Socially Robust AI

1 Course Information

Catalog: CSCE 689 / INFO 689 (cross-listed)
Schedule: Tu/Th, 3:10-4:55pm, HRBB 123
Credit Hours: 3
Special Course Designation: None

2 Instructor Details

Instructor

Name: Maria Teleki
Office: ABCD 123
Telephone: (123) 456-7890
Email: mariateleki@tamu.edu
Office Hours: Mondays from 3-5pm or by appointment.

TA

Name: A B
Office: ABCD 123
Telephone: (123) 456-7890
Email: ab@tamu.edu
Office Hours: Fridays from 3-5pm or by appointment.

Office Hours Office hours are a time for you to come ask questions, get help, and talk through whatever you are building! You are also very encouraged to use the *University Writing Center* for general writing and presentation skills.

Contact Us Email our instructional team at `sociallyrobustai@tamu.edu` and we will respond within 2 business days.

3 Course Description

This is a class about pushing AI to serve people better. AI systems are already deployed and millions of people are already on the other end of them. This course is about closing part of the gap between what the systems were trained on and who they actually have to serve. You will do that by combining social science and AI: social science tells you how people really behave, and how to measure it, and AI gives you the means to do something about it.

Concretely: you will find a **social robustness failure**, build **the metric** that captures it, and try **four different techniques** against that fixed metric to find out which one actually helps. Two artifacts ship at the end — your metric and the technique that came out ahead, both released publicly. We open with two weeks of that loop run at speed across six settings. My hope is that by Week 15 you will have built something you are proud to put your name on, and that someone real is better served because you built it.

Prerequisites: CSCE 310 or CSCE 603 or INFO 601 or approval of instructor; graduate classification. Programming experience in Python is assumed.

4 Course Structure

The Loop Everything in this course follows this pattern:

AI commits a social failure ⟶ **measure the social failure** ⟶ **try techniques to fix the social failure**

What Counts as a Social Robustness Failure Two criteria:

1. **A grounding.** Either a *population* — a group the system serves worse, named specifically enough that you could go find them — or a *social construct* with a citation. Politeness, epistemic authority, turn-taking rights, and perceived credibility all have social science literatures behind them.
2. **A measurement path.** One sentence on how you would know the failure is happening. What would you ask a person to judge? We may pull validated instruments from the social sciences for this part.

Two examples that would pass:

→ *"Assistants don't respond well to speakers who stutter. I would have annotators rate whether the response addresses what the speaker actually asked."*

→ *"Assistants misjudge when interruption is socially appropriate based on power dynamics [social science cite goes here]. I would have annotators rate appropriateness given the power dynamics across different power configurations."*

What would not pass is a failure that is only about latency, throughput, or aggregate accuracy on an unsocial task. That is a robustness failure with no *social* in it, and it belongs in a different course.

Every Metric in This Course Is a Judge Every metric in this course is built the same way, out of two parts:

- **Human annotations** define the construct. They are slow, expensive, and they very often disagree with each other — and that disagreement is often the most interesting thing in your project.
- **An LLM judge** carries it to scale: a rubric, worked examples, and a prompt. Judges have characteristic failure modes (position bias, verbosity bias, self-preference), so a judge is a measurement instrument like any other and has to be validated before you lean on it.

Your metric is a rubric, a judge, and an agreement number. The agreement between your judge and your human annotators, computed on your own data, is what earns you the right to report anything the judge produces.

Which Data You Look At, and When Every technique here is training-free, so it is tempting to think train/test discipline does not apply. It does. **Training-free does not mean fitting-free**: when you revise a rubric until agreement looks acceptable, or revise a prompt until the category scores improve, that loop *is* fitting — the optimizer is just *you* instead of Adam. So split your data in two before you look at any of it.

- **Train + dev** is your workshop. Rubric drafts, k-shot example selection, prompt revisions, threshold choices — anything you look at more than once lives here.
- **Test** is where you report.

Carve the split **before your annotation study**, and split on the right unit for your setup.

What Counts as a Technique A technique is any artifact someone else could pick up and run on their own problem:

- A prompt template with a documented rule for choosing its k-shot examples. If your contribution is knowing which four examples belong in the prompt and why, that is a technique, and in the right venue it could be publishable – it counts as specialized, contextual knowledge.
- A LoRA adapter, with the recipe for the data you trained it on.
- A data augmentation procedure: where to insert what, and the rule that decides.
- A routing policy.
- A reasoning scaffold: the intermediate steps you force a model through before it answers.

The Speedrun (Weeks 1–2) We run the loop six times, to first learn the shape of this type of work:

1. The AI commits a social robustness failure	2. The measure	3. The techniques
"Take me to the, uh, the one on Fifth — no, Sixth" — the assistant routes to Fifth, acting on the abandoned attempt instead of the repair **Construct: self-repair**	Did the reply act on the repaired version?	*Ex:* k-shot prompt of repair examples; Adapter tuned on synthetic repairs injected at natural positions
The same drug-interaction question gets a direct answer when the asker sounds like a clinician, and "consult a professional" when they sound like a patient **Construct: epistemic authority**	Rate completeness under each framing	*Ex:* Prompt that strips authority cues before answering
The assistant mishears a name, silently guesses, and books the wrong person — where a listener would have asked **Construct: repair initiation**	When unsure, does it ask or does it guess?	*Ex:* Adapter tuned on clarification-request data
A student asks about appealing past the deadline. The policy has a documented-hardship exception two sections later; the assistant reads the deadline clause and says no **Population: students**	Does the reading surface the exception that applies to this person?	*Ex:* Prompting to force an exceptions pass for students before answering
Speaker says *"I think it might be a fracture"*; the transcript reads *"it is a fracture"* — the hedge is gone and the commitment upgraded **Construct: epistemic stance**	Is the speaker's degree of commitment preserved?	*Ex:* k-shot prompt of hedge-preserving examples; Adapter tuned on hedged transcripts
The agent treats a thinking pause as an invitation and cuts in mid-thought **Construct: turn-taking rights**	Was speaking there appropriate, given who was talking to whom?	*Ex:* Tune a personalized per-annotator adapter; Reasoning scaffold over pause length, prosody, and relative status

The Three Parts After the speedrun, the semester runs in three parts, in the order of the loop.

- **Part I — Social Robustness Failures (Weeks 3–7).** Where systems fail people: who they are, how they disagree, what governs them, how they speak. We will learn how to find plausible failure cases.
- **Part II — Measurement (Weeks 8–10).** Construct validity, rubric design, judge validation, and the statistics that turn a number into a claim. You leave with a working metric.
- **Part III — Techniques (Weeks 11–14).** Four families — prompting and data, fine-tuning and adaptation, alignment under plurality, situated reasoning. You'll try a technique from each family and report on how they impact performance.

Readings Each week assigns a few relevant readings, listed in the schedule and linked in Canvas. They come from both technical and social-scientific literatures, because the phenomena we study are described in both. You do them on your own, and you take notes on them in the back pages of that week's Class Notes packet.

5 Learning Outcomes

By the end of this course, you will be able to:

1. Characterize a social robustness failure — who it fails, in what configuration, and what makes it social rather than just technical.
2. Ground a failure in a population or a cited social construct, and supply a measurement path for it.
3. Translate a social claim ("the AI doesn't understand disfluency") into a construct, a rubric, and a score — and say what the rubric drops.
4. Design an LLM judge, validate it against human annotations, and report the agreement that licenses its use.
5. Split a dataset at the correct unit, and explain why a training-free technique still overfits without a held-out set.
6. Apply the statistical tool that matches your design, and report effect sizes and CIs.
7. Build techniques from four different families and compare them against a fixed metric.
8. Defend a ranking: which technique came out ahead, by how much, with what uncertainty.
9. Run a small annotation study, quantify disagreement, and defend a decision to pool or preserve it.
10. Package and release an artifact a stranger can install, run, and extend.

6 Resource Policies

AI Coding Agents (Required) You are **required** to use an AI coding agent for the programming work in this course — Claude Code, Cursor, Codex, whatever you prefer. We want you to learn to direct an agent well, review what it returns,

and catch where it is confidently wrong, because all of this is now just a part of the job. You remain responsible for everything you submit; if an agent writes a function with a bug, that is your bug.

Collaboration Collaboration **is** encouraged for the Scavenger Hunt, Class Notes, the Project, the Position Paper, and Extra Credit. It **is not** allowed for Tests.

Class Repository Everything built in this course is released to a shared class repository, and prior semesters' repositories stay up. You may build on anything in a previous semester's repo as long as you cite it.

Discord There is a class Discord server, and the instructional team is on it. Ask questions, schedule meetings, and get help from classmates there.

Human Subjects Your Project collects annotations from real people, so this course runs under an approved class-level IRB protocol with me as PI. We complete IRB training together in Week 4, and you will write your own protocol for your project — recruitment language, consent script, risk assessment, data handling. You will not submit that one; it is practice, and it is a skill you will need the first time you run a study of your own.

What the Class Protocol Covers The protocol is written around a particular shape of study. Staying inside it is what lets you start collecting in Week 5 without waiting on an amendment:

- **Your stimuli come from sources that already exist.** Public corpora (Switchboard, CommonVoice), published datasets, or model outputs. If your social failure is population-based, you will not recruit the population your failure is about.
- **Your annotations come from general-population adults on Prolific**, doing rating tasks against your rubric.

Outside Resources They are allowed. Go look things up. Watch the YouTube explainer at 1.5x until it clicks, find the textbook chapter that says it a different way, ask a chatbot to walk you through the derivation, and talk to people — your classmates, us, the Discord, friends, and even the authors themselves (send an email, people are nice and often respond!). I want you to build up your skill set of knowing where to look and who to ask. So go find things, and bring them back to us. If a video made it click for you, say so, because it will probably click for someone else too.

7 Course Schedule & Grading

Materials

Readings and Videos Our textbook is *Jurafsky et al., Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models, 2026*, free online at `https://web.stanford.edu/~jurafsky/slp3/`; additional readings and videos will be specified in the Class Notes.

Annotation Budget Your Project pays real people to annotate real data, and that costs real money (about $100). Come to me ASAP if you need a scholarship and we will get it paid for. You do not need to explain anything, I do not need documentation, and it has no bearing on anything else in this course. Please just ask.

Assignments

Scavenger Hunt (1%) A Canvas quiz completed with others on the first day.

Class Notes (10%) Every lecture comes with a notes packet, posted to Canvas before we meet. Bring it (printed, or on a tablet you can write on, or set up for you to type over – whatever works best for your learning) and fill it in. Each packet has five parts:

- **Fill-in-the-blank notes.** The scaffolding is printed for you and some parts are left blank — the definition, the missing step of a derivation, the categories of a rubric, the line of code that does the actual work. You write those in as we go in lecture.
- **A Cornell Notes column.** The left margin of every page belongs to you: practice questions, cues, vocab words, connections to your own project, and the thing you did not quite follow and want to ask about on Tuesday. Fill it in during class or the same evening, while you still remember what confused you.
- **Reading pages.** Blank space for that week's readings, which you do on your own. This part has no scaffolding on purpose, because we want you to take notes the way you actually think: bullets, fragments, arrows between ideas, a box around the most important claim, a diagram of how the argument fits together, a sketch of the system being described, a two-column table putting one reading against another. The length of the notes is unimportant, the content being in a form that works for you is what matters.

- **Exercises.** A few per packet, to work at home — compute the α by hand on the toy set, run the lecture's judge prompt on three examples of your own, pick the right statistical test for a described design. They are short, and they are where the material moves from "I followed that" to "I can do that."
- **Summaries.** Regular boxes at the end of each section that you fill in yourself: in your own words, pull out what the big picture of the material in that section was.

Upload the completed packet — **all the parts, with every exercise worked** — as a single PDF in Canvas by Sunday at 11:59 PM. Handwrite it and photograph it if that is how you work, fill it in on your iPad, type it up, whatever works best for your learning. These will be graded on the exercises and on completion.

Project (40%, five checkpoints at 8%) One social robustness failure, one metric, four techniques. The project is individual, but talk to classmates, get advice, and pair program — Tuesdays are set aside for exactly this. Projects with sufficient research contribution may be submitted to workshops or conferences, and I would love to help you do that.

- **1. Failure (8%)**: Your grounding — a population, or a social construct with a citation — a documented instance of the failure, and your measurement path in a sentence. Transcripts, screenshots, and examples are plenty of evidence here. It might be the case that at scale this failure does not surface – that's fine, and that's why we run experiments! Your goal is to find a *plausible* entrypoint.
- **2. Annotation Study (8%)**: Split into train + dev and test before you look at anything. Write your instructions and your protocol, then run with 5–10 annotators on Prolific, once your config is approved. You might have full agreement, or disagreement – both are expected.
- **3. Your Metric (8%)**: Start with a naive judge — zero-shot, no rubric — and record its agreement with your annotators. That is your baseline (run on the test set). Then build the real thing: rubric, worked examples, judge prompt, and the agreement number. It must run and emit a score with a CI.
- **4. Comparison (8%)**: Run the four technique families against your failure, one per week, iterating as you go on train + dev. When you're happy with your technique, run it on the test set and add it to your results table. Say which came out ahead, by how much, with what uncertainty. "Nothing beat the baseline" is totally fine and will happen for some of you.
- **5. Ship It (8%)**: Release the metric and your experimental code (repository, README with install instructions) plus a paper and a poster in our final timeslot. A stranger should be able to install it and reproduce your headline number.

Tests (40% — 13%, 13%, 14% each) Individual, closed-notes, pencil and paper, 40 minutes. Each test has **four questions**, and every question has two parts:

- **(a) Technical.** Compute, derive, or write it. A κ by hand, a bootstrap interval, ten lines of scoring code, the right test for a described design for example.
- **(b) Sociotechnical.** Now say what that number means. What does it miss, who does it fail, what did the operationalization discard, was this the right instrument for the claim being made?

Your Class Notes are the best preparation for the tests, and they are the reason the packets exist. Engaging with your notes and practicing and studying the packet is your job. Tests cover previous weeks only, not the current week.

Position Paper (9%) A one page paper in LaTeX (template in Canvas), to practice forming an opinion in engineering and defending it with evidence. The topic changes yearly, based on live controversies in AI evaluation and governance. The rubric: do you have a position, articulate it clearly, give evidence, and structure the argument?

Extra Credit (5%) Help at least 5 people with their projects in class on Tuesdays. We will be walking around noting credit as we notice you helping others.

Schedule

Weekly Rhythm

- **Tuesday — workshop day** (except the first two weeks). Lecture videos are asynchronous and the room is a workshop: we (me and the TAs) will be walking around to help you and answer questions, so get help, help others, do corrections. Tests and Demos will also be on Tuesdays.
- **Thursday** — in-class lectures.
- **Sunday** — Checkpoints, Class Notes (packet and exercises), and the Position Paper due at 11:59 PM.

Semester Schedule *Note: This schedule is subject to change, check Canvas for up-to-date information.*

Week	Topics	Assignments Due
1	***The Speedrun (1 of 2)*** • Course introduction; how the project works • The loop: failure → measure → techniques • What counts as a social robustness failure, and what counts as a technique • Why every metric in this course is an LLM judge • Rows 1–2 of the speedrun table, start to finish • **Software:** Python; Anthropic/OpenAI APIs; class repo; your coding agent • **Readings:** *Koh et al., Wilds: A benchmark of in-the-wild distribution shifts, 2021*; *Selbst et al., Fairness and abstraction in sociotechnical systems, 2019*	• Scavenger Hunt • Class Notes
2	***The Speedrun (2 of 2)*** • Rows 3–6 of the speedrun table, start to finish • Splitting your data: train + dev against test • **Software:** scikit-learn `GroupShuffleSplit` — splitting by speaker or annotator, not by row • **Reading:** *Liao et al., Are we learning yet? a meta review of evaluation failures across machine learning, 2021*	• Class Notes
3	***Part I — Social Robustness Failures*** ***Finding a Failure*** • Where deployed systems fail people, and how those failures get found • Grounding in a population vs. grounding in a social construct • Personas and what an attribute list leaves out • **Software:** Hugging Face `datasets` • **Readings:** *Ribeiro et al., Beyond accuracy: Behavioral testing of NLP models with CheckList, 2020*; *Cheng et al., Marked personas: Using natural language prompts to measure stereotypes in language models, 2023*; *Ge et al., Scaling synthetic data creation with 1,000,000,000 personas, 2024*	• Class Notes • Test 1
4	***Annotation and Disagreement*** • Designing an annotation study: items, instructions, recruitment, validated instruments • Writing an IRB protocol • Measuring agreement: Cohen's κ, Fleiss' κ, Krippendorff's α • When annotators disagree, and why a gold standard can be a fiction • Why your $100 Prolific budget only buys you exploratory research • **Software:** Prolific; `krippendorff`, `statsmodels` for agreement • **Readings:** *Plank, The "problem"' of human label variation: On ground truth in data, modeling and evaluation, 2022*; *Sorensen et al., Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties, 2024*	• Class Notes • 1. Failure

Continued on next page...

Week	Topics	Assignments Due
5	***Policy and Governance*** • How AI systems read rules, and where the reading goes wrong • Rubrics and scorecards as measurement instruments • NIST AI RMF vs. the EU AI Act • **Software:** Fairlearn, Aequitas, Algorithm Audit • **Readings:** *Metcalf et al., Algorithmic impact assessments and accountability: The co-construction of impacts, 2021*; *Teleki, Choi, et al., The Due Process Deficit: Auditing AI Governance in U.S. Higher Education, 2026*	• Class Notes
6	***Speech Systems*** • Where speech AI shows up: voice modes, smart glasses, voice messaging • Two architectures: ASR into an LLM, against end-to-end speech models • Datasets and their transcription conventions: Switchboard, CommonVoice • One utterance, many valid transcripts, and what WER cannot see • **Software:** Whisper; Hugging Face ASR pipelines; `jiwer`; `pyannote.audio` • **Readings:** *Teleki, Dong, et al., Comparing ASR Systems, 2024*; *Choi et al., Beyond Single Ground Truth: Reference Monism as Epistemic Injustice in ASR Evaluation, 2026*; *Koenecke et al., Racial disparities in automated speech recognition, 2020*	• Class Notes • 2. Annotation Study
7	***Disfluency*** • Interjections, parentheticals, edited nodes, repairs, backchannels • Why models handle them badly • What a disfluency carries: confidence, emphasis, planning, memory/importance • Disfluency in translation *Züfle* et al., The Role of Disfluencies in Speech Translation, 2026* • **Software:** Switchboard disfluency annotations; Parselmouth • **Readings:** *Shriberg, Disfluencies in switchboard, 1996*; *Diachek et al., The effect of disfluency on memory for what was said. 2023*; *Lea et al., From user perceptions to technical improvement: Enabling people who stutter to better use speech recognition, 2023*	• Class Notes
8	***Part II — Measurement*** ***Building Your Metric*** • Construct, operationalization, and what gets discarded • Writing a rubric: categories, anchors, worked examples • Writing a judge prompt: pairwise against pointwise scoring • How judges fail: position bias, verbosity bias, self-preference • Validating a judge against your human annotations • **Readings:** *D. Li et al., From generation to judgment: Opportunities and challenges of llm-as-a-judge, 2025*; *Pan et al., Human-Centered Design Recommendations for LLM-as-a-judge, 2024*; *Shi et al., Judging the judges: A systematic study of position bias in llm-as-a-judge, 2025*	• Class Notes • Test 2

Continued on next page...

Week	Topics	Assignments Due
9	***Statistics I — What Kind of Experiment Did You Run?*** • Paired vs. unpaired designs; t-tests, Wilcoxon, Mann-Whitney • Bootstrap and permutation tests for small samples • Effect sizes and confidence intervals • Multiple comparisons: Holm-Bonferroni, Tukey's HSD • **Software:** SciPy, statsmodels, pingouin • **Readings & Videos:** *Dror et al., The hitchhiker's guide to testing statistical significance in natural language processing, 2018*, Paired vs. Unpaired t-tests, StatQuest on Bootstrapping, Effect Sizes	• Class Notes • 3. Your Metric
10	***Statistics II — Structure in Your Data*** • Scores nested within speakers, items, or annotators • Mixed-effects models: random intercepts and slopes • Choosing the right unit to split on • Power analysis: how many annotators and items you need • **Software:** `lme4`, `statsmodels.MixedLM`, `pymer4` • **Readings & Videos:** Simplistics LMM Video, Clappam LMM Video	• Class Notes
11	***Part III — Techniques*** ***Family 1 — Prompting & Data*** • Prompt templates: instructions, output format • Choosing k-shot examples — the selection rule is the contribution • Reading your category scores, then targeting the worst one *Teleki, Janjur, et al., Z-Scores: A Metric, 2026* • Generating synthetic training data • **Software:** Hugging Face `transformers` • **Readings:** *Schulhoff et al., The prompt report: A systematic survey of prompt engineering techniques, 2024*; *Min et al., Rethinking the role of demonstrations: What makes in-context learning work?, 2022*	• Class Notes • Test 3
12	***Family 2 — Fine-Tuning & Adaptation*** • Supervised fine-tuning; LoRA and adapters • Adapting to a speaker, a dialect, or a domain • Fine-tuning tradeoffs: cost, data requirements, and who can afford to build this way • **Software:** Hugging Face PEFT; `bitsandbytes` • **Readings:** *E. J. Hu et al., Lora: Low-rank adaptation of large language models, 2022*; *Z. Han et al., Parameter-efficient fine-tuning for large models: A comprehensive survey, 2024*	• Class Notes • Position Paper

Continued on next page...

Week	Topics	Assignments Due
13	***Family 3 — Alignment Under Plurality*** • RLHF and DPO; how a reward model learns from preferences • What gets lost when you pool annotators who disagree • Per-annotator reward models and routing • **In class:** fit a reward model pooled, then per annotator, and compare • **Software:** Hugging Face TRL (reward modeling, DPO) • **Readings:** *Z. Wang et al., A comprehensive survey of llm alignment techniques: RLHF, RLAIF, PPO, DPO and more, 2024*	• Class Notes
14	***Family 4 — Situated Reasoning*** • Reasoning scaffolds: the steps you force a model through before it answers • Reading the situation: power, relationship, and culture as inputs • Proactive agents in meetings and on wearables *Teleki* et al., A Cross-Community Agenda for Speech AI, 2026* • **Software:** Pipecat or LiveKit Agents for full-duplex voice • **Readings:** *Wei et al., Chain-of-thought prompting elicits reasoning in large language models, 2022*	• Class Notes • 4. Comparison
15	***Shipping*** • Packaging: README, install path, worked example, license • Reproducibility: seeds, pinned environments, judge model versions • Model cards and datasheets • **Software:** Hugging Face Hub model cards • **Readings:** *No reading this week*	• Class Notes • 5. Ship It

Grade Assignment **A**: $[90\% - 100\%]$, **B**: $[80\% - 90\%)$, **C**: $[70\% - 80\%)$, **D**: $[60\% - 70\%)$, **F**: $[0\% - 60\%)$. If you want to improve your grade: (1) You are allowed to do corrections for $\frac{1}{2}$ credit back, as described below; (2) There is an Extra Credit assignment, as described above.

Late Work Policy Makeup work for an excused absence is not late work and is exempt from this policy (Student Rule 7). For unexcused absences, late work is not permitted — but you may do corrections for $\frac{1}{2}$ credit back.

Corrections Policy You may do one set of corrections on any graded work for $\frac{1}{2}$ credit back. E.g. with a grade of 80, there were 20 points you did not get: fully correct corrections earn a 90 ($80 + (1.00 \times 20 \times \frac{1}{2})$), and corrections scored at 50% earn an 85 ($80 + (0.50 \times 20 \times \frac{1}{2})$). Requirements:

- Email the corrected PDF to `sociallyrobustai@tamu.edu` with the subject `CORRECTIONS:{ASSIGNMENT_NAME}`.
- For each question you missed: solve it correctly, then answer — why did you get it wrong the first time? Were you thinking about it a different way? What did you learn?
- You have 7 days (168 hours) from the original due date. *Corrections are not permitted for work due in the 15^{th} week.*

Regrade Policy Request a regrade if you spot an error or disagree with the grading. Email `sociallyrobustai@tamu.edu` with the subject `REGRADE:{ASSIGNMENT_NAME}` and include (i) the assignment, (ii) the specific items to regrade, and (iii) why each should be regraded. You have 7 days (168.0 hours) from receiving the grade; after that, there are no regrades.

8 References & Letters of Recommendation

References are generally needed for industry jobs, and **Letters of Recommendation** for academic contexts (e.g. graduate school, fellowships, scholarships). I would love to help, and I am happy to write both! (Provided you're doing your job as a good student.) I just need the following, so I have plenty of specific things to say on your behalf:

- 4 weeks before your letter is due, email `sociallyrobustai@tamu.edu` with the subject `LETTER/REFERENCE: {YOUR_NAME}`, attaching (i) your resume/CV and (ii) your goals — graduate school? what area? industry? what roles excite you? (iii) where am I sending this letter? (iv) give me a bullet point list of what you think I should write about in the letter (basically an outline).
- Do wonderful work on your project. Show me your best!
- Come to office hours regularly AND/OR talk to me on Tuesdays in class AND/OR talk to me before/after class for a 5 minute check-in: what did you learn, what are you stuck on, how is the project going? Your performance on the project is probably the best thing I can write about in your letter, so keep me in the loop on your project and your progress through the semester!

9 Attendance Policy

The university views class attendance and participation as an individual student responsibility. Students are expected to attend class and to complete all assignments. Please refer to Student Rule 7 in its entirety for information about excused absences, including definitions, and related documentation and timelines.

To notify us of an excused absence, email `sociallyrobustai@tamu.edu` with the subject `CLASS: EXCUSED ABSENCE` and attach relevant documentation as a PDF.

10 Makeup Work Policy

Students will be excused from attending class on the day of a graded activity or when attendance contributes to a student's grade, for the reasons stated in Student Rule 7, or other reason deemed appropriate by the instructor. Please refer to Student Rule 7 in its entirety for information about makeup work, including definitions, and related documentation and timelines. Absences related to Title IX of the Education Amendments of 1972 may necessitate a period of more than 30 days for make-up work, and the timeframe for make-up work should be agreed upon by the student and instructor" (Student Rule 7, Section 7.4.1). "The instructor is under no obligation to provide an opportunity for the student to make up work missed because of an unexcused absence" (Student Rule 7, Section 7.4.2). Students who request an excused absence are expected to uphold the Aggie Honor Code and Student Conduct Code. (See Student Rule 24.)

11 Academic Integrity Statement and Policy

"An Aggie does not lie, cheat or steal, or tolerate those who do."
"Texas A&M University students are responsible for authenticating all work submitted to an instructor. If asked, students must be able to produce proof that the item submitted is indeed the work of that student. Students must keep appropriate records at all times. The inability to authenticate one's work, should the instructor request it, may be sufficient grounds to initiate an academic misconduct case" (Section 20.1.2.3, Student Rule 20).
You can learn more about the Aggie Honor System Office Rules and Procedures, academic integrity, and your rights and responsibilities at aggiehonor.tamu.edu.

12 Americans with Disabilities Act (ADA) Policy

Texas A&M University is committed to providing equitable access to learning opportunities for all students. If you experience barriers to your education due to a disability or think you may have a disability, please contact the Disability Resources office on your campus (resources listed below) Disabilities may include, but are not limited to attentional, learning, mental health, sensory, physical, or chronic health conditions. All students are encouraged to discuss their disability related needs with Disability Resources and their instructors as soon as possible.
Disability Resources is located in the Student Services Building or at (979) 845-1637 or visit disability.tamu.edu.

13 Title IX and Statement on Limits to Confidentiality

Texas A&M University is committed to fostering a learning environment that is safe and productive for all. University policies and federal and state laws prohibit gender-based discrimination and sexual harassment, including sexual assault, sexual exploitation, domestic violence, dating violence, and stalking.
With the exception of some medical and mental health providers, all university employees (including full and part-time faculty, staff, paid graduate assistants, student workers, etc.) are Mandatory Reporters and must report to the Title IX Office if the employee experiences, observes, or becomes aware of an incident that meets the following conditions (see University Rule 08.01.01.M1):

- The incident is reasonably believed to be discrimination or harassment.
- The incident is alleged to have been committed by or against a person who, at the time of the incident, was (1) a student enrolled at the University or (2) an employee of the University.

Mandatory Reporters must file a report regardless of how the information comes to their attention – including but not limited to face-to-face conversations, a written class assignment or paper, class discussion, email, text, or social media post. Although Mandatory Reporters must file a report, in most instances, a person who is subjected to the alleged conduct will be able to control how the report is handled, including whether or not to pursue a formal investigation. The University's goal is to make sure you are aware of the range of options available to you and to ensure access to the resources you need.
Students wishing to discuss concerns related to mental and/or physical health in a confidential setting are encouraged to make an appointment with University Health Services or download the TELUS Health Student Support app for 24/7 access to professional counseling in multiple languages. Walk-in services for urgent, non-emergency needs are available during normal business hours at University Health Services locations; call 979.458.4584 for details.

14 Statement on Mental Health and Wellness

Texas A&M University recognizes that mental health and wellness are critical factors influencing a student's academic success and overall wellbeing. Students are encouraged to engage in healthy self-care practices by utilizing the resources and services available through University Health Services. Students needing a listening ear can call the Texas A&M Helpline (979.845.2700) from 4:00 p.m. to 8:00 a.m. weekdays and 24 hours on weekends for mental health peer support while classes are in session. The TELUS Health Student Support app provides access to professional counseling in multiple languages anytime, anywhere by phone or chat, and the 988 Suicide & Crisis Lifeline offers 24-hour emergency support at 988 or 988lifeline.org.
Students needing a listening ear can contact University Health Services (979.458.4584) or call the Texas A&M Helpline (979.845.2700) from 4:00 p.m. to 8:00 a.m. weekdays and 24 hours on weekends while classes are in session. 24-hour emergency help is also available through the 988 Suicide & Crisis Lifeline (988) or at 988lifeline.org.

References

Cheng, M., E. Durmus, and D. Jurafsky (2023). "Marked personas: Using natural language prompts to measure stereotypes in language models". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1504–1532.

Choi, A. S. G. et al. (2026). "Beyond Single Ground Truth: Reference Monism as Epistemic Injustice in ASR Evaluation". In: *In Prep. for ARR*. ACL Rolling Review. URL: https://mariateleki.github.io/pdf/2026WERrange.pdf. Cornell, Rev AI, Kangwon National University.

Diachek, E. and S. Brown-Schmidt (2023). "The effect of disfluency on memory for what was said." In: vol. 49. 8. American Psychological Association, p. 1306.

Dror, R. et al. (2018). "The hitchhiker's guide to testing statistical significance in natural language processing". In: *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 1383–1392.

Ge, T. et al. (2024). "Scaling synthetic data creation with 1,000,000,000 personas". In.

Han, Z. et al. (2024). "Parameter-efficient fine-tuning for large models: A comprehensive survey". In.

Hu, E. J. et al. (2022). "Lora: Low-rank adaptation of large language models". In: vol. 1. 2, p. 3.

Jurafsky, D. and J. H. Martin (2026). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models*. 3rd. Online manuscript released January 6, 2026. URL: https://web.stanford.edu/~jurafsky/slp3/.

Koenecke, A. et al. (2020). "Racial disparities in automated speech recognition". In: vol. 117. 14. National Academy of Sciences, pp. 7684–7689.

Koh, P. W. et al. (2021). "Wilds: A benchmark of in-the-wild distribution shifts". In: *International conference on machine learning*. PMLR, pp. 5637–5664.

Lea, C. et al. (2023). "From user perceptions to technical improvement: Enabling people who stutter to better use speech recognition". In: *Proceedings of the 2023 CHI conference on human factors in computing systems*, pp. 1–16.

Li, D. et al. (2025). "From generation to judgment: Opportunities and challenges of llm-as-a-judge". In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2757–2791.

Liao, T. et al. (2021). "Are we learning yet? a meta review of evaluation failures across machine learning". In: *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Metcalf, J. et al. (2021). "Algorithmic impact assessments and accountability: The co-construction of impacts". In: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 735–746.

Min, S. et al. (2022). "Rethinking the role of demonstrations: What makes in-context learning work?" In: *Proceedings of the 2022 conference on empirical methods in natural language processing*, pp. 11048–11064.

Pan, Q. et al. (2024). "Human-Centered Design Recommendations for LLM-as-a-judge". In: *Proceedings of the 1st Human-Centered Large Language Modeling Workshop*, pp. 16–29.

Plank, B. (2022). "The "problem"' of human label variation: On ground truth in data, modeling and evaluation". In: *Proceedings of the 2022 conference on empirical methods in natural language processing*, pp. 10671–10682.

Ribeiro, M. T. et al. (2020). "Beyond accuracy: Behavioral testing of NLP models with CheckList". In: *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4902–4912.

Schulhoff, S. et al. (2024). "The prompt report: A systematic survey of prompt engineering techniques". In.

Selbst, A. D. et al. (2019). "Fairness and abstraction in sociotechnical systems". In: *Proceedings of the conference on fairness, accountability, and transparency*, pp. 59–68.

Shi, L. et al. (2025). "Judging the judges: A systematic study of position bias in llm-as-a-judge". In: *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 292–314.

Shriberg, E. (1996). "Disfluencies in switchboard". In: *Proceedings of international conference on spoken language processing*. Vol. 96. 1. IEEE Philadelphia, PA, pp. 11–14.

Sorensen, T. et al. (2024). "Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 18, pp. 19937–19947.

Teleki, M., A. S. G. Choi, et al. (2026). "The Due Process Deficit: Auditing AI Governance in U.S. Higher Education". In: *FAccT '26*. The ACM Conference on Fairness, Accountability, and Transparency. URL: https://mariateleki.github.io/pdf/FAccT_26_ACAI.pdf. Cornell, AAAS Science and Technology Policy Fellow.

Teleki, M., X. Dong, et al. (2024). "Comparing ASR Systems in the Context of Speech Disfluencies". In: *INTERSPEECH '24*. Conference of the International Speech Communication Association. URL: https://www.isca-archive.org/interspeech_2024/teleki24_interspeech.pdf.

Teleki, M., S. Janjur, et al. (2026). "Z-Scores: A Metric for Linguistically Assessing Disfluency Removal". In: *ICASSP '26*. IEEE International Conference on Acoustics, Speech and Signal Processing. URL: https://arxiv.org/pdf/2509.20319.

Teleki*, M. et al. (2026). "A Cross-Community Agenda for Speech AI". In: *Under Submission at ARR*. ACL Rolling Review. URL: https://mariateleki.github.io/pdf/CrossCommunityAgenda.pdf.

Wang, Z. et al. (2024). "A comprehensive survey of llm alignment techniques: RLHF, RLAIF, PPO, DPO and more". In.

Wei, J. et al. (2022). "Chain-of-thought prompting elicits reasoning in large language models". In: vol. 35, pp. 24824–24837.

Züfle*, M. et al. (2026). "The Role of Disfluencies in Speech Translation". In: *Under Submission at TACL*. Transactions of the Association for Computational Linguistics. URL: https://arxiv.org/pdf/2608.02138. KIT, ETH Zürich.