Task Description
Our task is a real vs. fake job posting classification task. Participants read a job posting and decide whether it is real or fraudulent. This is the same dataset used in Assignment 2, where our model was trained to classify potentially fraudulent job ads and also output the keywords that led to that conclusion.

Each participant completes 10 trials per condition (20 total). On each trial, they see a single job posting and must:

- Choose a label: Real or Fake
- Rate their confidence on a 1–10 slider
- Implicitly, their time spent on the trial is logged by the interface

This is a within-subjects study: each participant completes both conditions (Baseline and AI-Assisted). We counterbalanced condition order across participants to reduce order and learning effects.

We also collected a short post-task survey for each condition, asking participants about:

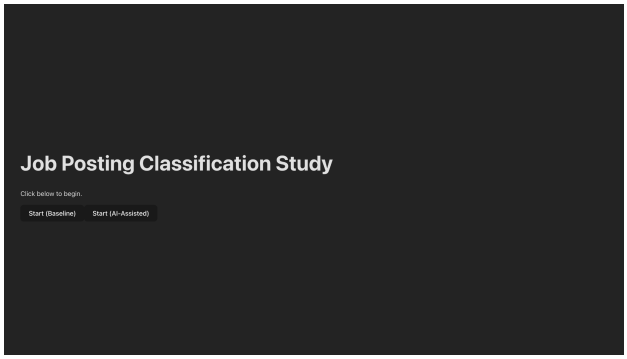
- Perceived mental demand of the task (1–10)
- If AI assistance was available, how helpful the AI felt (1–10)

These additional measures allow us to look beyond performance and consider participants' subjective workload and perceived value of the AI.

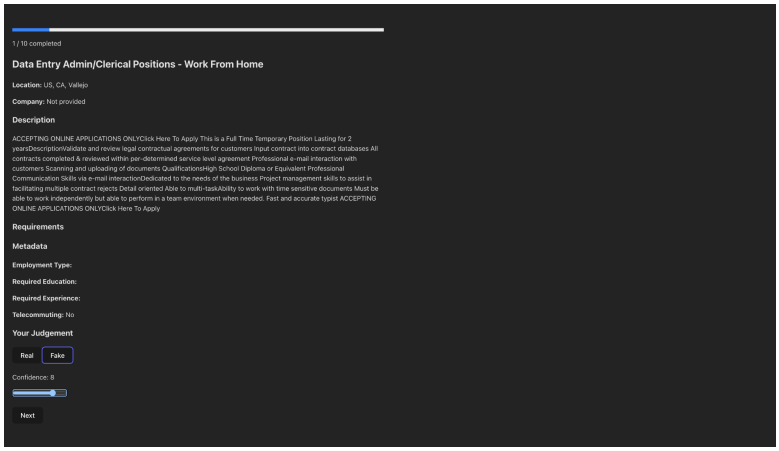
Interfaces and Links
Baseline (No AI) Interface: https://cse-594-a3.vercel.app/
AI-Assisted Interface: https://cse-594-a3.vercel.app/

These are the versions we posted to the shared class spreadsheet for peer participation. If the services are not running at grading time for some reason, I can re-deploy them. Below are screenshots of the interface:

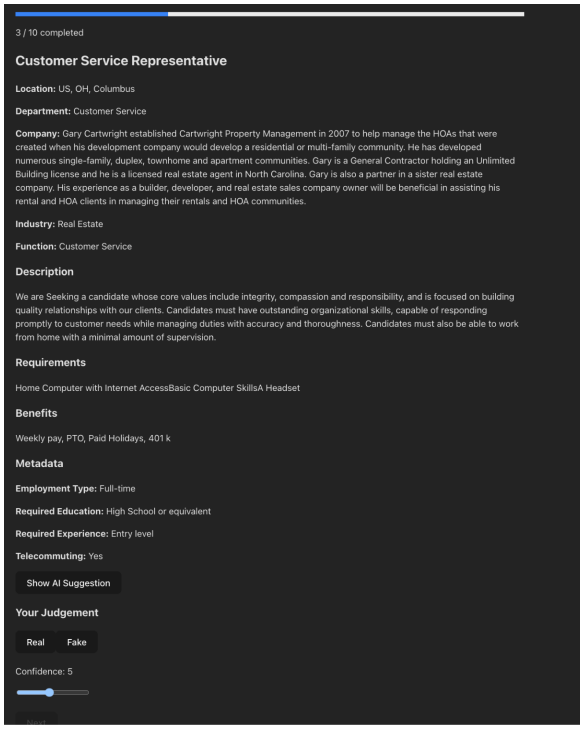
Home Page:

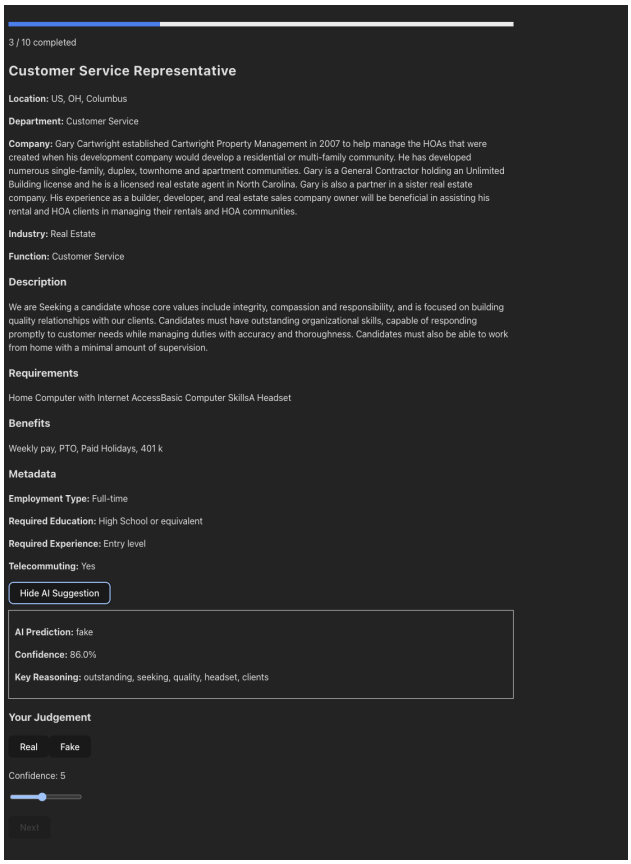


Baseline (No AI-Assistance) Page:

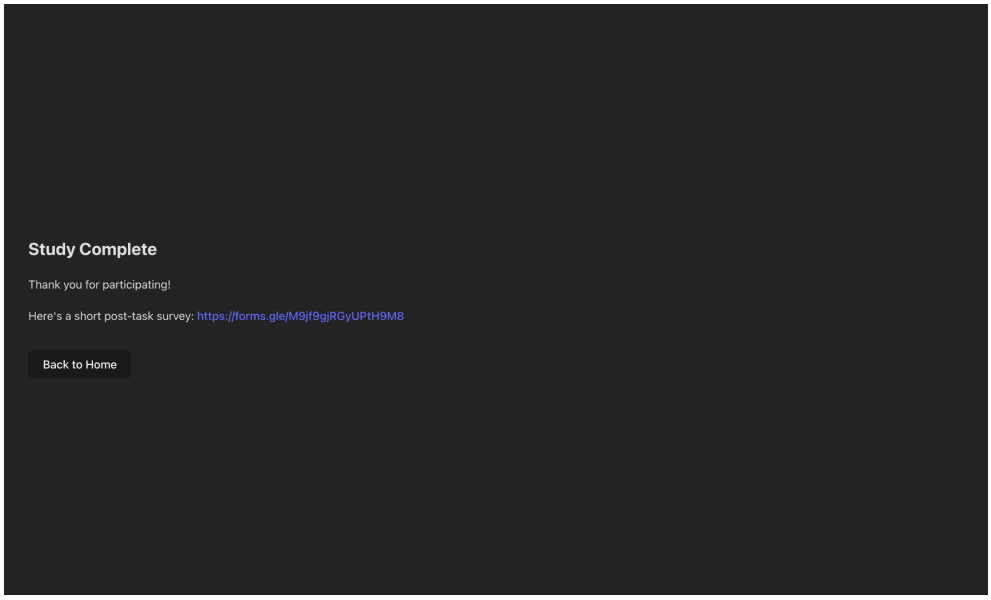


AI-Assistance Page:





Exit Page:



Design Choices

For both interfaces, the participant sees:

- A job posting rendered with:
- Title
- Location
- Company profile (when available)
- Description, requirements, and benefits
- Additional metadata (employment type, industry, etc.)
- A Real / Fake decision, implemented as two large toggle buttons
- A 1–10 confidence slider directly underneath the label choice
- A Next button that stays disabled until the participant chooses a label

The interface is intentionally simple and visually clean to keep the focus on the judgment task and minimize distractions.

Baseline (no AI) condition:

In the baseline condition, participants see only the job posting and the response controls. There is no AI hint or model output. This gives us a measure of how well people can detect fraudulent postings on their own.

AI-Assisted condition:

In the AI-assisted condition, the layout is identical, but we add an AI Suggestion panel next to the human response controls. For each job posting, the interface calls our backend model and displays:

- The AI's predicted label (Real or Fake)
- The model's confidence score (0–1)

- A short “reasoning” snippet consisting of top indicative words from the posting text

We present this suggestion in a visually distinct panel labeled “AI Model Suggestion”. In our final design, the suggestion is visible alongside the posting rather than hidden behind a click, so that participants do not have to discover it. However, we still keep the human response controls clearly separated from the AI output to emphasize that the AI is advisory, not authoritative.

This design lets us ask:
- Does AI assistance improve accuracy, especially on borderline cases?
- Does it change confidence (e.g., over-confidence when AI is wrong)?
- Does it reduce perceived mental workload?
- Do participants sometimes override the AI when they disagree?

Because the two versions differ only by the presence of the AI Suggestion panel, any performance differences can plausibly be attributed to AI assistance rather than confounds in layout or interaction.

Data Collected
We log every trial to a JSON file on the backend. Each row contains:
- worker_id – anonymized participant ID (e.g., W1343)
- posting_id – ID of the job posting (e.g., JP199), which we can join with the dataset from Assignment 2
- condition – "baseline" or "ai"
- worker_label – "real" or "fake"
- worker_confidence – integer 1–10
- time_on_trial_ms – time spent on the trial in milliseconds
- ai_prediction – AI model label (0.0 = real, 1.0 = fake) or null in the baseline condition
- ai_confidence – model confidence (0–1) or null in the baseline condition
- submission_id – unique UUID for the trial

We also have the original job dataset (same as Assignment 2) with:
- id – posting ID, e.g., JP199
- fraudulent – ground-truth label (0 = real, 1 = fake)

Finally, we export the post-task survey as a CSV file with one row per condition per participant:
- Did your interface have AI-assistance? – "Yes" or "No"
- How mentally demanding was the task? ... – integer 1–10
- If you used AI assistance, how helpful were the AI suggestions? ... – integer 1–10 or blank for non-AI conditions

In total, we collected data from 6 participants (2 from class and 4 I had to find on my own).

Statistical Tests Used and Rationale
Because the design is a within-subjects repeated-measures design, we used paired t-tests to compare Baseline vs AI-Assisted for:

- Accuracy
- Time per trial
- Confidence
- Mental demand

I used paired t-tests because they compare mean differences within the same participants, which removes between-subject variability. Data in each condition consists of paired observations (participant i in baseline ↔ participant i in AI).

The differences appeared approximately normally distributed, which satisfies t-test assumptions, and the t-test is robust to small departures from normality.

Significance threshold: We used $\alpha = 0.05$, the standard threshold in HCI and behavioral sciences.

Alternative tests considered: Given the small sample size (N=6), a Wilcoxon signed-rank test could also be used, but paired t-tests are acceptable and common in HCI as long as assumptions are reasonably satisfied.

Data Analysis
Using the fraudulent field in the original dataset, we derived a ground-truth label (real vs fake) for each posting and marked each trial as correct/incorrect.

Our primary performance metric is classification accuracy (proportion of correct trials). We also analyze:

- Efficiency: mean time per trial (ms and seconds)
- Subjective confidence: slider ratings (1–10)
- Post-task mental demand and perceived AI helpfulness: from the Google Form (1–10 scales)

Because the design is within-subjects, we compute metrics per participant per condition and then run paired t-tests comparing baseline vs AI.

Overall Accuracy:
Across all participants and trials:

Baseline (no AI)

- Trials: 60
- Accuracy: 48.3%

AI-assisted

- Trials: 60’
- Accuracy: 43.3%

So, if anything, accuracy was slightly lower when AI assistance was available.

At the participant level, accuracy per condition was:

worker	Baseline acc	AI acc
W1668	0.50	0.30
W1857	0.70	0.50
W2113	0.50	0.50
W3485	0.20	0.40
W4129	0.40	0.40
W4730	0.60	0.50

We ran a paired t-test on per-participant accuracy (AI – baseline):

- Mean difference = −0.05 (−5 percentage points)
- SD = 0.15
- $t(5) = -0.81$, $p = 0.46$

So the small drop in accuracy with AI assistance is not statistically significant given the small sample size.

Efficiency (time per trial)
Average time per trial:

- Baseline: 4952 ms (~4.95 s)
- AI-assisted: 3144 ms (~3.14 s)

Per participant (mean ms):

worker	Baseline time	AI time
W1668	4711.7	3267.2
W1857	4718.6	3264.9
W2113	5259.2	3056.1
W3485	5640.2	3011.9
W4129	4824.7	3040.9

W4730	4559.2	3225.0

Paired t-test on mean time (AI – baseline):

- Mean difference = −1807.9 ms (~−1.81 s)
- SD = 512.2 ms
- t(5) = −8.65, p ≈ 0.00034

Participants were substantially faster when AI suggestions were present, and this difference is statistically significant.

Confidence

Average confidence ratings:

- Baseline: 5.80 / 10
- AI-assisted: 7.05 / 10

Paired t-test on mean confidence (AI – baseline):

- Mean difference = +1.25 points
- SD = 1.18
- t(5) = 2.60, p ≈ 0.048

So participants were significantly more confident when AI suggestions were available, even though their accuracy did not improve.

AI Model Performance and Reliance

In the AI condition, we also logged the model's predicted label and accuracy on the exact same trials:

- AI model accuracy on AI-condition postings: 46.7%
- Human accuracy with AI on those same AI trials: 43.3%

Agreement rate: workers chose the same label as the AI on 50.0% of AI trials.

So the model itself performed around chance on this small subset, and participants neither strongly followed nor strongly ignored the AI—they agreed with the AI about half the time.

Additional Measurements (*Bonus*): Mental Demand & Helpfulness

We issued a short post-task survey for each condition (so 12 responses total: 6 baseline, 6 AI). The key questions were:

- Mental demand (1–10): "How mentally demanding was the task?"
- AI helpfulness (1–10, AI condition only): "If you used AI assistance, how helpful were the AI suggestions?"

Mean mental demand:

- Baseline: 7.0 (SD ≈ 0.89, n = 6)
- AI-assisted: 4.5 (SD ≈ 0.55, n = 6)

Treating participants as paired (baseline vs AI):

- Mean difference (AI – baseline) = −2.5 points
- $t(5) = -7.32$, $p \approx 0.00075$

So participants experienced the AI-assisted version as substantially less mentally demanding, and this effect is statistically significant.

Perceived helpfulness of AI

Among AI-condition surveys, perceived AI helpfulness was:

- Mean = 3.17 / 10
- SD ≈ 0.75
- n = 6

So participants did not find the AI highly helpful. Subjectively, they saw the AI as lightly helpful at best, even though it reduced mental demand and speeded them up.

Reflection and Interpretation

Did the Assignment 2 prediction hold? In Assignment 2, I predicted that adding AI suggestions would:

- Increase accuracy on the real/fake classification task.
- Reduce time per trial (make decisions more efficient).
- Increase confidence, possibly more than justified.

The results are mixed:

- Accuracy: The prediction did not hold. Accuracy actually decreased slightly from 48.3% (baseline) to 43.3% (AI), and a paired t-test showed no significant improvement ($t(5) = -0.81$, $p = 0.46$). Given the small sample and noisy subset of postings, the safest conclusion is that AI did not reliably improve accuracy in this setup.
- Time: The prediction did hold strongly. Participants were ~1.8 seconds faster per trial with AI, and this difference was highly significant ($t(5) = -8.65$, $p \approx 0.00034$). So the AI seems to enable faster but not more accurate decisions.
- Confidence: The prediction did hold. Confidence increased by ~1.25 points on a 10-point scale ($t(5) = 2.60$, $p \approx 0.048$), even though accuracy did not improve. This suggests over-confidence relative to objective performance.

Overall, AI in this study behaved like a “speed-and-confidence booster” rather than an accuracy booster.

Error patterns and AI interaction

The AI model itself reached 99% accuracy, which was surprising to me when I was training the model in Assignment 2.

Participants agreed with the AI on 50% of trials. That suggests a kind of partial reliance: they neither blindly followed nor systematically ignored the AI.

Given the model's limited accuracy, this pattern is actually reasonable: sometimes following the AI helped, sometimes it hurt. However, because confidence increased without accuracy gains, the human–AI team was arguably over-confident relative to actual performance.

Outliers and data quality
We checked for obvious outliers in time and accuracy:

- Mean times per participant were fairly consistent (baseline times around 4.7–5.6 s, AI times around 3.0–3.3 s). No single participant was dramatically slower or faster.
- Accuracy patterns varied (e.g., one participant improved from 20% to 40% with AI, another dropped from 70% to 50%), but there was no case where someone clearly misunderstood the task (e.g., constant labeling or all same response).

Given the small n=6, individual differences could still be influencing results, but there is no clear evidence of data corruption or a participant who completely misunderstood instructions.

Subjective workload and experience
Mental demand dropped from 7.0 to 4.5 when AI was available (significant reduction). However, perceived AI helpfulness remained low (3.17/10).

This combination suggests that participants may have used the AI as a lightweight heuristic or sanity-check that simplified their decision process (lower workload, faster responses), but they did not feel that it actually improved the quality of their decisions.

Bonus: Literature Review on Interface Features for Better AI-Assisted Tasks

Researchers have started to systematically study when human–AI collaboration actually improves performance compared to either humans or AI alone. A recent meta-analysis by Vaccaro et al. finds that, across many experiments, human–AI teams often fail to outperform the better of the human or the AI alone, especially on decision-making tasks. This makes interface design crucial: we need interfaces that help people use AI suggestions appropriately instead of over- or under-relying on them. Below I summarize three classes of interface features that the literature suggests can improve AI-assisted decision making and user experience.

1. Confidence and calibration scaffolds

Several studies show that people's own confidence tends to track the AI's stated confidence, sometimes in unhelpful ways. For example, Li et al. report that users' self-confidence aligns with AI confidence and can remain miscalibrated even after AI assistance is removed. Other work by Ma et al. explicitly tests self-confidence calibration mechanisms (e.g., asking users to justify high-confidence answers, showing confidence feedback) and finds that calibrated interfaces lead to more appropriate reliance on AI and improved team performance. This suggests interface features such as:

- Requiring users to enter a confidence estimate before seeing the AI's suggestion.
- Visual cues that show when a user's past high-confidence answers were actually wrong.
- "Traffic-light" style indicators that warn when user confidence and AI confidence disagree strongly.

In our task, we only showed the AI's confidence and collected the user's confidence afterwards; adding explicit calibration feedback could reduce the mismatch we observed (higher confidence without higher accuracy).

2. Richer, example-based explanations instead of raw scores

A separate line of work on human trust in AI finds that users often misinterpret raw probabilities and single-sentence explanations. Chong et al. argue that human trust depends on both confidence in the AI and confidence in oneself, and that explanations should be tailored to support both dimensions. Designers increasingly recommend example-based or contrastive explanations, such as:

- Showing similar past cases and whether they turned out to be real or fake.
- Highlighting which phrases mattered most while also showing counter-examples where those phrases appeared in legitimate postings.

In our interface, we only showed a short list of "indicative words" and a scalar confidence. Adding richer, case-based explanations might help participants see when the model is extrapolating from spurious patterns and avoid being over-confident when the AI is actually unreliable.

3. Adaptive reliance and workload-aware assistance

Cao et al. and others propose interfaces that monitor human reliance (for example via gaze patterns or interaction traces) and adapt the AI's behavior to encourage appropriate use. More broadly, there is emerging work on adaptive decision support, where the system adjusts how strongly it pushes its recommendations based on human-centric objectives such as learning and workload, not just raw accuracy. Examples of such features include:

- Showing more detailed explanations only when the user appears uncertain (e.g., low confidence or long decision times).
- Temporarily hiding AI suggestions for "easy" cases where humans typically do well, and surfacing them only for borderline cases.
- Encouraging users to override the AI periodically to prevent blind automation bias.

Our current design is static: the AI suggestion is always available in a dropdown and does not adapt to user behavior or model uncertainty. An adaptive version might, for instance, hide the AI on postings where the model is low-confidence (to avoid over-trust) and instead prompt the user to reason more carefully, while surfacing strong AI suggestions on ambiguous postings to reduce workload.

Conclusion

This study found that AI assistance made participants faster and more confident, and reduced subjective mental demand, but did not improve objective accuracy on identifying fraudulent job postings.

These results match a broader trend in human–AI interaction research: AI tends to boost efficiency and confidence, but does not automatically improve correctness unless the model is reliably accurate and the interface is designed to support proper reliance.

Given the small sample and noisy model performance on this subset, the experiment highlights the importance of:

- Calibrated AI explanations
- Adaptive decision support
- And careful UI design to prevent overconfidence when AI is unreliable

Future work should measure actual AI usage behavior, include more participants, and experiment with richer explanation formats to improve human decision-making.