frontiers | Frontiers in Communication

TYPE Original Research
PUBLISHED 07 August 2025
DOI 10.3389/fcomm.2025.1610404



OPEN ACCESS

EDITED BY
Anke van Kempen,
Munich University of Applied Sciences, Germany

REVIEWED BY
Florian Nafz,
Munich University of Applied Sciences, Germany
Matthew Laske,
University of North Texas, United States

*CORRESPONDENCE
Ralf Schmälzle
✉ schmaelz@msu.edu

RECEIVED 11 April 2025
ACCEPTED 08 July 2025
PUBLISHED 07 August 2025

CITATION
Schmälzle R, Lim S, Du Y and Bente G (2025) The art of audience engagement: LLM-based thin-slicing of scientific talks.
*Front. Commun.* 10:1610404.
doi: 10.3389/fcomm.2025.1610404



# The art of audience engagement: LLM-based thin-slicing of scientific talks

Ralf Schmälzle*, Sue Lim, Yuetong Du and Gary Bente

Department of Communication, Michigan State University, East Lansing, MI, United States

**Introduction:** This paper examines the thin-slicing approach—the ability to make accurate judgments based on minimal information—in the context of scientific presentations. Drawing on research from nonverbal communication and personality psychology, we show that brief excerpts (thin slices) of transcribed texts from real presentations reliably predict overall quality evaluations.

**Methods:** Using a novel corpus of over 100 real-life science talks, we employ Large Language Models (LLMs) to evaluate transcripts of full presentations and their thin slices. By correlating LLM-based evaluations of short excerpts with full-talk assessments, we determine how much information is needed for accurate predictions.

**Results:** Our results demonstrate that LLM-based evaluations align closely with human evaluations, proving their validity, reliability, and efficiency. Critically, even very short excerpts (<10% of a talk's transcript) strongly predict overall evaluations. This suggests that the first moments of a presentation convey relevant information that is used in quality evaluations and can shape lasting impressions. The findings are robust across different LLMs and prompting strategies.

**Discussion:** This work extends thin-slicing research to public speaking and connects theories of impression formation to LLMs and current research on AI communication. We discuss implications for communication and social cognition research on message reception. Lastly, we suggest an LLM-based thin-slicing framework as a scalable feedback tool to enhance human communication.

KEYWORDS

**public speaking, thin slices, science communication, audience engagement, impression formation**

## 1 Introduction

Humans instinctively form rapid impressions of others based on minimal cues. While research has demonstrated that such *thin slices of social behavior* (Ambady and Rosenthal, 1993) are surprisingly accurate across many domains, their applicability to evaluating complex scientific presentations remains less explored. Conventional wisdom and the popular science literature on public speaking suggests that we often judge a speaker's competence within seconds of them taking the stage (Ailes, 2012), but empirical research remains scarce. In this study, we investigate whether thin slices of scientific presentations—just like the talks given at academic conferences—can reliably forecast overall impressions of the presentations' quality. We introduce a novel dataset and new methods based on Large Language Models (LLMs) to analyze the predictive power of these fleeting initial moments.

This paper is structured as follows. First, we introduce the topic of thin-slicing and how it has been studied in nonverbal communication and social perception and cognition research. Then we zoom in on the topic of public speaking and discuss how research on social impression formation and thin-slicing relates to this domain. Third, we discuss how the advent of LLMs offers new ways to study thin-slicing productively and with unprecedented efficiency. We then introduce the current study and its hypotheses, which focus on the evaluation of thin-sliced transcripts[1] of a large corpus of science communication talks, followed by specific methods, results, and discussion.

## 1.1 Background

### 1.1.1 Thin slices of social behavior

A large body of research from social psychology and nonverbal communication suggests that we can often glean surprisingly accurate social information from only short glimpses into others' observable behaviors (Ambady and Rosenthal, 1992; Uleman, 2023). A classic example of social perception based on such thin slices comes from classroom teaching: Ambady and Rosenthal (1992, 1993) found that observers could predict a teacher's end-of-semester evaluations after watching only a brief, silent video clip of that teacher's classroom behavior. In their study, even 30 s of nonverbal interaction (without audio/verbal content) provided enough information for strangers to assess teaching effectiveness, forecasting evaluations given by the teacher's actual students months later.

This and similar studies launched the idea that rapid, minimal exposure to a person's behavior can reveal stable qualities (Ambady, 2010). In other words, relevant information (i) must be expressed and (ii) can be extracted from only a thin slice of the whole. Thus, the so-called thin-slicing paradigm typically involves presenting observers with brief excerpts of behavior—often just seconds-long videos of nonverbal behavior or even still pictures—and asking them to make judgments. Then, by correlating ratings that were made based on thin slices with ratings based on exposure to the whole interaction, it can be established how much information is needed to arrive at a stable judgment that is predictive of the whole. Meta-analyses show that across a broad range of social domains (Murphy and Hall, 2021; Slepian et al., 2014), thin slices can predict consequential outcomes like interpersonal warmth, personality characteristics, physician competence, relationship quality, or the outcome of legal proceedings (Carcone et al., 2015; Houser et al., 2007; Krumhansl, 2010; Nguyen and Gatica-Perez, 2015; Parrott et al., 2015).

While most existing thin-slicing literature focused on nonverbal behavior, this paradigm is also applicable to paraverbal and verbal domains (Hall et al., 2021; Slepian et al., 2014). For example, we can make rapid judgments from a voice or even a written email, like inferring a sender's gender and age during a phone call, or their emotional state or personality from their writing. Along similar lines, we may be able to infer a person's competence, confidence, or enthusiasm soon after they start speaking (DErrico et al., 2013; Gheorghiu et al., 2020; Rosenberg and Hirschberg, 2009).

### 1.1.2 Translating the thin-slicing approach to the public speaking domain

Evidently, the public speaking situation overlaps with that of classroom teaching, the domain in which thin-slicing research originated from. Both involve presentations by a speaker to an audience, i.e., a one-to-many form of communication that blends the interpersonal and mass communication domains (Berger et al., 2010). Although the thin-slicing approach stems from social psychology and education, almost all application contexts are communicative in nature (e.g., relationships, business, health, or legal interactions), focusing on the expression of social signals by senders and their perception by recipients. Thus, thin-slicing is very applicable to communication research in general and to public speaking in particular. This all suggests that thin-slicing research is highly relevant to public speaking. Indeed, in the popular literature on public speaking education, there is a widely cited notion of a "seven-second rule," suggesting that listeners decide within the first few moments of a talk whether the speaker is worth their attention (Ailes, 2012). Although closer inspection shows that this rule is based largely on anecdotal data, it is widely assumed and taught that a strong start matters greatly in presentations (Hey, 2024; Lucas, 2020). This aligns directly with thin-slicing research as well as work on first impressions more broadly (Todorov, 2017).

Some limited research has connected these domains, but they mostly address informal or non-scientific content (Chollet and Scherer, 2017; Cullen and Harte, 2017; Feng et al., 2019; Ismail, 2016). For instance, in an observational study of a meeting, Ewers (2018) found that the degree of dullness of a talk after 4 min predicted whether the speaker would "ponder on," i.e., go into overtime. Another related study by Cullen and Harte (2017) used the thin-slicing approach to TED talks by applying machine learning feature extraction techniques to visual and acoustic parameters. While this work is directly relevant to the current study, its focus on preselected TED talks, which tend to be optimized for entertaining popular audiences, sets it apart from our focus on scientific presentations. Other work by Chollet and Scherer (2017) also directly connects the thin-slicing literature to public speaking. In a study of 45 speakers giving informal impromptu presentations about the city of Los Angeles and a beauty product, Chollet and Scherer (2017) found that automatically assessed audio-visual features forecasted speech evaluations.

Perhaps the most directly related prior studies are the ones by Ismail (2016), Feng et al. (2019) and Biancardi et al. (2025). Ismail (2016) presents an elaborate proposal on how thin-slicing could be applied to public speaking evaluation, but without empirical data. Feng and colleagues, who are part of the Educational Testing Service Corporation, conducted a thin-slicing study with 17 speakers who were recruited via the Toastmasters organization and gave speeches about different pre-assigned topics. However, their study focused largely on the visual and nonverbal delivery factors, assessed via video-based thin-slicing and human ratings. Although they discuss speech content quality, it was not directly examined via thin-slicing. Lastly, a very recent study by Biancardi et al. (2025) presented a novel

1 Note: Whenever we refer to speech or presentation, we mean the speech/presentation transcript. We tried to make this clear throughout the paper but sometimes refer simply to speech or presentation for convenience. But we did not study the nonverbal or paraverbal factors of speeches in this study, rather the results refer only to the speech/presentation transcripts (also see Discussion).

corpus of scientific presentations along with crowd-sourced audience evaluations. These crowd-sourced evaluations annotate various speech and speaker characteristics (e.g., persuasiveness, engagement, confidence), which were carried out separately for the beginning, middle, end sections of the speech, as well as for the entire speech. This again resembles the gist of the thin-slicing methodology, and the results identify positive correlations between slices and the full speech.

Overall, while some promising work on using thin-slice style ratings for public speaking exists, there is a need for larger and more systematic investigations of science communication presentations leveraging the thin-slicing approach.

### 1.1.3 Potential of large-language models to enable thin-slicing studies of public speaking

Even though the thin-slicing paradigm originated from an inherently communicative situation (classroom teaching), the domains of thin-slicing research in social cognition and public speaking training and assessment have remained surprisingly distant. This gap between the thin-slicing literature and the literature on public speaking competency appears partly due to the practical challenges of studying realistic public speaking performances, which is labor-intensive and requires a large corpus of real speeches and many raters.[2] For example, Ambady and Rosenthal's (1993) work involved only 9 raters who had to watch and manually code all 39 video clips. Feng et al. (2019) highlighted the enormous burden these tasks place on raters. And the same challenges apply to the large body of research that uses human coders to study topics like social perception and impression formation (Grahe and Bernieri, 1999; Schmälzle et al., 2019; Wallbott and Scherer, 1986; Willis and Todorov, 2006), as well as more focused investigations of speaker ability, charisma, and similar topics (Cullen et al., 2018; Gheorghiu et al., 2020; Rosenberg and Hirschberg, 2009). In sum, the inherent challenges are clear: high expense and rater wear-out must be balanced with efficiency, representativeness, and other constraints.

LLMs pave the way for closing the gap between thin-slicing research and the perception and evaluation of speech performances. LLMs are advanced artificial intelligence systems trained on vast amounts of text data, enabling them to generate human-like text and perform various language-based tasks (Tunstall et al., 2022). LLMs have majorly shifted how we interact with and utilize AI, with models transforming industries and reshaping the future of communication and information processing (Bishop and Bishop, 2024; Mitchell, 2019).

LLMs offer a potential remedy for the challenges related to human raters discussed above (Argyle et al., 2023; Calderon et al., 2025; Gilardi et al., 2023; Markowitz and Hancock, 2024). LLMs can perform complex tasks, including evaluating speech transcripts in terms of basic tasks like word counts, and also higher-level impressions about social characteristics (Bubeck et al., 2023; Dillion et al., 2023). While it remains an empirical question whether LLM-based evaluations align with those made by humans, communication scholars are well-equipped to investigate it (Krippendorff, 2004; Neuendorf, 2017; Riff et al., 2014; Shrout and Fleiss, 1979). Thus, by demonstrating correlations between human and LLM evaluations of the same speeches, we can validate the use of LLMs for evaluating public speeches (i.e., their transcripts). This, in turn could scale up thin-slicing research far beyond what was previously feasible.

To be precise, we note that this type of LLM-based thin-slicing research is presently most applicable to analyze textual transcriptions of speeches. Although recent multimodal models could in theory also process entire speeches—including video and audio information conveying the speaker's nonverbal and paraverbal behaviors (all of which matter for impression formation)—the current study will be focused on evaluating the transcriptions of the speeches. An easy to use, accessible, fast, and valid way to assess the quality of presentations as reflected in their transcripts would likely open up new avenues of investigation for understanding and improving communication.

## 1.2 The current study

Building on the research streams summarized above, this study leverages LLMs to examine how much information is needed to predict science presentations' overall quality evaluations. We focus on science presentations because they constitute a setting where effective communication is critical (Doumont, 2009; Fischer et al., 2024; Gu and Bourne, 2007; Hey, 2024). Science presentations, like the ones given at conferences, require engaging an audience quickly on complex topics, keeping them attentive over a sustained period, and presenting information in such a way that it is comprehensible to audiences who are intellectually able and motivated to learn, but may not be familiar with the minutiae of the research. By testing the predictive power of early impressions, we thus aim to the literature on impression formation and practical public speaking assessment.

An overview of the study rationale is provided in Figure 1. In brief, to conduct the study, we first collected a large corpus of over 100 science presentations, which spanned a broad range of topics typical of the diverse research activity at a large Midwestern R01 university. Each presentation lasted between 8–12 min and consisted of actual research-based content that the speakers prepared themselves and were expert in.

The talks were transcribed (see footnote 1) (using OpenAI's Whisper followed by manual corrections, and we opted to keep filler sounds and words in the transcripts, e.g., Laserna et al., 2014; Laske and DiGennaro Reed, 2024). Then, we submitted the speech

2 First, while public talks, such as the popular TED Talk series, provide many examples of public presentations, they tend to represent only skilled speakers, which limits variability. Public speaking courses, on the other hand, often use impromptu speaking tasks rather than real-world presentations. This again limits variability and range, and it also produces speech contents that the speakers may not be expert in or care much about. Thus, when it comes to stimulus sampling, recordings of presentations may exist in large quantities, but sampling speeches that matter for science communication is more difficult than it may first seem. A second challenge is the need to obtain ratings from human observers. For example, if we wanted to assess the overall quality of talks, this would require having multiple evaluators listen to the entire talk (e.g., Feng et al., 2019). Next, if we wanted to evaluate a short slice (e.g., 1 min), then we would need a separate group of evaluators also listen to the slices. Then, if we wanted to test if even 30 s is enough, yet another group of evaluators is needed, and so forth. In summary, constraints regarding availability of appropriate speech material and the need to have them evaluated by many raters might make large-scale empirical studies of thin-slices of public speeches difficult, slow, and costly.

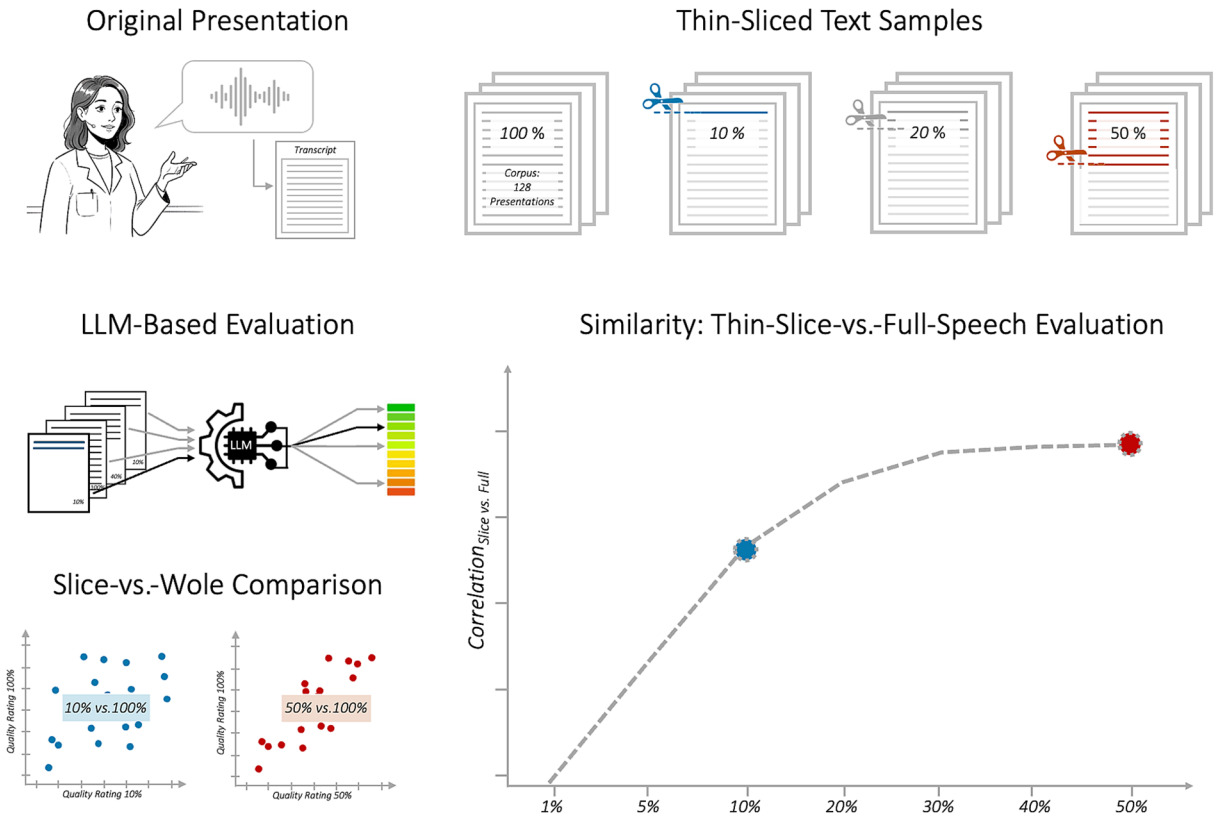


FIGURE 1
Logic of thin-slices evaluation of public speech performance in the context of science communication. Recordings of real-life talks about science topics are transcribed to text. Next, each transcript is thin-sliced into excerpts containing either the full speech text, or slices corresponding to 1, 5, 10%, etc. These slices are then submitted to LLM for quality assessment, leading to a table with ratings for all speeches and across all slices. Ratings are collected independently (no memory in the LLM) and evaluated via different prompts and multiple LLMs. Finally, we compare evaluations across slices to examine how much of the speech needs to be processed until stable quality predictions can be made.

transcripts to LLMs to rate the quality of presentation transcriptions. For each transcript, the LLMs rated thin-sliced subsets with slice lengths varying parametrically (e.g., the first 75% of the speech, 50, 40, 30, 20, 10, 5, 1%[3]) as well as the entire speech. We validated this LLM-based method using human raters.

Through this LLM-based thin-slicing approach, we address the following hypothesis and research questions. Based on the literature on thin-slicing summarized in the introduction, we predict that the quality ratings of thin-sliced subset of the presentations' text transcripts would positively correlate with the quality ratings of the entire transcript (i.e., show a thin-slicing effect; H1). Next, we examined how much content is required in the thin slice to best predict the quality ratings of the overall speech transcript (RQ1). We estimated that less than 20% of a speech (about 3 min) should be enough (Ewers, 2018) and aimed to pinpoint a more precise moment when additional information no longer becomes relevant. Finally, we explored whether the findings for H1 and RQ1 would differ by the specific LLM used and instructions provided to the LLMs (RQ2).

To our knowledge, no other study to-date has applied LLMs to evaluate the textual transcripts of public speeches or presentations; thus, our LLM-based thin-slicing approach contributes to the communication discipline as well as social cognition research. If the AI's thin-slice ratings correlate strongly with full-speech outcomes, then that could also lead to practical tools for providing automated feedback on presentations, such as feedback on clarity, coherence, or other aspects that could be inferred from the speech text or accumulated subsections thereof. Across the learning sciences, it is clear that feedback is a key ingredient of improvement (Domjan, 2020; OECD, 2010; Silver et al., 2021; Skinner, 1961). However, many existing feedback systems for public speaking does not provide real-time information (e.g., "content on slide 2 could be articulated more clearly") and generate arbitrary qualitative or composite scores that limit speaker improvement. LLM-based feedback systems can potentially address these limitations, giving speakers a quick ability to change course. Furthermore, comparing LLM and human ratings can also give us theoretical insights into how

3 Here we present %-sliced speeches, which are comparable across speakers due to varying speaking rates. We also examined the same set of questions using either time-based slicing (e.g., seconds) or word-based slicing (e.g., first 20 words), finding converging results. We therefore opted to present the %-based results, which are easy to grasp.

closely these models mirror human quality perceptions or perceived effectiveness.

# 2 Methods

## 2.1 Public speaking corpus

We assembled a corpus of 160 public science presentations for analysis. Eighty members of the scientific community presented two separate research-based talks about the area of their expertise. The speakers were mostly graduate students as well as some faculty. They were recruited from the academic community of Michigan State University with its more than 15,000 active researchers (postgraduates and academic staff). Recruitment strategies included flyers, posts to university-wide listservs, graduate programs, and academic units, as well as word-of-mouth promotion. Interested potential participants were informed about the scope of the study and asked to prepare the two talks and send their slides before the day of the study. Thus, the speakers had ample time to prepare their talks for professional science audiences and had a personal interest in the talks being of high quality. Reflecting this diversity, the given talks spanned across almost all subjects—from the challenges of social work to the frontiers of artificial intelligence (as well as business, biology, psychology, communication, statistics, zoology, neuroscience, and others).

The talks were presented in front of a large audience in a virtual reality (VR) environment that resembled a professional venue (conference-style hotel room). The talks were 8–12 min long and were recorded. Then we transcribed the presentations using OpenAI's Whisper model and manually checked the transcriptions for errors. We also screened the transcripts for quality and removed talks with incomplete recordings or poor audio. After this filtering, 128 speeches remained in the corpus. Overall, this corpus amounted to over 100,000 s and almost a quarter million of spoken words—about an entire conference days' worth of public speaking content. By working with text transcripts, we aimed to provide a consistent input to the language models and to enable human raters to evaluate content without being influenced by visual or audio cues.

## 2.2 LLM-based thin-slicing procedure

All analyses were conducted in Python, and we fully document the analysis pipeline online at https://github.com/nomcomm/thinslice_public_speaking. First, the text transcriptions were loaded and sliced (i.e., subsampled) into the first 1, 5, 10, 20, 30, 40, 50, 75, and 100% of the speech transcript. We also ran the same analyses using fixed numbers of words (because the percentages can contain different numbers of words). However, the conclusions hold, and thus we only report the percentage-based results here in the main text (see Supplementary materials for additional details).

### 2.2.1 LLM-based speech evaluation procedure: models and prompting strategies

We deployed two advanced language models, GPT-4o-mini and Gemini Flash 1.5, as AI evaluators. These models were chosen for their strong language understanding capabilities (Omar et al., 2025), which would enable them to judge coherence, clarity, engagement, and other qualities of the speeches.[4]

GPT-4o-mini and Gemini Flash 1.5 each evaluated all 128 speeches in all slices, yielding a set of AI-generated scores for every full speech and every excerpt. We experimented with five different prompt formulations for each model to ensure robustness of the AI's responses. Table 1 presents these different prompts. By using multiple prompts, we checked that the LLMs' ratings were not overly sensitive to prompt phrasing or context but rather circumscribed the semantic field of speech/presentation quality/public speaking evaluation in different facets. In total, this approach yielded 11,520 ratings, and the entire submission of speeches and retrieving the ratings took about 1 h.

### 2.2.2 Human evaluation and LLM validation

Because the use of LLMs in a way that mimics human raters is still evolving, we also conducted a human rating study to validate that LLM's ratings capture meaningful variation that can be perceived by humans. To this end, we recruited a group of 60 human raters ($mean_{age}$ = 38.2, $sd$ = 10.7, 24 self-identified males) to read through the speeches and rate their quality. The study was conducted online via Qualtrics, using participants recruited from the Prolific platform. The entire procedure was approved by the local institutional review board (IRB), all raters provided informed consent, and they received $4 for their evaluations, which took about 20 min.

Procedures were kept as parallel as possible as for the LLM-prompt. Because internal tests revealed that raters would have difficulties reading the entire speeches, we decided to evaluate only the 20% version. This amounted to about a half page to a page of text, which was most feasible in terms of attentional demand and study duration. The specific to which the raters responded was *"Please rate this transcript from a public presentation in terms of its quality on a scale from 1 (worst) to 10 (best). Consider factors such as clarity, engagement, and how easy it is to follow."* The 60 raters were split into two groups, and each evaluated a sample of 12 speeches drawn from the corpus of 128 speeches. A total of 24 speeches were evaluated.

#### 2.2.2.1 Statistical analysis methods

Once human ratings were collected, we examined interrater-agreement among the human raters based on intra-class methodology (Shrout and Fleiss, 1979). In parallel, we also assessed the agreement between different LLM models and prompting strategies. Next, we computed Pearson correlations between group-averaged perceived quality ratings and the LLM's ratings for each speech (Pearson, 1895). Lastly, following prior work on thin slice judgments, we computed Pearson correlations between the ratings for each slice and the rating for the entire speech. A high correlation between the slice and the entire speech suggests that the relevant information can be successfully extracted already within a much shorter slice. Our goal was to measure the threshold where this correlation reaches significance.

4 Based on reviewer feedback, we later also explored the potential of smaller and open models, such as the 8B and 70B versions of Meta's Llama, for this purpose, but the main *a priori* investigation was carried out with these two models. We provide results from these exploratory model tests in the supplement.

TABLE 1 Different prompts used to evaluate the speech transcripts.

| Prompt # | Prompt wording |
|---|---|
| #1 | *"Here is a transcript from a public presentation on a science/research topic. Please rate the speech quality on a scale from 1 (worst) to 10 (best). Consider factors such as clarity, engagement, and how easy it is to follow. Return only the single rating number as a plain integer, with no other text or characters. Here is the speech text:"* |
| #2 | *"You will receive a transcript of a science/research presentation. Rate the overall rhetorical quality on a scale from 1 (worst) to 10 (best), considering clarity, engagement, structure, and delivery. Return only the single rating number as a plain integer, with no other text or characters. Here is the speech text:"* |
| #3 | *"Given the following transcript of a science/research presentation, assess its overall speech quality. Focus on aspects such as clarity, engagement, and coherence. Provide only a single numerical rating from 1 (worst) to 10 (best), without any additional text. Here is the speech text:"* |
| #4 | *"Imagine you are an expert in public speaking evaluation. Below is a transcript from a science/research presentation. Please rate the effectiveness of the speech on a scale of 1 (worst) to 10 (best) based on clarity, engagement, and ease of understanding. Return only the single rating number as a plain integer, with no other text or characters. Here is the speech text:"* |
| #5 | *"Please evaluate the following transcript of a public science/research presentation. Assign a quality rating from 1 (worst) to 10 (best) based on your assessment. Return only a single rating number as a plain integer, with no other text or characters. Here is the speech text:"* |

# 3 Results

## 3.1 Interrater agreement and convergence of LLM and human ratings

Starting first from the human ratings (60 raters evaluating 24 speeches, split into two samples), we find that the human raters exhibited high consistency in their speech evaluations, as demonstrated by high-intra-class correlations $ICC_{2,1} = 0.92$ and 0.86 (Shrout and Fleiss, 1979). This demonstrates high levels of agreement among raters about speeches' perceived quality rankings; the high consistency also underscores that the group-averaged quality evaluations per speech are highly reliable (i.e., no benefits accrue from using additional raters; Kelley, 1925; Kim and Cappella, 2019; Kraemer, 1992).

Next, we conducted a parallel stream of analyses to assess agreement among different ways to elicit LLM evaluations (RQ2). In other words, we applied the same procedures as conducted for the human ratings to the data obtained for the different LLM models and the five prompts—essentially treating the model/prompt-instances as if they were 10 raters. This analysis yielded a high inter-model/prompt-agreement, $ICC = 0.93$—similar in magnitude as observed for the human raters.[5]

Having established that human raters as well as different LLM models with specific prompts each produce convergent ratings among each referent group, we proceeded to examine whether human and LLM-based evaluations also converge. To this end, we correlated the group-averaged speech evaluations from the human sample with the corresponding ratings from the LLM-based speech evaluations. We find that the correlation amounts to $r_{human\text{-}rating\text{-}vs.\text{-}LLM\text{-}rating} = 0.69$, which is highly significant ($t(22) = 4.47$, $p < 0.0001$). This shows that regardless of whether the speech transcripts are evaluated by humans or by LLMs, both evaluation modes yield very similar conclusions about which ones are considered high vs. low quality. This again underscores the promise of LLM-based evaluations in terms of validity, but with much higher efficiency (see Supplementary Figure S1).

## 3.2 Main analysis: thin slice correlations

The central question of our study was whether a thin slice of a public speech's transcript, like the first 10%, can predict quality evaluations for the entire speech transcript (H1 and RQ1). Our results suggest that it can. Table 2 presents the main results and Figure 2 plots the strength of the correlation between each part (slice) and the entire speech. As can be seen, correlations rise quickly across progressively thicker slices and converge at around 0.6/0.7. This effect is visible for both LLMs—whether from the Gemini or GPT4o-family (see Supplementary materials for an exploratory analysis of Llama Models). Furthermore, already at 10% of the entire speech, the plateau is basically reached, suggesting that from this point onwards, additional incoming speech content does not make much of a difference in terms of the overall evaluation.

Interestingly, even very thin slices, such as 5% or even 1% of the entire speech, show positive correlations. With the sample size of 128 speeches, the chance-level (i.e., $a = 0.05$) lies at $r = 0.146$ for a test of the hypothesis that the obtained correlation differs positively from zero. Even at the 1%-slice, most correlations (4 out of 5 for the GPT-based prompts and 2 out of 5 for the Gemini model) are above this threshold, and the threshold is passed for all (10 out of 10) model/prompt-configurations at the 5% slice. In the current sample, these slices correspond to just 15 (for 1%) and 60 (for 5%) words.

Inspection of Figure 2 suggests a dominant effect, namely that thin-slice correlations rise quickly and then plateau. Moreover, this effect is obtained regardless of the specific model or prompting strategy. However, as the lower-left panel in Figure 2 suggests, results for different models and prompts exhibited some variability. To examine this and potentially detect interactions or main effects of

5 Of note, to keep the inter-rater agreement analysis maximally comparable between human and LLM-ratings, we report results for the 20% slice speech transcripts. However, we also carried it out for other slices, finding even higher results for the entire speech ICC = 0.95 and similar values for all other slice conditions.

TABLE 2 Thin-slice to full-speech (part-to-whole) correlations for both LLMs.

| | | Slice thickness | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Model | Prompt | 0.01 | 0.05 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.75 |
| GPT | Prompt#1 | 0.32 | 0.54 | 0.65 | 0.72 | 0.71 | 0.74 | 0.74 | 0.67 |
| | Prompt#2 | 0.26 | 0.53 | 0.68 | 0.73 | 0.67 | 0.74 | 0.74 | 0.73 |
| | Prompt#3 | 0.33 | 0.54 | 0.69 | 0.72 | 0.75 | 0.73 | 0.70 | 0.75 |
| | Prompt#4 | 0.25 | 0.55 | 0.76 | 0.72 | 0.73 | 0.73 | 0.69 | 0.72 |
| | Prompt#5 | 0.10 | 0.44 | 0.65 | 0.70 | 0.68 | 0.57 | 0.65 | 0.72 |
| Gemini | Prompt#1 | 0.02 | 0.33 | 0.55 | 0.54 | 0.64 | 0.66 | 0.70 | 0.74 |
| | Prompt#2 | 0.27 | 0.44 | 0.60 | 0.61 | 0.66 | 0.66 | 0.78 | 0.80 |
| | Prompt#3 | 0.00 | 0.44 | 0.58 | 0.69 | 0.70 | 0.76 | 0.78 | 0.78 |
| | Prompt#4 | 0.25 | 0.31 | 0.56 | 0.45 | 0.51 | 0.56 | 0.58 | 0.64 |
| | Prompt#5 | 0.07 | 0.30 | 0.39 | 0.50 | 0.52 | 0.56 | 0.63 | 0.67 |
| | Average | 0.19 | 0.44 | 0.61 | 0.64 | 0.66 | 0.67 | 0.70 | 0.72 |

With 128 cases, correlation coefficients >0.146 are significant ($\alpha = 0.05$).

slice-condition, model, and prompts, we first compared all obtained correlation values for a given model and slice across the different prompts (see Supplementary materials and online code repository for full output). After correcting the resulting *p*-values for the large number of tests, there were no significant differences between prompts. Additionally, we also conducted a mixed effects analysis. This analysis revealed a small but significant effect of prompt as well as an interaction between the specific prompt, slice, and model. Further inspection of the amount of variance explained by the different factors (slice, model, or prompt), however, showed that by far the dominant effect was the rise of the correlations across successive slices (~50% variance explained by slice and model), whereas only 3% were added by including the prompt. In summary, this suggests that our findings are not an artifact of a peculiar prompt or an idiosyncrasy of one AI system. Instead, they reflect a stable AI-based assessment of speech/transcript quality that emerges regardless of how we queried the models. Therefore, while further prompt-tuning could offer marginal gains, the specific prompt used does not fundamentally alter the primary conclusion that performance is overwhelmingly driven by the rising slice-to-whole speech correlations.

## 4 Discussion

This study examined whether early impressions of public science presentations can predict the presentation's overall evaluation. We tested the potential of language models to reliably emulate speech evaluations in this context, thus focusing on the information expressed in transcripts of the originally spoken presentations. Our findings provide strong support for the thin-slicing effect, suggesting that a brief exposure to a presentations' transcript contains information that enables predictions of its overall quality ranking. Furthermore, we find that LLMs are accurate and efficient.

The results demonstrate that a thin slice of a presentation's written transcript allows forecasting its overall quality. In fact, even the very first few sentences contain predictive information that enabled correlations between 0.3 and above (see Figure 2). A plateau effect emerged starting at about 10% of the speech, suggesting that relevant information has been expressed at this point and evaluations do not change much from there on. Notably, even extremely brief excerpts—just 5% or even 1% of the full speech—exhibit positive correlations. In this dataset, these slices equate to roughly 15 words (1%) and 60 words (5%). Considering typical public speaking rates of 100–150 words per minute, this aligns remarkably well with the "7-s rule" (Ailes, 2012).

This main result was stable across different analysis methods and was seen similarly using different LLM models and prompt wordings. Moreover, it is important to highlight that the demonstrated correspondence between human evaluations and LLM-based evaluations validates the use of the latter in the first place. If there was no strong correlation between human and LLM-evaluations, or if the correlation was only moderate, then one could question the validity of LLMs. While the precise meaning of speech or presentation quality can be subject to debate—just like all verbal concepts in the social sciences and humanities (e.g., Wittgenstein, 1953),—our work approaches the issue empirically. Rather than defining quality *a priori* and sort of imposing a definition on participants (and AI), we demonstrate that a stable, shared understanding of it exists. The strong convergence between human and LLM evaluations, which is also consistent across multiple models and prompts, points to a robust commonality in evaluative impressions, which are a cornerstone of social cognition research (e.g., Fiske et al., 2007; Osgood et al., 1957).

Having established that LLMs' evaluations of the speeches converge with those of human raters, future researchers in this area can begin to use LLMs and thereby greatly increase the efficiency of thin-slicing studies, which are very time-consuming, costly, and taxing. In fact, the core of our LLM-based speech evaluation consists of ca. 15 lines of code in which a for-loop prompts the API of GPT4 and Gemini, respectively. With this pipeline, the independent evaluation of all 128 speeches across 2 models, 5 different prompts, and for the entire speech as well as 8 sub-slices (1–75%) took less than half an hour, costing less than $5. Compared to the ca. $150 we paid for the human evaluation study, which only comprised a small fraction of the volume (i.e., only one task instruction/prompt, one slice, and 24 speeches), it is easy to see the superiority of the

10.3389/fcomm.2025.1610404

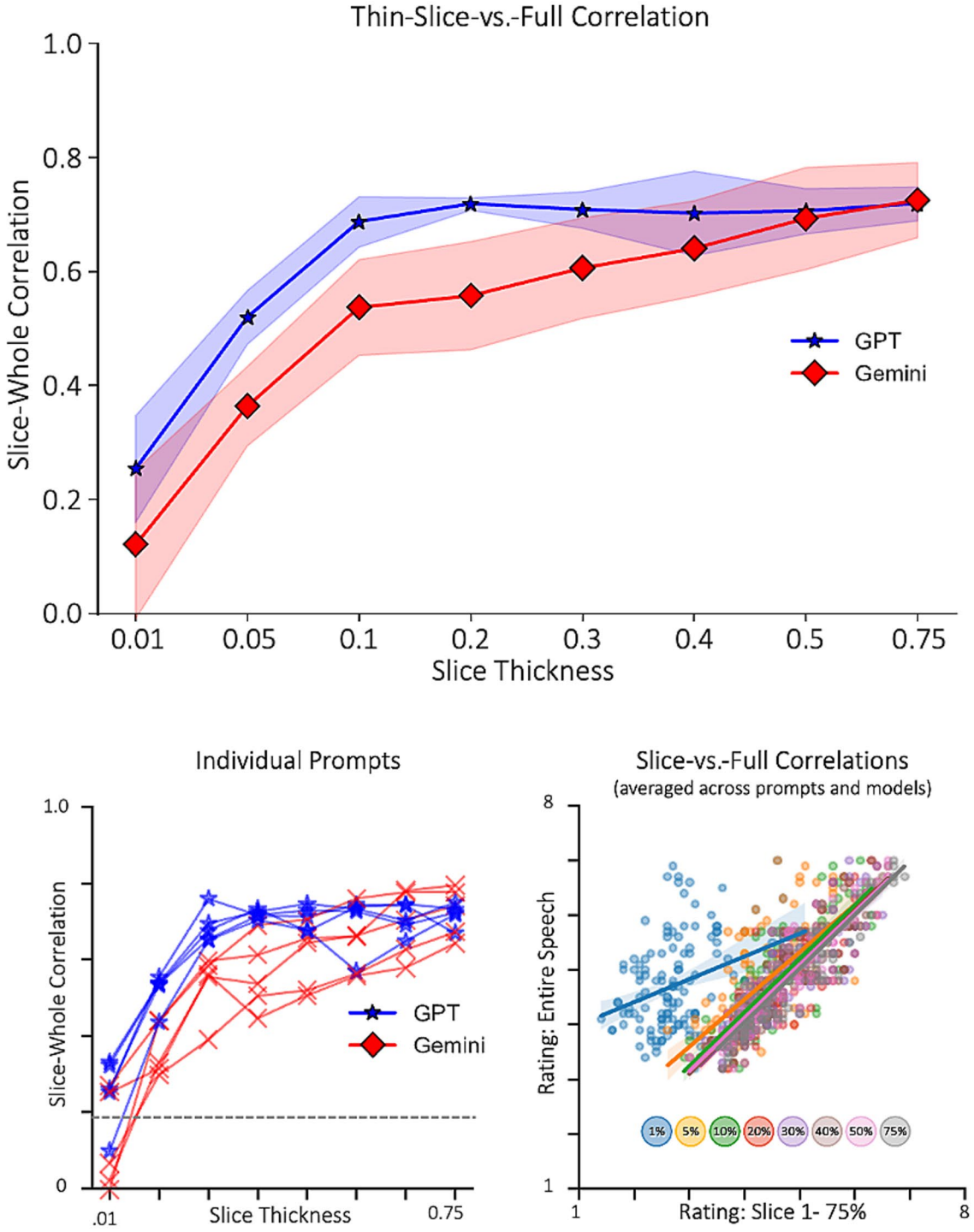


FIGURE 2
Thin-slice to full-speech (part-to-whole) correlations for both LLMs. Shaded corridors illustrate the variability across the five different prompts. Bottom panels: Left: Individual-prompt results for OpenAI's GPT (blue) and Google's Gemini (red) models. As can be seen, the same general pattern is present regardless of model family or prompt wording. Right: Scatter plots for all 128 speeches. As slice thickness increases, the predictions become progressively more aligned with the evaluation for the entire speech.

LLM-based assessment in terms of cost-effectiveness and scalability. This is not to say, however, that human evaluations are no longer necessary: it still needs to be demonstrated that LLM-based evaluations are consistent and converge with human impressions; but once that is established, as in the current context of evaluating speech transcripts, the pendulum swings clearly in favor of using LLMs (Argyle et al., 2023; Calderon et al., 2025; Dillion et al., 2023; Eger et al., 2025; Gilardi et al., 2023).

## 4.1 Theoretical implications and practical applications

In the following section, we first discuss the theoretical implications and connections of the current findings with the communication science literature and then point out practical applications.

The thin slices/first impressions literature provides empirical backing for Uncertainty Reduction Theory (URT; Berger and Calabrese, 1975). In particular, one of URT's core ideas is that people use limited salient information to make quick judgments to guide social interactions. To our knowledge, these connections have not been well articulated, and thus, the thin-slicing/first impressions literatures in psychology and uncertainty reduction theory in communication have evolved somewhat in parallel. However, by focusing on the initial encounter situation in which a new speaker presents themselves to an audience, it is natural to see how the two bodies of research converge: At the beginning of a talk, there is necessarily high uncertainty about what is going to happen next, what the talk will be about, and whether the speaker can get the point across. But this uncertainty is progressively resolved as the talk unfolds and audience members form impressions. The accuracy of these snap impressions has long been a debate in the social psychology literature (Jussim, 2017), but in the case of quality judgments, the current results suggest that it can indeed be quickly sensed how good a speaker/speech is.

Another relevant body of work, again unconnected to thin-slicing research, can be found in classical newspaper readership studies in journalism and mass communication. For example, Schramm (1947) showed that most readers of long-form newspaper articles stopped reading early on, as if they lost attention or found the text too long and dry. In fact, studies of article reading depths resulted in new media formats, such as the USA today newspaper with its brief articles; nowadays, similar evolutionary developments seem to unfold with online texts (Berger et al., 2023), simple choices like whether to read an article based on a headline (Scholz et al., 2017), but also with short video formats like TikTok and YouTube Shorts. Critically, the point is that readers make snap judgments about whether the content is interesting and whether it is worth to keep reading. Relatedly, there is also renewed interest in people's sequential media choices, i.e., how people choose between different songs, videos, books, and so on (Gong and Huskey, 2023). The work presented here connects these lines of inquiry insofar as it focuses on the choice within a given message (like a speech, but also a book, song, TV show, or newspaper article), i.e., how people make decisions implicitly about staying engaged. To avoid misunderstanding though, we have not yet studied here whether real audiences would "tune out" of some low-quality speeches after 10%, but recent work in neuroscience of audience response measurement suggests that this could be feasible (Schmälzle, 2022; Schmälzle et al., 2015), and promising results have already been obtained in assessing the impact of pitch and voice features on physiological responses (e.g., Rodero, 2022).

The current results are not only theoretically interesting regarding the nature of public speaking and how a speaker's skills are expressed as the speech unfolds, but they can also improve communication training and practice: By demonstrating the effectiveness of thin-slicing and the feasibility of LLMs for speech transcript analysis, we offer a pathway toward automated and scalable feedback and augmentation tools for speakers. Especially with automated public speaking training, such tools could offer valuable, immediate, and actionable feedback (Forghani et al., 2024; Valls-Ratés et al., 2023). For example, even within standard software tools like Microsoft PowerPoint, there is already a tool called "Speaker Coach" (Microsoft Corporation, 2025), which allows speakers to rehearse their slide shows and provides basic feedback about speech rate and overused filler phrases. However, this tool is very basic and does not give feedback about the content or organization of the presentation itself. These are areas where LLMs could help a lot to improve the speakers' notes, making them clearer, easier to understand, and ultimately more effective (Shulman et al., 2024). In the current study, however, the LLM only provided a numerical rating, which could provide feedback, but not yet actionable improvement suggestions although this is certainly possible.

Although the current study was focused on numerical ratings of quality, we did actually also peek into "how" the LLMs justified their ratings, finding that they were well able to identify key strengths and weaknesses regarding clarity, coherence, confidence, and other aspects (see Supplementary materials). Therefore, we foresee that software for public speaking training could incorporate AI-based methods to first transcribe the incoming speech into text as we did here or even use multimodal models to also capture the spoken language directly and then use a similar prompting-strategy to make judgments about key characteristics. These could be fed back to the speakers—either immediately after the speech or potentially even during the speech. Particularly with fine-tuned open-source LLMs (see Supplementary materials for an exploratory investigation of Meta's Llama model), this would seem very feasible and is already explored by industry. In essence, this would lead to systems like the popular writing-aide software Grammarly, which provides continuous feedback and suggestions about specific text sections.

In sum, the work presented here about LLM's capabilities to swiftly detect early warning signs of a talk that might be at risk of losing the audience could empower scientists and other professionals to refine their communication skills. This could lead to more effective dissemination of complex information to audiences. Given that science communication is crucial for public understanding of science, this could have great benefits. Moreover, the methodological framework developed here could be applied to other communication domains that build on public presentation skills, such as education, business, and politics.

## 4.2 Strengths, limitations, and avenues for future research

This study's strengths include its novel application of thin-slicing to public speaking, specifically focusing on the verbal communication channel. Also, the use of a large and high-quality corpus of science communication talks, and the exploration of multiple LLMs and prompts are positives. However, the study is limited by its focus on a specific type of communication (science talks) and a specific population of speakers (academics), although this homogeneity and expertise could also be viewed as an asset. Also, it is worth keeping in mind that the online raters only read and evaluated the 20% version of the transcribed speeches, not the full speeches. This was done because our pilot tests showed that reading the full transcriptions of

ca. 8 min long speeches could lead to coder wear out, especially if coders/raters were to evaluate multiple exemplars, which is desirable. However, given that the 20% speech evaluations from the human and AI agree, and the AI evaluations from 20% onwards show little change, we are confident that the human coding procedure is adequate.

Another point worth discussing is that the reliance on transcripts means that nonverbal and paraverbal cues, which are known to influence communication and impression formation, were not directly analyzed in this study. Integrating the verbal and nonverbal cues—as well as analyzing each channel's contribution separately—are valuable steps that we are working on next (Wolfe and Siegman, 2014). For example, advances in so-called multimodal LLMs could make it possible to submit not only speech transcripts, but the actual speech recording itself, or even a video. This would allow to study vocal and visual cues in much the same way as we did here using the speech text. Interestingly, Robert Rosenthal, the same researcher who also developed the thin-slicing methodology, has conducted work on the PONS-test (profile of nonverbal sensitivity) in which the goal is to quantify how much information from given channels (e.g., face, voice, body posture) gets conveyed during communication and how well observers can utilize or decode certain channels (see Hall, 2001). In a similar manner, it will be interesting to have observers evaluate, e.g., just a recording of the (silenced) speaker performance. This could provide insights into the relative weight of verbal, paraverbal, and nonverbal factors during impression formation (e.g., Mehrabian, 1972) as well as about the correlations between respective abilities within speakers.

Clearly, it needs to be kept in mind that the speeches were originally delivered verbally but then transcribed into written text. Spoken, conversational language differs from written expression, the former being more informal. This does not, however, challenge our findings because even if the LLM-based evaluations would punish against informal speech, this would equally apply to all speeches, keeping their relative ranking intact. But we do note that the act of transcribing itself could lead to a loss of information and that we noted some cases where we had to correct, e.g., transcription errors based on foreign speakers' accents, and other factors. On the other hand, by transcribing the speeches we also remove implicit cues about gender, speaking style, and so forth, which can also bias or interfere with the impression formation and speech evaluation process (e.g., Bavishi et al., 2010; Rodero et al., 2022; Schlamp et al., 2020).

Also, the influence of paralinguistic factors, such as "uhs," "ums," mumbling, or even stuttering is a topic worth mentioning. In our transcription process, we purposely kept filler words and occasional word repetitions in the corpus as these dysfluencies do contain diagnostic evidence (e.g., Laserna et al., 2014; Laske and DiGennaro Reed, 2024); but we corrected mumbling, thus making the transcripts clearer than the spoken speech would likely have been perceived. But this all points to the broader distinction of the domains of speech content/organization vs. speech delivery, which are core to public speaking skills (Aristotle, 2013; Cicero, 1942, 1949; Quintilian, 1920; Lucas, 2020). In sum, while the current work demonstrates the promise of LLM-based thin-slice-style evaluation of public speech transcripts, more work is needed to unpack the fine details of how scientists communicate their findings and how audience respond to them.

## 5 Summary and conclusion

This study demonstrates the power of thin-slicing for evaluating science presentations. Even brief textual excerpts from the start of a presentation are sufficient to predict overall quality. This aligns with thin-slicing effects in the nonverbal domain and extend them towards minimal linguistic cues. This approach, particularly when combined with LLMs, offers exciting possibilities for automated feedback and personalized, AI-augmented communication training.

## Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## Ethics statement

The studies involving humans were approved by MSU Institutional Review Board. The studies were conducted in accordance with the local legislation and institutional requirements. The ethics committee/institutional review board waived the requirement of written informed consent for participation from the participants or the participants' legal guardians/next of kin because the study is largely an LLM-based study (not involving humans). The other part is a human-based online study (prolific). IRB was obtained and determined that people can consent online by acknowledging the information provided (low risk).

## Author contributions

RS: Resources, Visualization, Funding acquisition, Writing – original draft, Project administration, Validation, Formal analysis, Supervision, Investigation, Data curation, Writing – review & editing, Conceptualization, Software, Methodology. SL: Writing – review & editing, Conceptualization, Investigation, Methodology, Writing – original draft, Data curation, Validation. YD: Investigation, Writing – original draft, Methodology, Writing – review & editing. GB: Funding acquisition, Conceptualization, Writing – review & editing, Investigation.

## Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This research was supported by NSF grant #2302608.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Generative AI statement

The authors declare that Gen AI was used in the creation of this manuscript. Generative AI was used for the work itself as a method tool (LLM-based speech evaluation). Gen AI was also used for wordsmithing. The idea generation was done by the researchers themselves.

## Publisher's note



## Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomm.2025.1610404/full#supplementary-material

## References

Ailes, R. (2012). You are the message. New York: Crown Business.

Ambady, N. (2010). The perils of pondering: intuition and thin slice judgments. *Psychol. Inq.* 21, 271–278. doi: 10.1080/1047840X.2010.524882

Ambady, N., and Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: a meta-analysis. *Psychol. Bull.* 111:256. doi: 10.1037/0033-2909.111.2.256

Ambady, N., and Rosenthal, R. (1993). Half a minute: predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *J. Pers. Soc. Psychol.* 64, 431–441. doi: 10.1037/0022-3514.64.3.431

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., and Wingate, D. (2023). Out of one, many: using language models to simulate human samples. *Polit. Anal.* 31, 337–351. doi: 10.1017/pan.2023.2

Aristotle (2013). Poetics. Oxford: OUP.

Bavishi, A., Madera, J. M., and Hebl, M. R. (2010). The effect of professor ethnicity and gender on student evaluations: judged before met. *J. Divers. High. Educ.* 3:245. doi: 10.1037/a0020763

Berger, C. R., and Calabrese, R. J. (1975). Some explorations in initial interaction and beyond: toward a developmental theory of interpersonal communication. *Hum. Commun. Res.* 1, 99–112. doi: 10.1111/j.1468-2958.1975.tb00258.x

Berger, J., Moe, W. W., and Schweidel, D. A. (2023). What holds attention? Linguistic drivers of engagement. *J. Mark.* 87, 793–809. doi: 10.1177/00222429231152880

Berger, C. R., Roloff, M. E., and Ewoldsen, D. R. (2010). The handbook of communication science. Thousand Oaks, California: Sage.

Biancardi, B., Chollet, M., and Clavel, C. (2025). Introducing the 3MT_French dataset to investigate the timing of public speaking judgements. *Lang. Resour. Eval.* 59, 371–390. doi: 10.1007/s10579-023-09709-5

Bishop, C. M., and Bishop, H. (2024). "The deep learning revolution" in Deep learning (Cham, Switzerland: Springer International Publishing), 1–22.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., et al. (2023) Sparks of artificial general intelligence: early experiments with GPT-4 arXiv. Available online at: http://arxiv.org/abs/2303.12712

Calderon, N., Reichart, R., and Dror, R. (2025) *The* alternative annotator test for LLM-as-a-judge: how to statistically justify replacing human annotators with LLMs. arXiv. Available online at: http://arxiv.org/abs/2501.10970

Carcone, A. I., Naar, S., Eggly, S., Foster, T., Albrecht, T. L., and Brogan, K. E. (2015). Comparing thin slices of verbal communication behavior of varying number and duration. *Patient Educ. Couns.* 98, 150–155. doi: 10.1016/j.pec.2014.09.008

Chollet, M., and Scherer, S. (2017) Assessing public speaking ability from thin slices of behavior. In 2017 12th IEEE international conference on automatic face gesture recognition (FG 2017), 310–316

Cicero, M. T. (1942). De oratore (E. W.Sutton & H.Rackham, Trans). Cambridge, Massachusetts: Harvard University Press Original work published ca. 55 B.C.E.

Cicero, M. T. (1949). De inventione (H. M.Hubbell, Trans). Cambridge, Massachusetts: Harvard University Press Original work published ca. 87 B.C.E.

Cullen, A., and Harte, N. (2017). Thin-slicing to predict viewer impressions of TED talks. In The 14th International Conference on Auditory-Visual Speech Processing, 58–63. doi: 10.21437/avsp.2017-12

Cullen, A., Hines, A., and Harte, N. (2018). Perception and prediction of speaker appeal – a single speaker study. *Comput. Speech Lang.* 52, 23–40. doi: 10.1016/j.csl.2018.04.004

DErrico, F., Signorello, R., Demolin, D., and Poggi, I. (2013). The perception of charisma from voice: a cross-cultural study. In 2013 Humaine association conference on affective computing and intelligent interaction. Geneva, Switzerland, 552–557. doi: 10.1109/acii.2013.97

Dillion, D., Tandon, N., Gu, Y., and Gray, K. (2023). Can AI language models replace human participants? *Trends Cogn. Sci.* 27, 597–600. doi: 10.1016/j.tics.2023.04.008

Domjan, M. (2020). The principles of learning and behavior. Belmont, California: Wadsworth Publishing.

Doumont, J. (2009). Trees, maps, and theorems. Brussels: Principiae.

Eger, S., Cao, Y., D'Souza, J., Geiger, A., Greisinger, C., Gross, S., et al. (2025) Transforming science with large language models arXiv. Available online at: http://arxiv.org/abs/2502.05151

Ewers, R. M. (2018). Do boring speakers really talk for longer? *Nature* 561, 464–465. doi: 10.1038/d41586-018-06817-z

Feng, G., Joe, J., Kitchen, C., Mao, L., Roohr, K. C., and Chen, L. (2019). A proof-of-concept study on scoring oral presentation videos in higher education. *ETS Res. Rep. Ser.* 1, 1–28.

Fischer, O., Jeitziner, L. T., and Wulff, D. U. (2024). Affect in science communication: a data-driven analysis of TED talks on YouTube. *Humanit. Soc. Sci. Commun.* 11, 1–9. doi: 10.1057/s41599-023-02247-z

Fiske, S. T., Cuddy, A. J., and Glick, P. (2007). Universal dimensions of social cognition: warmth and competence. *Trends Cogn. Sci.* 11, 77–83. doi: 10.1016/j.tics.2006.11.005

Forghani, D., Ghafurian, M., Rasouli, S., Nehaniv, C. L., and Dautenhahn, K. (2024). Evaluating people's perceptions of an agent as a public speaking coach. *Paladyn J. Behav. Robot.* 15, 1–24. doi: 10.1515/pjbr-2024-0004

Gheorghiu, A. I., Callan, M. J., and Skylark, W. J. (2020). A thin slice of science communication: are people's evaluations of TED talks predicted by superficial impressions of the speakers? *Soc. Psychol. Personal. Sci.* 11, 117–125. doi: 10.1177/1948550618810896

Gilardi, F., Alizadeh, M., and Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proc. Natl. Acad. Sci. USA* 120:e2305016120. doi: 10.1073/pnas.2305016120

Gong, X., and Huskey, R. (2023). Media selection is highly predictable, in principle. *Comput. Commun. Res.* 5:1. doi: 10.5117/CCR2023.1.15.GONG

Grahe, J. E., and Bernieri, F. J. (1999). The importance of nonverbal cues in judging rapport. *J. Nonverbal Behav.* 23, 253–269. doi: 10.1023/A:1021698725361

Gu, J., and Bourne, P. E. (2007). Ten simple rules for graduate students. *PLoS Comput. Biol.* 3:e229. doi: 10.1371/journal.pcbi.0030229

Hall, J. A. (2001). "The PONS Test and the psychometric approach to measuring interpersonal sensitivity," in *Interpersonal sensitivity: Theory and measurement*. ed. J. A. Hall and F. J. Bernieri (Lawrence Erlbaum Associates Publishers), 143–160.

Hall, J. A., Harvey, S. E., Johnson, K. E., and Colvin, C. R. (2021). Thin-slice accuracy for judging big five traits from personal narratives. *Pers. Individ. Differ.* 171:110392. doi: 10.1016/j.paid.2020.110392

Hey, B. (2024). Mastering scientific presentations: Unlocking your communication skills. Wiesbaden, Germany: Springer Fachmedien Wiesbaden.

Houser, M. L., Horan, S. M., and Furler, L. A. (2007). Predicting relational outcomes: an investigation of thin slice judgments in speed dating. *Hum. Commun.* 10, 69–81.

Ismail, M. (2016). Thin slices of public speaking: a look into speech thin slices and their effectiveness in accurately predicting whole-speech quality. *Commun. Cent. J.* 2, 18–38.

Jussim, L. (2017). Précis of social perception and social reality: why accuracy dominates bias and self-fulfilling prophecy. *Behav. Brain Sci.* 40:e1. doi: 10.1017/S0140525X1500062X

Kelley, T. L. (1925). The applicability of the spearman-Brown formula for the measurement of reliability. *J. Educ. Psychol.* 16, 300–303. doi: 10.1037/h0073506

Kim, M., and Cappella, J. N. (2019). Reliable, valid and efficient evaluation of media messages: developing a message testing protocol. *Int. J. Inf. Commun. Technol. Educ.* 23, 179–197. doi: 10.1108/JCOM-12-2018-0132

Kraemer, H. C. (1992). How many raters? Toward the most reliable diagnostic consensus. *Stat. Med.* 11, 317–331. doi: 10.1002/sim.4780110305

Krippendorff, K. (2004). Content analysis: An introduction to its methodology. Thousand Oaks, California: SAGE.

Krumhansl, C. (2010). Plink: "thin slices" of music. *Music. Percept.* 27, 337–354. doi: 10.1525/mp.2010.27.5.337

Laserna, C. M., Seih, Y. T., and Pennebaker, J. W. (2014). Um.. Who like says you know: filler word use as a function of age, gender, and personality. *J. Lang. Soc. Psychol.* 33, 328–338. doi: 10.1177/0261927X14526993

Laske, M. M., and DiGennaro Reed, F. D. (2024). Um, so, like, do speech disfluencies matter? A parametric evaluation of filler sounds and words. *J. Appl. Behav. Anal.* 57, 574–583. doi: 10.1002/jaba.1093

Lucas, E. (2020). The art of public speaking. New York, NY: McGraw Hill.

Markowitz, D. M., and Hancock, J. T. (2024). Generative AI are more truth-biased than humans: a replication and extension of core truth-default theory principles. *J. Lang. Soc. Psychol.* 43, 261–267. doi: 10.1177/0261927X231220404

Mehrabian, A. (1972). Nonverbal communication. New Brunswick: Aldine Transaction.

Microsoft Corporation (2025). Rehearse your slide show with speaker coach. Available online at: https://support.microsoft.com/en-us/office/rehearse-your-slide-show-with-speaker-coach-cd7fc941-5c3b-498c-a225-83ef3f64f07b (Accessed March 14, 2025)

Mitchell, M. (2019). Artificial intelligence: A guide for thinking humans. London, UK: Penguin UK.

Murphy, N. A., and Hall, J. A. (2021). Capturing behavior in small doses: a review of comparative research in evaluating thin slices for behavioral measurement. *Front. Psychol.* 12:667326. doi: 10.3389/fpsyg.2021.667326

Neuendorf, K. A. (2017). The content analysis guidebook. Thousand Oaks, California: SAGE.

Nguyen, L. S., and Gatica-Perez, D. (2015). I would hire you in a minute: thin slices of nonverbal behavior in job interviews. Proceedings of the 2015 ACM on international conference on multimodal interaction, Seattle, WA, USA, 51–58. doi: 10.1145/2818346.2820760

OECD (2010). Educational research and innovation inspired by technology, driven by pedagogy. Paris: Organization for Economic co-operation and Development (OECD).

Omar, M., Nassar, S., Hijazi, K., Glicksberg, B. S., Nadkarni, G. N., and Klang, E. (2025). Generating credible referenced medical research: a comparative study of openAI's GPT-4 and Google's Gemini. *Comput. Biol. Med.* 185:109545. doi: 10.1016/j.compbiomed.2024.109545

Osgood, C. E., Suci, G. J., and Tannenbaum, P. H. (1957). The measurement of meaning. Champaign, Illinois: University Illinois Press.

Parrott, C. T., Brodsky, S. L., and Wilson, J. K. (2015). Thin slice expert testimony and mock trial deliberations. *Int. J. Law Psychiatry* 42-43, 67–74. doi: 10.1016/j.ijlp.2015.08.009

Pearson, K. (1895). Vii. Note on regression and inheritance in the case of two parents. *Proc. R. Soc. Lond.* 58, 240–242.

Quintilian, M. F. (1920). Institutio oratoria (H. E. Butler, Trans.). Cambridge, Massachusetts: Harvard University Press.

Riff, D., Lacy, S., and Fico, F. (2014). Analyzing media messages: Using quantitative content analysis in research. London, UK: Routledge.

Rodero, E. (2022). Effectiveness, attractiveness, and emotional response to voice pitch and hand gestures in public speaking. *Front. Commun.* 7:869084. doi: 10.3389/fcomm.2022.869084

Rodero, E., Larrea, O., and Mas, L. (2022). Speakers' expressions before and in a public presentation. Pleasantness, emotional valence, credibility, and comprehension effects. *Prof. Inferm.* 31, 1–16. doi: 10.3145/epi.2022.jul.05

Rosenberg, A., and Hirschberg, J. (2009). Charisma perception from text and speech. *Speech Comm.* 51, 640–655. doi: 10.1016/j.specom.2008.11.001

Schlamp, S., Gerpott, F. H., and Voelpel, S. C. (2020). Same talk, different reaction? Communication, emergent leadership and gender. *J. Manage. Psychol.* 36, 51–74. doi: 10.1108/JMP-01-2019-0062

Schmälzle, R. (2022). Theory and method for studying how media messages prompt shared brain responses along the sensation-to-cognition continuum. *Commun. Theory* 32, 450–460. doi: 10.1093/ct/qtac009

Schmälzle, R., Häcker, F. E. K., Honey, C. J., and Hasson, U. (2015). Engaged listeners: shared neural processing of powerful political speeches. *Soc. Cogn. Affect. Neurosci.* 10, 1137–1143. doi: 10.1093/scan/nsu168

Schmälzle, R., Hartung, F.-M., Barth, A., Imhof, M. A., Kenter, A., Renner, B., et al. (2019). Visual cues that predict intuitive risk perception in the case of HIV. *PLoS One* 14:e0211770. doi: 10.1371/journal.pone.0211770

Scholz, C., Baek, E. C., O'Donnell, M. B., Kim, H. S., Cappella, J. N., and Falk, E. B. (2017). A neural model of valuation and information virality. *Proc. Natl. Acad. Sci. USA* 114, 2881–2886. doi: 10.1073/pnas.1615259114

Schramm, W. (1947). Measuring another dimension of newspaper readership. *Journalism Q.* 24, 293–306. doi: 10.1177/107769904702400401

Shrout, P. E., and Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychol. Bull.* 86, 420–428. doi: 10.1037/0033-2909.86.2.420

Shulman, H. C., Markowitz, D. M., and Rogers, T. (2024). Reading dies in complexity: online news consumers prefer simple writing. *Sci. Adv.* 10:eadn2555. doi: 10.1126/sciadv.adn2555

Silver, D., Singh, S., Precup, D., and Sutton, R. S. (2021). Reward is enough. *Artif. Intell.* 299:103535. doi: 10.1016/j.artint.2021.103535

Skinner, B. F. (1961). Teaching machines. *Sci. Am.* 205, 90–102. doi: 10.1038/scientificamerican1161-90

Slepian, M. L., Bogart, K. R., and Ambady, N. (2014). Thin-slice judgments in the clinical context. *Annu. Rev. Clin. Psychol.* 10, 131–153. doi: 10.1146/annurev-clinpsy-090413-123522

Todorov, A. (2017). Face value: The irresistible influence of first impressions. Princeton, NJ: Princeton University Press.

Tunstall, L., von Werra, L., and Wolf, T. (2022). Natural language processing with transformers. London, UK: O'Reilly Media, Inc.

Uleman, J. S. (2023). "Differences between spontaneous and intentional trait inferences," in *The handbook of impression formation: A social psychological approach*. eds E. Balcetis and G. B. Moskowitz (Thousand Oaks, California: Routledge).

Valls-Ratés, Ï., Niebuhr, O., and Prieto, P. (2023). Encouraging participant embodiment during VR-assisted public speaking training improves persuasiveness and charisma and reduces anxiety in secondary school students. *Front. Virtual Real.* 4:1074062. doi: 10.3389/frvir.2023.1074062

Wallbott, H. G., and Scherer, K. R. (1986). Cues and channels in emotion recognition. *J. Pers. Soc. Psychol.* 51, 690–699. doi: 10.1037/0022-3514.51.4.690

Willis, J., and Todorov, A. (2006). First impressions: making up your mind after a 100-ms exposure to a face. *Psychol. Sci.* 17, 592–598. doi: 10.1111/j.1467-9280.2006.01750.x

Wittgenstein, L. (1953). Philosophical investigations (Philosophische Untersuchungen). New York, NY: Macmillan.

Wolfe, A., and Siegman, S. (2014). Multichannel integrations of nonverbal behavior. East Sussex, UK: Psychology Press.