

ONLINE APPENDIX:
Gender Differences in Economics Seminars
American Economic Review

Pascaline Dupas Amy Handlan Alicia Sasser Modestino Muriel Niederle
Mateo Seré Haoyu Sheng Justin Wolfers*
and the Seminar Dynamics Collective

July 20, 2025

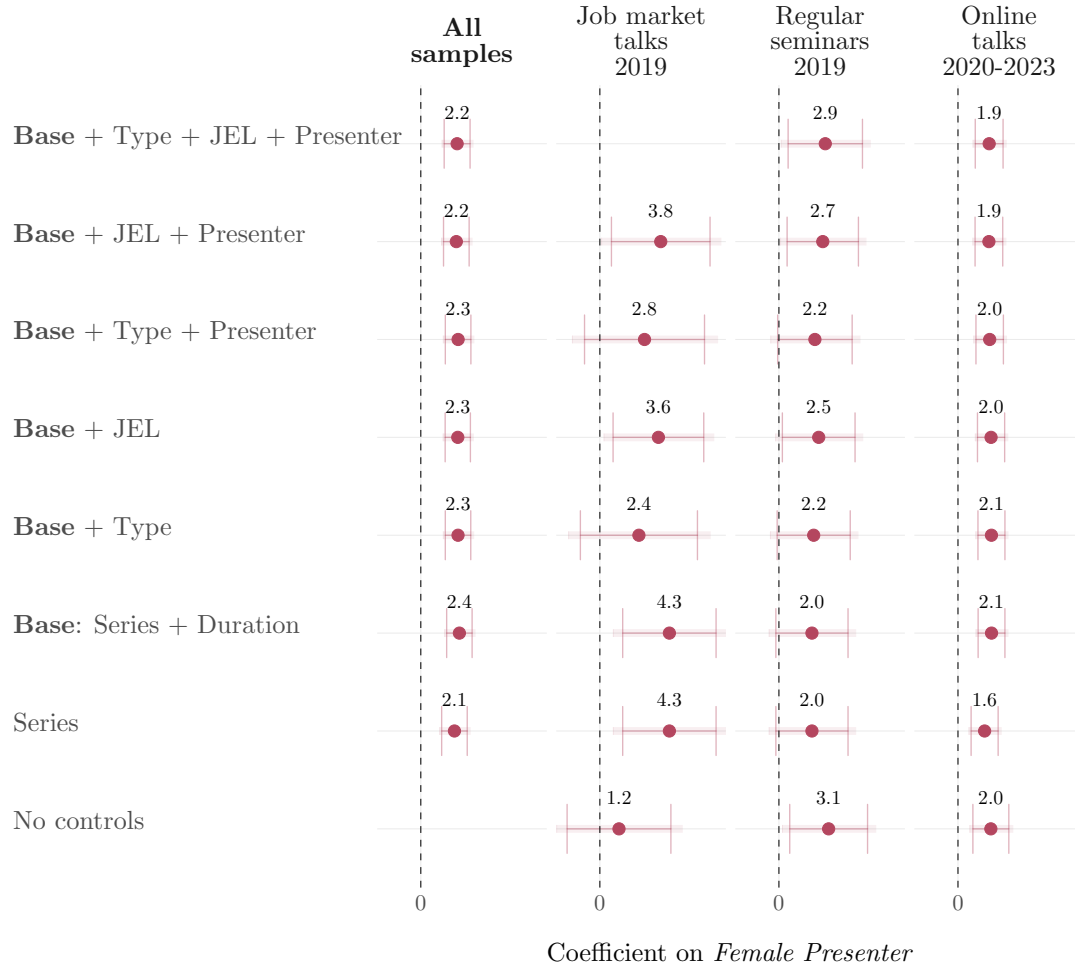
Contents

B Additional Figures and Tables	1
C Variable Descriptions	5
C.1 Number of Interruptions	5
C.2 Tenor and Tone of Interruptions	5
C.3 Types of Interruptions	6
C.4 Speaker Characteristics	7
C.5 Research Characteristics	9
C.6 Seminar Duration	11
C.7 Seminar series fixed effects	12
C.8 Human-Coder Characteristics	13
C.9 Attendance and Participation	13
D Details of Machine Learning Approach	14
D.1 Data Sources	14
D.2 Data Processing	15
D.3 Splitting Data into Training, Validation, and Testing Samples	15
D.4 Data Augmentation	17
D.5 Model Architecture	18
D.6 Model Training	19
D.7 Model Performance	20
D.8 Additional Validation with Alternative Methods	22
D.9 Additional Details on Computer-Coded Samples	27
D.10 Assignment of Research Types with LLMs	29

*Dupas: Princeton and NBER (pdupas@princeton.edu), Handlan: Brown (amy_handlan@brown.edu), Modestino: Northeastern (a.modestino@northeastern.edu), Sheng: Brown (haoyu_sheng@brown.edu), Niederle: Stanford and NBER (niederle@stanford.edu), Seré: University College London (m.sere@ucl.ac.uk), Wolfers: Michigan and NBER (jwolfers@umich.edu).

B Additional Figures and Tables

Figure B1: Gender Gap in Interruptions: Sensitivity to Controls with Standard Errors Clustered at the Presenter Level



Notes: This reproduces [Figure 1](#), with the additional of clustering at the presenter level, for the three samples for which we have individual identifiers and can cluster the standard errors at the presenter level.

Table B1: Main Outcomes for Job Market Talks: Reliability of Paper-Type Controls

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Job market talks JEL-only	Job market talks Type only	Job market talks JEL + Type	Job market talks Joint	Regular seminars JEL+Type	Regular seminars Joint	Pooled sample JEL+Type
Panel A: Total Attendance							
Female presenter	5.052 (1.570) {0.002}	3.624 (1.843) {0.051}	3.362 (1.836) {0.069}	4.740 (1.618) {0.004}	2.034 (1.092) {0.064}	2.151 (1.162) {0.065}	3.259 (0.952) {0.001}
Panel B: Total Interruptions							
Female presenter	3.768 (1.556) {0.016}	2.763 (1.828) {0.132}	3.458 (2.073) {0.097}	3.186 (1.617) {0.049}	2.870 (1.285) {0.027}	2.597 (1.241) {0.037}	2.855 (1.038) {0.006}
Panel C: Negative interruptions							
Female presenter	0.515 (0.278) {0.065}	0.365 (0.325) {0.262}	0.220 (0.315) {0.486}	0.574 (0.292) {0.050}	0.369 (0.308) {0.233}	0.384 (0.288) {0.183}	0.467 (0.206) {0.024}
Panel D: Positive interruptions							
Female presenter	-0.030 (0.374) {0.937}	-0.205 (0.419) {0.626}	-0.285 (0.490) {0.562}	-0.299 (0.535) {0.577}	-0.102 (0.615) {0.868}	-0.142 (0.567) {0.802}	-0.211 (0.419) {0.615}
Panel E: Types of Interruptions							
<i>Coef. on female presenter:</i>							
Type of interruption:							
Non-labeled	2.350 (1.312) {0.075}	1.550 (1.592) {0.331}	1.681 (1.599) {0.294}	2.195 (1.421) {0.123}	0.555 (0.999) {0.579}	0.251 (0.995) {0.801}	1.101 (0.853) {0.198}
Suggestions	0.751 (0.400) {0.062}	0.493 (0.451) {0.276}	0.709 (0.449) {0.116}	0.505 (0.430) {0.242}	0.673 (0.382) {0.080}	0.627 (0.356) {0.079}	0.573 (0.281) {0.042}
Comments	0.046 (0.638) {0.942}	-0.340 (0.769) {0.658}	-0.029 (0.849) {0.972}	-0.497 (0.683) {0.468}	0.131 (0.599) {0.828}	0.227 (0.568) {0.689}	-0.089 (0.457) {0.846}
Criticism	-0.417 (0.449) {0.354}	-0.373 (0.523) {0.476}	-0.101 (0.517) {0.845}	-0.302 (0.468) {0.520}	0.158 (0.385) {0.682}	0.160 (0.377) {0.671}	-0.042 (0.308) {0.893}
Clarifications	1.646 (1.079) {0.129}	2.104 (1.138) {0.066}	1.965 (1.330) {0.141}	1.839 (1.101) {0.096}	1.119 (0.745) {0.135}	1.135 (0.743) {0.128}	1.442 (0.650) {0.027}
Follow-ups	-0.199 (0.509) {0.696}	-0.200 (0.503) {0.691}	-0.204 (0.539) {0.705}	-0.240 (0.483) {0.620}	0.390 (0.359) {0.278}	0.398 (0.331) {0.230}	0.119 (0.287) {0.678}

Notes: Col. (1) is the preferred specification for the job market sample from Table 2. Col. (2) includes estimation for the job market sample with controls for JEL and Type controls. Col. (4) is the preferred specification for the regular seminar sample with all controls, including JEL and Type controls, from Table 2. “Joint” cols. (3) and (5), are estimated together. The control variables (including JEL and Type) have coefficients that are estimated jointly, and separate coefficients are estimated for the job market sample and for the regular seminar sample. Col. (6) estimates one female presenter coefficient for the pooled job market and regular seminar samples and uses both JEL and type controls.

There are only 39 unique job market papers by female presenters, limiting our ability to estimate on all coefficients reliably. To show that “type” controls do not overturn the results for the job market sample, we pool the job market sample and the regular seminar sample to jointly estimate the coefficients on controls while estimating sample-specific coefficients for presenter gender. We find almost the same results (column 4) as when we only control for JEL and not type (column 1). In contrast, the JEL+Type specification estimated on the job market sample alone tends to deviate from the results in the jointly estimated specification (column 3 vs. column 4). For this reason, we favor the JEL-only specification for the job market sample. For the regular seminar sample, the strong similarity in results between columns 5 and 6 indicates that the JEL+Type specification is appropriate.

Table B2: Number of Interruptions: Sensitivity to Controls

	(1) Job market talks 2019	(2) Regular seminars 2019	(3) NBER SI 2019	(4) Online talks 2020-2023	(5) NBER SI 2022	(6) All Samples (cols 1–5)
Panel A: No controls						
Female presenter	1.184 (1.581) {0.455}	3.077 (1.418) {0.031}	-1.295 (1.001) {0.196}	2.034 (0.656) {0.002}	-0.532 (0.751) {0.479}	0.970 (0.420) {0.021}
<i>Controls:</i>	None	None	None	None	None	None
Panel B: Most comparable controls						
Female presenter	4.303 (1.403) {0.002}	2.044 (1.219) {0.095}	0.820 (0.732) {0.263}	2.076 (0.492) {<0.001}	0.227 (0.602) {0.706}	1.660 (0.326) {<0.001}
<i>Controls:</i>						
Seminar series	Yes	Yes	Yes	Yes	Yes	Yes
Seminar duration	Yes	Yes	Yes	Yes	Yes	Yes
Mean (male presenter)	34.39	25.35	14.68	10.93	13.74	15.15
Number of observations	240	292	447	994	467	2,440
Number of unique talks	176	245	442	994	467	2,324

Notes: This table replicates [Table 2](#) with different control specifications. See [Figure 1](#) for full sensitivity analysis.

Table B3: Presenter Gender and Number of Interruptions, by Broad Field

	(1) Job market talks 2019	(2) Regular seminars 2019	(3) NBER SI 2019	(4) Online talks 2020–2023	(5) NBER SI 2022	(6) All Samples (cols. 1–5)
Female presenter						
x Macro	5.308 (2.809) {0.060}	-1.777 (4.862) {0.715}	4.128 (1.872) {0.028}	1.477 (0.691) {0.033}	0.759 (1.357) {0.577}	1.647 (0.547) {0.003}
x Micro	1.220 (2.056) {0.554}	3.774 (1.368) {0.006}	-0.750 (0.745) {0.315}	1.478 (0.694) {0.034}	0.349 (1.164) {0.764}	0.940 (0.432) {0.029}
x Other	7.662 (2.926) {0.010}	0.838 (2.745) {0.760}	0.440 (0.861) {0.610}	3.258 (1.294) {0.012}		3.146 (0.968) {0.001}
<i>Mean for male presenter:</i>						
Macro	35.75	29.31	17.26	11.73	15.53	15.66
Micro	37.59	25.76	13.95	10.02	11.72	15.75
Other	29.60	15.03	8.73	10.77	n.a	12.72
<i>Number of observations:</i>						
Macro	51	41	215	419	240	924
Micro	124	224	173	294	227	1,084
Other	65	27	59	281		432

Notes: Standard errors in parentheses and p-values in brackets. Specifications identical to those shown in [Table 2](#). As per the data use agreement, all talks in the NBER 2022 sample were labeled as either “Macro” or “Micro”. Col. 5: For the smaller sample used in [Table 4](#), the coefficient for female x micro is 0.349, with a se of 1.163, and the coefficient for female x macro is 0.759, with a se of 1.357. The control mean for Macro seminars is 16.22, and 13.27 for Micro.

C Variable Descriptions

This section includes detailed descriptions of variables used in this paper. Additional details on human coders (including tests for coder biases) can be found in [Dupas et al. \(2021\)](#). Additional details on machine coding can be found in [Appendix D](#).

C.1 Number of Interruptions

Number of interruptions is the number of times that people other than the presenter spoke during the presentation.

Job market and departmental seminars: The number of interruptions was recorded by human coders using the Qualtrics tool developed by [Dupas et al. \(2021\)](#). Each time the presenter was interrupted, the coder registered it in the software.

NBER SI 2019: Same as for the job market and departmental seminars.

Online seminars: The number of interruptions is based on the number of times non-presenters spoke in the presentation recordings. We identify when there are changes between speakers on this large sample using the automatic, diarization procedure from [Seré \(2023\)](#). Then the dominant speaker is labeled as the presenter and the remaining speakers are labeled as audience. We drop extremely short interruptions and start counting interruptions once the presenter starts giving the presentation (after the introduction). We additionally identify moderators based on the non-presenter speaking in the beginning and drop seminars that are dominated by the moderator asking questions (representing reading questions from an online chat). The full analysis is in [Seré \(2023\)](#) and is qualitatively consistent with the results in this paper.

NBER SI 2022: The number of interruptions is based on the number of times non-presenter and non-discussants spoke in presentation recordings. We create time-stamps for when each person is speaking using Trint. Manually, we label if that is the presenter, discussant, or an audience member talking. We drop interruptions that are extremely short. We also do not include discussions in the room of the presentation after the formal talk is concluded (sometimes the recording continued after the presentation ended). Additionally, we drop introductions from the moderator at the beginning of the talk to introduce the speaker.

C.2 Tenor and Tone of Interruptions

Job market and departmental seminars: Tenor of interruptions for this sample was assessed by the human coders using the Qualtrics tool. Coders could leave an interruption unlabeled (in fact, this was the default option), or could be labeled as patronizing, disruptive, demeaning, hostile, or supportive. We consider the first four labels as indicative of “Negative” tenor, and the latter one as “Positive” tenor. The labels were assigned according to the coder’s perception of the interruption.

The specific written instructions to the coders were: “You should think of this as assessing the interaction towards the speaker. We aren’t asking you to code the intention of the person making the comment, nor how it was taken by the speaker. But rather, it’s your assessment of the tenor of the comment in a scientific setting. *You may leave this blank for many interjections, only filling this up if you think it is warranted.* You can click as many buttons as you deem appropriate, so, an interaction can be, e.g. just supportive, just patronizing or both. Or nothing. Use your judgment.”

NBER SI 2019: Tenor of interruptions for this sample was assessed by the human coders using the Qualtrics tool. The default was that the question was neutral (no label). Positive interruptions were interruptions labeled as particularly valuable, constructive or collegial according to the coder’s perception of the interruption. Due to the data-use agreement with NBER, we were not able to count negative-tenor interruptions. Coders received the same instructions as for the two other human-coded samples mentioned above.

Online seminars: Tone of interruptions is imputed using a deep neural network to predict tone from audio recordings. Negative and positive tone-of-voice here represent “how it sounds” to the machine compared to the tones of voice actors used to train the algorithm. Tone from the machine does not include “what it is said,” body language, or context from earlier in the seminar. Negative interruptions are those the machine predicted to be either negative-aggressive (angry or disgusted) or negative-passive (sad or fearful). Positive interruptions are those the machine predicted as positive (happy or positively surprised). We train separate tone-prediction models for male and female speakers due to average differences in voice pitch across gender. We require the deep neural network to assign over a 50 percent probability of a positive or negative tone, otherwise the interruption is labeled as neutral. A more detailed description of training and robustness is in [Appendix D](#).

NBER SI 2022: Same as for the online seminars.

C.3 Types of Interruptions

Job market and departmental seminars: Type of interruptions for this sample was assessed by the human coders using the Qualtrics tool. Coders had the option to record if an interruption had any of the following (non-exclusive) characteristics: clarification, criticism, suggestion, comment, or follow-up. A clarification is asking the presenter to clarify a previous point or a something shown on the slides. A criticism is a direct critique or challenge to the presenter’s work. A suggestion is suggesting that the presenter conducts additional analyses or modifies their analysis in a specific way. A comment is a broader category capturing a non-question interruption that does not neatly fall under criticism or suggestion. A follow-up is an interruption that picks up on a previous interruption by the same or another audience member.

NBER SI 2019: Same as for job market and departmental seminars.

Online seminars: There are two classification schemes for interruption types in this sample: first, whether an interruption is a question (interrogative) or not (declarative). Second, whether an interruption involves the audience talking over the presenter or not.

Interruptions are categorized as either interrogative or declarative depending on whether the transcript of the interruption includes a question mark (?). We feed the presentation transcript through standard text-cleaning software to ensure appropriate punctuation is included to match the words used. [Seré \(2023\)](#) includes two alternative specifications of this: one uses pattern-matching approach that scans the text of each interruption for specific phrases and structures commonly associated with questions, such as "do you think," "what is," or "how can." The other uses a text-prediction algorithm to take text as input and predict where it was a question or not using a BERT classifier. All methods return similar results.

Interruptions are also categorized by whether the audience talking over the presenter or not. Using the audio recording and diarization algorithm, we are able to see if audio wavelengths are best modeled as a univariate distribution (reflecting a single speaker), or as a multivariate distribution (reflecting the influence of multiple voices). Therefore, audience member speech (an interruption) and is best modeled with a multivariate distribution is labeled as an interruption where the audience member talking over the presenter.

NBER SI 2022: We were not allowed to save transcripts of the seminars due to our data-use agreement with NBER for this sample. Therefore we were not able to generate data on interruption types.

C.4 Speaker Characteristics

C.4.1 Speaker's Gender

Job market and departmental seminars: Speaker gender for the presenter, discussant, or audience member for this sample was recorded by the human coders using the Qualtrics tool. This was based on the coder's observation from the seminar and name of the presenter, discussant, or audience member.

NBER SI 2019: Same as for job market and departmental seminars.

Online seminars: Presenter, discussant, and audience gender is inferred from an audio processing algorithm that compares the similarity of the speaker's voice to those of the voices of male and female voice actors in a variety of datasets. [Appendix D](#) includes more detail on the algorithm structure and evaluation.

NBER SI 2022: Same as for the online seminars.

C.4.2 Presenter's Home Institution

Fixed effects for the presenter's home institution use the institution's type, rank, and/or country.

Job market and departmental seminars: Presenter home institution for this sample was recorded by the human coders using the Qualtrics tool. We have too many singletons to include institution-specific fixed effects for each presenter’s home departments so we instead group home institutions into coarser clusters by type and rank, yielding 5 institution-rank fixed effects for each of: Top 1-6 economics departments; Top 7-20 economics departments; Other U.S. academic institutions; Other institutional academic institutions; Non-academic institutions. University ranks for top 10 and 20 institutions is from the US News and World Report 2017 Rankings.

NBER SI 2019: Presenter home institution for this sample was observed by the human coders. Coders then referenced an excel sheet that listed the University ranks for top 10 and 20 institutions is from the US News and World Report 2017 Rankings. Coders then recorded whether the presenter’s home institution was a top 20 institution, a non-top 20 institution, or a non-academic institution using the Qualtrics tool.

Online seminars: The online seminars sample includes 7 fixed effects: Top 1-6 economics departments; Top 7-20 economics departments; Other U.S. academic institutions; Other EU academic institutions; Other UK academic institutions; Non-US/EU/UK institutions; and Central Banks, IMF, World Bank, etc. The additional fixed effects, compared to the human-collected sample, are due to the increased heterogeneity of this sample, which encompasses a broader set of presentations and conferences from a more diverse international context, not limited to the U.S.

NBER SI 2022: Information on individual presenters could not be collected for this sample, in agreement with NBER.

C.4.3 Presenter’s Citation Count

We collect information on citations using Google Scholar profiles when we were allowed to collect the presenter’s identity. For the NBER SI 2019 and NBER SI 2022 samples, our data-use agreements with NBER prevented us collecting the name of presenters and therefore we were not allowed to collect citations. When a new video was detected, it was downloaded and processed, and the presenter’s name—extracted from the video—was used to scrape their Google Scholar profile. This process was conducted in multiple waves between 2022 and 2023, so citation counts reflect the moment the video was downloaded. As a result, presenters who appeared more than once may have different citation counts across observations, reflecting values close to the time of each seminar.

Job market and departmental seminars: Citation counts were obtained from the Google Scholar profile of the presenter for the 2019 departmental seminars. For the job market talks, it was not meaningful to collect citation data, as presenters are typically completing their PhDs when entering the job market and few have any citations.

Online seminars: Citation counts were obtained from the Google Scholar profile of the presenter.

C.4.4 Presenter’s Rank and Seniority

Job market and departmental seminars: We do not collect seniority measures for presenters in job talks or departmental seminars. For job talks it was not meaningful to collect seniority data, as presenters are typically completing their PhDs and entering the job market.

NBER SI 2019: Presenter’s seniority (categorical) for this sample was observed by the human coders and recorded using the Qualtrics tool for three broad tiers per our data use agreement with the NBER. The three tiers were senior (tenured) faculty member, junior (untenured) faculty member, and other (student / post-doc / non-academic).

Online seminars: Presenter’s seniority (continuous) was computed as the difference between the year in which the presenter published their first paper (obtained from Google Scholar) and the year in which the seminar took place (2020 to 2023).

NBER SI 2022: Our data-use agreement with NBER for this sample prevented us collecting the name of presenters and therefore we were not able to look up presenters’ seniority.

C.5 Research Characteristics

C.5.1 Paper/Presentation Topic (JEL)

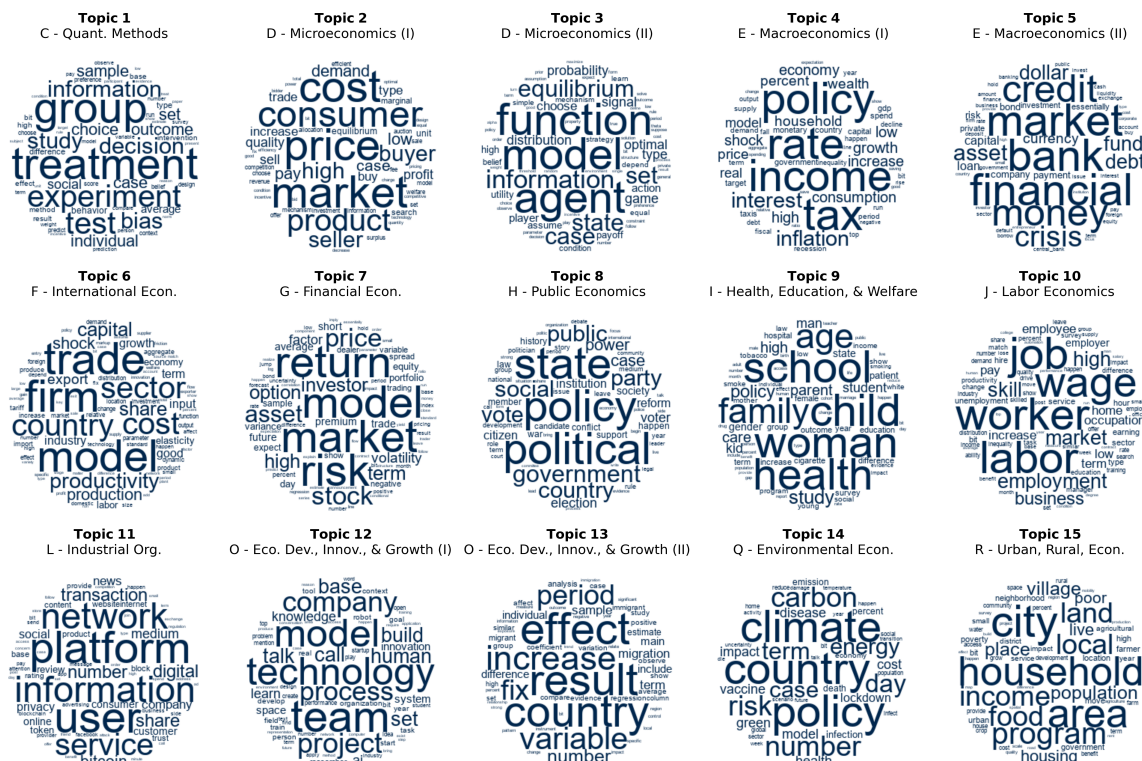
Job market and departmental seminars: JEL codes were typically drawn from the title page of the paper, and if multiple codes were listed we chose the most frequent, and in the event of a tie, chose the most relevant. If JEL codes were not listed, we chose the most relevant based on the content of the paper. We then constructed 7 fixed effects corresponding to single-digit JEL codes: C (math and quant methods), D (micro), E+F+G (macro, international, and financial), H+I (public, health, education and welfare), J (labor), L (industrial organization), Other codes.

NBER SI 2019: For this sample we did not collect paper topics since we were not allowed to record identifying information about the presenter, their paper, or the program.

Online seminars: Presentation topics are identified using a topic modeling approach applied to the audio transcripts of the seminars and are matched to JEL codes. For that, a pre-processed transcripts were analyzed as the corpus. A “topic” in this context is defined as a cluster of words that frequently appear together. The analysis was conducted using Mallet, a toolkit for topic modeling that includes sampling-based implementations of Latent Dirichlet Allocation ([McCallum, 2002](#)). The number of topics was selected using the coherence-score of models with varying number of topics, and the one with 15 topics achieved the highest coherence value before a plateau in the score occurs. The word clouds in [Figure C2](#) display the most probable words for each topic. We match the word clouds to the closest JEL code using a natural language processing tool. Then we group the 15 topics into five groups of JEL codes: macro, micro, finance, econometrics, and general.

NBER SI 2022: For this sample we did not collect paper topics, and due to our data-use agreement we were unable to match papers to recordings after our initial data processing had completed (which included deleting recordings).

Figure C2: Paper Topics in Online Seminars



C.5.2 Research Field

For our analysis in grouping seminar series by field, we create the broader research field categories of micro and macro. For some samples we also categorize into theory/econometrics and others.

Job market and departmental seminars: For presentations in regular departmental seminars, fields are based on seminar series. For example, all papers presented in the Harvard macroeconomics seminar are classified as macro. We grouped seminars into three broad categories: micro, macro, and theory/econometrics. For the job market talks, we coded field ex-post based on the paper's research topic (as described above).

NBER SI 2019: Human coders observed which program the paper was presented in and then used an excel spreadsheet listing the research field based on the manual labels of programs from [Chari and Goldsmith-Pinkham \(2017\)](#). We categorized fields into the broad categories of macro, micro, and finance, as per our data use agreement with NBER for this sample.

Online seminars: The research field for the seminar series is an aggregation from the paper topics into three categories: macro, micro, and others. This aggregation process was assisted by a natural language processing tool. Specifications run with series' field labeled by a research assistant return similar results as those from the topic model.

NBER SI 2022: Research field comes from manual labels of programs from [Chari and Goldsmith-Pinkham \(2017\)](#). Due to our data-use agreement for this sample, we categorized programs as either macro or micro and were unable to have a third category for finance or more granular fields. [Handlan and Sheng \(2023\)](#) include a table which shows the program assignment to different fields.

C.5.3 Paper/Presentation Types

We code four additional characteristics for paper/presentation topics using an LLM: (1) Theory: Is the primary purpose of this paper developing economic theory? (2) Data: Is the primary purpose of this paper analyzing empirical data? (3) Experimental: Does this paper involve a field or lab experiment? (4) Policy: Is this paper substantially engaged with analyzing economic policies? Details on this process are in the appendix [Section D.10](#).

Job market and departmental seminars: We use the entire paper text for the paper corresponding to the presentation for this sample with the LLM to code the the four type categories (theory/-data/experimental/policy).

NBER SI 2019: By agreement with the NBER for this sample, individual presentation titles (and corresponding papers) were not allowed to be coded.

Online seminars: We use the transcripts from the online seminar presentations with the LLM to code the the four type categories (theory/data/experimental/policy).

NBER SI 2022: As with the NBER SI 2019 sample, we were not allowed to code the presentation title due to our agreement.

C.6 Seminar Duration

All measures of seminar duration are measured in minutes. There are some decisions we made in selecting our samples that affect the average duration measures. For the NBER SI samples, we drop very short presentations, including the lightning rounds and egg timers which are under 15 minutes. For the NBER SI 2022 sample, we also drop presentations where the recording was interrupted due to technical issues. For the online seminar sample, we exclude seminars shorter than 20 minutes and longer than 120 minutes.

Job market and departmental seminars: Seminar duration was recorded by the human coder using the Qualtrics tool. It is measured as the actual number of minutes between the seminar's start time and end time based on when coders clicked the start and end buttons, rather than the scheduled time.

NBER SI 2019: Same as for job market and departmental seminars.

Online seminars: Seminar duration was measured from the presentation recordings. The start of the presentation starts when the main presenter begins speaking.

NBER SI 2022: Seminar duration was measured from the presentation recordings. The start of the presentation starts when the main presenter begins speaking and ends once the presentation, discussion, and formal questions conclude. We manually verified when the presentation ends and people break for small conversations, and do not include measures for after the end of the presentation.

C.7 Seminar series fixed effects

We use seminar series fixed effects to control for differences in seminar culture across series. For the NBER SI samples, we were unable to record the exact program a presentation took place in due to our data-use agreement. As a second best, we construct a proxy version of fixed effects using data permitted in our agreements.

Job market and departmental seminars: Seminar series was recorded by human coders using the Qualtrics tool. We subsequently cleaned the seminar names to remove typos and fix differences in capitalization. Recognizing that job market talks bring a different audience and norms, we code all job market talks within a department as one series (eg “Harvard job market seminars”) while other seminars are coded the usual way (eg “Harvard Labor Seminar”).

NBER SI 2019: By agreement with the NBER for this sample, individual program meetings (which correspond to conference tracks) were not allowed to be coded. We use a combination of the conference track field (e.g., micro, macro, and finance) (following [Chari and Goldsmith-Pinkham \(2017\)](#)) as well as the format-style to create a “proxy” seminar series fixed effect. Format style was recorded by the coders and included attributes such as whether the session was joint between multiple programs and whether it included a discussant, a Q&A session at the end, and a moratorium on questions at the beginning. Using both field and format, we then constructed a set of saturated “Field $\times$ Format” fixed effects which yields 15 categories (broad field $\times$ format combinations).

Online seminars: Seminar series were identified directly with the YouTube channel from which the recordings were sourced. However, in instances where this was not sufficient, natural language processing (NLP) techniques were applied to both the title and description of each recording within the same YouTube channel. For example, applying these techniques to the titles and descriptions of seminars from the CEPR channel, originally yielding 168 seminars, enabled their subsequent categorization into 12 distinct seminar series.

NBER SI 2022: Similar to above, individual program meetings were not coded due to a data-use agreement with NBER. We use a combination of the field and format-style to create a “proxy” seminar series fixed effect. Each program was coded into two broad fields: micro and macro (see

Handlan and Sheng (2023)). We also document the presentation format as whether there was a discussant and how long the discussant presented. While we do not directly observe whether a presentation has a Q&A session at the end and a moratorium on questions at the beginning, we use the distributional attributes of audience interruptions to proxy for Q&A and moratorium. We label seminars with more than 90 percent of the interruptions occurring in the last 25 percent of the total duration as having a Q&A session, and seminars with no interruptions in the first ten minutes as having a moratorium. We then constructed a set of saturated “Field $\times$ Format” fixed effects that is comparable to what we have for the NBER SI 2019 sample.

C.8 Human-Coder Characteristics

Job market and departmental seminars: There were 77 coders who collected our seminar samples, of whom 73 percent were female, 73 percent were in an applied micro field, and 36 percent were in the fourth year or higher in their Ph.D. program.

NBER SI 2019: There were 29 coders involved in this data collection effort (four of whom had participated in the earlier department seminar study). Roughly half (53 percent) of the NBER coders were female, most (83 percent) specialized in an applied micro field, and 31 percent were in their fourth year or higher in their doctoral program.

C.9 Attendance and Participation

Job market and departmental seminars: Attendance was recorded by human coders using the QQualtrics tool. Coders counted the number of people in attendance as well as the number of female and male attendees in the audience. Finally, they also record whether attendance was higher than normal.

NBER SI 2019: Same as for job market and departmental seminars.

Online seminars: Participation is measured using the number of unique audience members who speak in the presentation. From the recording, we do not know attendance but can proxy this by the diversity of speakers from the audience. Audience members are tracked throughout a presentation by algorithmically matching different interruptions together based on the similarity of speech patterns.

NBER SI 2022: We were unable to measure attendance in-person from the recordings of presentations in this sample because the camera only points at the presenter. We also did not track audience members throughout the presentation.

D Details of Machine Learning Approach

In this appendix section, we provide more context, details, and validations for the machine learning approaches of this paper.

Economic research applying machine learning approaches to analyzing audio data is both relatively new and expanding into new sub-fields. One set of studies analyzes the voices of macroeconomic policymakers, focusing on the tone of Federal Reserve officials in speeches (Alexopoulos et al., 2024; Bisbee et al., 2025; Gorodnichenko et al., 2023). Other studies focus on the workplace, and researchers have investigated how audio analysis of job interviews can aid recruiters in screening candidates (Liem et al., 2018; Naim et al., 2015; Nguyen et al., 2014; Teodorescu et al., 2022). In educational settings, audio recordings of teacher and student behavior have been used to assess classroom climate and learning James et al. (2018). And some have analyzed vocal cues in Supreme Court arguments (Knox and Lucas, 2021). Given that advances in computing power paired with machine learning methods have made audio-as-data accessible to the research community, one of the contributions of our paper is to provide some guidelines for economists in the appropriate use of these methods.

D.1 Data Sources

There are three data sources used for training and developing the model from the computer science literature: Mozilla Common Voice, RAVDESS, and CREMA-D.

Mozilla Common Voice (V4) is a dataset with short audio clips (3-5 seconds) of people reading standardized sentences with corresponding labels for whether the speaker is female, male, or other. We restrict our sample to clips that are agreed to be female or male speakers by multiple crowd-sourced ratings. Our starting sample is 507,184 audio clips that we downsample to get balanced data. There are many versions of the Common Voice dataset that grows over time. We use this data to train the gender classification algorithm.

RAVDESS is a database of audio clips (3-5 seconds) of 12 female and 12 male actors reading standardized sentences multiple times using different directed emotions (angry, disgust, fear, sad, neutral, calm, surprised, and happy) and intensity (normal or strong intensity). There are 1440 clips total. Speaker gender is self-identified and provided by speakers themselves. We group together emotions into four tones: angry and disgust make negative-aggressive, fear and sad make negative-passive, neutral and calm make neutral, and surprised and happy make positive. We use this data to train the tone-of-voice classification algorithm with CREMA-D.

CREMA-D is an audio database with short audio clips (3-5 seconds) with 91 diverse actors reading standardized sentences using different directed emotions (angry, disgust, fear, sad, neutral, and happy). There are 7,442 clips total, which we downsample to balance across tones. Gender is self-identified by the speaker, and tone is validated by crowd-sourced ratings. All speakers are speaking English in the recordings. We group together emotions into four tones: angry and disgust make negative-aggressive, fear and sad make negative-passive, neutral is neutral, and happy is

positive. From this dataset, we leverage audio clips and corresponding labels for speaker gender and tone-of-voice. We use this data to train the tone-of-voice classification algorithm with RAVDESS.

D.2 Data Processing

Each audio clip needs to be processed and converted from sound waves into numerical vectors that we can input into our deep learning model. Audio signals are complex—they comprise many frequencies, with each frequency registering at a different intensity, and these intensities change over time. As a result, we follow the literature on voice recognition (including [Gorodnichenko et al. \(2023\)](#), [Bai and Zhang \(2021\)](#), [Sell et al. \(2018\)](#)) and this section describes that process.

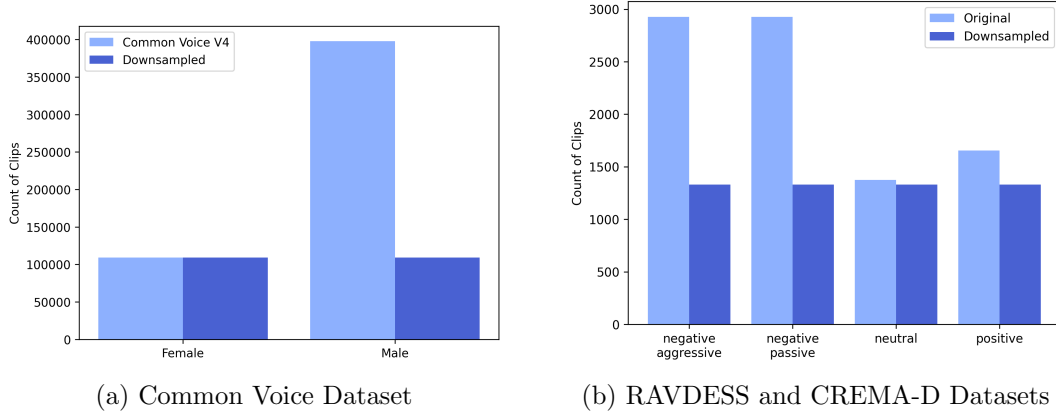
For each short audio clip (about 5 seconds in length), we use a python package called Librosa to extract three types of audio measures: Mel spectrogram, MFCC, and chroma scale values. We have a separate number to measure the intensity on different frequencies (pitches or notes) in the Mel, MFCC, and Chroma. These three types of scale capture slightly different qualities: (1) “Mel Spectro-gram Frequencies” capture the volume (in decibels) of each pitch within a clip; (2) “Chromatic Scale Frequencies” capture the musical scale (but not octave) of each clip; and (3) “Mel Frequency Cepstral Coefficients” capture the timbre of each clip—the elements of sound that make trumpet sound different from a human singing the same musical note. In total, our audio vector is 180 dimensions. The richness of this summary is necessary because gendered differences in speech patterns reflect a combination of frequencies and intensities over intervals of time—and relies on stable acoustic features such as pitch and timber—rather than a single feature.

On a technical note, we set the sampling rate to 16 kHz for all our data. This rate is appropriate for speech data, as it captures the relevant frequencies for human voice while minimizing unnecessary data. The CREMA-D dataset, which is used in our study, is saved at 16 kHz, so we adopt this as the standard rate for consistency. We also use mono-channel sampling.

D.3 Splitting Data into Training, Validation, and Testing Samples

To train our machine learning model to predict gender or tone labels from measures of audio, we need to first set up training, validation, and testing subsets of our data. The testing set is a leave-out set that is not used in developing the model in any way and serves as our out-of-sample accuracy benchmark. The remainder of our sample is split into different training and validation sets. The training subset is used to estimate the parameters of the network via an iterative process known as gradient descent which adjusts model parameters to minimize the distance between predicted gender attributions and the known gender of the speaker. The validation subset determines when to stop updating parameters to prevent overfitting. We employ an ensemble learning approach called “bagging” which trains multiple classification models based on different subsets of the data and then averages their predictions to assign a final label. We train five different models each with unique subsets of the data and stop training when no further improvement is observed on a validation subset. How we estimate multiple versions of our algorithm is explained below, but first we discuss creating the different training, validation, and testing sets in more detail.

Figure D1: Gender Balance Through Downsampling



To balance categories in our dataset, we apply down sampling. For the Common Voice data there are many more male observations than the 109,290 female observations. Therefore, we randomly select 109,290 male clips to create balance. This decreases the overall sample size by about half, but 218,580 is still sufficiently large for our analysis. For the combined RAVDESS and CREMA-D dataset, we have imbalance across the tones. Here we down sample the CREMA-D data and keep all the RAVDESS data such that each tone-of-voice category has 1330 observations (also balanced by gender) before data augmentation.

We then remove a random 10 percent of observations from the sample to make our testing or leave-out sample that we use to evaluate the out-of-sample accuracy of our models. For the Common Voice dataset, we require there to be an equal number of male and female clips in the testing sample. For the RAVDESS and CREMA-D datasets, we require the testing sample to have equal representation across the tones and to not split actors between training and testing sets. In these later datasets the same speaker supplies multiple clips. By restricting speakers to not span the training and testing sets, we are preventing artificially high accuracy coming from the algorithm learning how one speaker talks (called “leakage”).¹

The remaining 90 percent of our data is then used for training our model. We use an ensemble learning technique called “bagging” where our final prediction is actually the majority prediction across multiple, separately trained models. In our case, we train the same model architecture five different times using different subsamples data for training and validation. In short, training data is used to improve the fit of the model while validation data tells us when to stop iteratively updating parameters. First we will explain how the data is divided and then we will explain the training and validation producer more below.

This data is split into five different subsets (called folds) that are representative (balanced labels). As before, we do not allow speakers to be present in multiple folds. We then create one

¹Data leakage can make an algorithm seem more accurate and generalizable than it truly is. With data leakage, the researcher may conclude their algorithm performs very well out of sample, when in fact, the observations in the testing set are from the same actor and tone as that in the training set. The validation and testing sets should not be used directly to shift model weights.

Figure D2: Training, Validation, and Testing Samples for Gender and Tone

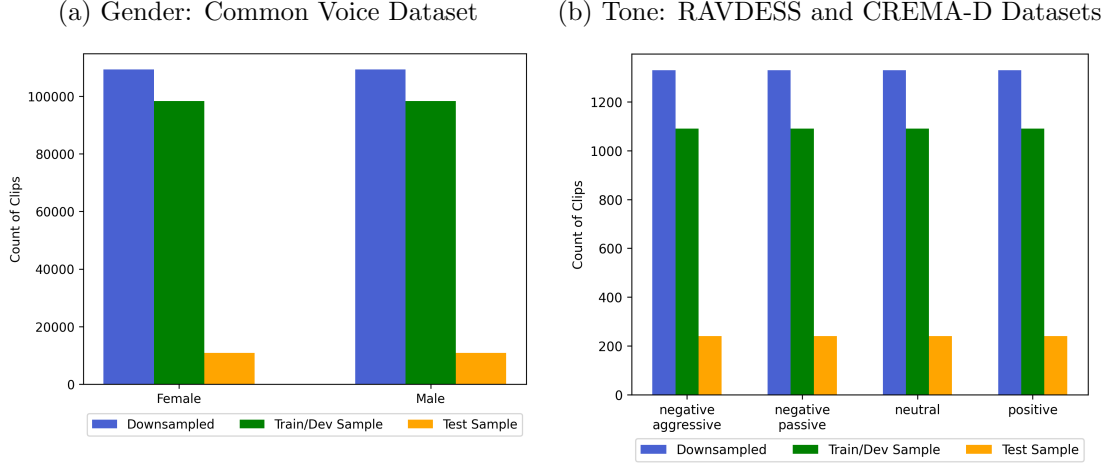
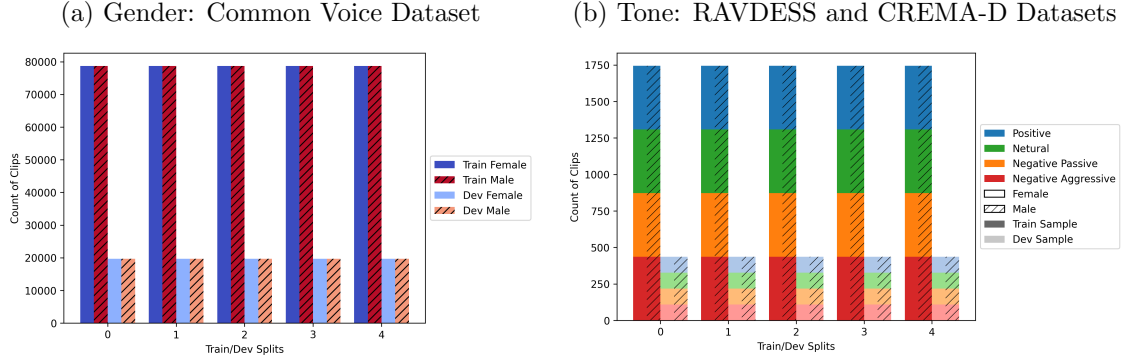


Figure D3: Five-Fold Cross-Validation Splits for Gender and Tone



“split” of the data by having one of the five folds be the validation/development sample and the remaining four folds act as the training sample. Below are graphs showing that in each of the five ways of splitting the data (having one fold be a validation/development set and the rest allocated to training) still has a balance in labels across the samples. We then train a separate model for each of these subsets, which we will later aggregate in our ensemble learning approach.

D.4 Data Augmentation

The RAVDESS and CREMA-D datasets alone provide a small sample to train a tone-classifier. Accordingly, we apply a technique called data augmentation, where we minimally perturb the audio data but assign it the same label as the original clip. This provides us with synthetic data observations we can use for training our models. We do not apply data augmentation for the gender classifier because we already have a large dataset with the Common Voice dataset.

This technique has its roots in image classification. For example, if you are trying to train a model to identify pictures of cats, you can rotate a photo of a cat by 90 degrees, and it should still be labeled as a cat. While this is obvious to humans, the machine benefits by seeing different

angles and rotations. For each rotation, color shift, or stretching of an image, the researcher can produce a “new” observation that helps the model learn how to classify images of cats. We apply the audio analysis version of augmentation where we slightly speed up, slow down, pitch up, pitch down, and add white noise to an audio clip. The underlying component of the audio clip remains the same, but the numerical measure of that audio clip varies slightly. This helps the algorithm by observing more data and becoming more robust to overfitting.

Specifically, we perform time-stretching where the speed of the audio is adjusted to 0.81 times (slow down) and 1.07 times (speed up) the original rate. Similarly, we apply pitch-shifting by raising the pitch by +1 semitone (pitch up) and lowering it by -3.5 semitones (pitch down). Gaussian white noise is also added to simulate real-world environmental noise. Moreover, we trim the first and last second of each audio file to remove non-speech sections while preserving the core speech content. These augmentations increase our sample size for training and development by *six* times. Each augmented clip is assigned to the same training and development set that the original clip belonged to. We ensure that the leave-out testing sample is not subjected to any data augmentation to maintain the integrity of testing performance.

D.5 Model Architecture

For this model we adapt the architecture from [Bhattacharya et al. \(2022\)](#). The authors are also working with tone-of-voice classification and are able to achieve a very accurate classification model. We adapt the model in four clear ways: first we use different audio feature inputs (totaling 180 features instead of their 193). Second, we are classifying audio into four tone categories rather than six emotion categories. Third, we use audio of only English speakers and their project focuses on comparing tones across languages by working with English, German, and Italian recordings. Fourth, we require strict sample-splitting procedures to prevent data leakage. We apply the same architecture for our tone model for gender classification and simply change the number of classes from four to two. [Table D1](#) details the hyperparameters we used for training the audio-classification models.

Table D1: Hyperparamters

Parameter	Value
Loss Criterion	Cross Entropy Loss
Optimizer	Adam
Learning Rate	1e-4
Decay Rate	1e-6
Momentum	0.1
Random Seed	2024
Training Epoch	200
Patience	20
Batch	64

D.6 Model Training

For both gender and tone classification tasks, we trained separate sets of models using five-fold cross-validation to ensure robust performance and reduce overfitting. Each model’s training process was carried out with a consistent set of hyperparameters and the same overall architecture, allowing us to maintain comparability across models and leverage ensemble predictions for greater accuracy.

Data Loading and Preprocessing Pipelines: Each dataset was preprocessed through a standardized pipeline to ensure consistency across samples. The data preparation routine includes balancing the data by downsampling categories and filtering. For instance, subsets were created to train models on gender-specific samples or to control for emotion intensity. Additionally, for tone classification, we maintained balance across the categories by applying downsampling and using multiple datasets in tandem to ensure representation across tones and genders.

Feature Standardization and Scaling: To ensure uniform input distributions and facilitate model convergence, each audio feature vector was normalized. This step adjusts all features to have a mean of zero and a standard deviation of one, which is essential for models with deep architectures. Normalization minimizes the risk of gradient vanishing or exploding during training, especially for neural networks that are sensitive to unscaled inputs.

Augmentation-Specific Filters and Conditions: For tone classification, data augmentation was selectively applied to increase robustness without compromising the dataset’s original structure. As mentioned before, augmentations included time-stretching (speed adjustments to 0.81 and 1.07 times the original rate), pitch-shifting (+1 and -3.5 semitones), and adding Gaussian white noise to simulate real-world environmental conditions. Each augmented clip retained its original label and was assigned to the same fold as its non-augmented counterpart to prevent cross-fold leakage. Additionally, we filter samples to include only original, non-augmented clips in some models, allowing control over augmentation intensity per task.

Specific Subset Selection for Training Samples: Subsets were created where clips were restricted to actors with ID numbers below a set threshold, allowing control over sample composition. Similarly, we filtered out specific emotion types and intensity levels, such as strong or surprised tones, depending on the experimental conditions. This approach helped maintain a high level of control over sample characteristics, balancing variation in the data without introducing artificial biases across model versions.

Cross-Validation and Ensemble Prediction: We applied five-fold cross-validation to ensure balanced representation across labels and mitigate overfitting. Each fold served as a validation set once, while the remaining four folds were used for training. To prevent data leakage, no individual speaker appeared in both training and validation sets for any fold. For each task, an ensemble approach was used: we trained five models with different cross-validation splits, and the final prediction for each sample was determined by a majority vote across these models, enhancing prediction reliability and minimizing potential biases.

Batch Training and Hyperparameters: Each model was trained in batches of 64 samples using a learning rate of 1×10^{-4} , weight decay of 1×10^{-6} , and a momentum of 0.9. Training

continued for up to 200 epochs, with early stopping applied after 20 epochs of non-improvement in validation loss. This setup, consistent across both gender and tone classification, provided a balance between model flexibility and convergence stability.

Library and Environment Specifications: The training and analysis were implemented in Python using torch for deep learning, pandas for data manipulation, and sklearn for scaling and performance metrics.

Evaluation Metrics and Performance Logging: During training, we monitored validation performance using metrics such as accuracy, precision, recall, and F1 score, providing a comprehensive assessment of model robustness. These metrics helped optimize model selection and confirmed performance consistency across folds. All evaluation results were logged per epoch to track model behavior over time and identify optimal stopping points.

By including these steps and conditions, we ensure that the model training approach is fully documented and reproducible. Each step reflects a commitment to rigorously maintaining dataset balance, preventing data leakage, and achieving model robustness suitable for both gender and tone classification tasks.

D.7 Model Performance

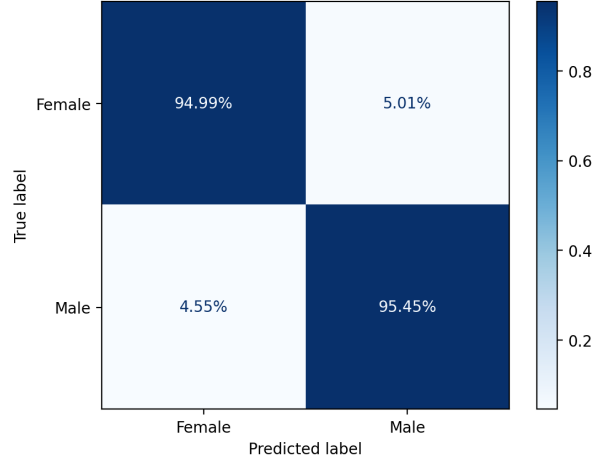
In this section we report the out-of-sample evaluation of our trained machine learning models. Recall, we set aside a sample of voice actor recordings that are not used in estimating parameters or determining when to stop training the model. This set is the leave-out sample. Below, we describe our evaluations of the gender-classifier and the tone-classifier models. The next section includes additional validation exercises with alternative methods, including direct human validation.

The gender classification model is trained to sort audio recordings into male or female speakers. The training data is constructed by volunteers who identify as male or female, described earlier in the appendix. We find a very high accuracy of 95 percent, indicating a reliable model for labeling male and female speakers. [Figure D4](#) is the confusion matrix for this algorithm on the leave-out sample. It shows that 94.99 percent of female recordings are correctly labeled as being female speakers, and 95.45 percent of male recordings are predicted to be male.

The tone classification model is trained to categorize audio into four distinct tones: Negative Aggressive (anger and disgust), Negative Passive (sadness and fear), Neutral (neutral and calm), and Positive (happiness and surprise). Three versions of the ensemble model were trained, each using a different subset of the data: one trained exclusively on female speakers, one on male speakers, and a combined model using both female and male data. Performance was evaluated on a hold-out sample to measure out-of-sample accuracy. The model trained on female-only data achieved an accuracy of 62.5 percent, while the model trained on male-only data achieved 53.3 percent. The combined model, which included data from both genders, achieved an accuracy of 56.3 percent, indicating some performance variation based on the gender-specific data used for training. The gender-specific models are used in the main specifications of the paper.

Now let us discuss the accuracy metrics by tone groups we use in our analysis: negative, positive,

Figure D4: Confusion Matrix for Gender Classification (Leave-out Sample)



Notes: The true label is the label provided by the voice actor, and the predicted label comes from the trained algorithm applied to this leave-out sample that was not used in training. Values shows accuracies within rows (e.g. to top row includes the percent of female clips predicted as female or male).

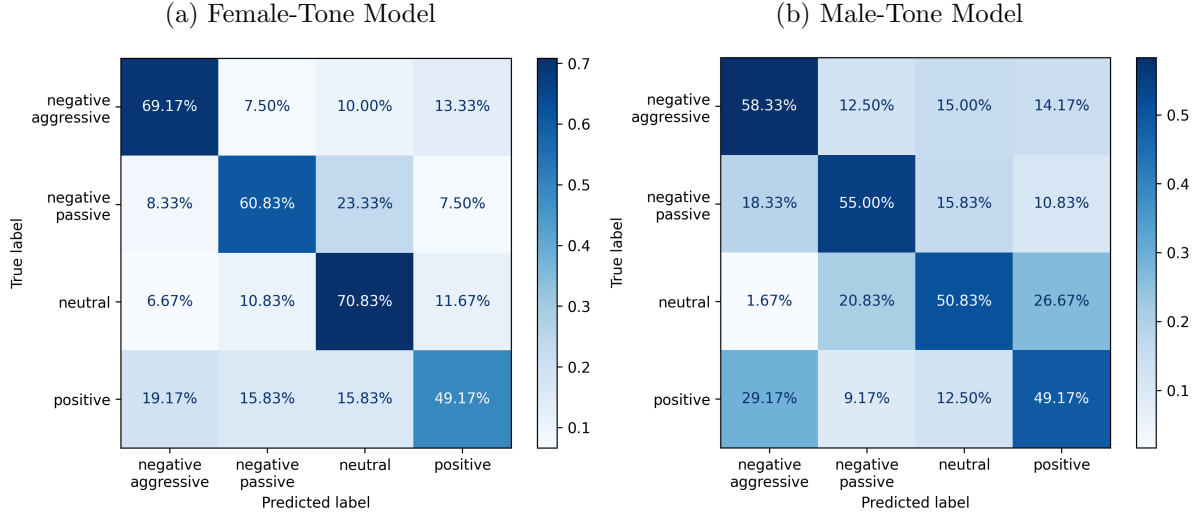
and neutral. In the paper we group the negative tone categories together so Negative-Aggressive and Negative-Passive are jointly called Negative. Of the Negative observations in the testing sample, the female model correctly predicts Negative 72.9 percent of the time and the male model does so 72.1 percent. Because we add the two negative categories, the random benchmark for Negative would be 50 percent. For Positive labels the random benchmark is still 25 percent. Both the male and female models correctly identify Positive audio clips 49.2 percent of the time. Again, these accuracy benchmarks are much better than random and perform similarly to our human validation set as described in the next section.

Figure D5 below displays the confusion matrices for the tone classification model trained on separate gender-based datasets. The first figure shows the performance of the model trained exclusively on female data, while the second figure corresponds to the model trained on male data. Each confusion matrix provides insights into the model’s classification accuracy across four tone categories: Negative Aggressive, Negative Passive, Neutral, and Positive.

In both figures, the diagonal entries represent the proportion of correctly classified samples. The off-diagonal entries capture misclassifications, where a tone category was incorrectly predicted as another. For instance, the model trained on male data some confusion between Negative Aggressive and Positive tones (mostly driven by actors talking very loudly).

These confusion matrices highlight that while both gender-specific tone models achieve reasonable performance, there are variations in classification accuracy across tone categories, likely due to inherent differences in the tone expression between genders. However, for negative and positive categories we use in the paper, we find similar accuracies across speaker gender. The matrices provide a useful diagnostic for identifying which tones are more challenging to classify accurately.

Figure D5: Confusion Matrices for Tone Classification (Leave-out Sample)



Notes: The true label is the intended tone from the voice actor, and the predicted label comes from the trained algorithm. Each matrix shows the model trained with data from a one gender, and tested on the leave-out sample for that same gender. The leave-out samples were not used in training the models. Values shows accuracies within rows (e.g. to top row includes the percent of negative-aggressive clips predicted as the four different tones).

D.8 Additional Validation with Alternative Methods

In addition to evaluating the machine learning models with their accuracy on leave-out samples, we have performed substantial robustness checks with human coders and alternative coding specifications on both the data from voice actors and the economics seminar data. Overall, we find that the algorithm does as well as human coders in classifying gender and does reasonably well in classifying tone.

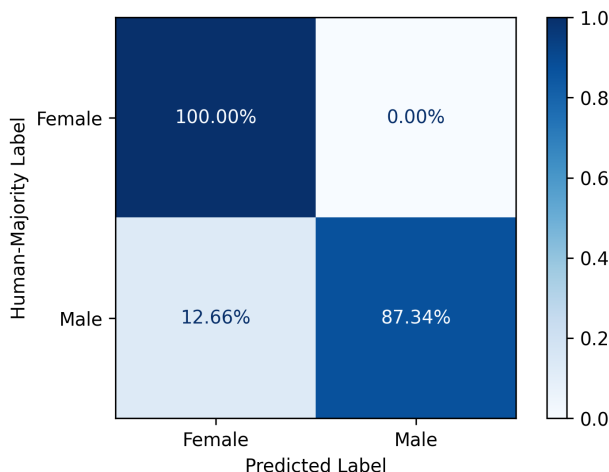
D.8.1 Human Validation

We performed three different validation exercises for the gender and tone classification with a team of human coders where they listened to samples of online economic seminars and labeled voice actor recordings. For each exercise, we have between four to six humans providing labels for audio data that we use to benchmark our algorithm. A common metric we report is the “human-majority label” which requires that a majority of the human coders agree on either a gender or tone label. Additionally, the range of accuracies of individual human coders provides an informative benchmark for the difficulty of our classification tasks, and we can compare our algorithm metrics to that range.

In the first task, we compare human-coded gender labels with the computer-coded gender labels for 140 recordings from the online seminar sample. [Figure D6](#) summarizes the results. For this task we had five coders. The recording samples were each 5 seconds long and included a mix of male and female speakers. For all but one recording, there was a clear majority label. Our gender-classification neural network perfectly predicts 100 percent of the recordings that human-coded as

female. For speakers labeled as male, our algorithm also predicts male 87.3 percent of the time. The extremely high accuracy relative to the actor data presented in [Figure D4](#), along with additional validations in the next section using names, suggests that the gender classification algorithm is highly accurate.

Figure D6: Confusion Matrix for Gender Classification (Human Validation)



Notes: The online seminar subsample for human validation includes 140 short recordings of a mix of presenters and audience members. The human-majority label is the label agreed on by at least three of the five human coders. The human-majority labeled 60 clips as female and 79 clips as male. One clip had no majority label, and the algorithm predicted that to be male. Values show the agreement percentages within rows (e.g. the top row shows the share of clips predicted to be female and male among the clips labeled as female by human coders).

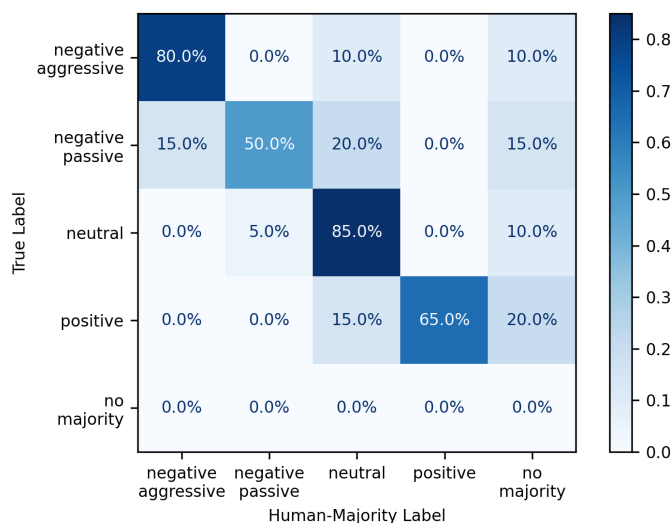
For the second and third tasks, we asked our human-coders to classify the tone (negative-aggressive, negative-passive, neutral, and positive) for a sample of voice actor recordings and a sample of the online seminar recordings. When training the human-coders how to label “like the machine does,” we presented them with a total of 12 training examples from voice actors across the four tone groups and two genders. Importantly, we gave them the following instructions: “Classify tone based on speech patterns and sounds, not on the words that are being said. Again, we are classifying based on what someone sounds like, not what they are saying.”

Classifying tone is a more difficult task than classifying gender, due to both the increase in categories and the nuances of tone itself. Accordingly, the first exercise is to evaluate how well the human coders do in labeling the voice actor data. [Figure D7](#) shows the confusion matrix comparing the human-majority tone label to the actor’s intended tone for a sample of 80 recordings. The accuracies in the different categories range from 50 percent to 85 percent, and the overall accuracy is 70 percent. We can see the fifth column represents cases when the human coders dramatically disagreed and highlights the challenges of classifying tone. For individual coders, the range of their individual total accuracy scores was between 58.8 percent and 75 percent accuracy. When splitting the sample into female and male actors, we find similar accuracy ranges.

The human coders were able to predict the actors’ tone with an accuracy that was on par with

the machine-predictions on the leave-out sample, albeit slightly higher (reported in Figure D5). There is only about a 10 percentage point difference between the machine and the human coders (57.9 percent compared to 70 percent), where the machine’s accuracy out of sample is on par with the lowest accuracies from the individual coders (58.8 percent). Our interpretation is that the machine is labeling tone about as well as some humans.

Figure D7: Confusion Matrices for Tone Classification (Human Benchmark)



Notes: This sample includes 80 recordings of voice actors balanced by gender and tone. The true label is the intended tone from the voice actor. The human-majority label is the label agreed on by at least three of the four human coders. There were many observations where there was no majority. Values shows accuracies within rows (e.g. to top row includes the percent of negative-aggressive clips predicted as the four different tones).

We take this a step further and ask our human coders to also listen to a set of 160 recordings from online seminars. Of those, 80 observations are short, five-second clips like the inputs to the algorithm and 80 observations are longer audio recordings (at the utterance level where we choose the tone based on average tone probabilities over the five-second intervals of the utterance). As an internal check, a random number of clips did not include speech to ensure the coders correctly labeled them as non speech (in the end there were 20 and all were correctly identified).

We look to the remaining 140 recordings to compare between human labels and the machine predictions. At least one human coder agrees with the machine prediction for 55.5 percent of this sample (52.5 percent for the 51 five-second clips, and 58.8 percent for the 59 longer utterance recordings). Human-majority labels only exist for 78.6 percent of the 140 recordings of economists (meaning there is substantial heterogeneity in opinions from the human coders on how to label tones). The machine algorithm agrees with the human-majority 30 percent of the time on the 110 clips where a human-majority exists.

Given the human-coders reported challenges in analyzing tone like the machine (based on their accuracy with the actor data and comments from the coders), it is difficult to attribute the disagreement to something with the algorithm, the small sample, or how humans understand tone.

We interpret this as a disagreement between the machine and the human coders on the intensity thresholds between tones. The human coders are much more likely to code “neutral” tones, especially on the economic seminar sample. The human-majority labels 44 percent of recordings as “neutral” tone, comparatively the algorithm labels “neutral” to 27 percent. This pattern also shows up in [Figure D7](#), but is less pronounced.

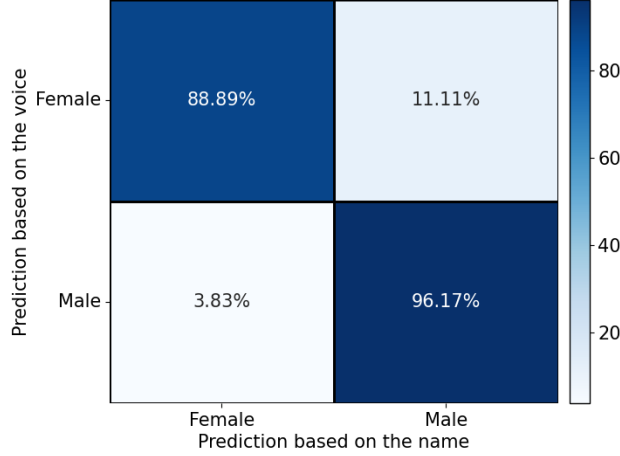
Finally, we note that this additional human validation of tone was not possible for the NBER SI 2022 sample because we were not able to save and share presentation recordings due to our data-use agreement with NBER. Nevertheless, we see the validation out-of-sample on the online seminar data encouraging about the accuracies of the models compared to how humans hear gender and tone.

D.8.2 Gender Classification Using Names

Besides using audio data to classify speakers’ gender, another popular method is to use the speakers’ names. We use a large dictionary, called Genderize, that maps names to genders. This mapping is probability-weighted to account for names that are used for both men and women. We use this on the names of speakers in the online seminar sample directly. For the NBER SI 2022 sample, we were not allowed to record presenter names. Accordingly, we use the NBER SI 2022 program to create bounds for the number of female presenters using female authorship.

The confusion matrix in [Figure D8](#) compares the name-based gender prediction to the audio-based prediction that is the main method of the paper. There is a high degree of agreement between the approaches. Because is a risk of error with the name approach - for example, the dictionary associates “Ariel” as a common female name but it can also be a male name - we do not expect a perfect match even for a perfect algorithm. However, the large amount of agreement across our robustness approaches gives us preponderance of the evidence that the algorithm does well in capturing female and male speaking patterns.

Figure D8: Gender Prediction with Voice vs. Name (Online Seminars)



Notes: This is based on the online seminar sample where we have data on presenter name. There is large agreement between the two approaches. Values show the agreement percentages within rows (e.g. the top row shows the percent of presenters predicted to be female who have names that are labeled as female or male).

We also perform a name-gender classification verification for the NBER SI 2022 sample, but we are limited to a looser validation using the publicly available NBER SI programs (which include participants' names). We are able to do a direct comparison with discussants (where we find 30.9 percent of discussants are classified as female by name, and 31 percent when classified by voice). For presenters, we have to do an indirect comparison using author composition .

We create two bounds to benchmark the number of female presenters we impute for the NBER SI 2022. The upper bound is the share of papers with at least one female author, representing the extreme case where the female coauthor for every paper with at least one female author was presenting. The lower bound is the share of papers with only female authors, representing the other extreme, where female coauthors don't present when there are other male coauthors.

Using author names from the NBER SI 2022 programs, we apply the name-gender classifier to measure the upper and lower bounds for female presenter share. 9.6 percent of papers have only female authors and therefore must have a female presenter. 51.5 percent of papers have at least one female author and therefore *could* have a female presenter. This is a fairly wide interval. Nevertheless, our finding that 33.8 percent of presenters at NBER SI 2022 were female is safely in the middle of that interval and therefore seems reasonable.

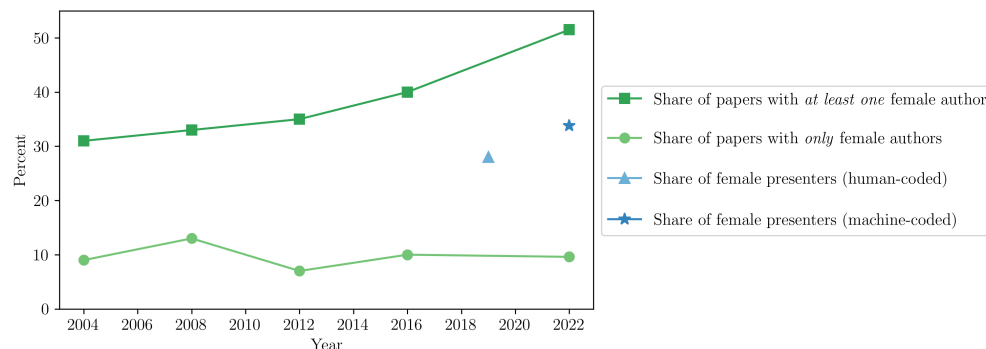
For additional context, we plot additional years of information for the NBER SI author and presenter female shares in [Figure D9](#). We plot the share of female presenters in the NBER SI 2022 sample compared to the NBER SI 2019 sample. For the upper and lower bounds using female authorship, we plot data points prior to 2022 directly from tables in [Chari and Goldsmith-Pinkham \(2017\)](#) and we extend those measures to 2022.

Ultimately, this plot shows that the gender-classification we have for NBER SI 2022 produces a reasonable share of female presenters. Both the numbers from 2022 and the human-coded share

from 2019 are around 30 percent. Additionally, it is within the upper and lower bounds we estimate from female authorship.

Finally, this figure also shows a continued increase in the share of papers at NBER SI that have a female author. This new result is an extension to one table of [Chari and Goldsmith-Pinkham \(2017\)](#) and shows encouraging progress towards the profession’s goal of inclusivity at top conferences, like the NBER SI.

Figure D9: NBER Summer Institute Shares of Female Authors and Presenters



Notes: The share of papers with at least one female author and with only female authors (green squares and circles) is from [Chari and Goldsmith-Pinkham \(2017\)](#) for 2004, 2008, 2012, and 2016 and is based on the NBER SI programs. We extend this to NBER SI 2022 using the online posted programs and the name-gender classifier. The share of female presenters for NBER SI 2019 (blue triangle) comes from the human-coders in the audience. The share of female presenters for NBER SI 2022 (blue star) comes from the gender-prediction algorithm and presenters’ voices.

D.9 Additional Details on Computer-Coded Samples

In this section we provide additional detail for the selection of seminars in the online seminar and NBER SI 2022 samples.

D.9.1 Online Seminars

To be included in the sample, each seminar had to be part of a seminar series organized or sponsored by an economics department based in either the United States or Europe. Seminars sponsored by highly regarded research institutions such as the American Economic Association (AEA) and the Centre for Economic Policy Research (CEPR) were also included.² Once a set of YouTube channels hosting these seminars was identified, all available seminars were downloaded. This process was carried out periodically between early 2022 and early 2023, ensuring that the complete archive of each channel was collected, resulting in a total of 2,398 online seminars. Several filtering steps were subsequently applied to obtain the final dataset.

Although all seminars took place during the COVID-19 pandemic, a random manual inspection—particularly of the final seminars in each series—was conducted to ensure that no hybrid

²For example, the American Economic Association provides an extensive list of online seminars at <https://www.aeaweb.org/resources/online-seminars>.

seminars were included in the sample.

To focus on standard academic presentations in economics, seminars shorter than 20 minutes or longer than 110 minutes were excluded. The presenter of each seminar was defined as the individual with the longest total speaking time. If a second speaker had more than 15 minutes of speaking time, the seminar was excluded to ensure that only events with a clearly defined presenter were retained. The moderator was identified using a three-step heuristic based on typical seminar structure. First, if the first speaker was not the presenter, that speaker was designated as the moderator. Second, if no moderator was identified using this rule and the seminar had only two participants, the non-presenter was assigned as the moderator. These two rules allowed identification of a moderator in 90 percent of the seminars. For the remaining cases, the last non-presenter speaker was designated as the moderator, under the assumption that moderators often both introduce and conclude the seminar. A seminar was classified as moderator-dominated if the moderator accounted for at least 51 percent of the interruptions; such seminars were excluded from the analysis as they are unlikely to reflect the typical interaction dynamics of economics seminars.

Since the analysis includes presenter-related variables such as citation counts, only seminars where the presenter’s Google Scholar profile could be reliably identified were retained. As explained in the main text, this was done using a combination of NLP and web scraping methods, followed by manual verification to ensure the matched profile corresponded to the actual presenter, avoiding potential name confusions. In addition, two seminar series—originally included through the automated procedure—were discarded as they mostly featured computer science presentations. While including them did not alter the main results, they were removed to ensure disciplinary consistency. The final dataset consists of 994 seminars. This includes 30 seminars where the speakers were identified as belonging to other fields (mostly computer science or biostatistics) but were retained because they presented in economics seminars. As these speakers tend to exhibit different characteristics (e.g., citation counts five times higher than the sample average), a control variable was added to indicate their presence.

D.9.2 NBER SI 2022

To focus on standard presentations in economics, we drop the short presentations (the egg timers and lightning rounds) and panels. We also do not include measures from programs where participants opt-ed out of the study. Finally, we also do not include analysis of audience speech for samples with poor audio quality. This produces a sample of 257 presentations.

The presentations at NBER SI 2022 were in hybrid format and required audience members and the presenter to speak into microphones. Therefore, we run a pass over the sample to select talks where the audio quality of audience members is sufficiently good in terms of timing and volume. We validate the timing manually and we identify a volume threshold, under which we say the audience audio is insufficient. We use the root mean square of the audio waveform as the volume threshold, because it directly reflects how well the audio was captured by the microphone, and it is mathematically related to many other standard sound measures, such as peak-to-peak amplitude

and intensity.

D.10 Assignment of Research Types with LLMs

Standard JEL codes describe the substantive field within which each research paper lies. In addition, we seek to code a second dimension, coding the research style or method. This matters because, for instance, a public finance developing theory about the asymmetric information between a taxpayer and the government may elicit a different audience with different engagement than a public finance paper analyzing empirical results from a field experiment. As such, we create the following four variables for each presentation.

1. Theory: Is the primary purpose of this paper presented developing economic theory?
2. Empirical: Is the primary purpose of this paper analyzing empirical data?
3. Experiment: Is this paper focused on a field or lab experiment?
4. Policy analysis: Is the primary purpose of this paper evaluating a specific program or policy, studying a specific law or regulation change, or simulating policy responses.

These groups are not mutually exclusive, and it is not unusual for a paper to engage in several of these dimensions at once.

We use a large language model to code these variables. For the seminar samples, we feed it the underlying research papers, and for the online seminars, we feed it a transcript of the entire seminar. (We cannot create these variables for the NBER samples where we were not permitted to code identifying information about the paper being presented.)

D.10.1 LLM Details

Here we provide the details of the LLM approach, including model, parameters, and prompts. We use OpenAI’s GPT-4o model, and feed the data through the Assistants API. The Assistants API provides a straightforward form of “Retrieval Augmented Generation” in which the LLM has access not only to the instructions (below), but also the uploaded file when making evaluations. Each paper is uploaded and evaluated independently (“zero shot” learning), so that the evaluation of no one paper affects the interpretation of another. As is standard, we use zero temperature to minimize the variation in GPT responses. And because we use the API, none of the data is reincorporated into GPT’s training. Specifically, for each paper or transcript, we:

1. Upload the transcript or the paper (using a readable rather than scanned pdf) to Open AI using its File API.
2. Set the instructions (listed below) for this project-specific Assistant, and give the Assistant access to that specific paper (or transcript).

3. Ask the Assistant to “Please read the attached economics research paper carefully, and assess its characteristics using the rubric I’ve given you.” This prompt also includes a request for structured outputs so that the output is machine-readable.
4. Extract the variable values and validation explanations

The instructions that guide the LLM Assistant are as follows:

Context: You are a professor of economics at a leading economics department, and you study the economics research process. As part of your research, you read recent papers closely, and categorize them according to a well defined rubric. You are careful and thorough as you read, and thoughtful about how you categorize each new research paper, considering them within the context of the sort of research papers you would read in top-tier economics journals, like the AER, QJE, JPE, ReStud, or Econometrica.

Task: Please read the attached research paper carefully, paying attention not only to what it analyzes, but how the analysis is conducted. I will guide you through a series of questions, asking you to describe in a structured way the attributes of that research paper. Be prepared to explain your reason for each answer.

Instructions: First, collect the following data about each paper:

1. Title
2. Author(s)
3. Abstract or whole paper: Was this file the entire research paper, or just an abstract?
4. Number of pages
5. JEL codes: The JEL codes should be listed on the first page, but if you can’t find any, create your own based on your reading of the paper (be sure to put an asterisk next to any you create).
6. Keywords: The keywords should be listed on the first page, but if you can’t find any, create your own based on your reading of the paper (be sure to put an asterisk next to any you create).

Second, and most importantly, please answer the following questions. In each case, provide a simple yes or no answer, and an indication of your confidence in your assessment (on a scale of 1= "very unsure", 2= "quite unsure", 3= "somewhat unsure" 4 "neutral" 5="somewhat confident" 6="quite confident" 7="very confident"). Provide a few relevant bullet points explaining the reasons for your assessment, as well a few relevant bullet points providing the most important reasons why a colleague might disagree with or doubt your assessment. These questions are not mutually exclusive:

1. Is the primary purpose of this paper developing economic theory?
2. Is the primary purpose of this paper analyzing empirical data?
3. Is this paper focused on a field or lab experiment? (This category does not include natural experiments.)
4. Is the primary purpose of this paper evaluating a specific program or policy, studying a specific law or regulation change, or simulating policy responses in ways that

someone might cite as evidence in support or opposition to a policy. (Please look for more than peripheral references to policy implications.)

Our evaluation of online seminars uses the transcript of the presenter’s speech, excluding interruptions, rather than the underlying research paper as input.. As such, we tweak these instructions, so that we tell the LLM that it is not reading “the attached research paper”, but rather instruct the LLM that it is being provided with “the transcripts of research seminars” and asking about “this paper” we asked about “the paper being presented.”

D.10.2 LLM Evaluation

To evaluate the accuracy of our labels, we created a human-labeled validation set. We randomly selected 75 papers for the validation set: 20 from the job market sample, 15 from the regular seminar sample, and 40 from the online seminar sample. Each paper is independently labeled by two of the authors. The pair resolved any disagreements and submitted consensus labels. A separate author developed the LLM prompt and was blinded to these labels. [Table D2](#) summarizes the LLM accuracy for each label. Overall, the pooled evaluation shows very high accuracy for the LLM, ranging from 82.7 percent to 97.3 percent agreement. The precision for policy-labels implies the LLM was more likely to label a paper/presentation as policy compared to the human-labels – likely driven by papers discussing policy implications.

Table D2: LLM Accuracy

Sample	Metric	Theory	Empirical	Experimental	Policy	Sample Size
Job Talk	Accuracy	0.750	0.900	0.950	0.800	20
	Precision	1.000	0.944	0.667	0.692	
	Recall	0.667	0.944	1.000	1.000	
Regular Seminar	Accuracy	0.800	0.933	1.000	0.867	15
	Precision	1.000	1.000	1.000	0.778	
	Recall	0.700	0.923	1.000	1.000	
Online Seminars	Accuracy	0.875	0.925	0.975	0.775	40
	Precision	0.917	0.960	0.833	0.533	
	Recall	0.880	0.923	1.000	0.800	
Pooled	Accuracy	0.827	0.920	0.973	0.800	75
	Precision	0.951	0.964	0.846	0.649	
	Recall	0.780	0.930	1.000	0.923	

Notes: Table shows accuracy of the LLM predictions for a paper’s type. Ground-truth labels were manually created by the authors. Accuracy is defined as the number correctly predicted, divided by the total. Precision is the number of true positives, divided by the number of true and false positives. Recall is the number of true positives, divided by the number of true positives and false negatives.

References

- Alexopoulos, M., Han, X., Kryvtsov, O., and Zhang, X. (2024). More than words: Fed chairs’ communication during congressional testimonies. *Journal of Monetary Economics*, 142:103515.
- Bai, Z. and Zhang, X.-L. (2021). Speaker recognition based on deep learning: An overview. *Neural Networks*, 140:65–99.
- Bhattacharya, S., Borah, S., Mishra, B. K., and Mondal, A. (2022). Emotion detection from multilingual audio using deep analysis. *Multimedia Tools and Applications*, 81(28):41309–41338.
- Bisbee, J., Fraccaroli, N., and Kern, A. (2025). Yellin’ at yellin: Hostile sexism in the federal reserve congressional hearings. *The Journal of Politics*.
- Chari, A. and Goldsmith-Pinkham, P. (2017). Gender representation in economics across topics and time: Evidence from the nber summer institute. Working Paper 23953, National Bureau of Economic Research.
- Dupas, P., Modestino, A. S., Niederle, M., Wolfers, J., and Collective, T. S. D. (2021). Gender and the dynamics of economics seminars. Working Paper 28494, National Bureau of Economic Research.
- Gorodnichenko, Y., Pham, T., and Talavera, O. (2023). The voice of monetary policy. *American Economic Review*, 113(2):548–84.
- Handlan, A. and Sheng, H. (2023). Gender and tone in recorded economics presentations: Audio analysis with machine learning. *Unpublished*.
- James, A., Kashyap, M., Victoria Chua, Y. H., Maszczyk, T., Núñez, A. M., Bull, R., and Dauwels, J. (2018). Inferring the Climate in Classrooms from Audio and Video Recordings: A Machine Learning Approach. In *2018 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, pages 983–988. ISSN: 2470-6698.
- Knox, D. and Lucas, C. (2021). A dynamic model of speech for the social sciences. *American Political Science Review*, 115(2):649–666.
- Liem, C. C., Langer, M., Demetriou, A., Hiemstra, A. M., Sukma Wicaksana, A., Born, M. P., and König, C. J. (2018). Psychology meets machine learning: Interdisciplinary perspectives on algorithmic job candidate screening. In *Explainable and interpretable models in computer vision and machine learning*, pages 197–253.
- McCallum, A. (2002). Mallet: A machine learning for language toolkit. <http://mallet.cs.umass.edu>.
- Naim, I., Tanveer, M. I., Gildea, D., and Hoque, M. E. (2015). Automated prediction and analysis of job interview performance: The role of what you say and how you say it. In *2015 11th IEEE*

International Conference and Workshops on Automatic Face and Gesture Recognition (FG), volume 1, pages 1–6.

Nguyen, L. S., Fraundorfer, D., Mast, M. S., and Gatica-Perez, D. (2014). Hire me: Computational inference of hirability in employment interviews based on nonverbal behavior. *IEEE Transactions on Multimedia*, 16(4):1018–1031.

Sell, G., Snyder, D., McCree, A., Garcia-Romero, D., Villalba, J., Maciejewski, M., Manohar, V., Dehak, N., Povey, D., Watanabe, S., et al. (2018). Diarization is hard: Some experiences and lessons learned for the jhu team in the inaugural dihard challenge. In *Interspeech*, pages 2808–2812.

Seré, M. (2023). Don’t stop me now: Gender attitudes in academic seminars through machine learning. *Unpublished*.

Teodorescu, M., Ordabayeva, N., Kokkodis, M., Unnam, A., and Aggarwal, V. (2022). Determining systematic differences in human graders for machine learning-based automated hiring. *Brookings Working Paper Series*.