Decoding Persuasiveness in Eloquence Competitions: An Investigation into the LLM's Ability to Assess Public Speaking

Alisa Barkar[1][a], Mathieu Chollet[2,3][b], Matthieu Labeau[1][c], Beatrice Biancardi[4][d] and Chloé Clavel[5][e]

[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, U.K.*
[3] *IMT Atlantique, LS2N, UMR CNRS 6004, 44307 Nantes, France*
[4] *CESI LINEACT, Nanterre, France*
[5] *ALMAnaCH, INRIA, Paris, France*

Keywords: Public Speaking Assessment, Large Language Models (LLMs), Persuasiveness Prediction, Interpretable Features, Textual Modality, Automatic Speech Evaluation, Open-Source Models.

Abstract: The increasing importance of public speaking (PS) skills has fueled the development of automated assessment systems, yet the integration of large language models (LLMs) in this domain remains underexplored. This study investigates the application of LLMs for assessing PS by predicting persuasiveness. We propose a novel framework where LLMs evaluate criteria derived from educational literature and feedback from PS coaches, offering new interpretable textual features. We demonstrate that persuasiveness predictions of a regression model with the new features achieve a Root Mean Squared Error (RMSE) of 0.6, underperforming approach with hand-crafted lexical features (RMSE 0.51) and outperforming direct zero-shot LLM persuasiveness predictions (RMSE of 0.8). Furthermore, we find that only LLM-evaluated criteria of language level is predictable from lexical features (F1-score of 0.56), disapproving relations between these features. Based on our findings, we criticise the abilities of LLMs to analyze PS accurately. To ensure reproducibility and adaptability to emerging models, all source code and materials are publicly available on GitHub.

1 INTRODUCTION

Public speaking (PS) is a vital skill for professional success and self-confidence (Schreiber and Hartranft, 2017), thus automated assessment systems have emerged to provide affordable training solutions for students and young professionals (Rodero and Larrea, 2022; Schneider et al., 2015; Kurihara et al., 2007). Recently Large Language Models (LLMs) have also penetrated the field of automatic PS assessment. Commercial systems, such as Poised[1], Yoodli[2], and GPT-based services[3],[4], have embraced this trend. Meanwhile, academic research has primarily focused on classification/regression models (Chen et al., 2015) and neural networks (Ashwin and Rajendran, 2022; Tun et al., 2023), resulting in a significant gap between research and commercial practice. **Our study addresses this gap by employing LLMs for a subjective dimension prediction (*e.g.*, persuasiveness, engagement, etc.)**, a commonly used approach in the literature on automatic PS assessment. Specifically, we concentrate on the persuasiveness dimension, which has been demonstrated to have a strong correlation with textual characteristics. For instance, (Larrimore et al., 2011) analyzed the relationship between word count and language types in the context of their persuasiveness in securing funding, while (Park

[a] https://orcid.org/0009-0002-7459-8917
[b] https://orcid.org/0000-0001-9858-6844
[c] https://orcid.org/0000-0003-1890-2814
[d] https://orcid.org/0000-0002-6664-6117
[e] https://orcid.org/0000-0003-4850-3398
[1] https://www.poised.com/about-us
[2] https://app.yoodli.ai/
[3] https://bit.ly/48eqllR
[4] https://bit.ly/3YeKnYy

Barkar, A., Chollet, M., Labeau, M., Biancardi, B. and Clavel, C.
Decoding Persuasiveness in Eloquence Competitions: An Investigation into the LLM's Ability to Assess Public Speaking.
DOI: 10.5220/0013158400003890

In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence (ICAART 2025) - Volume 3*, pages 538-546
ISBN: 978-989-758-737-5; ISSN: 2184-433X

et al., 2014) showed that textual features outperform visual ones in predicting persuasiveness.

Our objective is to evaluate the ability of LLMs to analyze public performance, focusing on the textual modality, which has received less attention compared to the extensively studied audio and visual modalities (Eyben et al., 2016; Nguyen et al., 2012; Chen et al., 2015). Existing approaches to textual aspect analysis often lack interpretability and usefulness for feedback. For example, prior research has relied on non-interpretable embeddings (Das et al., 2021) or hand-crafted features, such as uni-/bi-grams (Park et al., 2014), word rate, pause length, fillers (Dinkar et al., 2020), sentiment lexicons (Chen et al., 2015), and LIWC-based emotional features (Pennebaker et al., 2022), to assess speaker charisma (Yang et al., 2020). To address this problem with interpretability and usefulness, building on PS coaches' interviews and established assessment grids (Chollet and Lefebvre, 2022), **we developed a list of textual criteria essential for successful PS performance.** We then created an LLM-based method to evaluate these criteria and analyzed their relationship with existing hand-crafted features as well as their effectiveness in persuasiveness prediction.

Building on LLMs' success in tasks like political speech identification (Gilardi et al., 2023) and sentiment classification (Zhu et al., 2023; Latif et al., 2023), and their challenges in justifying automatic story evaluations (Chhun et al., 2024; Binz and Schulz, 2022; Lamprinidis, 2023), we hypothesized they could reliably analyze text and provide features but struggle with accurately evaluating persuasiveness. To enhance accessibility and reproducibility while reducing environmental impact, we focused on smaller, open-source LLMs that show a smaller carbon footprint (*e.g.* 539 tons for Llama 2 (AI, 2023) against 552 tons for GPT models (Patterson et al., 2021)). **Recognizing the rapid model advancements, we highlight our methodological contributions and offer an open-source implementation with results adaptable to newer models as well as additional analyses not covered in this paper on GitHub[5] .** Based on the identified gaps we formulate the following research questions (RQs):

RQ-1: How accurately can LLMs directly predict the persuasiveness of a speech based on its transcript?

RQ-2: Could LLM-evaluated criteria be a high-level representation of lexical features?

2 3MT_FRENCH DATASET

Motivations. Our long-term goal is to create an open-source system available for French students, therefore, we focus on the educational context. However, most available PS datasets are not relevant to educational settings. For example, the MIT Interview dataset (Courgeon et al., 2014) covers job interviews, POM (Park et al., 2014) focuses on film reviews, and AVSpeech (Ephrat et al., 2018) and YouTube-8M (Abu-El-Haija et al., 2016) target YouTube videos, SAC-LAD (Song et al., 2023) and CREMA-D (Cao et al., 2014) focus on anxiety and emotion, and NUSMSP (Gan et al., 2017) lack persuasiveness annotations. We, therefore, use the 3MT_French dataset (Biancardi et al., 2024), which offers crowd-sourced persuasiveness evaluations of French PhD students' presentations:

- 3-minute presentations from "Ma Thèse en 180 secondes" (*135 female, 113 male*) on diverse research topics.
- Each presentation was rated on a 5-point Likert scale for persuasiveness, adapted from the Public Speaking Competence Rubric (Schreiber et al., 2012). Three annotators per sample assessed how effectively the speaker constructed a credible, convincing message. For details on the annotation protocol, refer to (Biancardi et al., 2024).
- The dataset does not include transcripts. We generated them using Whisper (Radford et al., 2022), a state-of-the-art Automatic Speech Recognition model. After manual verification and removal of corrupted samples, we retained 227 transcripts.

3 METHODOLOGIES FOR PUBLIC SPEAKING EVALUATION

3.1 Zero-Shot Persuasiveness Prediction (RQ-1)

Prompting Techniques. This paper does not focus on prompt engineering, so we avoid advanced techniques such as Chain of Thought (Wei et al., 2023) or Contrastive Explanations (Paranjape et al., 2021). Instead, following natural language prompts from (Korini and Bizer, 2023), (Huang et al., 2023), and (Cheng et al., 2023), we use simple Instruction-based Zero-shot Prompting, providing transcripts and questions in separate sections. Our goal was to avoid

[5] https://github.com/abarkar/PublicSpeakingAnalysisWithLLMs

Table 1: Example of prompt structure for persuasiveness evaluation.

Prompt
Analysis description: **TASK**=[Evaluate transcript of the public performance given in the TRANSCRIPT section. To understand how to evaluate it, use the description of the dimension described in the DIMENSION section. Give ONLY one number as an answer - this number is a score and it has to be given based on the scale described in the SCALE section.] **SCALE**=[Scale from 1=not at all to 5=very much.] **DIMENSION**=[In your opinion, how persuasive is the speech in this transcript, *i.e.*, do the person effectively craft a convincing message? Is their reasoning rigorous?] **TRANSCRIPT**=[*transcript content*]

long transcripts with few-shot prompting and keep the task simple and concise for the small models. Given the strong influence of prompting on model outcomes (Zhao et al., 2021), (Webson and Pavlick, 2022), we test several prompts in English, French, and Multilingual settings[6]. Table 1 shows an example used in Section 5 for results presentation. To address **RQ-1**, we structure prompts into sections: TASK – contains instructions for the model, DIMENSION – provides dimension-related question, SCALE – defines the grading scale, and TRANSCRIPT – holds the transcript. Prompts are generated automatically for each transcript. We refer to individual prompt as p and the set of tested prompts as P.

Persuasiveness Definition for LLMs. To align with human scores from the 3MT_French dataset we adapt the question from (Biancardi et al., 2024) originally formulated in French. In Table 1 we provide an English translation.

Result Post-Processing and Notations. We automatically extract scores using the spaCy (Honnibal and Montani, 2017) library and filter out inconsistent LLM outputs (e.g., multiple predicted scores, grading scale changes, or text generation instead of scoring). Each result is manually reviewed to confirm inconsistencies. For any model referred to as m from the set M, and for any prompt $p \in P$, the LLM-predicted persuasiveness scores are denoted as $\{\hat{y}_{i_{p,m}}\}_{i \in N}$, where N represents the dataset samples (either full or test set).

3.2 New Interpretable Features vs Lexical Features (RQ-2)

Lexical Features. In order to validate the usefulness of the new interpretable features, we extracted several types of hand-crafted lexical features from (Barkar et al., 2023) that are categorized into three main groups: **Language Level** (lexical and syntactic properties of the text), **Richness of Vocabulary and Transitions** (variety and transitions within the vocabulary) and **Affective, Cognitive, and Perceptual Processes** (quantify the presence of affective (emotions), cognitive (thought processes), and perceptual (sensory and experiential) elements of LIWC (Pennebaker et al., 2022) dictionary)[7]. For details on the features refer to (Barkar et al., 2023). We use spaCy (Honnibal and Montani, 2017) for part-of-speech tagging, the French tagger (Labrak and Dufour, 2022) for fine-grained tags, and LIWC (Pennebaker et al., 2022) with the French dictionary (Piolat et al., 2011) for LIWC features.

Criteria Collection and Multiple-Choice Question Composition. To address **RQ-2**, we analyze speech using established French evaluation grids (*e.g.* (EPF, 2013), *"Ma thèse en 180 secondes" evaluation grid*). These criteria were validated through interviews with 11 French PS coaches (Chollet and Lefebvre, 2022) and supplemented with criteria from the international literature on persuasiveness and clarity. Each criteria was encoded as A, B or C[8]. While criteria may vary by culture, we aim for broad applicability, acknowledging cultural limitations. We used multiple-choice questions to create an evaluation prompt[9] for each criteria and transcript. Scores are referred to as LLM-evaluated criteria. The identified criteria are listed below and their relation to lexical features is discussed.

Topic Presentation. Rate the presentation of the subject based on clarity and originality. We drew upon (Chollet and Lefebvre, 2022).

Structure. Evaluate the structure of the speech based on clarity, organisation, and effectiveness of transitions. We drew upon (Chollet and Lefebvre, 2022). We hypothesise the structure criteria can be related to *richness of vocabulary and transitions* category of lexical features.

Language Level. Assesses the appropriateness of the language level in the speech based on clarity, avoidance of jargon, slang, or offensive terms, and richness of vocabulary and expression. We drew upon (Chollet and Lefebvre, 2022). We hypothesise that language level will be related to *language level* category.

Passive Voice. Rates the use of passive voice in the speech based on clarity and appropriateness, with consideration for whether it enhances clarity/objectivity or hinders understanding. Motivated by (Inzunza, 2020) who demonstrated the correlation

[6] Examples of the prompts can be found in our open-source GitHub repository

[7] The full feature list and formulas are available on our GitHub

[8] Questions available on our GitHub

[9] Examples available in our GitHub

between text perception (objective, ambiguous, clear) and the utilisation of passive voice.

Conciseness. Rates the length of the sentences in the speech based on clarity and readability, considering whether they are too short, appropriately varied, or excessively long and unclear. We drew upon (Melloni et al., 2017), who measure conciseness based on the length of the full text. However, since all examples in our dataset exhibited approximately the same length, we opted to use the length of sentences as the criteria. Thus, we assess the conciseness of each sentence. We hypothesise that this criteria can be related to *language level* category.

Redundancy. Evaluates the redundancy of the speech content, considering whether it enhances clarity, contains some repetition for emphasis, or includes excessive redundancy that obscures the message. According to (Cao and Zhuge, 2019), conciseness is defined by the absence of redundancy between sentences. We hypothesise connections to *language level* category.

Negative Language. Rates the use of negative words/expressions in the speech, considering whether they are effectively used, noticeable but not impactful, or excessively negative and creating a poor impression. We rely on the research of (Martin, 2017) who illustrated the differing effects of positive and negative word-of-mouth. We hypothesise connections will be observable with *language in relation to the Affective, Cognitive and Perception processes* category.

Metaphor. Rates the use of metaphors in the discourse, considering whether they enhance explanation, are present but could be improved, or are absent. Drawing upon evidence from the literature, such as (Goatly, 1997), which underscores the significance of metaphoric language, and studies like (Ortony, 1993), which elucidate its connection to memorable and efficient explanations, we incorporated criteria for metaphoric language usage.

Storytelling. Rates the narrative in the speech, considering whether it's absent, present but irrelevant, or effectively used and relevant to the content. The literature proposes several studies on the relationship between storytelling and language proficiency (Zuhriyah, 2017; Natasia and Angelianawati, 2022). Storytelling also is related to persuasiveness in marketing (Zubiel-Kasprowicz, 2016) and in the development of teaching skills (Morrison and Lorusso, 2023). Therefore, we integrated an evaluation of storytelling into our criteria.

4 EXPERIMENTAL DETAILS

Performance Evaluation Metrics. We compare persuasiveness evaluations from LLMs against ground truth (GT) using Root Mean Squared Error (RMSE), Coefficient of Determination (R^2), and Median Absolute Error (MedAE)[10]. We denote metrics as functions $A(\cdot,\cdot)$ between predicted and ground-truth persuasiveness scores.

Self-Consistency Evaluation Metrics. We evaluate LLM models' self-agreement using Intraclass Correlation Coefficients (ICC) following (Koo and Li, 2016). Applying the Two-Way Mixed Effects Model with consistency, we calculate $ICC_{3,1}$ for single rater/measurement, following (McGraw and Wong, 1996). This method rates each sample from a fixed set N using three runs of model m on the same sample i and prompt p.

Choice of Open-Source Models. We evaluate three top open-source LLMs [11]: Llama2 (Touvron et al., 2023), Mistral (Jiang et al., 2023), and Llama3[12], chosen for their strong performance and ability to handle our 900-token prompts. Llama2 excels in long-context tasks (Xu et al., 2023), Mistral 7B outperforms Llama2-13B (Jiang et al., 2023), and Llama3 surpasses Mistral on GPQA and GSM8K. All rank among the top four on the HuggingFace Open LLM Leaderboard[13]. Models are accessed via Ollama[14] using 4-bit quantization.

Commercial Baseline Model. In addition to the open-source models, we use GPT-4o-mini. GPT-4o-minis' performance and cost-efficiency make it a valuable comparison point for open-source models (OpenAI, 2024b).

Model temperature and Top_p. Given the diversity of model families and the parameter sensitivity of LLMs, we opted to use the recommended default temperature and top_p for each model.

5 EXPERIMENTS AND RESULTS

5.1 RQ-1

To address **RQ-1**, we compare LLM-predicted persuasiveness scores to ground-truth (GT) scores (*i.e.*, human annotations from (Biancardi et al., 2024)

[10] Additional metrics can be found on our GitHub

[11] For the moment of writing the paper

[12] https://ai.meta.com/blog/meta-llama-3/

[13] https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

[14] https://ollama.com/blog

Table 2: AM_m of LLM-predicted persuasiveness on full data. Lower RMSE and MedAE and higher R^2 indicate better quality of prediction. The best metrics of each column are highlighted in bold.

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

on a 5-point Likert scale). Each model $m \in M = \{\text{LLaMA2, LLaMA3, Mistral, GPT-4o-mini}\}$ is tested three times with the same prompt $p \in P$, with runs indexed by $id \in \{1,2,3\}$. The output of run id for model m with prompt p is denoted as $\{\hat{y}_{i\,p,m,id}\}_{i \in N}$, where N is the set of data samples. For the evaluation metric $A(\{\hat{y}_{i\,p,m,id}\}_{i\in N}, \{y_i\}_{i\in N})$ that compares predictions to GT scores we compute the average metric across runs for each model-prompt pair (m, p) (due to observed prompt dependency[15], we report metrics averaged over all prompts $p \in P$ for each model):

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

Table 2 presents results for all metrics $A(\cdot,\cdot)$ using the full dataset (N).

We note that the average results in Table 2 also reflect the relative behaviours of the models in each prompt. RMSE, in the same units as the target, reveals a large average error of about 0.9 points for all open-source models, which is significant for a 5-point Likert scale. Negative R^2 values across all models highlight a clear discrepancy between LLM-predicted and ground-truth persuasiveness scores for all three open-source models. MedAE highlights typical errors of 0.61 for LLaMA2, 0.6 for Mistral, and 0.48 for LLaMA3, showing LLaMA3's lower errors. GPT-4o-mini underperforms across all metrics. Additional Kolmogorov-Smirnov test (K-S) and linear regression show significant differences in distribution between GPT-4o-mini and open-source models, with K-S statistics of 0.8 for Mistral and 0.5 for LLaMA2 ($p < 0.001$). Regression slopes deviate from 1, suggesting GPT-4o-mini's poor performance results from a different distribution, not just a shift. Its low **R**^{2} confirms its difficulty in predicting persuasiveness. Additionally, we measure the quality of model evaluation in terms of self-agreements[16]. To that end, for each pair of model m and prompt p, we calculate the agreement between ratings in three runs:

Table 3: Elastic Net for persuasiveness prediction using different features against LLaMA3 and baseline on test data N. Best metric in bold.

Input	RMSE	R^2	MedAE
LLM-evaluated criteria	0.59	**-0.02**	**0.35**
Lexical features	**0.51**	-0.03	0.40
LLaMA3	0.80	-0.80	0.38
Baseline: mean prediction	0.60	-0.012	0.36

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

For the prompt, presented in Table 1 we observed high self-agreement for Mistral model ($ICC_{3,1} = 0.73$ with 95% confidential interval (CI) $[0.61, 0.83]$), while LLaMA2, LLaMa3 and GPT-4o-mini showed the low agreements ($ICC_{3,1}$ equals $0.39, 0.34$ and 0.22 respectively) other prompts demonstrated the same tendencies.

5.2 RQ-2

In order to address **RQ-2** and compare LLM-evaluated criteria as the new textual features to the lexical features, we use both feature sets as input for ElasticNet[17] to predict persuasiveness. We used an 80/20% train/test split and compared the results to the zero-shot persuasiveness prediction by the LLaMA3 model (which showed the best results in **RQ-1**) and to the baseline model that predicted simply the mean of the persuasiveness.

From Table 3, we observe that ElasticNet with lexical features slightly outperforms the baseline mean prediction, while using criteria as features achieves results close to the mean prediction, though slightly worse than lexical features. RMSE indicates an average prediction error of about $0.5 - 0.6$ for regression models against 0.8 points for LLaMA3 (AM_m on test data N), showing that the feature-based approach is more accurate. Negative R^2 values indicate poor alignment with GT persuasiveness. MedAE value of 0.35 shows a smaller typical error for LLM-evaluated criteria compared to lexical features (0.40) and zero-shot LLM-based persuasiveness prediction (0.38).

Finally, we studied whether the LLM-evaluated criteria can be predicted based on lexical features, i.e. whether there is some correspondence between these features. Using Random Forest with class weighting for imbalanced data, we predict LLM-evaluated criteria from lexical features. We assess performance with balanced accuracy, precision, recall, and F1-score. Additionally, we compute Cohen's Kappa (Cohen, 1960) between LLM-evaluated criteria and crite-

[15] Detailed results for each prompt are available on GitHub.

[16] ICCs for each prompt are available on GitHub

[17] Other regression models were also tested and results can be found on our GitHub.

Table 4: Random Forest results for criteria prediction with results for majority-vote classifier in brackets. Metrics of Balanced Accuracy (BA), Precision(P), Recall(R), F1-score, Cohen's Kappa(κ) and Class Distribution (CD).

Criteria	BA	P	R	F1	κ	CD
Redundancy	0.21 (0.33)	0.21 (0.17)	0.22 (0.41)	0.21 (0.24)	-0.19	A: 14, B: 17, C: 10
Lang. Level	0.52 (0.5)	0.57 (0.4)	0.61 (0.63)	0.56 (0.49)	0.05	A: 15, B: 26
Passive Voice	0.27 (0.33)	0.41 (0.43)	0.54 (0.66)	0.47 (0.52)	-0.12	A: 8, B: 6, C: 27

ria classified from lexical features. As a baseline, we use a majority-vote classifier. We report results for the LL-evaluated criteria with balanced classes (A, B and C having a close number of samples) in Table 4 with baseline results in brackets.

Overall, the results were fairly poor, suggesting that LLM-evaluated criteria are difficult to predict from lexical features. Only the language level criteria classification performs slightly better than the baseline. Besides, Cohen's κ also shows a slight agreement (according to (Landis and Koch, 1977)) only between the prediction of the language level criteria and its LLM evaluation, which indicates that the handcrafted features share the most information with the language level LLM-evaluated criteria. However, the poor performance in classifying criteria raises concerns about the reliability of LLMs in evaluating these criteria, which we intend to explore in future work.

6 CONCLUSIONS

Our experiments have led us to draw the following findings:

1. **Open-Source Models Outperform Commercial Models but Struggle with Zero-Shot Persuasiveness Prediction.** The best-performing model, LLaMa3, achieved an RMSE of 0.89, compared to 1.48 by GPT-4o-mini. These results support our hypothesis that LLMs perform poorly in subjective dimension evaluation. Furthermore, all models except Mistral exhibit **low self-consistency**, though it exceeds the inter-rater agreement in the 3MT_French dataset (11%).
2. **LLM-Evaluated Criteria Improve Direct Persuasiveness Prediction by LLMs but Underperform Compared to a Regression Model with Hand-Crafted Features.** While using LLM-evaluated criteria as features for persuasiveness evaluation (RMSE 0.6) improves upon direct LLM prediction (RMSE 0.8), it still lags behind classical ML models (RMSE of 0.5), confirming that hand-crafted features remain the most accurate approach for predicting persuasiveness.
3. **LLM-Evaluated Criteria Are not Effective as High-Level Representations of Lexical Features.** Only language-level criteria classified from lexical features (F1-score 0.56) outperform the baseline majority-vote classifier (F1-score 0.49), with positive Cohen's κ (0.05) showing slightly similar to ground truth distribution. No other criteria show such connections, indicating they cannot serve as high-level interpretable representations of lexical features.

Compared to prior systems (Schneider et al., 2015; Kurihara et al., 2007) we provide a new framework for future research to capture the performance characteristics of LLMs in automatic public speech evaluation using well-formulated criteria based on educational literature and the expertise of PS coaches. Notably, LLMs were not trained on the data, highlighting that the poor performance is not due to data quality but rather the LLMs' inability to leverage general knowledge for assessing a subjective task. These results should be validated with other PS datasets to investigate whether similar findings emerge in different contexts (e.g., vlogs or TED talks in languages other than French or with speakers of more heterogeneous levels). **Differences in LLM-based prediction outcomes with prior research (*e.g.* 66% accuracy (Park et al., 2014)) raise questions about the reliability of systems relying on generative models such as Yoodli[18].** Further research is needed to assess the impact of different prompting techniques and parameter choices on LLM performance. Finally, comparing LLM-annotated criteria with expert annotations will be a key next step.

ACKNOWLEDGEMENTS

This research was partly funded under the ANR REVITALISE grant ANR-21-CE33-0016-02 and the ANR SINnet grant ANR-23-CE23-0033-01. We note that ChatGPT-4o (OpenAI, 2024a) was used for grammar correction in all sections of this paper. We thank the reviewers for their detailed feedback that helped us to present the most interesting insights of our work.

REFERENCES

Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., and 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.

[18] https://app.yoodli.ai/

Ashwin, T. S. and Rajendran, R. (2022). Audio feature based monotone detection and affect analysis for teachers. *2022 IEEE Region 10 Symposium (TENSYMP)*, pages 1–6.

Barkar, A., Chollet, M., Biancardi, B., and 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*, pages 51–55. Association for Computing Machinery.

Biancardi, B., Chollet, M., and Clavel, C. (2024). Introducing the 3mt_french dataset to investigate the timing of public speaking judgements. In *Language Resources and Evaluation*, pages 1–20. Springer.

Binz, M. and 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., and Verma, R. (2014). CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390.

Cao, M. and Zhuge, H. (2019). Automatic evaluation of text summarization based on semantic link network. In *2019 15th International Conference on Semantics, Knowledge and Grids (SKG)*, pages 107–114.

Chen, L., Leong, C. W., Feng, G., Lee, C. M., and Somasundaran, S. (2015). Utilizing multimodal cues to automatically evaluate public speaking performance. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 394–400.

Cheng, M., Durmus, E., and Jurafsky, D. (2023). Marked personas: Using natural language prompts to measure stereotypes in language models.

Chhun, C., Suchanek, F. M., and Clavel, C. (2024). Do language models enjoy their own stories? prompting large language models for automatic story evaluation.

Chollet, M. and Lefebvre, L. (2022). Livrable REVITALISE - D1.2 Grille d'évaluation de la prise de parole en public. Technical report, IMT Atlantique, Département Automatique, Productique et Informatique (DAPI), Campus de Nantes, 4 rue Alfred Kastler, 44300 Nantes Cedex 3.

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., and Hoque, M. (2014). Mach: My automated conversation coach.

Das, S., Haque, S. A., and Tanveer, M. I. (2021). Persistence homology of tedtalk: Do sentence embeddings have a topological shape? *ArXiv*, abs/2103.14131.

Dinkar, T., Vasilescu, I., Pelachaud, C., and Clavel, C. (2020). How confident are you? exploring the role of fillers in the automatic prediction of a speaker's confidence. In *ICASSP*, pages 8104–8108.

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. Accessed: 2024-10-18.

Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W. T., and 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.

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., and Truong, K. P. (2016). The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Trans. Affect. Comput.*, 7(2):190–202.

Gan, T., Wong, Y. K., Mandal, B., Li, J., Chandrasekhar, V., and 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., and 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.

Goatly, A. (1997). *The Language of Metaphors*. Routledge, 1st edition.

Honnibal, M. and 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., and Lyu, M. R. (2023). Revisiting the reliability of psychological scales on large language models.

Inzunza, E. R. (2020). Reconsidering the use of the passive voice in scientific writing. *The American Biology Teacher*, 82(8):563–565.

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., and El Sayed, W. (2023). Mistral 7b. *ArXiv*, abs/2310.06825.

Koo, T. K. and Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. In *Journal of Chiropractic Medicine*, volume 15, pages 155–163.

Korini, K. and Bizer, C. (2023). Column type annotation using chatgpt. In *arXiv preprint*.

Kurihara, K., Goto, M., Ogata, J., Matsusaka, Y., and Igarashi, T. (2007). Presentation sensei: a presentation training system using speech and image processing. pages 358–365.

Labrak, Y. and 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.

Lamprinidis, S. (2023). Llm cognitive judgements differ from human. *ArXiv*, abs/2307.11787.

Landis, J. R. and 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., and 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., and Schuller, B. (2023). Can large language models aid in annotating speech emotional data? uncovering new frontiers. *ArXiv*, abs/2307.06090.

Martin, W. C. (2017). Positive versus negative word-of-mouth: Effects on receivers. *Academy of Marketing Studies Journal*, 21:1.

McGraw, K. O. and Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. In *Psychological Methods*, volume 1, page 30. American Psychological Association.

Melloni, G., Caglio, A., and Perego, P. (2017). Saying more with less? disclosure conciseness, completeness and balance in integrated reports. In *SRPN: Corporate Reporting (Topic)*.

Morrison, H. J. and Lorusso, J. R. (2023). Developing teacher candidates' professional advocacy skills through persuasive storytelling. In *Journal of Physical Education, Recreation & Dance*, volume 94, pages 6 – 11.

Natasia, G. and Angelianawati, L. (2022). Students' perception of using storytelling technique to improve speaking performance at smpn 143 jakarta utara. *JET (Journal of English Teaching)*.

Nguyen, A.-T., Chen, W., and Rauterberg, G. (2012). Online feedback system for public speakers. In *2012 IEEE Symposium on E-Learning, E-Management and E-Services*, pages 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.

Ortony, A., editor (1993). *Metaphor and Thought*. Cambridge University Press, 2nd edition.

Paranjape, B., Michael, J., Ghazvininejad, M., Zettlemoyer, L., and Hajishirzi, H. (2021). Prompting contrastive explanations for commonsense reasoning tasks. In *Findings*.

Park, S., Shim, H. S., Chatterjee, M., Sagae, K., and 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*, pages 50–57, New York, NY, USA. Association for Computing Machinery.

Patterson, D. A., Gonzalez, J., Le, Q. V., Liang, C., Munguía, L.-M., Rothchild, D., So, D. R., Texier, M., and Dean, J. (2021). Carbon emissions and large neural network training. *ArXiv*, abs/2104.10350.

Pennebaker, J., Boyd, R., Booth, R., Ashokkumar, A., and Francis, M. (2022). Linguistic inquiry and word count: Liwc-22. In *Pennebaker Conglomerates*.

Piolat, A., Booth, R., Chung, C., Davids, M., and Pennebaker, J. (2011). La version française du dictionnaire pour le liwc : Modalités de construction et exemples d'utilisation. In *Psychologie Française*, volume 56, pages 145–159.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision.

Rodero, E. and Larrea, O. (2022). Virtual reality with distractors to overcome public speaking anxiety in university students.

Schneider, J., Börner, D., van Rosmalen, P., and Specht, M. M. (2015). Presentation trainer, your public speaking multimodal coach. *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*.

Schreiber, L. and Hartranft, M. (2017). Introduction to public speaking. In Rice, T. S., editor, *Fundamentals of Public Speaking*. College of the Canyons, California.

Schreiber, L. M., Paul, G. D., and 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., and Zhang, H. (2023). Detection of public speaking anxiety: A new dataset and algorithm. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 2633–2638.

Touvron, H., Martin, L., Stone, K., Albert, P., et al. (2023). Llama 2: Open foundation and fine-tuned chat models.

Tun, S. S. Y., Okada, S., Huang, H.-H., and Leong, C. W. (2023). Multimodal transfer learning for oral presentation assessment. *IEEE Access*, 11:84013–84026.

Webson, A. and Pavlick, E. (2022). Do prompt-based models really understand the meaning of their prompts? In Carpuat, M., de Marneffe, M.-C., and Ruiz, I. V. M., editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344, Seattle, United States. Association for Computational Linguistics.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. (2023). Chain-of-thought prompting elicits reasoning in large language models.

Xu, P., Ping, W., Wu, X., McAfee, L. C., Zhu, C., Liu, Z., Subramanian, S., Bakhturina, E., Shoeybi, M., and Catanzaro, B. (2023). Retrieval meets long context large language models. *ArXiv*, abs/2310.03025.

Yang, Z., Huynh, J., Tabata, R., Cestero, N., Aharoni, T., and 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.

Zhu, Y., Zhang, P., Haq, E.-U., Hui, P., and Tyson, G. (2023). Can chatgpt reproduce human-generated labels? a study of social computing tasks. *ArXiv*, abs/2304.10145.

Zubiel-Kasprowicz, M. (2016). Storytelling as modern architecture of narration in marketing communication.

Zuhriyah, M. (2017). Storytelling to improve students' speaking skill.