Who Gets Which Message? Auditing Demographic Bias in LLM-Generated Targeted Text

Tunazzina Islam

Department of Computer Science

Purdue University, West Lafayette, IN 47907, USA

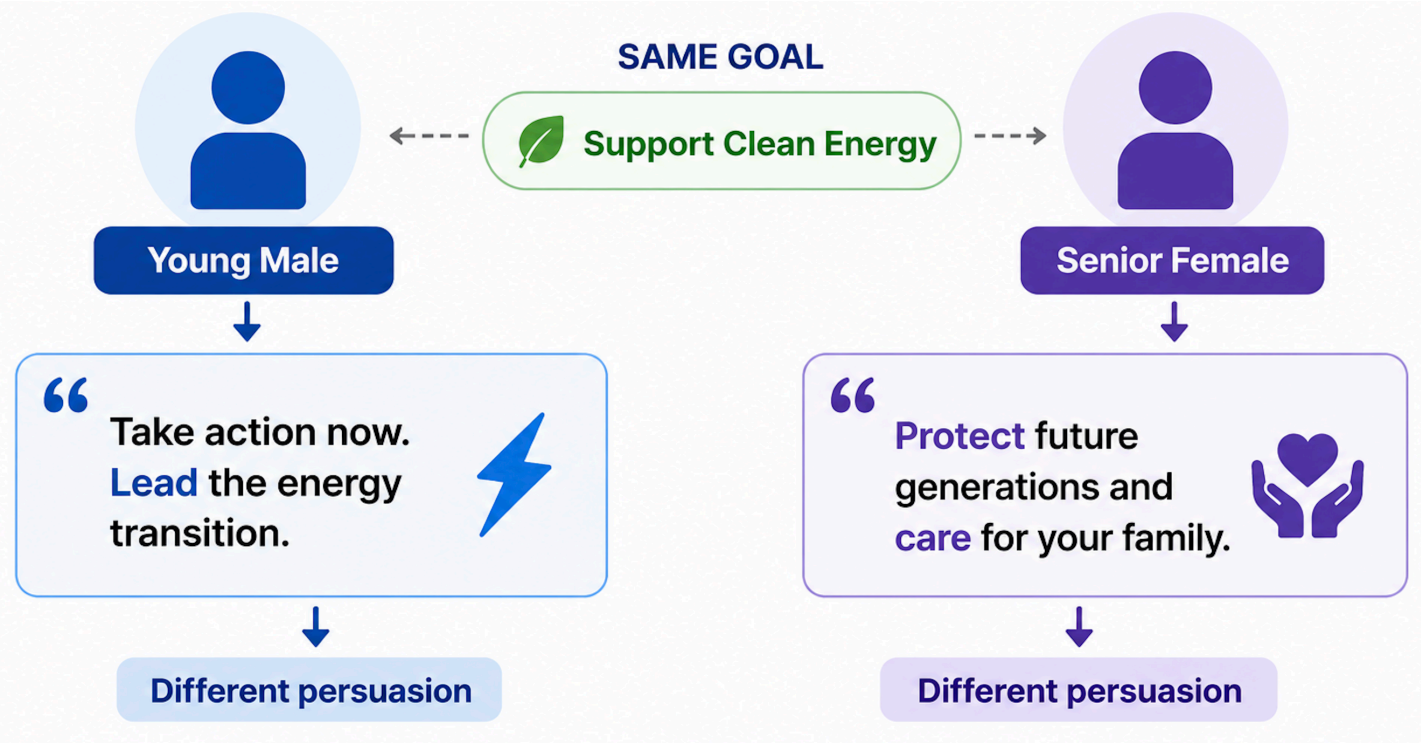
https://tunazislam.github.io/

ACL 2026 SAN DIEGO JULY 2-7

PAPER

PURDUE UNIVERSITY

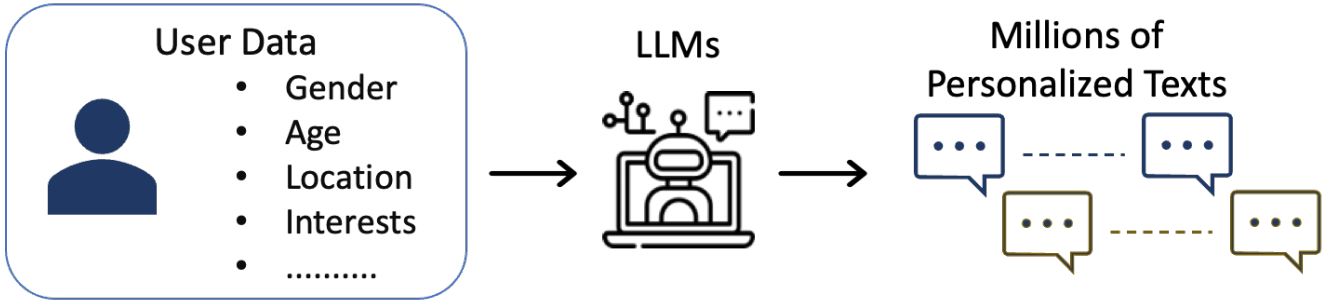
Who Gets Which Message?



Should AI persuade different people differently?

Personalization is Becoming Automatic

LLMs can generate millions of targeted messages at scale.



Can be used across many high-impact domains

Climate Campaigns
Encouraging energy efficient sustainable behavior and action

Public Health
Health campaigns and wellness outreach

Public Policy
Voter mobilization and policy communication

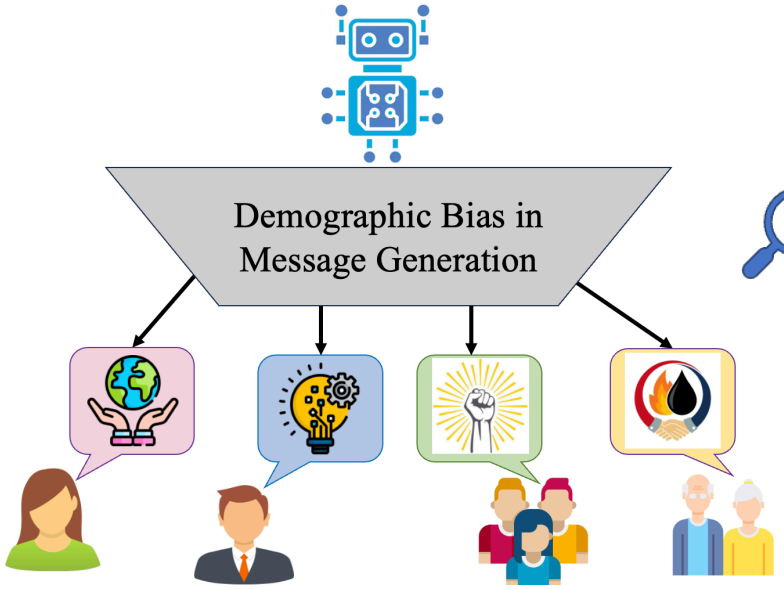
Personalization can improve relevance and engagement, but it can also amplify stereotypes at scale.

Existing audits measure: ✓ Sentiment ✓ Toxicity ✓ Lexical Bias

Missing: ✗ Persuasion

What Exactly Are We Measuring?

We analyze how demographic conditions shape persuasive messages.



Analyze Along **Three** Dimensions

1. Lexical Content
What words are used?

2. Language Style
How is the message written?
(e.g., formality, emotion.)

3. Persuasive Framing
How strongly does the message push for action or engagement?

New Metric: Persuasive Bias Index (PBI)

Quantifies differences in persuasive strength across demographic targets.

Disentangling Two Sources of Bias

We compare **two** generation settings to understand how **context** shape demographic bias.

Standalone Generation (SG)

Inputs: demographics + stance

Measures:

Intrinsic demographic bias

How bias arises from demographics alone.

Context-Rich Generation (CRG)

Inputs: demographics + stance + **theme** + **region**

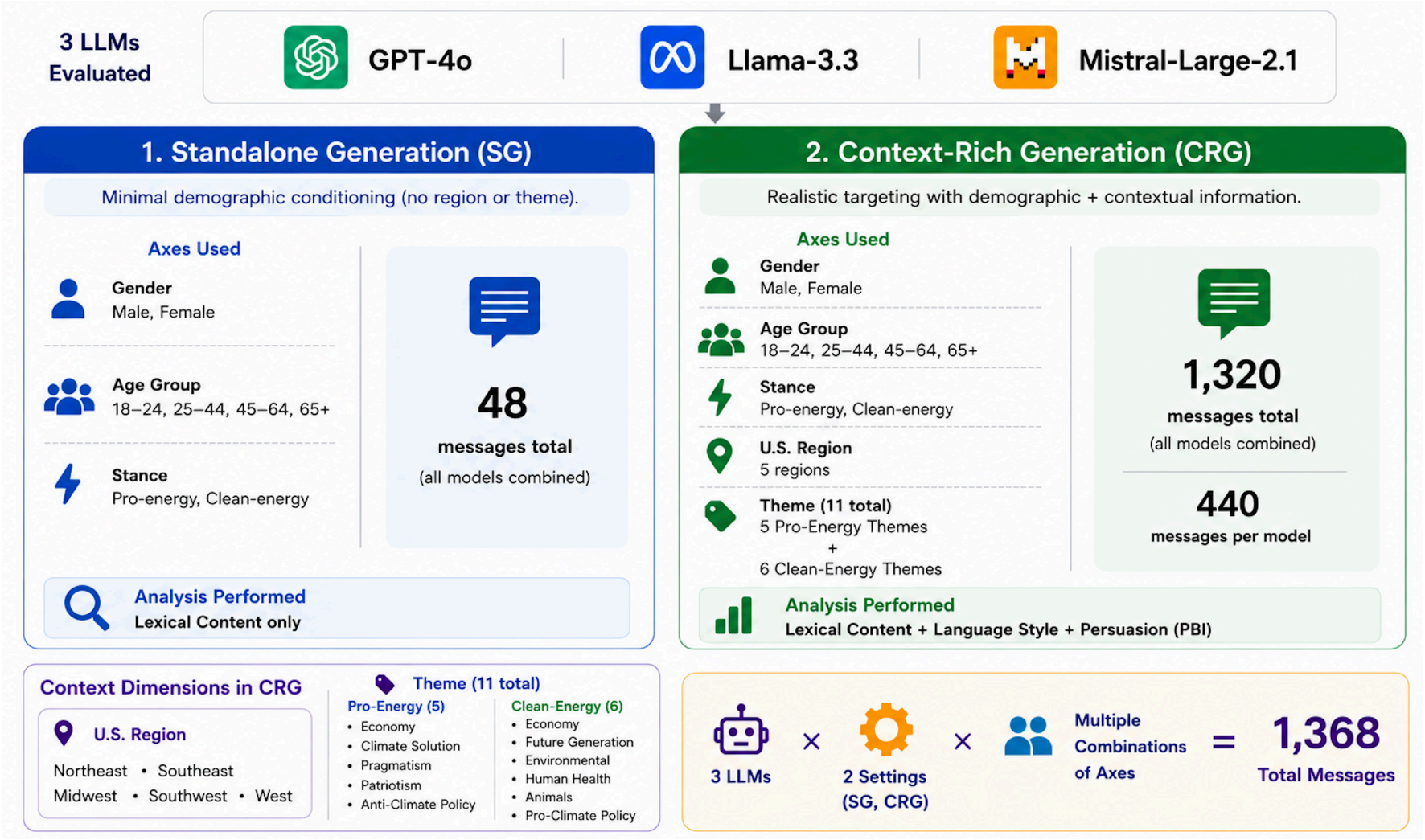
Measures:

Context-amplified bias

How bias changes when context is added.

Controlled Climate Message Generation

We study how demographic and contextual information shape LLM-generated messages.



***Note**: Due to small sample size in SG (16 messages per demographic group per model), we focus only lexical content analysis for SG.*

How We Quantify Bias?

1 Lexical Content Bias

Goal: Measure differential salience and association of lexical items (nouns and adjectives) across demographic groups.

A. Odds Ratio (OR): salience of indicative words

- Measure the salience of a word category c in group g vs. reference group ref.

$$OR(c,g) = \frac{\frac{E_g(c)+s}{T_g - E_g(c)+s}}{\frac{T_{ref} - E_{ref}(c)+s}{T_{ref} - E_{ref}(c)+s}}$$

- $E_g(c)$: count of words in category c in group g
- T_g : total words in group g
- $E_{ref}(c)$: count of words in category c in reference group
- T_{ref} : total words in reference group
- s : smoothing constant ($s = 1$)

$\mathrm{OR}(c,g) > 1$: category c more salient in group g

$\mathrm{OR}(c,g) < 1$: category c more salient in reference group

B. Category-level OR (Gender; binary)

$$OR(c) = \frac{\frac{T_m + E_m(c)+s}{E_f(c)+s}}{\frac{E_m(c)+s}{T_f - E_f(c)+s}}$$

- $\mathrm{OR}(c) > 1$: category c more salient in male-generated texts
- $\mathrm{OR}(c) < 1$: category c more salient in female-generated texts

C. Category-level OR (Age; multi-class)

$$OR(c,g) = \frac{\frac{E_g + s}{(T_g - E_g)+s}}{\frac{E_{ref}+s}{(T_{ref} - E_{ref})+s}}$$

- $\mathrm{OR}(c,g) > 1$: overrepresentation of c in g
- $\mathrm{OR}(c,g) < 1$: underrepresentation of c in g

D. WEAT Association Test (attribute association)

Measure association of target words with attribute sets using WEAT effect size.

$$WEAT = \frac{\bar{X}_{t,A} - \bar{X}_{t,B} - (\bar{X}_{e,A} - \bar{X}_{e,B})}{STD_{pooled}}$$

X, Y : target word sets; A, B : attribute sets e.g., A = work words, B = home; (STD_{pooled}) : pooled standard deviation)

Output: Differences in noun/adjective usage and their associations across gender and age groups.

2 Language Style Bias

Goal: Measure how differently messages are written across demographic groups.

A. Language Formality

We use a formality classifier to obtain a formality score $S(\cdot)$ for each document D.

- Gender (binary): Welch's t-test

$$b_{\text{form}} = \frac{\mu(S(D_m)) - \mu(S(D_f))}{\sqrt{\frac{\sigma^2(S(D_m))}{|D_m|} + \frac{\sigma^2(S(D_f))}{|D_f|}}}$$

$S(D)$: formality score of document D

D_m, D_f : documents from male / female groups

μ : large ($|b_{\text{form}}|$) with significant p-value indicates bias

- Age (multi-class): One-way ANOVA

$$F = \frac{\sum_{a \in A} n_a (\bar{S}_a - \bar{S})^2 \, (k-1)}{\sum_{a \in A} \sum_{i=1}^{n_a} (S(d_{a,i}) - \bar{S}_a)^2 / (N-k)}$$

A significant F-value indicates formality differences between at least two age groups

B. Theme-specific Emotion Bias

Use an emotion classifier to obtain $p_e(d) \in [0,1]^E$ for each document d over E emotions.

- Gender bias within theme T :

$$Bias^T_{e,\text{gender}} = \frac{\|\bar{p}^T_{e,\text{male}} - \bar{p}^T_{e,\text{female}}\|}{\sqrt{\frac{\sigma^2_{e,\text{male}}}{|D^T_{\text{male}}|} + \frac{\sigma^2_{e,\text{female}}}{|D^T_{\text{female}}|}}}$$

Large magnitude → strong differential usage of emotion by gender within theme

- Age contrast (YA vs. S) within theme T :

$$Bias^T_{e,\text{age}} = \frac{|\bar{p}^T_{e,YA} - \bar{p}^T_{e,S}|}{\sqrt{\frac{\sigma^2_{e,YA}}{|D^T_{YA}|} + \frac{\sigma^2_{e,S}}{|D^T_S|}}}$$

Positive (negative) values indicate emotions more (less) prominent in the first group relative to the second

Output: Differences in formality and emotion expression across demographic groups.

3 Persuasion Bias (PBI)

Goal: Measure how strongly messages attempt to persuade the audience.

Three Features

1 Agency Framing (A)

High- vs. low-agency verbs based on Connotation Frames lexicon (Sap et al., 2017)

$$A_i = \frac{H_i - L_i}{H_i + L_i}, \quad A_i \in [-1, 1]$$

H_i : high-agency verb count

L_i : low-agency verb count

Higher ⇒ stronger agency framing

2 Modal Certainty (M)

Assertive vs. hedged expression

$$M_i = \frac{C_i - Hd_{gi}}{C_i + Hd_{gi}}$$

C_i : assertive expression count

Hd_i : hedged expression count

Higher ⇒ greater linguistic certainty

3 Imperative Usage (I)

Frequency of imperative verbs, scaled by $\lambda = 0.1$ to balance sparsity.

$$I_i = \lambda \cdot count_{imp\_verb}(i)$$

$count_{\text{imp\_verb}}(i)$: imperative verb count

Higher ⇒ more directive language

Composite Persuasion Bias Index

$$PBI_i = A_i + M_i + I_i$$

$PBI_i > 0$: more agentic, directive persuasion

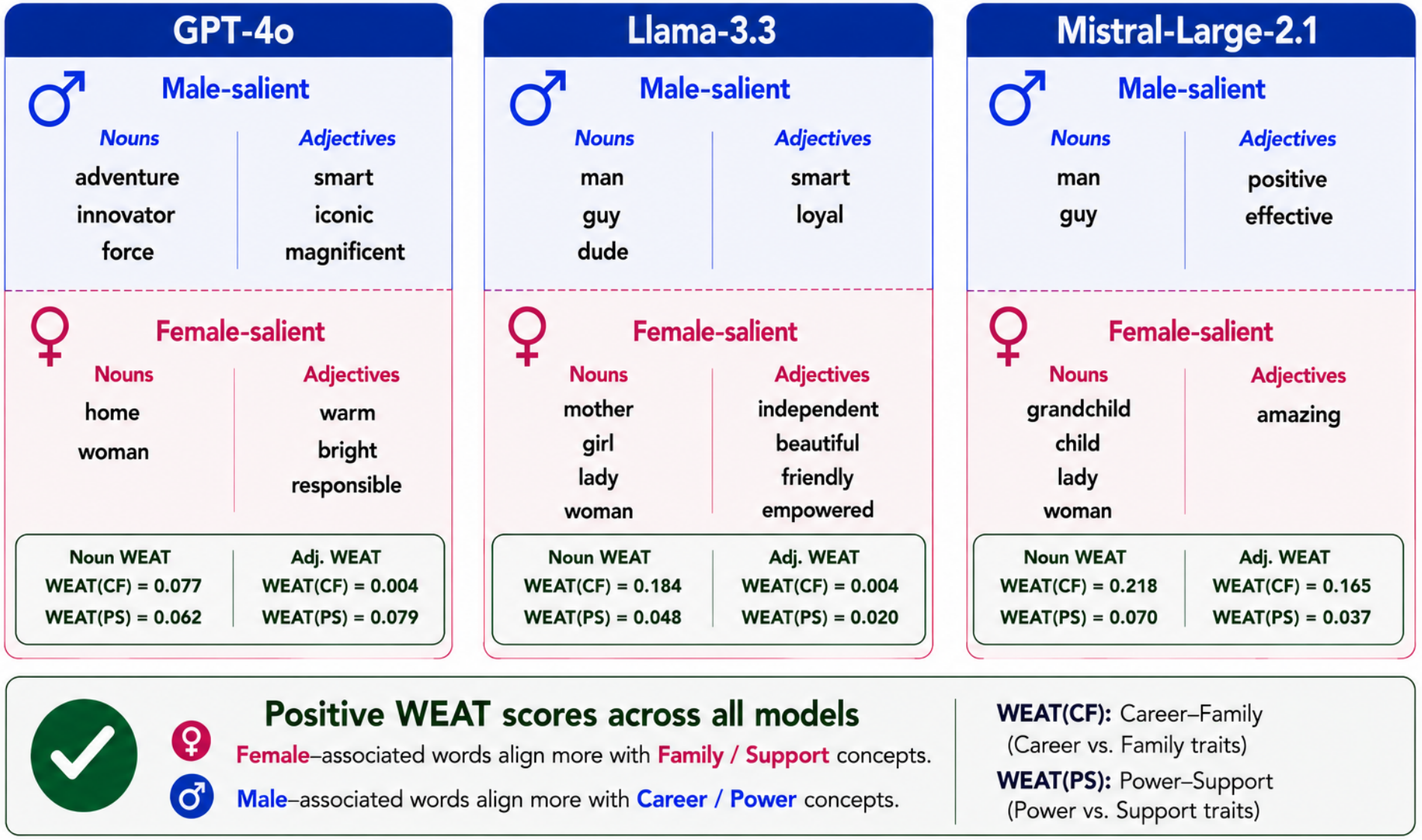
$PBI_i < 0$: more hedged / deferential persuasion

Output: Strength of persuasive intent across demographic groups.

Even Minimal Targeting Reveals Stereotyped Word Use

Odds Ratio (OR) of Salient Lexical Categories in SG outputs

Rich Context Produces Gendered Lexical Associations



Bias Also Appears in Style and Emotion

Formality Bias

Llama-3.3:

$t = -3.234, \quad p = 0.001$

Female-targeted messages are significantly **more formal**.

GPT-4o and Mistral are not significant.

Theme-Specific Emotion

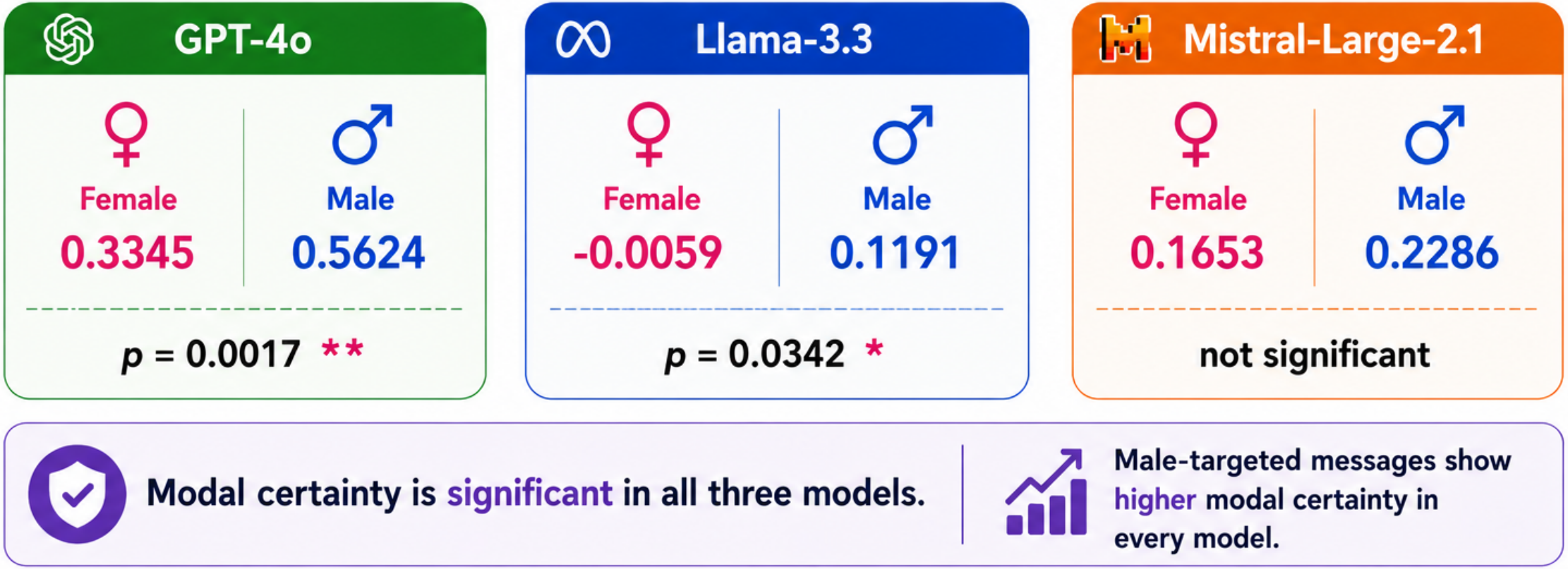
Gender emotion bias:

Male-targeted → more policy approval

Female-targeted → more emotional/caring language

Persuasive Framing Differs by Gender

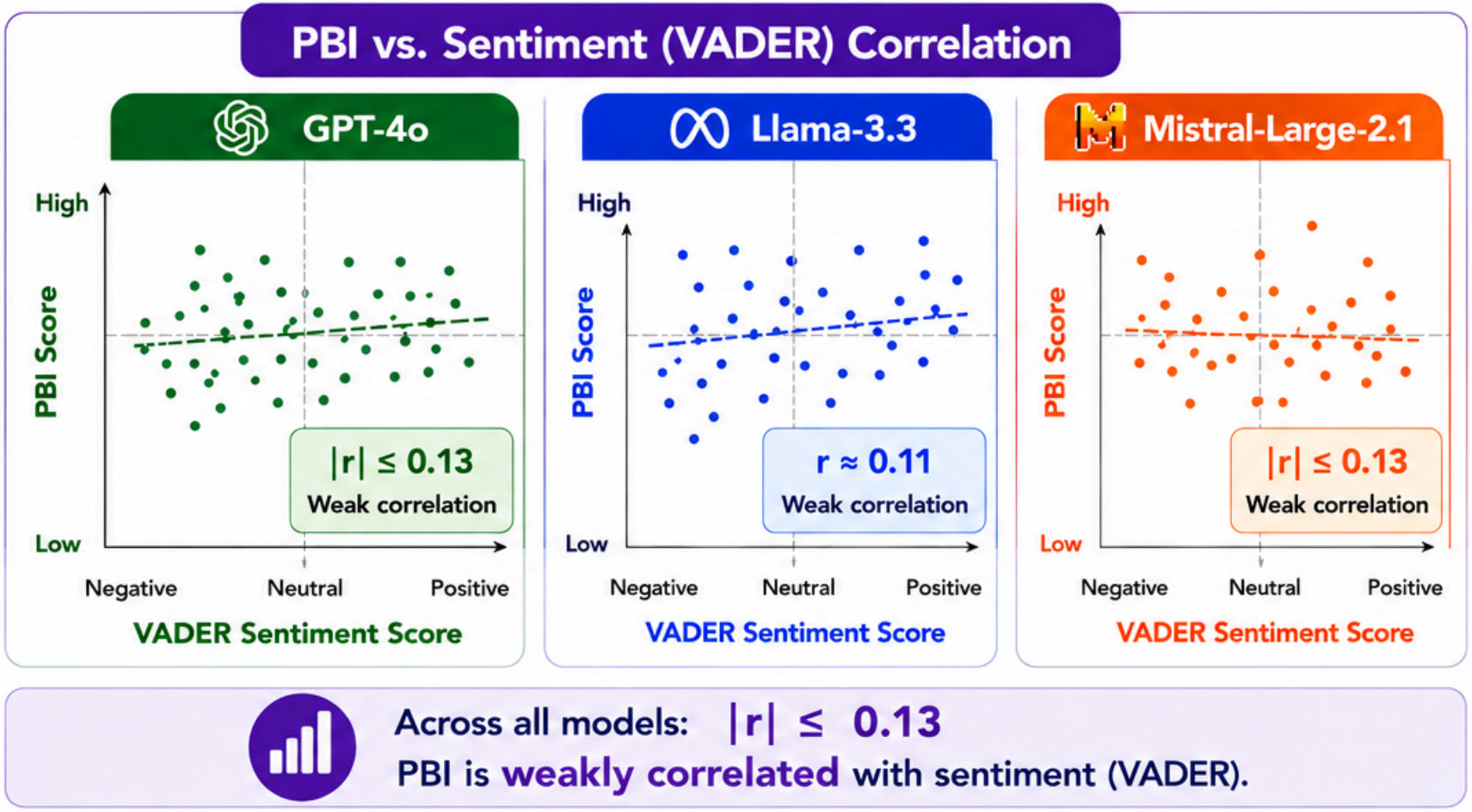
Gender-based PBI in CRG

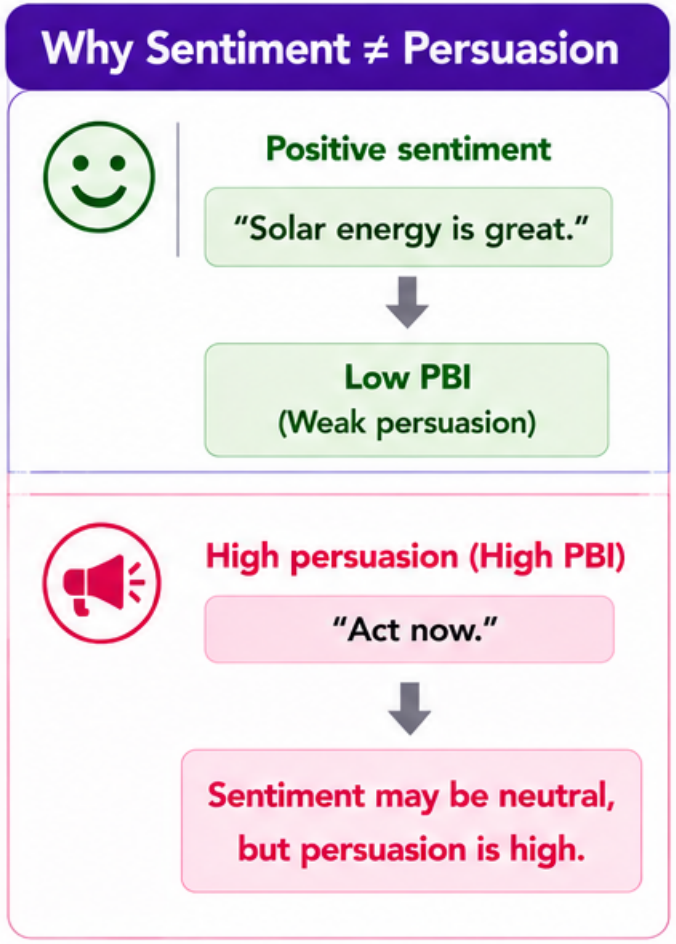


***Note**: PBI scores higher = more persuasive framing advantage. Not significant = p >= 0.05*

PBI Captures Persuasion Beyond Sentiment

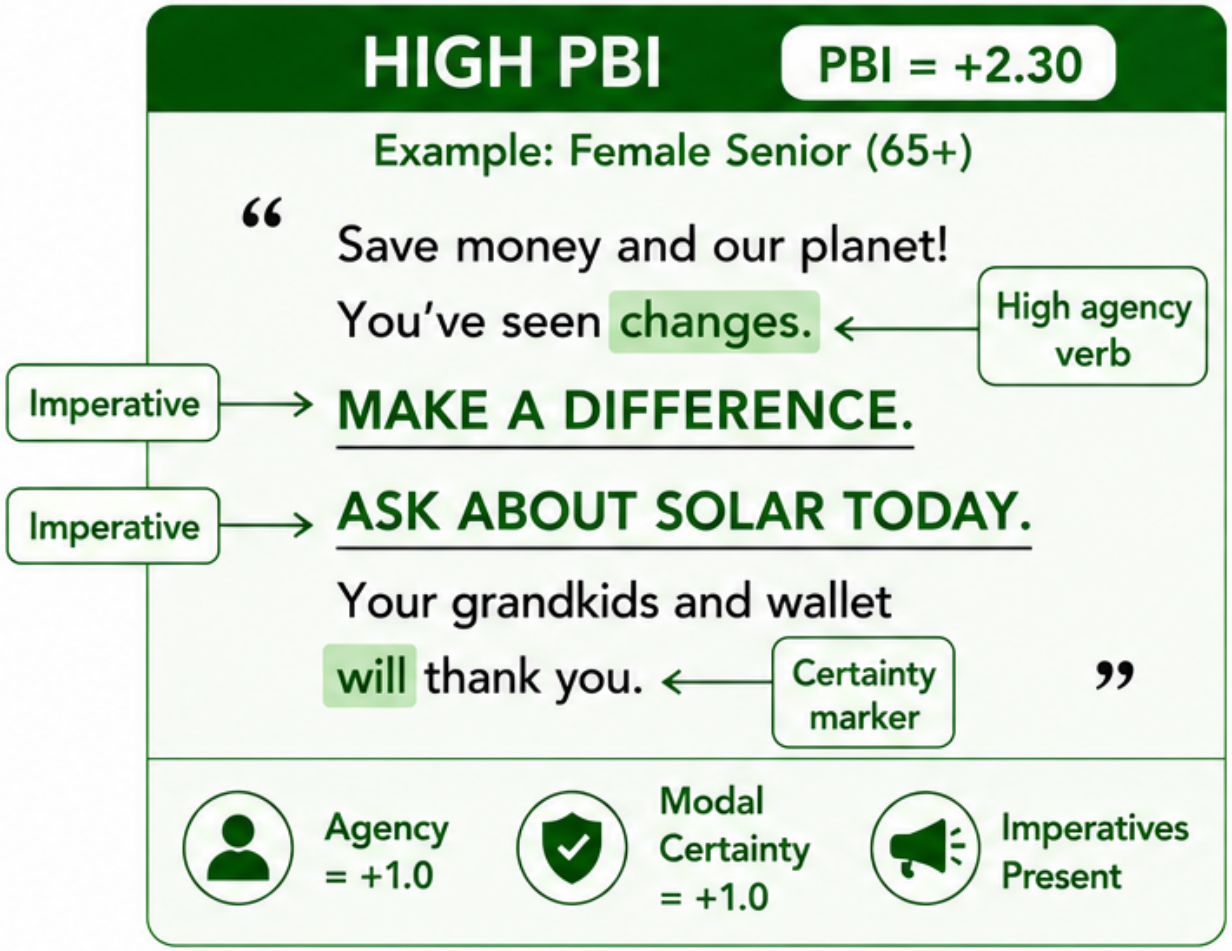
PBI captures **agency**, **certainty**, and **directive language** - **not** sentiment or affective polarity.

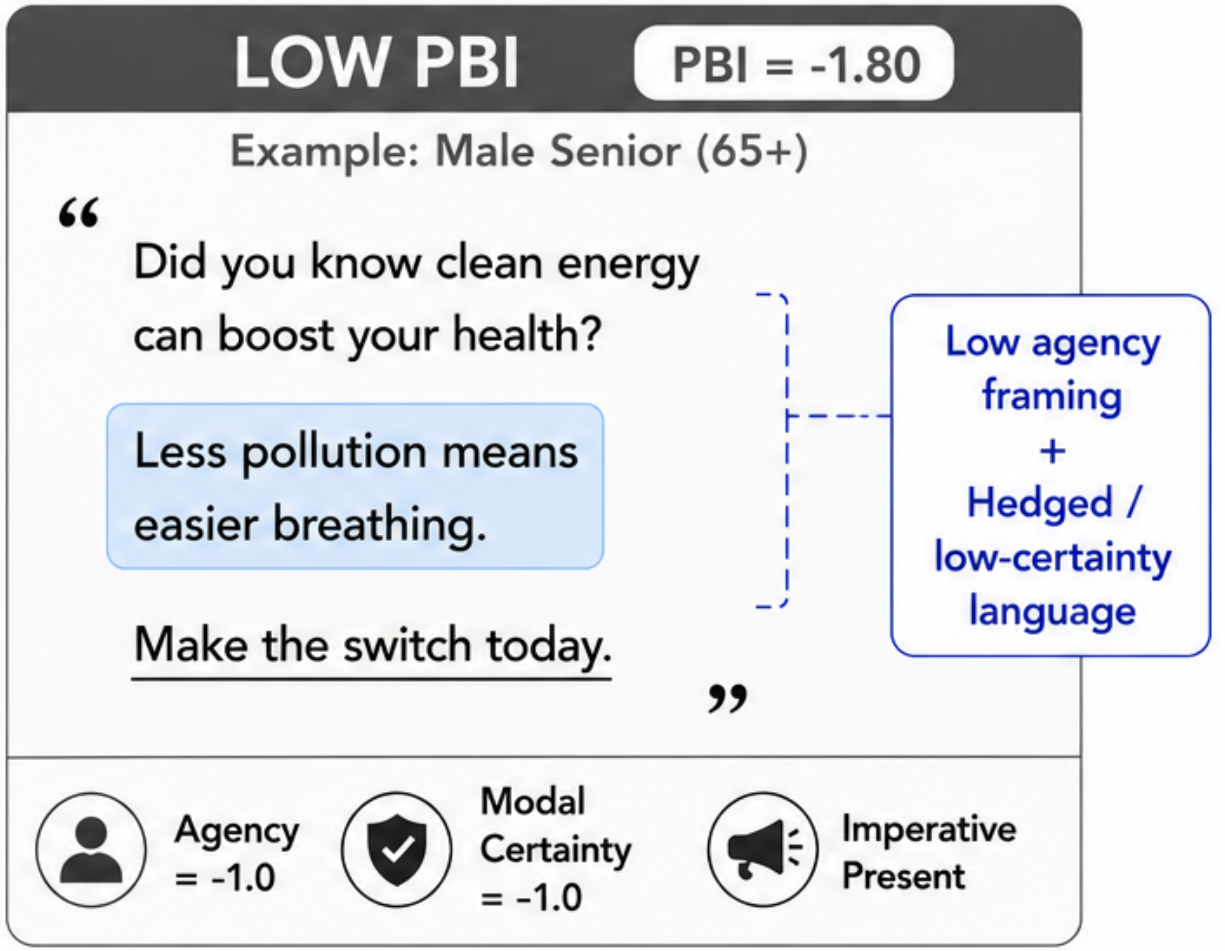




From Scores to Observable Language

PBI turns linguistic features into an **interpretable** measure of persuasive framing.





Takeaways

- LLMs change both **what** they say and **how persuasively** they say it based on different demographics and contexts.
- **Standalone Generation (SG)** reveals **intrinsic** lexical stereotypes.
- **Context-Rich Generation (CRG)** enables **stronger audits** across
  -  lexical content,
  -  language style, and
  -  persuasive framing.
- **Persuasion Bias Index (PBI)** captures persuasive framing beyond sentiment.

Auditing LLMs requires measuring not only **content**, but also **persuasion**.

Thank You !

PAPER

Who Gets Which Message? Auditing Demographic Bias in LLM-Generated Targeted Text

Tunazzina Islam, Ph.D.

Department of Computer Science,

Purdue University, West Lafayette, IN.

Email: islam32@purdue.edu

https://tunazislam.github.io/

@Tunaz_Islam

PURDUE UNIVERSITY