelianawati, 2022). La narration est également liée à le pouvoir de persuasion en marketing (Zubiel-Kasprowicz, 2016) ainsi qu'au développement des compétences pédagogiques (Morrison & Lorusso, 2023). Par conséquent, nous avons intégré une évaluation de la narration dans nos critères.

4 Détails expérimentaux

Mesures d'évaluation des performances. Nous comparons les évaluations du pouvoir de persuasion générées par les LLMs à la vérité de terrain (GT) à l'aide de l'Erreur Quadratique Moyenne (RMSE), du Coefficient de Détermination (R^2) et de l'Erreur Médiane Absolue (MedAE) [6]. Nous notons les métriques comme des fonctions $A(\cdot, \cdot)$ entre les scores du pouvoir de persuasion prédits et ceux de référence.

6. Des métriques supplémentaires sont disponibles sur notre GitHub

Métriques d'évaluation de l'auto-cohérence. Nous évaluons l'auto-cohérence des LLMs à l'aide du Coefficient de Corrélation Intraclasse (ICC), conformément à (Koo & Li, 2016). En appliquant le modèle à effets mixtes à deux voies (juges fixes, sujets aléatoires) avec estimation de la cohérence ($ICC_{3,1}$), nous mesurons la fidélité des évaluations individuelles du modèle, selon (McGraw & Wong, 1996). Cette méthode consiste à évaluer chaque échantillon d'un ensemble fixe N à l'aide de trois exécutions du même modèle m sur le même échantillon i et le même prompt p.

Choix des modèles open source. Nous évaluons trois des meilleurs LLM open source : Llama2 (Touvron *et al.*, 2023), Mistral (Jiang *et al.*, 2023) et Llama3, sélectionnés pour leurs performances élevées et leur capacité à traiter nos prompts de 900 tokens. Llama2 est particulièrement performant pour les tâches à long contexte (Xu *et al.*, 2023), Mistral 7B surpasse Llama2-13B (Jiang *et al.*, 2023), et Llama3 dépasse Mistral sur les benchmarks GPQA et GSM8K. Tous ces modèles figurent parmi les quatre premiers du classement HuggingFace Open LLM Leaderboard. Les modèles sont utilisés via Ollama, avec une quantification en 4 bits [7].

Modèle commercial de référence. En complément des modèles open source, nous utilisons GPT-4o-mini. Les performances et l'efficacité économique de GPT-4o-mini en font un point de comparaison pertinent avec les modèles open source (OpenAI, 2024b).

Température et top_p du modèle. Étant donné la diversité des familles de modèles et la sensibilité des LLMs à leurs paramètres, nous avons choisi d'utiliser les valeurs par défaut recommandées pour la température et le top_p de chaque modèle.

5 Expériences et résultats

5.1 QR-1

Pour répondre à la **QR-1**, nous comparons les scores du pouvoir de persuasion prédits par les LLMs aux scores de référence (*c.-à-d.*, les annotations humaines issues de (Biancardi *et al.*, 2024) sur une échelle de Likert à 5 points). Chaque modèle $m \in M = \{\text{LLaMA2, LLaMA3, Mistral, GPT-4o-mini}\}$ est testé trois fois avec le même prompt $p \in P$, les essais étant indexés par $id \in \{1, 2, 3\}$. La sortie de l'essai id pour le modèle m avec le prompt p est notée $\{\hat{y}_{i_{p,m,id}}\}_{i \in N}$, où N est l'ensemble des échantillons de données. Pour la métrique d'évaluation $A(\{\hat{y}_{i_{p,m,id}}\}_{i \in N}, \{y_i\}_{i \in N})$ qui compare les prédictions aux scores de référence, nous calculons la moyenne de la métrique sur les trois essais pour chaque paire modèle-prompt (m, p) (en raison d'une dépendance observée aux prompts [8], nous rapportons les métriques moyennées sur l'ensemble des prompts $p \in P$ pour chaque modèle) :

$$AM_m = \frac{1}{|P|} \sum_{p \in P} \frac{1}{3} \sum_{id \in \{1,2,3\}} A(\{\hat{y}_{i_{p,m,id}}\}_{i \in N}, \{y_i\}_{i \in N}) \quad (1)$$

Nous reconnaissons que l'on pourrait également moyenner les sorties du LLM avant de calculer la mesure $A(\cdot, \cdot)$. Cependant, dans une tâche subjective telle que la prédiction de la persuasion,

7. Llama3 : https://ai.meta.com/blog/meta-llama-3/; Open LLM Leaderboard : https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard; Ollama : https://ollama.com/blog

8. Les résultats détaillés pour chaque prompt sont disponibles sur GitHub.

où les sorties du modèle peuvent présenter une forte variabilité et un désaccord important entre générations, il est plus pertinent de calculer $A(\cdot,\cdot)$ pour chaque sortie, puis d'en faire la moyenne, afin de mieux refléter la distribution réelle des jugements produits par le modèle. Le Tableau 2a présente les résultats pour l'ensemble des métriques $A(\cdot,\cdot)$ en utilisant l'intégralité du jeu de données (N). Nous notons que ces résultats reflètent également les comportements relatifs des modèles pour chaque prompt. La RMSE, exprimée dans les mêmes unités que la variable cible, révèle une erreur moyenne importante d'environ $0,9$ points pour tous les modèles open-source, ce qui est significatif pour une échelle de Likert à 5 points. Les valeurs négatives de R^2 pour l'ensemble des modèles soulignent une divergence nette entre les scores du pouvoir de persuasion prédits par les LLMs et les scores de référence, pour les trois modèles open-source. La MedAE met en évidence des erreurs typiques de $0,61$ pour LLaMA2, $0,6$ pour Mistral, et $0,48$ pour LLaMA3, indiquant des erreurs plus faibles pour LLaMA3. GPT-4o-mini affiche des performances inférieures sur l'ensemble des métriques. Des tests supplémentaires de Kolmogorov-Smirnov (K-S) ainsi qu'une régression linéaire montrent des différences significatives de distribution entre GPT-4o-mini et les modèles open-source, avec des statistiques K-S de $0,8$ pour Mistral et $0,5$ pour LLaMA2 ($p < 0,001$). Les pentes de régression s'écartent de 1, suggérant que les performances faibles de GPT-4o-mini résultent d'une distribution différente, et non d'un simple décalage. Sa valeur faible de $\mathbf{R}^2$ confirme la difficulté du modèle à prédire le pouvoir de persuasion. Par ailleurs, nous mesurons la qualité de l'évaluation des modèles en termes d'auto-cohérence [9]. À cette fin, pour chaque paire modèle m et prompt p, nous calculons l'accord entre les évaluations issues de trois exécutions :

$$ICC_{3,1 \text{ or } k}(\{\hat{y}_{ip,m,id=1}\}_{i\in N}, \{\hat{y}_{ip,m,id=2}\}_{i\in N}, \{\hat{y}_{ip,m,id=3}\}_{i\in N}) \tag{2}$$

Pour le prompt présentée dans le Tableau 1, nous avons observé une forte auto-cohérence pour le modèle Mistral ($ICC_{3,1} = 0,73$ avec un intervalle de confiance (IC) à 95% $[0,61; 0,83]$), tandis que LLaMA2, LLaMA3 et GPT-4o-mini ont montré de niveaux faibles de cohérence ($ICC_{3,1}$ égal à $0,39$, $0,34$ et $0,22$ respectivement). Les autres prompts ont démontré les mêmes tendances.

LLM	RMSE	R^2	MedAE
LLaMA2	**0.89**	-1.72	0.61
Mistral	0.9	-1.52	0.6
LLaMA3	**0.89**	-1.59	**0.48**
GPT-4o-mini	1.48	-7.05	1.34

(a) Prédiction du pouvoir de persuasion par différents LLMs.

Entrée	RMSE	R^2	MedAE
Critères évalués par LLM	0.59	**-0.02**	**0.35**
Caractéristiques lexicales	**0.51**	-0.03	0.40
LLaMA3	0.80	-0.80	0.38
Ligne de base : moyenne	0.60	-0.012	0.36

(b) ElasticNet pour la prédiction du pouvoir de persuasion.

TABLEAU 2 – Évaluation de la prédiction du pouvoir de persuasion. Les meilleures métriques sont indiquées en gras.

5.2 QR-2

Afin de répondre à la **QR-2** et de comparer les critères évalués par les LLM, considérés comme de nouvelles caractéristiques lexicales, aux caractéristiques lexicales, nous utilisons les deux ensembles de variables comme entrées du modèle ElasticNet [10] pour prédire le pouvoir de persuasion. Nous

9. Les coefficients de corrélation intra-classe (ICC) pour chaque prompt sont disponibles sur GitHub
10. D'autres modèles de régression ont également été testés ; les résultats sont disponibles sur notre GitHub.

avons appliqué une division des données en 80/20% pour l'entraînement et le test, et avons comparé les résultats à ceux de la prédiction du pouvoir de persuasion en zéro-shot par le modèle LLaMA3 (ayant obtenu les meilleurs résultats dans la **QR-1**), ainsi qu'au modèle de référence qui prédit simplement la moyenne du pouvoir de persuasion.

À partir du Tableau 2b, nous observons que le modèle ElasticNet utilisant des caractéristiques lexicales surpasse légèrement la prédiction moyenne de référence, tandis que l'utilisation des critères comme variables explicatives donne des résultats proches de cette prédiction moyenne, bien que légèrement inférieurs à ceux des caractéristiques lexicales. La RMSE indique une erreur de prédiction moyenne d'environ 0.5 à 0.6 pour les modèles de régression, contre 0.8 points pour LLaMA3 (AM_m sur les données de test N), ce qui montre que l'approche fondée sur les caractéristiques est plus précise. Les valeurs négatives de R^2 traduisent une concordance faible avec les scores du pouvoir de persuasion de référence. Une valeur de MedAE de 0.35 indique une erreur typique plus faible pour les critères évalués par modèle de langue par rapport aux caractéristiques lexicales (0.40) et à la prédiction du pouvoir de persuasion en zéro-shot par LLM (0.38).

Enfin, nous avons étudié si les critères évalués par le LLM pouvaient être prédits à partir de caractéristiques lexicales, c'est-à-dire s'il existe une correspondance entre ces deux types de données. En utilisant une forêt aléatoire avec pondération des classes pour gérer les données déséquilibrées, nous prédisons les critères évalués par le LLM à partir des caractéristiques lexicales. La performance est évaluée à l'aide de la précision équilibrée, de la précision (precision), du rappel (recall) et du score F1. De plus, nous calculons le Kappa de Cohen (Cohen, 1960) entre les critères évalués par le LLM et ceux classifiés à partir des caractéristiques lexicales. En guise de référence, nous utilisons un classifieur à vote majoritaire. Nous présentons les résultats pour les critères évalués par le LLM dont les classes sont équilibrées (A, B et C ayant un nombre d'exemples similaire) dans le Tableau 3, avec les résultats de référence indiqués entre parenthèses.

Dans l'ensemble, les résultats se sont révélés assez faibles, ce qui suggère que les critères évalués par les LLMs sont difficiles à prédire à partir des seules caractéristiques lexicales. Seule la classification des critères liés au niveau de langue dépasse légèrement la performance de la ligne de base. Par ailleurs, le κ de Cohen indique également un léger accord (selon (Landis & Koch, 1977)) uniquement entre la prédiction des critères liés au niveau de langue et leur évaluation par le LLM, ce qui suggère que les caractéristiques lexicales partagent le plus d'information avec ces critères. Toutefois, les performances modestes obtenues dans la classification des critères soulèvent des questions quant à la fiabilité des LLM pour leur évaluation, une problématique que nous envisageons d'approfondir dans des travaux futurs.

Critère	PE	P	R	F1	κ	CD
Redondance	0.21 (0.33)	0.21 (0.17)	0.22 (0.41)	0.21 (0.24)	-0.19	A : 14, B : 17, C : 10
Niveau de langue	0.52 (0.5)	0.57 (0.4)	0.61 (0.63)	0.56 (0.49)	0.05	A : 15, B : 26
Voix Passive	0.27 (0.33)	0.41 (0.43)	0.54 (0.66)	0.47 (0.52)	-0.12	A : 8, B : 6, C : 27

TABLEAU 3 – Résultats du modèle forêt aléatoire pour la prédiction des critères, avec les résultats du classificateur par vote majoritaire indiqués entre parenthèses. Les métriques incluent l'la Précision Équilibrée (PE), la Précision (P), le Rappel (R), le Score F1, le Kappa de Cohen (κ) et la Distribution des Classes (CD).

6 Conclusions

Nos expériences nous ont permis de dégager les constats suivants :

1. **Les modèles open-source surpassent les modèles commerciaux mais présentent des difficultés dans la prédiction du pouvoir de persuasion en zero-shot** : Le meilleur modèle, LLaMa3, atteint une RMSE de $0,89$, contre $1,48$ pour GPT-4o-mini. Ces résultats confirment notre hypothèse selon laquelle les grandes modèles de langue obtiennent de performances faibles dans l'évaluation des dimensions subjectives. Par ailleurs, tous les modèles sauf Mistral présentent une **auto-cohérence faible**, bien qu'elle dépasse l'accord inter-juges du jeu de données 3MT_French (11%).
2. **Les critères évalués par les LLM améliorent la prédiction directe du pouvoir de persuasion par les LLM, mais restent moins performants qu'un modèle de régression avec des caractéristiques lexicales** : L'utilisation des critères évalués par les LLM comme variables explicatives pour la prédiction du pouvoir de persuasion (RMSE $0,6$) permet d'améliorer les performances par rapport à la prédiction directe par LLM (RMSE $0,8$), mais reste inférieure à celle des modèles d'apprentissage automatique classiques (RMSE de $0,5$), confirmant que les caractéristiques lexicales demeurent la méthode la plus précise pour prédire le pouvoir de persuasion.
3. **Les critères évalués par les LLM ne constituent pas des représentations de haut niveau efficaces des caractéristiques lexicales** : Seuls les critères de niveau linguistique classifiés à partir des caractéristiques lexicales (F1-score $0,56$) surpassent le classificateur de vote majoritaire (F1-score $0,49$), avec un coefficient κ de Cohen positif ($0,05$) indiquant une distribution légèrement similaire à la vérité terrain. Aucun autre critère ne montre de telles corrélations, ce qui suggère qu'ils ne peuvent pas être utilisés comme représentations interprétables de haut niveau des caractéristiques lexicales.

Comparé aux systèmes précédents (Schneider *et al.*, 2015; Kurihara *et al.*, 2007), nous proposons un nouveau cadre pour les recherches futures visant à caractériser les performances des grandes modèles de langue dans l'évaluation automatique des discours publics, en s'appuyant sur des critères bien définis issus de la littérature pédagogique et de l'expertise de coachs en prise de parole. Il est important de noter que les LLMs n'ont pas été entraînés sur les données utilisées, ce qui souligne que leurs performances faibles ne sont pas dues à la qualité des données, mais plutôt à leur incapacité à mobiliser des connaissances générales pour évaluer une tâche subjective. Les écarts entre les résultats prédictifs obtenus avec les LLMs et ceux de travaux antérieurs (*e.g.* 66 % de précision (Park *et al.*, 2014)) soulèvent des interrogations quant à la fiabilité des systèmes reposant sur des modèles génératifs tels que Yoodli.

Plusieurs limites doivent toutefois être prises en compte. La qualité imparfaite des transcriptions Whisper, notamment en français, peut affecter l'extraction des caractéristiques lexicales et la fiabilité des évaluations. Le corpus *3MT_French*, bien qu'adapté au contexte éducatif exploré ici, reste limité en taille, biaisé vers des discours bien notés, et insuffisant pour des approches de type fine-tuning. Les critères utilisés, fondés sur des retours d'experts en art oratoire, sont interprétables mais ne couvrent pas toute la complexité de la concept de persuasion. Enfin, le choix de modèles, combiné à une quantification agressive et un prompting simplifié, restreint les capacités de raisonnement des LLMs. Des travaux futurs devront intégrer des modèles plus récents, optimisés pour le raisonnement, et évaluer notre approche dans des contextes discursifs plus variés et enrichis d'annotations humaines.

Remerciements

Cette recherche a été partiellement financée par la subvention ANR REVITALISE ANR-21-CE33-0016-02 et la subvention ANR SINnet ANR-23-CE23-0033-01. Nous signalons que ChatGPT-4o (OpenAI, 2024a) a été utilisé pour la correction grammaticale dans l'ensemble des sections de cet article. Nous remercions les évaluateurs D'ICAART et de l'atelier EvalLLM2025 pour leurs commentaires détaillés, qui nous ont permis de mettre en valeur les aspects les plus pertinents de notre travail.

Références

ABU-EL-HAIJA S., KOTHARI N., LEE J., NATSEV P., TODERICI G., VARADARAJAN B. & VIJAYANARASIMHAN S. (2016). Youtube-8m : A large-scale video classification benchmark.

AI M. (2023). Llama 2 : Open foundation and fine-tuned chat models. Accessed : September 17, 2024.

ASHWIN T. S. & RAJENDRAN R. (2022). Audio feature based monotone detection and affect analysis for teachers. In *2022 IEEE Region 10 Symposium (TENSYMP)*, p. 1–6.

BARKAR A., CHOLLET M., BIANCARDI B. & CLAVEL C. (2023). Insights into the importance of linguistic textual features on the persuasiveness of public speaking. In *Companion Publication of the 25th International Conference on Multimodal Interaction*, p. 51–55 : Association for Computing Machinery. DOI : 10.1145/3610661.3617161.

BIANCARDI B., CHOLLET M. & CLAVEL C. (2024). Introducing the 3mt_french dataset to investigate the timing of public speaking judgements. In *Language Resources and Evaluation*, p. 1–20 : Springer.

BINZ M. & SCHULZ E. (2022). Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences of the United States of America*, **120**.

CAO H., COOPER D. G., KEUTMANN M. K., GUR R. C., NENKOVA A. & VERMA R. (2014). Crema-d : Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, **5**(4), 377–390. DOI : 10.1109/TAFFC.2014.2336244.

CAO M. & ZHUGE H. (2019). Automatic evaluation of text summarization based on semantic link network. In *15th International Conference on Semantics, Knowledge and Grids (SKG)*, p. 107–114. DOI : 10.1109/SKG49510.2019.00026.

CHEN L., LEONG C. W., FENG G., LEE C. M. & SOMASUNDARAN S. (2015). Utilizing multimodal cues to automatically evaluate public speaking performance. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, p. 394–400.

CHENG M., DURMUS E. & JURAFSKY D. (2023). Marked personas : Using natural language prompts to measure stereotypes in language models. arXiv : 2305.18189.

CHHUN C., SUCHANEK F. M. & CLAVEL C. (2024). Do language models enjoy their own stories ? prompting large language models for automatic story evaluation. arXiv : 2405.13769.

CHOLLET M. & LEFEBVRE L. (2022). *Livrable REVITALISE - D1.2 Grille d'évaluation de la prise de parole en public*. Rapport interne, IMT Atlantique, Département Automatique, Productique et Informatique (DAPI), Campus de Nantes. https://hal.science/hal-04529070.

COHEN J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**, 37–46.

COURGEON M., MARTIN J.-C., MUTLU B., PICARD R. & HOQUE M. (2014). Mach : My automated conversation coach. In *UbiComp 2013 - Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. DOI : 10.1145/2493432.2493502.

DAS S., HAQUE S. A. & TANVEER M. I. (2021). Persistence homology of tedtalk : Do sentence embeddings have a topological shape ? arXiv preprint arXiv :2103.14131. https://api.semanticscholar.org/CorpusID:232380031.

DINKAR T., VASILESCU I., PELACHAUD C. & CLAVEL C. (2020). How confident are you ? exploring the role of fillers in the automatic prediction of a speaker's confidence. In *ICASSP*, p. 8104–8108. DOI : 10.1109/ICASSP40776.2020.9054374.

EPF (2013). Grille critériée pour l'évaluation des exposés oraux en phy 111 & 112. https://ilearn.epf.fr/formation-enseignants/grille-evaluation/docs-compleementaires-1.pdf. Accès le 18 octobre 2024.

EPHRAT A., MOSSERI I., LANG O., DEKEL T., WILSON K., HASSIDIM A., FREEMAN W. T. & RUBINSTEIN M. (2018). Looking to listen at the cocktail party : a speaker-independent audio-visual model for speech separation. *ACM Transactions on Graphics*, **37**(4), 1–11. DOI : 10.1145/3197517.3201357.

EYBEN F., SCHERER K. R., SCHULLER B. W., SUNDBERG J., ANDRÉ E., BUSSO C., DEVILLERS L. Y., EPPS J., LAUKKA P., NARAYANAN S. S. & TRUONG K. P. (2016). The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, **7**(2), 190–202. DOI : 10.1109/TAFFC.2015.2457417.

GAN T., WONG Y. K., MANDAL B., LI J., CHANDRASEKHAR V. & KANKANHALLI M. S. (2017). Nus multi-sensor presentation (nusmsp) dataset. http://mmas.comp.nus.edu.sg/NUSMSP.html. Dataset from the National University of Singapore (NUS).

GILARDI F., ALIZADEH M. & KUBLI M. (2023). Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences of the United States of America*, **120**. https://api.semanticscholar.org/CorpusID:257766307.

GOATLY A. (1997). *The Language of Metaphors*. Routledge, 1 édition. DOI : 10.4324/9780203210000.

HONNIBAL M. & MONTANI I. (2017). spaCy 2 : Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

HUANG J. T., WANG W., LAM M. H., LI E. J., JIAO W. & LYU M. R. (2023). Revisiting the reliability of psychological scales on large language models. arXiv preprint arXiv :2305.19926.

JIANG A. Q., SABLAYROLLES A., MENSCH A., BAMFORD C., CHAPLOT D. S., DE LAS CASAS D., BRESSAND F., LENGYEL G., LAMPLE G., SAULNIER L., LAVAUD L. R., LACHAUX M.-A., STOCK P., LE SCAO T., LAVRIL T., WANG T., LACROIX T. & EL SAYED W. (2023). Mistral 7b. *arXiv*, **abs/2310.06825**. arXiv : 2310.06825.

KOO T. K. & LI M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, **15**(2), 155–163.

KORINI K. & BIZER C. (2023). Column type annotation using chatgpt. arXiv preprint. arXiv : 2306.00745.

KURIHARA K., GOTO M., OGATA J., MATSUSAKA Y. & IGARASHI T. (2007). Presentation sensei : a presentation training system using speech and image processing. In *Proceedings of the 9th International Conference on Multimodal Interfaces*, p. 358–365. DOI : 10.1145/1322192.1322256.

LABRAK Y. & DUFOUR R. (2022). Antilles : An open french linguistically enriched part-of-speech corpus. In *25th International Conference on Text, Speech and Dialogue (TSD)*, Brno, Czech Republic : Springer. https://hal.archives-ouvertes.fr/hal-03696042.

LAMPRINIDIS S. (2023). Llm cognitive judgements differ from human. *arXiv*, **abs/2307.11787**. arXiv : 2307.11787.

LANDIS J. R. & KOCH G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, **33**(1), 159–174. Kappa Interpretation Table retrieved from ResearchGate : https://www.researchgate.net/figure/Criteria-for-the-Interpretation-of-Kappa-values-by-Landis-Koch-1977_tbl1_259976499.

LARRIMORE L., JIANG L., LARRIMORE J., MARKOWITZ D. M. & GORSKI S. (2011). Peer to peer lending : The relationship between language features, trustworthiness, and persuasion success. *Journal of Applied Communication Research*, **39**, 19–37.

LATIF S., USAMA M., MALIK M. I. & SCHULLER B. (2023). Can large language models aid in annotating speech emotional data ? uncovering new frontiers. *ArXiv*, **abs/2307.06090**. https://api.semanticscholar.org/CorpusID:259837093.

MARTIN W. C. (2017). Positive versus negative word-of-mouth : Effects on receivers. *Academy of Marketing Studies Journal*, **21**, 1. https://api.semanticscholar.org/CorpusID:148681912.

MCGRAW K. O. & WONG S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, **1**(1), 30.

MELLONI G., CAGLIO A. & PEREGO P. (2017). Saying more with less ? disclosure conciseness, completeness and balance in integrated reports. SRPN : Corporate Reporting (Topic). CorpusID:29162676.

MORRISON H. J. & LORUSSO J. R. (2023). Developing teacher candidates' professional advocacy skills through persuasive storytelling. *Journal of Physical Education, Recreation & Dance*, **94**, 6–11.

NATASIA G. & ANGELIANAWATI L. (2022). Students' perception of using storytelling technique to improve speaking performance at smpn 143 jakarta utara. *JET (Journal of English Teaching)*. CorpusID:253017067.

NGUYEN A.-T., CHEN W. & RAUTERBERG G. W. M. (2012). Online feedback system for public speakers. In *2012 IEEE Symposium on E-Learning, E-Management and E-Services*, p. 1–5.

OPENAI (2024a). Gpt-4o-mini : A commercial large language model for public speaking analysis. https://openai.com. Accessed : 2024-10-21.

OPENAI (2024b). Gpt-4o-mini : Advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence.

ORTONY A., Éd. (1993). *Metaphor and Thought*. Cambridge University Press, 2 édition.

PARANJAPE B., MICHAEL J., GHAZVININEJAD M., ZETTLEMOYER L. & HAJISHIRZI H. (2021). Prompting contrastive explanations for commonsense reasoning tasks. In *Findings of the Association for Computational Linguistics : EMNLP 2021*.

PARK S., SHIM H. S., CHATTERJEE M., SAGAE K. & MORENCY L.-P. (2014). Computational analysis of persuasiveness in social multimedia : A novel dataset and multimodal prediction approach. In *Proceedings of the 16th International Conference on Multimodal Interaction*, p. 50–57, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/2663204.2663260.

PATTERSON D. A., GONZALEZ J., LE Q. V., LIANG C., MUNGUÍA L.-M., ROTHCHILD D., SO D. R., TEXIER M. & DEAN J. (2021). Carbon emissions and large neural network training. arXiv preprint arXiv :2104.10350.

PENNEBAKER J., BOYD R., BOOTH R., ASHOKKUMAR A. & FRANCIS M. (2022). Linguistic inquiry and word count : Liwc-22. In *Pennebaker Conglomerates*.

PIOLAT A., BOOTH R., CHUNG C., DAVIDS M. & PENNEBAKER J. (2011). La version française du dictionnaire pour le liwc : modalités de construction et exemples d'utilisation. *Psychologie Française*, **56**, 145–159. DOI : 10.1016/j.psfr.2011.07.002.

RADFORD A., KIM J. W., XU T., BROCKMAN G., MCLEAVEY C. & SUTSKEVER I. (2022). Robust speech recognition via large-scale weak supervision.

RODERO E. & LARREA O. (2022). Virtual reality with distractors to overcome public speaking anxiety in university students.

RUELAS INZUNZA E. (2020). Reconsidering the use of the passive voice in scientific writing. *The American Biology Teacher*, **82**(8), 563–565. JSTOR 48734696.

SCHNEIDER J., BÖRNER D., VAN ROSMALEN P. & SPECHT M. M. (2015). Presentation trainer, your public speaking multimodal coach. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, p. Presentation Trainer, your Public Speaking Multimodal Coach.

SCHREIBER L. & HARTRANFT M. (2017). Introduction to public speaking. In T. S. RICE, Éd., *Fundamentals of Public Speaking*. California : College of the Canyons.

SCHREIBER L. M., PAUL G. D. & SHIBLEY L. R. (2012). The development and test of the public speaking competence rubric. *Communication Education*, **61**, 205–233.

SONG W., WU B., ZHENG C. & ZHANG H. (2023). Detection of public speaking anxiety : A new dataset and algorithm. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, p. 2633–2638. DOI : 10.1109/ICME55011.2023.00448.

TOUVRON H., MARTIN L., STONE K., ALBERT P. *et al.* (2023). Llama 2 : Open foundation and fine-tuned chat models. arXiv : 2307.09288.

TUN S. S. Y., OKADA S., HUANG H.-H. & LEONG C. W. (2023). Multimodal transfer learning for oral presentation assessment. *IEEE Access*, **11**, 84013–84026.

WEBSON A. & PAVLICK E. (2022). Do prompt-based models really understand the meaning of their prompts ? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 2300–2344, Seattle, United States : Association for Computational Linguistics. DOI : 10.18653/v1/2022.naacl-main.167.

WEI J., WANG X., SCHUURMANS D., BOSMA M., ICHTER B., XIA F., CHI E., LE Q. & ZHOU D. (2023). Chain-of-thought prompting elicits reasoning in large language models. arXiv preprint arXiv :2201.11903.

XU P., PING W., WU X., MCAFEE L. C., ZHU C., LIU Z., SUBRAMANIAN S., BAKHTURINA E., SHOEYBI M. & CATANZARO B. (2023). Retrieval meets long context large language models. *ArXiv*, **abs/2310.03025**.

YANG Z., HUYNH J., TABATA R., CESTERO N., AHARONI T. & HIRSCHBERG J. (2020). What makes a speaker charismatic ? producing and perceiving charismatic speech. In *Speech Prosody 2020*.

ZHAO T. Z., WALLACE E., FENG S. *et al.* (2021). Calibrate before use : Improving few-shot performance of language models. arXiv preprint arXiv :2102.09690.

ZHU Y., ZHANG P., HAQ E.-U., HUI P. & TYSON G. (2023). Can chatgpt reproduce human-generated labels ? a study of social computing tasks. *ArXiv*, **abs/2304.10145**. https://api.semanticscholar.org/CorpusID:258236184.

ZUBIEL-KASPROWICZ M. (2016). Storytelling as modern architecture of narration in marketing communication. CorpusID:151719224.

ZUHRIYAH M. (2017). Storytelling to improve students' speaking skill. CorpusID:149116911.