Proceedings of the Nineteenth International AAAI Conference on Web and Social Media (ICWSM 2025)

GNN and LLM Insights: Multimodal Cues and Gender Disparities in Video Conversations

Dimosthenis Stefanidis,[1] George Pallis, [1] Marios Dikaiakos [1] Nicos Nicolaou [2]

[1] University of Cyprus
[2] University of Warwick
stefanidis.dimosthenis@ucy.ac.cy, pallis.george@ucy.ac.cy, dikaiakos.marios@ucy.ac.cy, Nicos.Nicolaou@wbs.ac.uk

Abstract

As video content on online platforms continues to increase, understanding the complex aspects of interpersonal communication becomes crucial. Central to this exploration is the pressing issue of gender bias, which manifests in multimodal interactions through visual, vocal, or verbal cues. These interactions present challenges in extracting and interpreting the subtle cues that may point to underlying biases. To tackle these challenges, we introduce a semi-automatic extraction of features and knowledge from user-generated content on video web platforms. Using 1,091 unstructured multi-participant video conversations from Shark Tank, we examine whether the multimodal cues (e.g., emotions) of a conversational participant (e.g., entrepreneur) affect another participant (e.g., investor) differently due to gender biases. Our methodology employs advanced deep learning algorithms for cues extraction and leverages Graph Neural Networks to model the multi-participant conversations. To complement our findings, we utilize textual features extracted through our methodology and employ GPT-4 to simulate decision-making scenarios, thereby assessing its analytical capabilities and potential gender biases.

Introduction

In an era marked by a significant rise in video content across web platforms like YouTube (Che, Ip, and Lin 2015), technological advancements continually struggle with new techniques to process, analyze, and interpret this vast amount of data (Pu et al. 2021). A particularly pressing task in this landscape is the extraction of features and knowledge from user-generated content, which often consists of rich interactions and complex interpersonal communications.

Delving deeper, a significant subset of this challenge lies in understanding multimodal interactions within these videos—particularly in extracting vocal, visual, and verbal cues. Such cues, stemming from complex and intertwined layers of information, are crucial in deciphering human interactions and are especially complex when multiple participants are involved. As dialogues flow, visual and auditory cues overlap, requiring sophisticated computational methodologies for accurate interpretation (Zhang et al. 2019b).

Given these advanced computational capabilities, our study ventures into a domain where such multimodal cues play a decisive role: start-up investment pitches (Liebregts et al. 2020). These pitches, characterized by complex interactions, epitomize persuasive scenarios and serve as a pivotal use case that enables the exploration and validation of our methodologies within a context where communicative dynamics critically influence outcomes. Amidst this context, an ongoing discussion around gender bias in investment decisions has gained prominence (Poczter and Shapsis 2018; Pistilli et al. 2022; Jetter and Stockley 2023). Yet, current state-of-the-art research provides a limited perspective, often overlooking the multifaceted factors that dictate investment outcomes. Hence, several studies simultaneously advocate for a deeper dive into the effects of facial expressions, vocal cues, and linguistic patterns, especially concerning their interplay with gender (Allison et al. 2022; Khurana and Lee 2023). With these considerations in mind, our research hinges on the focal question: *Do multimodal cues of males and females affect funding decisions differently?*

To answer this question, we collect unstructured video data from YouTube, specifically focusing on investment pitches from the renowned Shark Tank competition. The nature of this competition offers real-world context for scrutinizing gender bias in conversations. We then introduce a semi-automatic extraction methodology of features and knowledge from this user-generated video content, capturing rich interactions from interpersonal communications. To achieve this, we construct a video processing pipeline that extracts and integrates visual, vocal, and verbal cues from the various participants present in each video. On top of the pre-processed data, we employ statistical analysis, Graph Neural Networks (GNNs) and Large Language Models (LLMs) to: (i) discern multimodal cues' impact on investment outcomes with a focus on gender bias; (ii) effectively model the dynamic, multi-participant nature of conversations within these pitches; and (iii) examine the investment decisions and gender bias in state-of-the-art LLMs.

In summary, the contributions of this work are:

- An **exploratory analysis** that delves into the association between multimodal cues of entrepreneurs and the funding decisions made by angel investors. Central to our inquiry is the profound influence of gender bias. Further, we examine whether investor emotions are differentially swayed based on an entrepreneur's gender.
- A **methodology for multimodal cues extraction and**

conversation modeling that leverages deep learning algorithms and GNNs on unstructured video conversations. This approach introduces a novel application within the computational social science and entrepreneurship community, where dynamic, multimodal interaction modeling between entrepreneurs and investors as graphs have not been extensively explored.

- An **evaluation study for gender bias encoded in state-of-the-art generative models**, focusing on a comparative analysis between GPT-4 and our domain-specific trained GNN models. This comparison is central to our study as it assesses overall performance and examines the extent of gender bias in investment decisions between domain-specific trained GNN models and the GPT-4, thus shedding light on the potential biases introduced by state-of-the-art AI-driven decision-making systems.

Related Work

The success of entrepreneurs in securing investment is influenced significantly by how they present themselves and their ventures, with various personal and communicative characteristics playing pivotal roles. Exploring the association between voice characteristics and funding, Allison (Allison et al. 2022) found that voice intensity is significantly related to perceived passion and, subsequently, to funding success. Similarly, visual cues also play a crucial role; studies like those by Tsay (Tsay 2021) and Jiang (Jiang, Yin, and Liu 2019) show that displaying higher levels of joy during pitches enhances funding prospects. Moreover, research by Davis (Davis et al. 2021) indicates that facial expressions impact funding differently across genders, with women benefiting from expressions like anger and disgust, while men benefit from sadness and happiness. Additionally, while studies such as Clarke (Clarke, Cornelissen, and Healey 2019) and Ren (Ren et al. 2021) explore how literal and figurative language in pitches affects funding outcomes, findings suggest that the type of language used has limited impact, except for arousal words which are positively associated with project success. *Although there are multiple studies that examined diverse communication signals, there is no study that examines the success of entrepreneurs combining vocal, facial and verbal characteristics.*

Further complicating the entrepreneurial landscape, gender bias pervades various domains, influencing perceptions and outcomes through widespread discriminatory practices (Way, Larremore, and Clauset 2016). This societal bias is particularly pronounced in entrepreneurship, where studies have consistently shown that female entrepreneurs encounter formidable challenges in securing funding for their businesses due to prevailing investor perceptions and gender stereotypes (Poczter and Shapsis 2018; Pistilli et al. 2022). These stereotypes typically favor male entrepreneurs, who are perceived as more assertive and competent due to societal norms (Rudman and Phelan 2008). *Despite the extensive research, there remains a significant gap in understanding how gender and associated multimodal cues influence interpersonal communication during entrepreneurial pitches.*

The manifestation of gender bias does not confine itself merely to conventional societal frameworks but also extends into the area of AI, where biases from training data infiltrate the predictive models (Roselli, Matthews, and Talagala 2019). This extension of bias is particularly notable in LLM models like ChatGPT, which has been revealed to inadvertently reinforce gender stereotypes through its responses (Gross 2023). This phenomenon not only reflects but can also amplify existing societal biases. Further explorations into the ethical and social challenges posed by ChatGPT, contribute to a multifaceted discourse around the responsible deployment of AI (Ray 2023; Van Dis et al. 2023). *Although research exploring gender biases in LLMs exists, it often does not specifically address how these biases influence concrete decision-making processes such as investment decisions (Ray 2023). Thus, our study seeks to fill this gap by specifically examining how these biases affect investment decisions made by ChatGPT 4.0, testing the model's response to gender-disclosed versus non-disclosed pitches.*

Amid these considerations, the application of advanced technologies such as deep learning and GNNs in conversation modeling presents new opportunities. GNNs, in particular, have shown promise in capturing complex interaction dynamics in multi-speaker conversations, effectively modeling relationships that traditional models might miss (Zhang et al. 2019a; Zhou et al. 2020; Liang et al. 2022). *However, the potential of GNNs to detect and analyze gender biases within conversations remains largely unexplored.*

Methodology

Exploring conversational dynamics in videos, our research constructs a comprehensive methodology to reveal potential gender biases in multimodal interactions (Figure 1). Initially, we collected a dataset of 1,091 unstructured multi-participant video conversations from Shark Tank and developed a sophisticated data preprocessing phase to handle voice, face, and text data. From the processed data, we extracted 30 multimodal features (e.g., vocal emotions, lexical diversity) to serve as independent variables (IVs) in our analysis, thereby exploring complexities within the conversational dynamics. To ensure the reliability and validity of our findings, we integrated a set of control variables, accounting for potential external factors that might influence the outcomes related to the primary IVs of interest. Subsequently, we combined traditional statistical methods with advanced models like GNNs to analyze conversational dynamics and uncover hidden biases. Lastly, we utilized GPT-4 to evaluate its predictive capabilities and investigate inherent gender biases, thus providing a thorough inquiry into the ethical considerations of using AI in decision-making.

Data Preprocessing

Dataset Collection. We collected a dataset of 1,091 unstructured multi-participant video conversations from YouTube, specifically focusing on investment pitches from the popular U.S. pitch competition, Shark Tank, where entrepreneurs present their ventures to a panel of five investors for potential investment. This competition has been the subject of various academic inquiries due to its authentic depic-

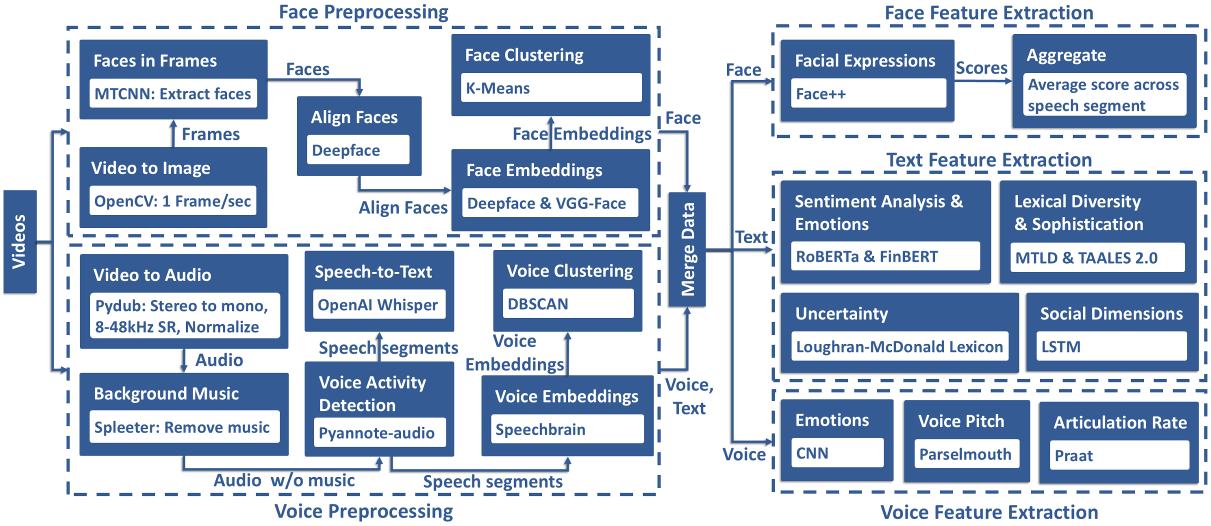


Figure 1: Video Processing Methodology

tion of entrepreneurial pitching (Khurana and Lee 2023; Jetter and Stockley 2023). Our dataset, with each video averaging a length of 10 minutes, was assembled starting with the collection of foundational information from Wikipedia. This information contained the name of each start-up and the specific season-episode they appeared in. With this data at hand, we executed structured searches on YouTube using the query ”Shark Tank season X episode Y COMPANY_NAME”, enabling the location of the corresponding videos.

Voice & Text Processing. First, using the library Pydub, we extracted the audio from the video files, converted the stereo audio to mono with sample rate 8kHz-48kHz, and normalized the audio (based on EBU R128 standard). To minimize musical distortions, we utilized Spleeter, a pre-trained deep learning library, and removed any background music (Hennequin et al. 2020). Then, we used the library Pyannote-audio (Bredin et al. 2020) to divide the audio into distinct speech segments (Voice Activity Detection and Speaker Change Detection). Each speech segment is a continuous utterance by a speaker and possesses an ID indicating its sequence (e.g., 1), raw audio data, and its time-frame within the video (e.g., start=23s, end=40s). From the 1091 videos, we obtained over 60600 speech segments, averaging 55 segments per video, each with an average duration of 8.8 seconds. To identify the various speakers within a video, voice embeddings for each speech segment were created using Speechbrain toolkit (Ravanelli et al. 2021). Subsequently, we employed the DBSCAN algorithm from Scikit-learn to cluster similar voice embeddings, ensuring that each cluster represents a unique speaker. By determining the number of clusters, we ascertained the total number of participants engaged in each video conversation. Finally, we converted each speech segment to text using the OpenAI Whisper tool (Radford et al. 2023).

Face Processing. We then processed the images of the videos. Specifically, using OpenCV (Bradski and Kaehler 2008), we first cut each video into frames, with each frame lasting approximately one second. In total, we have over 670000 frames from the 1091 videos, with each video providing 615 frames on average. Then, we extracted and aligned (i.e., rotate face according to the eyes) the face of each individual using Deepface tool and the MTCNN algorithm (Serengil and Ozpinar 2020). For each extracted face, we created their face (vector) embeddings using Deepface and VGG-Face (Parkhi, Vedaldi, and Zisserman 2015). Then, we grouped together similar face embeddings using Scikit-Learn and K-Means (Pedregosa et al. 2011) and created N clusters for each video (N = #speakers identified by voice processing). Each point in a cluster represents a facial image, while a cluster represents a unique participant.

Merge Voice, Face and Text. Due to the asynchronous nature of the multimodal data (e.g., the voice of a speaker could be accompanied by the visual data from another speaker), we developed a methodology for aligning the speaker-specific data across different speech segments. Initially, associating the text with voice was straightforward, given that the text was directly transcribed from speech segments, forming a natural linkage. Subsequently, to link the vocal and facial data, we introduced a mapping methodology represented by Formula (1). Let $S = \{s_1, s_2, \ldots, s_n\}$ be the set of speech segments, $F = \{f_1, f_2, \ldots, f_m\}$ be the set of facial clusters, and $T(f_j, s_i)$ be the number of faces in facial cluster f_j that appear during the timeframe of speech segment s_i. Our objective was to map each speech segment s_i to the facial cluster f_j that maximizes the count of face occurrences T during s_i.

$$M(s_i) = \arg\max_{f_j \in F} T(f_j, s_i) \tag{1}$$

Domain-Specific Processing. To expand our domain-specific dataset, we manually coded 11 additional variables from the videos (e.g., business model, #investors). We opted for manual coding given the nuanced nature of these variables and the potential inaccuracies of automated methods in such context-sensitive tasks. We hired 2 experts and provided them with specific coding guidelines and training prior to manual coding (Allison et al. 2022). Then, we provided videos from the 1091 pitches to the first expert to code the variables. After that, we provided 450 pitches, randomly selected from the whole sample to the second expert to validate the coded variables. The interrater agreement between the experts was acceptable with a Krippendorff's a = 0.94. In the case of disagreements between the two experts, a third expert validated the coded values.

Next, we categorized participants based on their roles, such as investors and entrepreneurs. To achieve this, we calculated the similarity between participants (Formula 2). Specifically, for each participant i in pitch X, we computed the cosine similarity of face (F_{ij}) and voice embeddings (V_{ij}) with every other participant j in pitch Y, where N is the number of participants in all pitches and $Y \neq X$. Then, we ranked each participant based on the magnitude of their cosine similarities (R_i), indicating the degree of resemblance across pitches. Finally, based on the known number of investors I in each pitch, we categorize the top I ranked participants as investors, and the remaining participants as entrepreneurs (Formula 3).

$$R_i = \text{rank}\left(\sum_{j=1, j\neq i}^{N} F_{ij} + V_{ij}\right) \tag{2}$$

$$\text{Category}_i = \begin{cases} \text{Investor} & \text{if } R_i \leq I, \\ \text{Entrepreneur} & \text{otherwise.} \end{cases} \tag{3}$$

Feature Extraction

Independent Variables (IVs). From the preprocessed data, we extract 30 features, grouped into 9 categories:

Voice Emotions: For identifying voice emotions, we use a deep neural network (CNN) classification model which predicts the emotions of a human speaker encoded in an audio file (80% f1 score) (neutral, calm, happy, sad and angry) (de Pinto et al. 2020).

Voice Pitch: refers to "the relative highness or lowness of a tone as perceived by the ear, which depends on the number of vibrations per second produced by the vocal cords" (Britannica 2020). For measuring the voice pitch, we use Parselmouth tool (Jadoul, Thompson, and De Boer 2018).

Voice Articulation Rate: Articulation rate is a prosodic feature defined as "a measure of rate of speaking in which all pauses are excluded from the calculation" (Goldman-Eisler 1961). For measuring the articulation rate of a speaker, we use Praat tool (Boersma and Weenink 2018).

Facial Expressions: For identifying facial expressions, we use Face++. It captures 7 expressions (happiness, sadness, surprise, anger, fear, disgust and neutral) and is used widely in the research community (Wang, Otto, and Jain 2016) as it has 90.6% accuracy for face detection (Jung et al. 2018).

Text Emotions: For measuring sentiment and text emotions, we utilize two different tools based on BERT (Devlin et al. 2018). First, we use RoBERTa, a language model for sentiment analysis (95% accuracy) and emotion detection (sadness, happiness, love, anger, fear and surprise; 93.95% accuracy) (Liu et al. 2019). Then, we use FinBERT, a language model for financial sentiment analysis with an accuracy score of 86% (e.g. "Pre-tax loss totaled euro 0.3 million") (Liu et al. 2021).

Social Dimensions in Conversation: Social dimensions refer to various aspects of social interactions and relationships that can be identified and measured through the analysis of conversations. These dimensions can include power dynamics, emotional states, social identities, and norms and expectations around communication styles. To measure these social dimensions, we use a deep neural network that predicts the type of the relationship that is expressed in text (trust, conflict, knowledge, power, status, support, romance, similarity, identity and fun; 0.85 AUC) (Choi et al. 2020).

Lexical Diversity: Lexical diversity refers to the range of different words used in a text, with a greater range indicating a higher diversity (McCarthy and Jarvis 2010). To measure lexical diversity, we use the state-of-the-art algorithm called MTLD (McCarthy and Jarvis 2010).

Lexical Sophistication: The construct of lexical sophistication includes both the depth and breadth of lexical knowledge (i.e. range of advanced words used) (Read 2013). For measuring lexical sophistication, we utilize TAALES 2.0 tool (Kyle, Crossley, and Berger 2018).

Uncertainty: To measure the "uncertainty" of speakers, we filter their speech transcripts using the lexicon of Loughran and McDonald (Loughran and McDonald 2011), and count the occurrences of the "uncertain" words.

Control Variables. We control for several factors that could bias our data analysis. First, we control for the season of each pitch (e.g. dummy/binary variables for the 12 seasons of Shark Tank) to account for the impact of the global investment climate during the year of the filming, as well as slight variations in the show's format (e.g. in season 5 they removed a fee in the form of equity or royalties paid simply for appearing on the show) (Jetter and Stockley 2023). We control for the total number of presenters in a pitch, their ethnicity ("Black", "White", "Asian", "Mixed") and their age (as given by Face++, MAE=7.65) (Lavanchy, Reichert, and Joshi 2022; Allison et al. 2022). Furthermore, we control for whether the presenters have a patent for their product ("Patent Obtained", "Patent Filed/In-Progress", "No patent effort") (Lavanchy, Reichert, and Joshi 2022) and whether they have a loan/debt (1 or 0) related to their business. In addition, we control for whether a product/service is seasonal (1 or 0) (e.g., a product/service gets most of their sales in specific periods like summer or Christmas), its distribution channels ("Physical", "Web", "Both") and the revenue model of the venture ("Production/Transactional model", "Rental/Leasing model", "Subscription model", "Other"). We also control for the following industries, "Children/Education", "Fashion/Beauty", "Fitness/Sports/Outdoors", "Food and Beverage", "Health/Wellness/Cleaning", "Lifestyle/Home", "Pet Products", "Software/Tech", "Other" (Jetter and Stockley 2023). Finally, we control for the revenue of the venture in the previous year (Lavanchy, Reichert, and Joshi 2022).

Dependent Variable (DV). Following prior entrepreneurial research, funding was operationalized as a binary variable indicating whether a venture has received an offer (1) or not (0) (Jetter and Stockley 2023).

Modeling and Explainability

Statistical Modeling. To analyze our dataset, we initially adopt logistic regression, a widely-used technique in data analysis literature. This approach necessitates aggregating information from individual speech segments within each video. Specifically, we extract the relevant features from each segment and compute a weighted average score for the entire video. This weighted average takes into account the time contribution of each team member's speech segment within the pitch. Separate weighted averages are calculated for entrepreneurs and investors, respectively, using the time proportion of their individual speech segments. Formula (4) and (5) elucidates how these weighted averages are determined. This aggregated data serves as the foundation for the logistic regression analysis carried out in our study.

$$\text{Entrepreneurs avg score}(F) = \sum_{i=1}^{SE} \frac{\text{Secs}(i) \cdot \text{Score}(F, i)}{\sum_{j=1}^{SE} \text{Secs}(j)} \quad (4)$$

$$\text{Investors avg score}(F) = \sum_{i=1}^{SI} \frac{\text{Secs}(i) \cdot \text{Score}(F, i)}{\sum_{j=1}^{SI} \text{Secs}(j)} \quad (5)$$

SE, SI = Speech Segments of Entrepreneurs or Investors, F = Feature

GNN Modeling and Explainable AI. While the statistical modeling can successfully identify the differences of the multimodal cues between speakers, it has a limitation. Specifically, it uses a static approach that averages speaker features across the entire video, failing to capture the dynamic nature and temporal patterns of the conversation. Consequently, it is unable to fully account for the temporal dependencies and relationships between speakers throughout the conversation.

To address the aforementioned limitations and drawing on insights from previous studies, we introduce a GNN model to capture the dynamic interactions between participants (e.g., entrepreneurs and investors) as a graph. By transforming each conversation into a graph representation, we not only capture the temporal dependencies inherent in the dialogue but also identify pivotal nodes and features (e.g., key conversational turns) that significantly influence the outcome of a conversation. Our preference for GNNs over other models (e.g., RNN, transformers) is due to their superior ability to handle conversational data and track the evolving significance of conversational segments (Liang et al. 2022; Chen et al. 2022). Furthermore, research supports that GNNs excel in tasks where the graph's connectivity and complexity are critical to the task (Di Massa et al. 2006; Abadal et al. 2021). Thus, through this approach, we aim to provide a more comprehensive understanding of the multimodal cues and their role in predicting the success or failure of pitches in competitions.

To implement this approach, we first extract the multimodal cues from each speech segment per conversation (Figure 2). Subsequently, we create a graph representation of each conversation, where speech segments serve as nodes and conversational flow as edges. Following graph construction, we employ a GNN architecture which consists of a 4-layer structure: two GCN layers to effectively perceive localized node features, a linear layer for mapping learned representations, and an LSTM layer to grasp temporal dependencies within the dialogues. The model includes a global mean pooling step and a dropout layer with a rate of 0.5 for regularization before classification through a log-softmax layer. Importantly, the model's design ensures that hidden layers halve their size successively, starting with an initial dimension of 128 for the first layer. Also, we use the Adam optimizer, defined by a learning rate of 0.01, and train utilizing Cross-Entropy Loss. For hyperparameter optimization, we employ grid search for maximizing the F1 score. Two specific GNN models are trained, each tailored to a gender demographic: one utilizing graphs from females, and the other from males, thereby enabling the models to decipher feature importance and predict the likelihood of pitch funding within gendered contexts.

To interpret and explain the predictions made by our models, we utilize Explainable AI tools such as Pytorch (Paszke et al. 2019), CAPTUM (feature importance) (Kokhlikyan et al. 2020), and GNNExplainer (node importance) (Ying et al. 2019). Based on these tools, we extract the feature importance per graph/pitch. Our resulting dataset comprises 897 rows, with 272 rows for the female pitches and 625 rows for the male pitches. The dataset has N columns, representing the number of features, and each cell in the dataset contains the importance score of a specific feature for a given pitch. These scores help identify the key factors that contribute to the success or failure of a pitch.

To compare the importance of features and nodes between female and male entrepreneurs, we employ unpaired t-tests. This analysis allows us to assess any differences in feature and node importance based on gender. Additionally, we evaluate the predictive performance of our trained GNN models using stratified 5-fold cross-validation to ensure robustness across different subsets of the data.

LLM Modeling. In light of the advancements in LLMs, we incorporate GPT-4 into our analysis to evaluate its analytical capabilities and potential gender biases. Utilizing text transcripts derived through our computational methodology, we first anonymize the text transcripts of Shark Tank pitches by removing identifiers such as the entrepreneurs' name, product name, and company name to prevent data leakage. We then use these anonymized transcripts to present GPT-4 with two distinct versions, one with and one without gender identifiers, allowing us to explore how gender disclosure influences the model's investment predictions.

Utilizing t-tests and regression analysis, we assess how the concealment or disclosure of gender influences GPT-4's decision-making processes. This approach allows us to probe the latent biases that may be intrinsic to AI systems and highlights the ethical dilemmas of employing AI in crit-

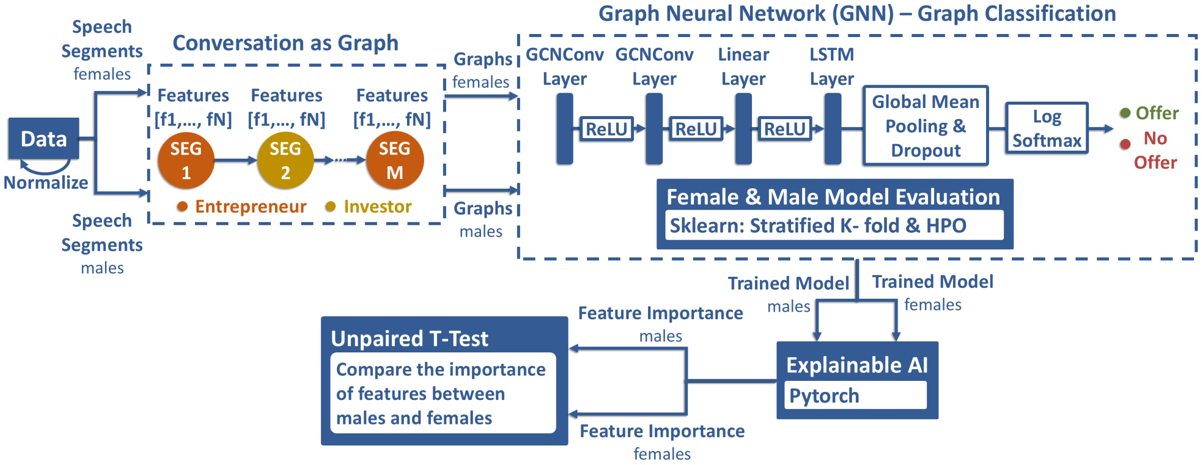


Figure 2: GNN Methodology

ical decision-making scenarios, such as entrepreneurial investment settings (Zhou et al. 2023; Rivas and Zhao 2023). Moreover, we conduct a comparative analysis between GPT-4's performance and that of our trained GNN models. This comparison is pivotal, as it not only evaluates the performance but also the extent of gender bias in the predictions of these advanced AI models, thereby elucidating the strengths and limitations of state-of-the-art LLMs versus custom-trained neural networks in real-world investment scenarios.

Experiments

System Performance

In the empirical evaluation of the proposed system, computational performance was assessed using a hardware configuration of 2 x Intel(R) Xeon(R) Gold 6230 CPUs @ 2.10GHz, a Nvidia Tesla T4 GPU 16GB GDDR6 VRAM, and 96GB RAM. Our performance analysis reveals a significant linear relationship ($p < .001$) between the video duration and processing time, underscoring the computational efficiency and scalability of the system across various video lengths. Moreover, the system maintained a consistent performance profile across all processing categories (Figure 3), thus enabling accurate forecasting of resource utilization and processing time for different video durations.

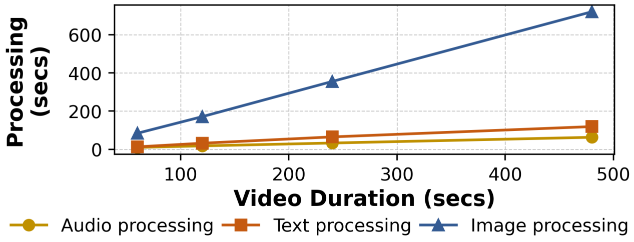
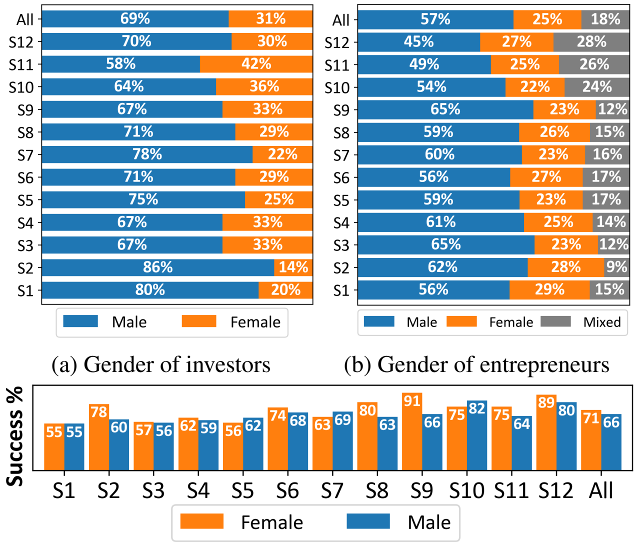


Figure 3: Processing Time vs Video Duration



(a) Gender of investors

(b) Gender of entrepreneurs

(c) Percentage of offers received per gender group

Figure 4: Statistics of Shark Tank

Descriptive Statistics

As we set the stage for our experiments, understanding the underlying trends within our dataset becomes crucial. Hence, in Figure 4, we provide a snapshot of gender distribution among both investors and entrepreneurs in Shark Tank. These figures show a notable predominance of male investors (69%) and male-led start-ups (57%), yet also indicate that female-led start-ups tend to secure funding more frequently (71%) than their male-only counterparts (66%).

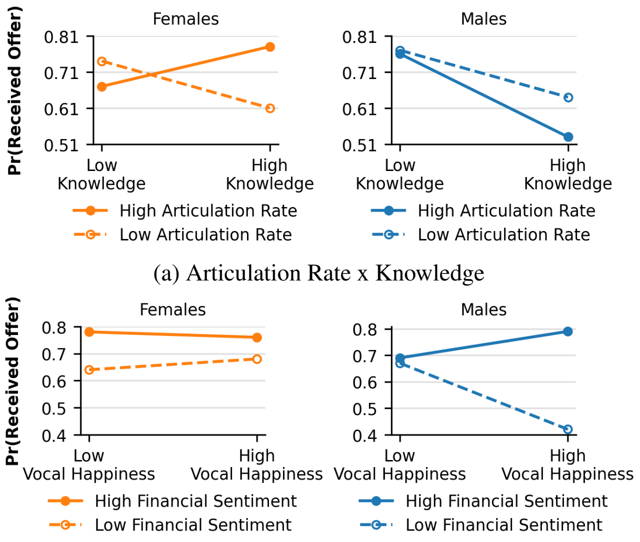
Data Analysis Using Logistic Regression

Entrepreneurs' Cues: Our first line of inquiry focuses on the differential effects of entrepreneurs' vocal, visual, and verbal attributes on funding between males and females. Through separate regression analyses for each gen-

Feature	Females	Males	Test Diff.
(E) Trust	-0.010	0.105***	3.97*
(E) Conflict	-0.093**	0.005	7.54**
(E) Knowledge	0.012	-0.055***	6.46*
(E) Lexical Diversity	1.363	-2.238*	5.17*
(E) Uncertainty	0.300***	0.094*	5.07*
(I) Conflict	-0.108***	-0.051***	5.41*
(I) Power	-0.012	0.034***	6.70**
(E) (INT) Articulation Rate x Knowledge	0.181	-0.060	5.19*
(E) (INT) Vocal Happiness x Financial Sentiment	-0.130	0.568**	6.07*
(E) (INT) Voice Pitch x Facial Sadness x Conflict	-0.001	0.001	8.99**
(E) (INT) Articulation Rate x Smiling x Lexical Sophistication	0.000*	0.000	4.81*
(E) (INT) Articulation Rate x Facial Anger x Lexical Sophistication	0.000**	-0.000	9.06**

Notes. $***p < .001, **p < .01, *p < .05$; (I): Investor's Feature, (E): Entrepreneur's Feature, (INT): Interaction Term

Table 1: Logistic regression - Compare the importance features between males and females



(a) Articulation Rate x Knowledge

(b) Vocal Happiness x Financial Sentiment

Figure 5: Two-way Interactions of Entrepreneurs' Characteristics (Low = -1SD, High = +1SD)

der, we examined 30 independent variables (IVs). Our primary findings, summarized in Table 1, spotlight five key attributes—trust, conflict, knowledge, lexical diversity, and uncertainty—where the impact on funding varies significantly between the two genders.

For instance, showing trust in investors appears to bolster funding opportunities for males, while for females, the same trust can be neutral or even detrimental. Conversely, conflict with investors harms females but has minimal effect on males, highlighting gender biases in conflict perceptions. Showing in-depth knowledge or increased lexical diversity can dissuade investment for males; however, these attributes don't pose the same disadvantage for females, suggesting gender-specific interpretations. Uncertainty stands out as the only attribute that is beneficial for both genders, but more so for females, possibly viewed as greater self-awareness. These results, backed by statistical evidence, underscore the multifaceted complexities in interpersonal communication.

Investors' Cues: Building on our earlier examination of entrepreneurs' cues, we turned our lens toward investors' vocal, visual, and verbal reactions to explore their role in the funding process, and more importantly, how they affect male and female entrepreneurs differently. Our analysis, separated by gender and exploring 30 investors' cues, zeroes in on two key investor indices—conflict and power scores—as these exhibited significant gender-based variations in funding.

In summary, our findings reveal that investors' behavior significantly contributes to gender disparities in funding (Table 1). Specifically, higher conflict scores from investors disproportionately disadvantage female entrepreneurs. This suggests that females, more than males, might be penalized for any perceived differences or disagreements with potential investors. Conversely, a higher power score from investors significantly benefits male entrepreneurs but not females. This implies that investors, perhaps subconsciously, lean toward supporting pitches where they can exercise control, especially if presented by male entrepreneurs.

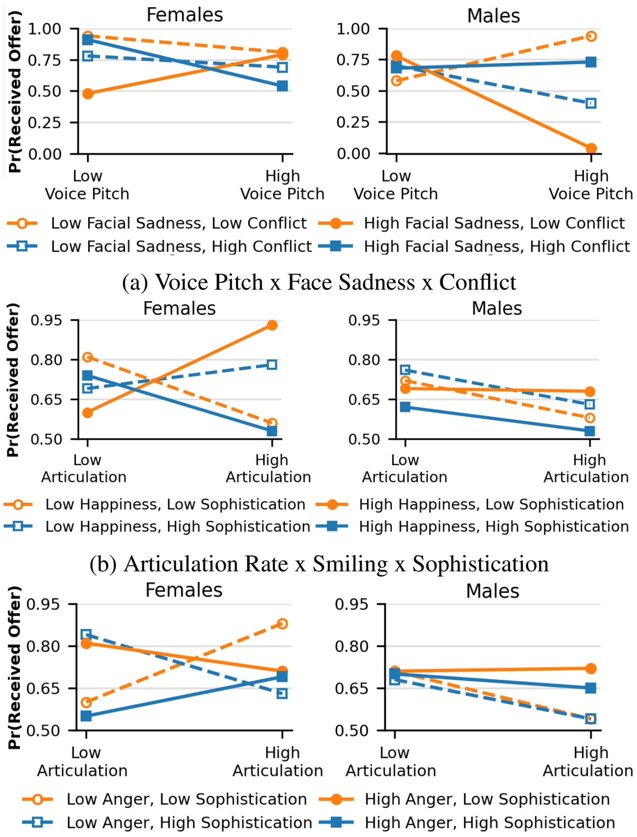
Two-Way Interactions: To delve deeper into our analysis, we introduce two-way interaction terms (Jaccard and Turrisi 2003). These involve creating a product term between two variables to gauge if the effect of one variable is modulated by the level of a second variable. In the context of our study, this allows us to observe how paired cues of entrepreneurs jointly affect funding decisions, and if the combined effects differ between males and females. Our investigation uncovers two significant gender-based differences. First, in the interaction between 'Articulation Rate' and 'Knowledge' (Figure 5a), we find a statistically significant disparity between males and females ($x^2(1) = 5.19$, $p < .05$). Notably, rapid speech combined with in-depth knowledge substantially boosts funding chances for females ($p < .001$). Second, the interaction between 'Vocal Happiness' and 'Financial Sentiment' (Figure 5b) diverges significantly between genders ($x^2(1) = 6.07$, $p < .05$). While negative financial sentiment diminishes funding prospects for both genders, a cheerful tone mitigates this effect significantly more for females than for males ($p < .001$).

Three-Way Interactions: Further, we introduce three-way interactions to assess if the combined effect of two variables on an outcome is influenced by a third variable's level. For our study, this means assessing how the interplay among three cues of entrepreneurs affect funding and discerning gender-specific patterns. Our analysis reveals three gender-based differences. First, the interaction between 'Voice Pitch', 'Facial Sadness', and 'Conflict' (Fig-

Features	Males (N=625)		Females (N=272)		t	p-value
	Mean	SD	Mean	SD		
(E) Trust	0.002	0.006	-0.000	0.001	5.255	0.000
(E) Conflict	0.001	0.002	-0.000	0.015	2.472	0.014
(E) Knowledge	0.002	0.014	0.004	0.006	-2.348	0.019
(E) Uncertainty	-0.002	0.001	0.001	0.013	-2.746	0.006
(I) Conflict	0.002	0.003	-0.000	0.023	2.639	0.009
(I) Power	0.002	0.007	0.001	0.006	2.388	0.017
(E) (INT) Articulation Rate x Knowledge	-0.001	0.009	0.004	0.009	-6.619	0.000
(E) (INT) Vocal Happiness x Financial Sentiment	0.001	0.003	0.000	0.002	3.066	0.002
(E) (INT) Voice Pitch x Facial Sadness x Conflict	0.000	0.000	-0.000	0.000	2.547	0.011
(E) (INT) Articulation Rate x Smiling x Lexical sophistication	-0.000	0.001	-0.000	0.001	-2.268	0.024
(E) (INT) Articulation Rate x Facial Anger x Lexical Sophistication	-0.001	0.001	-0.000	0.000	-3.833	0.000

Notes. (I): Investor's Feature, (E): Entrepreneur's Feature, (INT): Interaction Term

Table 2: GNN - Two-sample t-test to compare the importance features between males and females



(a) Voice Pitch x Face Sadness x Conflict

(b) Articulation Rate x Smiling x Sophistication

(c) Articulation Rate x Facial Anger x Sophistication

Figure 6: Three-way Interactions of Entrepreneurs' Characteristics (Low = -1SD, High = +1SD)

ure 6a) significantly differs between the genders ($x^2(1) = 8.99$, $p < .01$). Female entrepreneurs are more likely to secure funding when they have a high-pitched voice, exhibit sadness, and avoid conflict ($p < .001$). Second, the interaction among 'Articulation Rate', 'Smiling', and 'Lexical Sophistication' (Figure 6b) diverges significantly between males and females ($x^2(1) = 4.81$, $p < .05$). Females are more likely to receive funding when speaking quickly, refraining from excessive smiling, and using straightforward language ($p < .01$). Lastly, the interaction between 'Articulation Rate', 'Facial Anger', and 'Lexical Sophistication' (Figure 6c) is also significantly different between genders ($x^2(1) = 9.06$, $p < .01$). Specifically, females are more likely to secure funding when they speak quickly, appear less angry, and use simpler language ($p < .001$).

GNN

After conducting initial analyses with logistic regression, we sought to validate and potentially refine our findings using GNN models. Specifically, our first point of investigation was to re-examine the differential effects of entrepreneurs' cues on funding based on gender. For brevity and focus, we concentrated on the 5 cues that previously showed gender-based disparities in their effect on funding: trust, conflict, knowledge, lexical diversity, and uncertainty. Table 2 presents the independent t-tests, comparing the significance of these feature importances for both genders. The results reaffirm our earlier findings, indicating that all examined features have significantly different impacts on male and female entrepreneurs.

Our subsequent analysis with the GNN models centered on the 2-way and 3-way interaction effects. Again, we aimed to validate if the combined effects of certain attributes on funding varied significantly between male and female entrepreneurs. Table 2 provides the independent t-tests comparing these interaction effects for both genders. Consistent with our logistic regression analyses, the results from the GNN models corroborate that all inspected 2-way and 3-way interactions exhibit significant gender-based variations.

Next, we conduct a structural analysis of the graphs/pitches and extract the most important node/segment for

Model	Precision	Recall	F1
GNN - Vocal (F)	0.70±0.07	0.63±0.03	0.64±0.03
GNN - Vocal (M)	0.64±0.04	0.60±0.03	0.60±0.04
GNN - Facial (F)	0.65±0.07	0.61±0.05	0.61±0.07
GNN - Facial (M)	0.64±0.03	0.60±0.02	0.60±0.03
GNN - Verbal (F)	0.69±0.05	0.66±0.04	0.67±0.04
GNN - Verbal (M)	0.61±0.14	0.57±0.11	0.57±0.11
GNN - All (F)	**0.72±0.05**	**0.69±0.06**	**0.70±0.06**
GNN - All (M)	**0.69±0.04**	**0.67±0.02**	**0.66±0.02**

F: Females, M: Males

Table 3: Performance of GNNs using Stratified 5-fold CV

each pitch. Our findings reveal a significant gender-based difference in the timing of the most pivotal speech segment for entrepreneurs who successfully received funding ($p <$ .001). Specifically, we observe that this critical segment occurred notably earlier for females, around 33% into the conversation, while for males, this segment emerged around 50% into the pitch. Additionally, this specific segment was found to be significantly more influential in predicting pitch success for males ($p <$ .001). On the other hand, for entrepreneurs who did not secure funding, the timing of the most crucial speech segment showed no significant difference between genders ($p >$ .1). These insights could indicate that investors' information processing may vary depending on the entrepreneur's gender. This might also suggest variations in strategic pitching approaches between genders. For instance, in the case of female entrepreneurs, the findings could suggest that investors are persuaded earlier in the pitch, potentially because essential information is presented sooner. Alternatively, it might point to an underlying gender bias in investors' decision-making processes.

We next evaluate the predictive performance of the GNN model for male and female entrepreneurs (Table 3). We find that the models trained with all types of features (vocal, facial, verbal) outperform those trained solely on vocal, facial, or verbal features ($p <$ .000, $p <$ .000 and $p <$ .000, respectively). Furthermore, the GNN model, trained with all types of features, achieved an F1 score of 0.70 for females and 0.66 for males, while the difference in predictive performance between the genders is statistically significant ($p <$ .000). Interestingly, despite having approximately 2.3 times more data for males than for females, our results suggest that female entrepreneurs may be more expressive and easier to predict than their male counterparts.

Finally, to test for gender bias within our model, we conduct a t-test comparing the predictions for male and female entrepreneurs (Table 4). We find that the difference in predictive performance between the genders is not statistically significant ($p >$ .1), indicating no evident gender bias.

GPT-4

In our evaluation process of GPT-4, we analyzed its performance using real data from Shark Tank pitches to explore its decision-making capabilities and potential gender biases, comparing it to our trained GNN models (multimodal and text-only) to benchmark its effectiveness. Initially, we anonymized the pitches by hiding identifiers such as the entrepreneurs' name, product name, and company name using GPT-4 to prevent data leakage (Prompt 1). Then, we presented the same anonymized pitch twice: once without revealing the gender of the entrepreneurs (Prompt 2) and once with gender disclosure (Prompt 3). By presenting the same pitch twice—once without and once with gender information—we were able to directly compare the model's responses and identify any potential gender bias. These tests were conducted using the April 9, 2024, release of GPT-4.0 Turbo, and at a temperature setting of 0 to ensure deterministic and predictable response patterns, crucial for evaluating consistent patterns in AI behavior.

	Males		Females		t	p-value
	Mean	SD	Mean	SD		
GNN-All	0.736	0.441	0.739	0.440	-0.093	0.926

Table 4: Two-sample unpaired t-test to compare the predictions of GNN between males and females

Prompt 1: *From the provided pitch, please conceal the names of the entrepreneurs, the product, and the company. DO NOT CHANGE ANYTHING ELSE. PITCH: "Text"*

Prompt 2: *Predict whether the startup pitch is likely to be funded by the investors on the US version of the TV show 'Shark Tank'. Consider the show's historical funding patterns, investor personalities, and the entertainment value of the pitch, alongside traditional business evaluation criteria such as Problem and Solution, Business Model, Market Potential, Sales and Financials, Team Motivation, Risks and Concerns, and Valuation. PROVIDE ONLY ONE CHOICE FOR YOUR DECISION. The output should have the following format: 'Funded' or 'Not Funded' Also, provide a few keywords for your chain of thought. PITCH: "Text"*

Prompt 3: *PROMPT 2 + The following pitch belongs to a team of female/male entrepreneurs.*

First, we evaluate the predictive performance of GPT-4 (Table 5) by comparing it against our domain-specific trained GNN models, focusing initially on textual data to ensure a fair comparison. In this context, our textual GNN model outperformed GPT-4 by 10.8% in F1 score for female entrepreneurs and by 4.9% for male entrepreneurs. To further enhance our analysis, we also assessed GPT-4 against our multimodal GNN model, which includes verbal, vocal, and facial cues. Here, the GNN model demonstrated superior performance, outperforming GPT-4 by 13.8% for females and 13.9% for males in F1 score. These results underscore the effectiveness of integrating multiple data modalities, which significantly boosts prediction performance. Overall, while GPT-4 is known for its robust textual data processing capabilities, our findings highlight that domain-specific trained models can achieve higher performance.

Next, to investigate whether GPT-4 exhibits gender bias, we conducted unpaired t-tests comparing the GPT-4 predictions for male and female entrepreneurs (Table 6). Initially, with gender not disclosed, there was no significant

Model	Precision	Recall	F1
GPT-4.0 Turbo (F)	0.794	0.435	0.562
GPT-4.0 Turbo (M)	0.736	0.404	0.521

F: Females, M: Males

Table 5: GPT-4.0 Performance in Predicting 'Shark Tank' Investment Outcomes

	Males		Females		t	p-value
	Mean	SD	Mean	SD		
W/O Gender	0.359	0.480	0.391	0.489	-0.877	0.381
W/ Gender	0.423	0.494	0.556	0.498	-3.619	0.000

Table 6: Two-sample unpaired t-test to compare the predictions of GPT-4 between males and females

difference in predictions between genders ($p > .1$). However, after disclosing gender, predictions for female entrepreneurs were significantly more favorable than for male entrepreneurs ($p < .001$), suggesting a gender bias. To further validate these findings, we performed a logistic regression analysis with the DV being the predictions of GPT-4 and the main IV being the interaction term between 'isFemale' x 'Gender Revealed'. This analysis confirmed that the differences in predictions became statistically significant ($p < .001$) and favored females once gender was disclosed (Figure 7). These results substantiate the presence of gender bias in GPT-4 investment decisions when gender is known.

To validate the reasoning and output of GPT-4 in our investment prediction evaluation, we analyzed the keywords derived from Prompt 2, which highlight its chain-of-thought for each prediction (Figure 8). Our findings reveal distinct patterns in the reasoning process of GPT-4. For pitches predicted as likely to be funded, the five most frequently cited reasons were "Innovative Product", "High Market Potential", "Clear Problem/Solution", "Entertainment Value", and "Engaging Presentation". In contrast, for pitches deemed unlikely to be funded, the reasons included "Niche Market/Product", "Financial Concerns", "Unclear Presentation", "Early Stage Product", and "Scalability Concerns". These insights provide a deeper understanding of the factors GPT-4 considers significant in determining the potential success or challenges of entrepreneurial ventures.

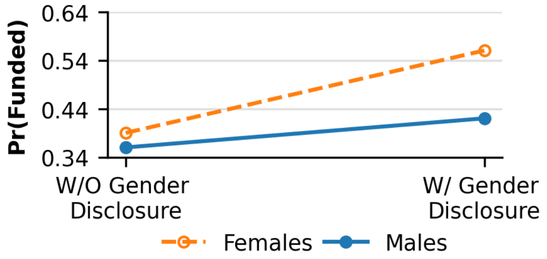


Figure 7: Two-way Interaction - Gender Disclosure Effects on GPT-4's Funding Decisions

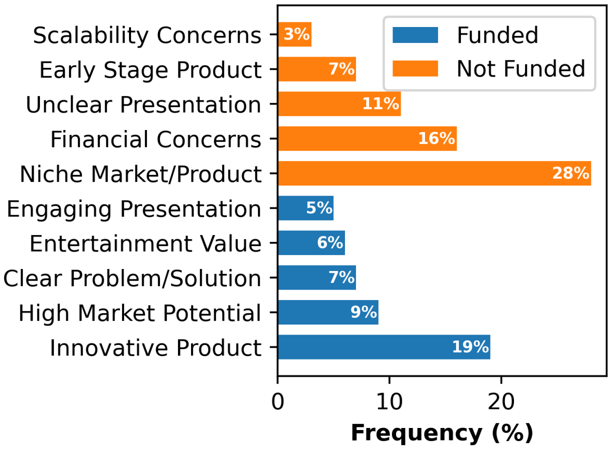


Figure 8: Top 10 Reasons for Funding Decisions of GPT-4

Limitations and Future Work

This study, while comprehensive in its approach, encounters certain limitations that must be acknowledged. Firstly, the accuracy and inherent limitations of the AI tools that were employed to extract 30 distinct characteristics of entrepreneurs and investors, can significantly impact our overall findings. These tools, while advanced, are not flawless and their performance can vary, potentially affecting the reliability of the feature extraction process. Specifically, the CNN model used for voice emotion detection may vary in effectiveness across different accents and audio qualities, possibly skewing emotion recognition results. Similarly, Parselmouth, which measures voice pitch, and Praat, which assesses articulation rate, could both be affected by audio quality and background noise, impacting their accuracy. Face++, used for facial expressions, may also exhibit inconsistencies due to different lighting conditions or image qualities. Additionally, text analysis tools like RoBERTa and FinBERT may carry biases from their training data, which could affect their accuracy, particularly in specialized contexts like financial sentiment analysis. Tools such as MTLD and TAALES 2.0, which measure lexical diversity and sophistication, may not fully capture the range of vocabulary if speakers use specialized jargon or non-standard language forms. Finally, the use of the Loughran and McDonald lexicon to measure uncertainty might also fail to capture the contextual nuances of how uncertainty is expressed, potentially limiting the breadth of detected uncertainties.

Another limitation of our research is that focuses exclusively on U.S.-based episodes of Shark Tank, which may introduce cultural biases. This geographical concentration means the findings might not be fully representative or applicable to entrepreneurial dynamics in different cultural or business contexts. As such, the results should be interpreted with an understanding of these potential limitations and the specific context of the U.S. business environment. Additionally, our methodological choice to model conversations with a linear, unidirectional graph might oversimplify the complex dynamics of real-life interactions.

For future research, we aim to investigate more complex graph structures to better understand non-linear discourse relationships (Chen et al. 2022), enhancing our modeling capabilities for real conversations. This exploration will complement our ongoing efforts to explore various demographic biases, including those based on ethnicity or age, thereby enhancing our understanding of communication biases. Extending this approach further, significant insights could be gained by applying our methodology to areas such as online job interviews, law videos, and media interviews. Investigating biases in hiring decisions through online job interviews, uncovering conviction biases related to gender or ethnicity in law videos, and exploring how gender biases influence interviewer and guest interactions are just a few examples. These applications could not only broaden the scope of our methodology but also deepen our understanding of communicative processes in various professional settings.

Conclusion

Our research offers a pioneering exploration into the complexities of multimodal communication and its intersection with gender bias. By leveraging deep learning algorithms, we construct a methodology for extracting multimodal cues from unstructured video conversations, subsequently modeled using a GNN architecture. Our approach efficiently captures the complex dynamics of multi-participant dialogues. Central to our findings is the elucidation of how various multimodal cues, within the domain of investment decisions, vary considerably based on gender. Moreover, our utilization of GPT-4 in decision-making simulations underscores both the potential and caution needed with AI systems, especially concerning biases, emphasizing the imperative for rigorous fairness assessments in AI-driven outcomes.

Data and Code Availability

Data and code are available in:
https://github.com/dstefa02/GNN-and-LLM-Insights-Multimodal-Cues-and-Gender-Disparities-in-Video-Conversations.

Ethics Statement

All results presented in this paper are based on aggregated estimates and do not contain any individual information. The participants in 'US Shark Tank' are aware of the public broadcast and wide accessibility of the content they appear in. Episodes utilized for analysis were accessible on YouTube, reflecting their availability in the public domain.

References

Abadal, S.; et al. 2021. Computing graph neural networks: A survey from algorithms to accelerators. *ACM Computing Surveys (CSUR)*, 54(9): 1–38.

Allison, T. H.; et al. 2022. Can you hear me now? Engendering passion and preparedness perceptions with vocal expressions in crowdfunding pitches. *JBV*, 37(3): 106193.

Boersma, P.; and Weenink, D. 2018. Praat: doing phonetics by computer [Computer program]. Version 6.0. 37. *Retrieved February*, 3: 2018.

Bradski, G.; and Kaehler, A. 2008. *Learning OpenCV: Computer vision with the OpenCV library.* " O'Reilly Media, Inc.".

Bredin, H.; et al. 2020. Pyannote. audio: neural building blocks for speaker diarization. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7124–7128. IEEE.

Britannica, T. 2020. Editors of encyclopaedia. *Argon. Encyclopedia Britannica.*

Che, X.; Ip, B.; and Lin, L. 2015. A survey of current YouTube video characteristics. *IEEE MultiMedia*, 22(2): 56–63.

Chen, Z.; et al. 2022. Structured Hierarchical Dialogue Policy with Graph Neural Networks. In *National Conference on Man-Machine Speech Communication*, 264–277. Springer.

Choi, M.; et al. 2020. Ten social dimensions of conversations and relationships. In *Proceedings of The Web Conference 2020*, 1514–1525.

Clarke, J. S.; Cornelissen, J. P.; and Healey, M. P. 2019. Actions speak louder than words: How figurative language and gesturing in entrepreneurial pitches influences investment judgments. *AMJ*, 62(2): 335–360.

Davis, B. C.; et al. 2021. Gender and counterstereotypical facial expressions of emotion in crowdfunded microlending. *Entrepreneurship Theory and Practice*, 45(6): 1339–1365.

de Pinto, M. G.; et al. 2020. Emotions understanding model from spoken language using deep neural networks and mel-frequency cepstral coefficients. In *2020 IEEE conference on evolving and adaptive intelligent systems (EAIS)*, 1–5. IEEE.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*

Di Massa, V.; et al. 2006. A comparison between recursive neural networks and graph neural networks. In *The 2006 IEEE International Joint Conference on Neural Network Proceedings*, 778–785. IEEE.

Goldman-Eisler, F. 1961. The significance of changes in the rate of articulation. *Language and Speech*, 4(3): 171–174.

Gross, N. 2023. What ChatGPT Tells Us about Gender: A Cautionary Tale about Performativity and Gender Biases in AI. *Social Sciences*, 12(8): 435.

Hennequin, R.; et al. 2020. Spleeter: a fast and efficient music source separation tool with pre-trained models. *Journal of Open Source Software*, 5(50): 2154.

Jaccard, J.; and Turrisi, R. 2003. *Interaction effects in multiple regression.* 72. Sage.

Jadoul, Y.; Thompson, B.; and De Boer, B. 2018. Introducing parselmouth: A python interface to praat. *Journal of Phonetics*, 71: 1–15.

Jetter, M.; and Stockley, K. 2023. Gender match and negotiation: evidence from angel investment on Shark Tank. *Empirical Economics*, 64(4): 1947–1977.

Jiang, L.; Yin, D.; and Liu, D. 2019. Can joy buy you money? The impact of the strength, duration, and phases of an entrepreneur's peak displayed joy on funding performance. *AMJ*, 62(6): 1848–1871.

Jung, S.-G.; et al. 2018. Assessing the accuracy of four popular face recognition tools for inferring gender, age, and race. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Khurana, I.; and Lee, D. J. 2023. Gender bias in high stakes pitching: an NLP approach. *Small Business Economics*, 60(2): 485–502.

Kokhlikyan, N.; et al. 2020. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*.

Kyle, K.; Crossley, S.; and Berger, C. 2018. The tool for the automatic analysis of lexical sophistication (TAALES): version 2.0. *Behavior research methods*, 50: 1030–1046.

Lavanchy, M.; Reichert, P.; and Joshi, A. 2022. Blood in the water: An abductive approach to startup valuation on ABC's Shark Tank. *J. Bus. Ventur.*, 17: e00305.

Liang, Y.; et al. 2022. Emotional conversation generation with heterogeneous graph neural network. *Artificial Intelligence*, 308: 103714.

Liebregts, W.; et al. 2020. The promise of social signal processing for research on decision-making in entrepreneurial contexts. *Small business economics*, 55(3): 589–605.

Liu, Y.; et al. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Liu, Z.; et al. 2021. Finbert: A pre-trained financial language representation model for financial text mining. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence*, 4513–4519.

Loughran, T.; and McDonald, B. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of finance*, 66(1): 35–65.

McCarthy, P. M.; and Jarvis, S. 2010. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2): 381–392.

Parkhi, O. M.; Vedaldi, A.; and Zisserman, A. 2015. Deep face recognition.(2015).

Paszke, A.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Pedregosa, F.; et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, 12: 2825–2830.

Pistilli, L.; et al. 2022. Gender Bias in Entrepreneurship: What is the Role of the Founders' Entrepreneurial Background? *Journal of Business Ethics*, 1–22.

Poczter, S.; and Shapsis, M. 2018. Gender disparity in angel financing. *Small Business Economics*, 51: 31–55.

Pu, J.; et al. 2021. Deepfake videos in the wild: Analysis and detection. In *Proceedings of the Web Conference 2021*, 981–992.

Radford, A.; et al. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, 28492–28518. PMLR.

Ravanelli, M.; et al. 2021. SpeechBrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624*.

Ray, P. P. 2023. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*.

Read, J. 2013. Validating a test to measure depth of vocabulary knowledge. In *Validation in language assessment*, 41–60. Routledge.

Ren, J.; et al. 2021. Exploring the subjective nature of crowdfunding decisions. *J. Bus. Ventur.*, 15: e00233.

Rivas, P.; and Zhao, L. 2023. Marketing with chatgpt: Navigating the ethical terrain of gpt-based chatbot technology. *AI*, 4(2): 375–384.

Roselli, D.; Matthews, J.; and Talagala, N. 2019. Managing bias in AI. In *Companion Proceedings of The 2019 World Wide Web Conference*, 539–544.

Rudman, L. A.; and Phelan, J. E. 2008. Backlash effects for disconfirming gender stereotypes in organizations. *Research in organizational behavior*, 28: 61–79.

Serengil, S. I.; and Ozpinar, A. 2020. Lightface: A hybrid deep face recognition framework. In *2020 innovations in intelligent systems and applications conference*, 1–5. IEEE.

Tsay, C.-J. 2021. Visuals dominate investor decisions about entrepreneurial pitches. *Academy of Management Discoveries*, 7(3): 343–366.

Van Dis, E. A.; et al. 2023. ChatGPT: five priorities for research. *Nature*, 614(7947): 224–226.

Wang, D.; Otto, C.; and Jain, A. K. 2016. Face search at scale. *IEEE transactions on pattern analysis and machine intelligence*, 39(6): 1122–1136.

Way, S. F.; Larremore, D. B.; and Clauset, A. 2016. Gender, productivity, and prestige in computer science faculty hiring networks. In *Proceedings of the 25th international conference on world wide web*, 1169–1179.

Ying, Z.; et al. 2019. Gnnexplainer: Generating explanations for graph neural networks. *Advances in neural information processing systems*, 32.

Zhang, D.; et al. 2019a. Modeling both Context-and Speaker-Sensitive Dependence for Emotion Detection in Multi-speaker Conversations. In *IJCAI*, 5415–5421.

Zhang, Z.; et al. 2019b. Neural multimodal belief tracker with adaptive attention for dialogue systems. In *The world wide web conference*, 2401–2412.

Zhou, J.; et al. 2020. Graph neural networks: A review of methods and applications. *AI open*, 1: 57–81.

Zhou, J.; et al. 2023. Ethical ChatGPT: Concerns, challenges, and commandments. *arXiv preprint arXiv:2305.10646*.

Ethics Checklist

1. For most authors...

(a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes. Our research does not violate social contracts such as privacy norms or perpetuate unfair profiling. We analyzed publicly available data from 'US Shark Tank', and ensured that all results are based on aggregated estimates without disclosing any individual information (refer to Ethics Statement).

(b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes

(c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, we clarify the appropriateness of our methodological approach for the claims made. The methodology section details the use of AI tools, GNNs, and LLMs for analyzing multimodal cues in investment decision-making scenarios (refer to Methodology section).

(d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, we clarify possible artifacts in the data used, given population-specific distributions. Our dataset primarily includes U.S.-based episodes of Shark Tank, which might introduce cultural biases (refer to Limitations and Future Work section).

(e) Did you describe the limitations of your work? Yes, we described the limitations of our work, including the reliance on AI tools for feature extraction and the focus on U.S.-centric Shark Tank episodes (refer to Limitations and Future Work section).

(f) Did you discuss any potential negative societal impacts of your work? Yes, we discussed potential negative societal impacts of our work, particularly in the context of gender bias and AI ethics (refer to Conclusion section).

(g) Did you discuss any potential misuse of your work? Yes, we discussed the potential misuse of our work, especially concerning the biases inherent in AI models like GPT-4 and the need for responsible deployment of AI systems (refer to Conclusion section).

(h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, we described steps taken to prevent or mitigate potential negative outcomes, such as ethical considerations in using AI tools, data anonymization, and ensuring the reproducibility of findings (refer to Ethics Statement and Data and Code Availability sections).

(i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes, we have read the ethics review guidelines and ensured that our paper conforms to them, as evidenced by our detailed Ethics Statement and the transparency in our methodology and data handling practices (refer to Ethics Statement and Methodology sections).

2. Additionally, if your study involves hypotheses testing...

   (a) Did you clearly state the assumptions underlying all theoretical results? NA

   (b) Have you provided justifications for all theoretical results? NA

   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA

   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA

   (e) Did you address potential biases or limitations in your theoretical framework? NA

   (f) Have you related your theoretical results to the existing literature in social science? NA

   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA

3. Additionally, if you are including theoretical proofs...

   (a) Did you state the full set of assumptions of all theoretical results? NA

   (b) Did you include complete proofs of all theoretical results? NA

4. Additionally, if you ran machine learning experiments...

   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes (refer to Methodology, Experiments and Data and Code Availability sections)

   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes (refer to Methodology and Experiments sections)

   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes (refer to Experiments section)

   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes (refer to System Performance subsection

   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes (refer to Methodology and Experiments sections)

   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

   (a) If your work uses existing assets, did you cite the creators? Yes, we cited the creators of all existing assets used in our study, including the AI tools and data sources like YouTube videos and the associated tools for data processing (refer to Methodology section and References).

(b) Did you mention the license of the assets? Yes (see Data and Code Availability section)

(c) Did you include any new assets in the supplemental material or as a URL? Yes (refer to Data and Code Availability section)

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, as the data used was sourced from publicly available YouTube videos, consent from individuals was not directly obtained. However, the use of such publicly available data falls within ethical research practices (refer to Ethics Statement).

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, the data does not contain personally identifiable information or offensive content. Our study focused on aggregated data from public sources without targeting individual identities (refer to Ethics Statement).

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? No, while our study involved the creation of a new dataset, we have not yet implemented the FAIR principles in its curation and release. Future efforts could focus on enhancing the dataset's compliance with these principles, including improving data findability, accessibility, interoperability, and reusability.

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? No, we did not create a Datasheet for the Dataset as per Gebru et al.'s guidelines in the current phase of our research. However, we recognize the importance of such documentation for transparency and ethical usage of datasets. We plan to consider this aspect in future updates or releases of the dataset to provide comprehensive documentation regarding its creation, usage, and limitations.

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? No. While the study involved researchers manually annotating the dataset rather than typical study participants, the full text of instructions given to these researchers for data annotation is available online (see "Data and Code Availability" section).

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA, because the manual annotation by researchers did not involve risks typically associated with human subjects research. Furthermore, our study's focus was on the analysis of publicly available data, and the annotators/researchers were not subjects of research themselves (refer to Ethics Statement). Based on our institution's policies and IRB, this type of internal research activity is out of scope and, therefore, approval was not required."

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? No, because the researchers involved in data annotation were not study participants in the traditional sense. They were part of the research team, and their compensation would fall under standard research labor, not participant compensation as in crowdsourced or experimental studies.

(d) Did you discuss how data is stored, shared, and deidentified? Yes, we ensured that all manually annotated data were stored securely and shared responsibly. While our dataset primarily consisted of publicly available information, any manual annotations were handled in a way that respected privacy and ethical guidelines. Personal identifiers were not part of the dataset, aligning with deidentification practices (refer to Ethics Statement and Data and Code Availability sections).