Let The Jury Decide: Fair Demonstration Selection for In-Context Learning through Incremental Greedy Evaluation

Sadaf MD Halim, Chen Zhao, Xintao Wu, Latifur Khan, Christan Earl Grant, Fariha Ishrat Rahman, Feng Chen

Fair In-Context Learning

Pretrained LLMs often encode demographic, societal, or linguistic preferences.

These issues can cause:

- Toxicity
- Stereotypical completions
- Irregularities in classification tasks

In-Context Learning (ICL)

- ICL enables models to learn from a few examples at inference time.
- No fine-tuning required – examples are provided as part of the input prompt.

Used in:

- - Few-shot classification
- - Question answering
- - Summarization

- Input Prompt = {Demo 1, Demo 2, …, Demo k} + Test Query → Model Prediction

Why Demonstration Selection Is Critical

- Demonstrations are the only supervision LLMs receive during inference.
- Poor selection can amplify irresponsible outputs and reduce accuracy.

Impacts of demo selection:

- Prediction accuracy
- Generalization across subgroups
- Unfair outcomes for certain groups

Responsible In-Context Classification

Goal: Maximize metrics like Demographic Parity, Equalized Odds ⚖️

Challenges:

- - Select k (typically 5 or 10) from n demonstrations (where n is a large number)
- - Ensure responsible outputs
- - Retain predictive utility (accuracy, F1)

JUDGE (**JU**ry-based **D**emonstration Selection via **G**reedy **E**valuation)

Our approach, JUDGE addresses demonstration selection through a multi-step process.

1. **Jury Set Selection:** Creating a set of examples for greedy evaluation.
2. **Candidate Pruning:** Reducing the pool of candidates to a much smaller pool.
3. **Iterative Greedy Selection:** Building the final demonstration set greedily by using performance on the jury set as a heuristic.

The Jury Set, $\mathcal{J}$

A carefully constructed group of examples, providing a balanced representation across all combinations of groups and labels.

$$\mathcal{C} = \{(g, y) : g \in \mathcal{G}, y \in \mathcal{Y}\}$$

Each subset $\mathcal{J}_{(g,\, y)}$ consists of $|\mathcal{J}| / |C|$ examples

Uses SentenceBERT embeddings and the cosine similarity measure:

$$\text{sim}(x_i, x_j) = \frac{e(x_i) \cdot e(x_j)}{\|e(x_i)\| \|e(x_j)\|}$$

The Jury Set, $\mathcal{J}$

Each example is chosen such that it maximizes distance from existing examples.

$$\mathcal{J}_{g,y} = \{x_1, \ldots, x_m\} \text{ where}$$
$$x_i = \arg \min_{x \in \mathcal{D}_{g,y} \setminus \{x_1, \ldots, x_{i-1}\}} \max_{j<i} \text{sim}(x, x_j)$$

Finally, we have:

$$\mathcal{J} = \bigcup_{(g,y) \in \mathcal{C}} \mathcal{J}_{g,y}$$

Candidate Pruning

We similarly prune the space of candidates (down to ~3%)

$$\mathcal{D}_{reduced} = \{x_1, \ldots, x_n\} \text{ where}$$

$$x_i = \arg \min_{x \in \mathcal{D}_{candidate} \setminus \{x_1, \ldots, x_{i-1}\}} \max_{j<i} \text{sim}(x, x_j)$$

Objective and Score Functions

Our objective provides a balance between performance in terms of accuracy (denoted by a) and metrics like Demographic Parity (denoted by f)

$$\mathcal{S}^* = \text{argmax}_{\mathcal{S} \subseteq \mathcal{D}_{reduced}, |\mathcal{S}|=k} \, \text{score}(\mathcal{S}, \mathcal{J})$$

$$\text{score}(\mathcal{S}, \mathcal{J}) = \omega \cdot f(\mathcal{S}, \mathcal{J}) + (1 - \omega) \cdot a(\mathcal{S}, \mathcal{J})$$

Greedy Selection

1. We start with the empty set, $S_0 = \emptyset$
2. Select the first example that maximizes score on the jury set.

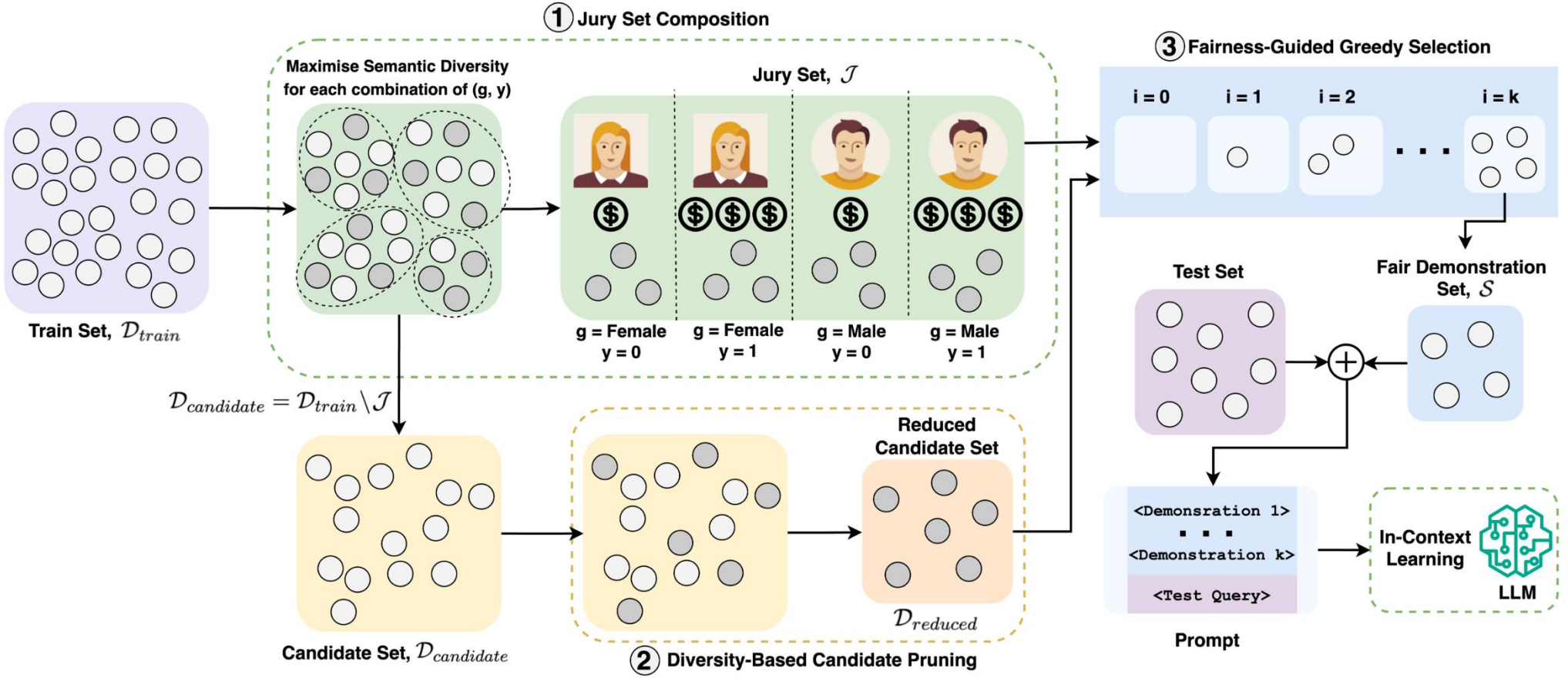
$$x_1 = \arg \max_{x \in \mathcal{D}_{\text{reduced}}} \text{score}(\{x\}, \mathcal{J})$$

1. Add the selected example to the demonstration set. $S_1 = \{x_1\}$
2. At each step t, select the candidate that maximizes score when added to the current set.

$$x_t = \arg \max_{x \in \mathcal{D}_{\text{reduced}} - S_{t-1}} \text{score}(S_{t-1} \cup \{x\}, \mathcal{J})$$

1. Update the selected set with the newly chosen example. $S_t = S_{t-1} \cup \{x_t\}$
2. Repeat the process until k examples are selected. $|S_t| = k$

Overview


① Jury Set Composition
Maximise Semantic Diversity for each combination of (g, y)
Jury Set, $\mathcal{J}$
Train Set, $\mathcal{D}_{train}$
g = Female y = 0
g = Female y = 1
g = Male y = 0
g = Male y = 1
$\mathcal{D}_{candidate} = \mathcal{D}_{train} \setminus \mathcal{J}$
Candidate Set, $\mathcal{D}_{candidate}$
② Diversity-Based Candidate Pruning
Reduced Candidate Set
$\mathcal{D}_{reduced}$
③ Fairness-Guided Greedy Selection
i = 0
i = 1
i = 2
i = k
Test Set
Fair Demonstration Set, $\mathcal{S}$
<Demonsration 1>
<Demonstration k>
<Test Query>
Prompt
In-Context Learning
LLM

Datasets and Baselines

Datasets:

Adult
COMPAS
Law School
ACS-Income

Baselines:

Random
Balanced
Counterfactual
Instruct
FairICL
FCG
FADS

Results

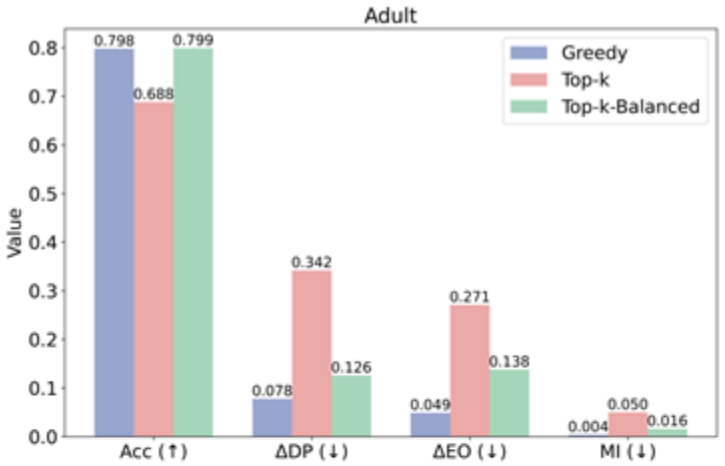
Table 1: Results for Adult with 5 demonstrations, across 4 LLMs. Each cell shows $Mean_{S.D.}$

	Method	Acc. ↑	ΔDP ↓	ΔEO ↓	MI ↓
LLAMA-3-8B	Random	$0.772_{0.008}$	$0.185_{0.004}$	$0.191_{0.006}$	$0.023_{0.002}$
	Balanced	$0.706_{0.015}$	$0.216_{0.011}$	$0.146_{0.014}$	$0.022_{0.001}$
	Cfact.	$0.731_{0.017}$	$0.185_{0.019}$	$0.158_{0.023}$	$0.018_{0.003}$
	Instruct	$0.753_{0.013}$	$0.299_{0.011}$	$0.308_{0.012}$	$0.052_{0.006}$
	FairICL	$0.764_{0.009}$	$0.170_{0.004}$	$0.097_{0.008}$	$0.016_{0.002}$
	FCG	$0.795_{0.011}$	$0.097_{0.009}$	$0.157_{0.006}$	$0.011_{0.001}$
	FADS	$0.743_{0.015}$	$0.157_{0.012}$	$0.114_{0.014}$	$0.019_{0.003}$
	JUDGE	$\mathbf{0.798}_{0.012}$	$\mathbf{0.078}_{0.011}$	$\mathbf{0.049}_{0.012}$	$\mathbf{0.004}_{0.001}$
MISTRAL-7B	Random	$0.709_{0.013}$	$0.201_{0.010}$	$0.124_{0.009}$	$0.019_{0.003}$
	Balanced	$0.594_{0.014}$	$0.230_{0.011}$	$0.185_{0.012}$	$0.025_{0.004}$
	Cfact.	$0.722_{0.011}$	$0.143_{0.008}$	$0.193_{0.013}$	$0.011_{0.003}$
	Instruct	$0.729_{0.021}$	$0.162_{0.019}$	$0.171_{0.023}$	$0.015_{0.004}$
	FairICL	$0.761_{0.006}$	$0.151_{0.011}$	$0.159_{0.007}$	$0.012_{0.002}$
	FCG	$0.752_{0.015}$	$0.132_{0.014}$	$0.093_{0.019}$	$\mathbf{0.006}_{0.001}$
	FADS	$\mathbf{0.769}_{0.009}$	$0.180_{0.008}$	$0.129_{0.005}$	$0.021_{0.002}$
	JUDGE	$0.767_{0.012}$	$\mathbf{0.101}_{0.009}$	$\mathbf{0.024}_{0.005}$	$\mathbf{0.006}_{0.001}$
GEMMA-2-9B	Random	$0.754_{0.006}$	$0.394_{0.008}$	$0.423_{0.013}$	$0.091_{0.005}$
	Balanced	$0.701_{0.014}$	$0.482_{0.023}$	$0.413_{0.026}$	$0.113_{0.021}$
	Cfact.	$0.752_{0.015}$	$0.311_{0.015}$	$0.372_{0.011}$	$0.087_{0.016}$
	Instruct	$0.742_{0.011}$	$0.428_{0.009}$	$0.479_{0.013}$	$0.108_{0.008}$
	FairICL	$0.753_{0.014}$	$0.318_{0.019}$	$0.392_{0.026}$	$0.089_{0.013}$
	FCG	$0.755_{0.017}$	$0.233_{0.025}$	$0.192_{0.018}$	$\mathbf{0.013}_{0.003}$
	FADS	$0.759_{0.013}$	$0.353_{0.011}$	$0.387_{0.016}$	$0.072_{0.006}$
	JUDGE	$\mathbf{0.769}_{0.012}$	$\mathbf{0.177}_{0.018}$	$\mathbf{0.101}_{0.009}$	$0.018_{0.003}$
QWEN-2.5-32B	Random	$0.745_{0.012}$	$0.215_{0.010}$	$0.132_{0.010}$	$0.023_{0.004}$
	Balanced	$0.708_{0.014}$	$0.245_{0.013}$	$0.165_{0.012}$	$0.027_{0.003}$
	Cfact.	$0.748_{0.014}$	$0.225_{0.014}$	$0.143_{0.011}$	$0.025_{0.003}$
	Instruct	$0.733_{0.007}$	$0.239_{0.013}$	$0.161_{0.009}$	$0.026_{0.005}$
	FairICL	$0.743_{0.009}$	$0.192_{0.012}$	$0.147_{0.015}$	$0.027_{0.009}$
	FCG	$0.762_{0.013}$	$0.111_{0.014}$	$0.098_{0.013}$	$0.007_{0.002}$
	FADS	$0.712_{0.009}$	$0.220_{0.007}$	$0.141_{0.006}$	$0.023_{.003}$
	JUDGE	$\mathbf{0.771}_{0.008}$	$\mathbf{0.096}_{0.005}$	$\mathbf{0.062}_{0.004}$	$\mathbf{0.005}_{0.001}$

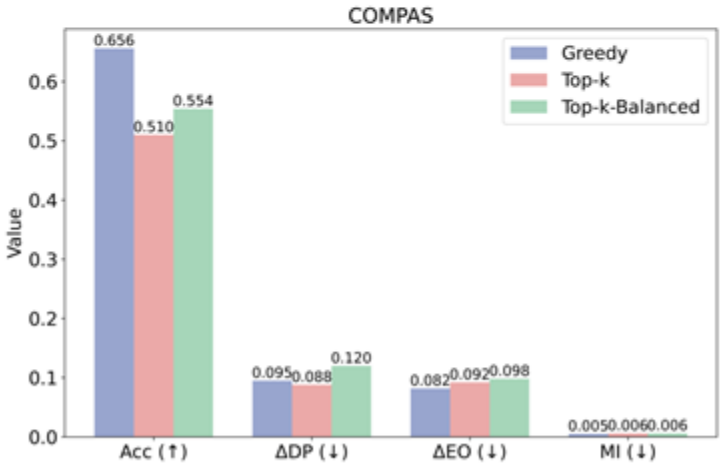
Table 2: Results for COMPAS with 5 demonstrations, across 4 LLMs. Each cell shows $Mean_{S.D.}$

	Method	Acc. ↑	ΔDP ↓	ΔEO ↓	MI ↓
LLAMA-3-8B	Random	$0.617_{0.011}$	$0.209_{0.009}$	$0.199_{0.008}$	$0.021_{0.003}$
	Balanced	$0.620_{0.012}$	$0.235_{0.011}$	$0.218_{0.013}$	$0.027_{0.002}$
	Cfact.	$0.582_{0.009}$	$0.187_{0.006}$	$0.193_{0.007}$	$0.017_{0.001}$
	Instruct	$0.566_{0.010}$	$0.135_{0.009}$	$0.164_{0.010}$	$0.015_{0.001}$
	FairICL	$0.621_{0.009}$	$0.192_{0.007}$	$0.188_{0.006}$	$0.020_{0.002}$
	FCG	$0.614_{0.007}$	$0.182_{0.005}$	$0.197_{0.005}$	$0.019_{0.001}$
	FADS	$0.575_{0.008}$	$0.167_{0.006}$	$0.160_{0.005}$	$0.014_{0.002}$
	JUDGE	$\mathbf{0.656}_{0.010}$	$\mathbf{0.105}_{0.008}$	$\mathbf{0.082}_{0.007}$	$\mathbf{0.006}_{0.001}$
MISTRAL-7B	Random	$0.513_{0.012}$	$0.097_{0.008}$	$0.120_{0.009}$	$0.016_{0.002}$
	Balanced	$0.512_{0.007}$	$0.079_{0.005}$	$0.083_{0.004}$	$0.013_{0.003}$
	Cfact.	$0.487_{0.010}$	$0.059_{0.009}$	$\mathbf{0.062}_{0.009}$	$0.015_{0.004}$
	Instruct	$0.497_{0.012}$	$0.082_{0.010}$	$0.105_{0.008}$	$0.014_{0.002}$
	FairICL	$0.515_{0.006}$	$0.082_{0.005}$	$0.098_{0.005}$	$0.017_{0.004}$
	FCG	$0.489_{0.009}$	$0.074_{0.004}$	$0.108_{0.006}$	$0.013_{0.003}$
	FADS	$0.531_{0.010}$	$0.091_{0.005}$	$0.117_{0.007}$	$0.015_{0.009}$
	JUDGE	$\mathbf{0.541}_{0.007}$	$\mathbf{0.055}_{0.004}$	$0.075_{0.004}$	$\mathbf{0.002}_{0.000}$
GEMMA-2-9B	Random	$0.615_{0.008}$	$0.310_{0.005}$	$0.314_{0.006}$	$0.049_{0.003}$
	Balanced	$0.601_{0.009}$	$0.359_{0.006}$	$0.348_{0.005}$	$0.067_{0.004}$
	Cfact.	$0.604_{0.007}$	$0.261_{0.004}$	$0.272_{0.005}$	$0.044_{0.005}$
	Instruct	$0.609_{0.011}$	$0.291_{0.009}$	$0.309_{0.012}$	$0.047_{0.006}$
	FairICL	$0.622_{0.010}$	$0.265_{0.011}$	$0.282_{0.012}$	$0.040_{0.005}$
	FCG	$0.648_{0.007}$	$0.099_{0.003}$	$0.091_{0.005}$	$0.008_{0.003}$
	FADS	$0.621_{0.014}$	$0.307_{0.011}$	$0.303_{0.09}$	$0.053_{0.009}$
	JUDGE	$\mathbf{0.665}_{0.006}$	$\mathbf{0.062}_{0.002}$	$\mathbf{0.039}_{0.003}$	$\mathbf{0.002}_{0.000}$
QWEN-2.5-32B	Random	$0.637_{0.007}$	$0.242_{0.005}$	$0.221_{0.006}$	$0.029_{0.003}$
	Balanced	$\mathbf{0.652}_{0.008}$	$0.248_{0.007}$	$0.240_{0.011}$	$0.031_{0.005}$
	Cfact.	$0.611_{0.008}$	$0.244_{0.006}$	$0.228_{0.006}$	$0.031_{0.004}$
	Instruct	$0.633_{0.006}$	$0.234_{0.003}$	$0.214_{0.004}$	$0.026_{0.002}$
	FairICL	$0.639_{0.008}$	$0.211_{0.005}$	$0.218_{0.005}$	$0.025_{0.003}$
	FCG	$0.623_{0.006}$	$0.149_{0.004}$	$0.144_{0.003}$	$0.018_{0.003}$
	FADS	$0.645_{0.008}$	$0.224_{0.006}$	$0.207_{0.004}$	$0.025_{0.003}$
	JUDGE	$0.649_{0.004}$	$\mathbf{0.138}_{0.005}$	$\mathbf{0.134}_{0.003}$	$\mathbf{0.010}_{0.001}$

Greedy vs Top-K



(a) Adult dataset



(b) COMPAS dataset

Effect of Jury Set Size