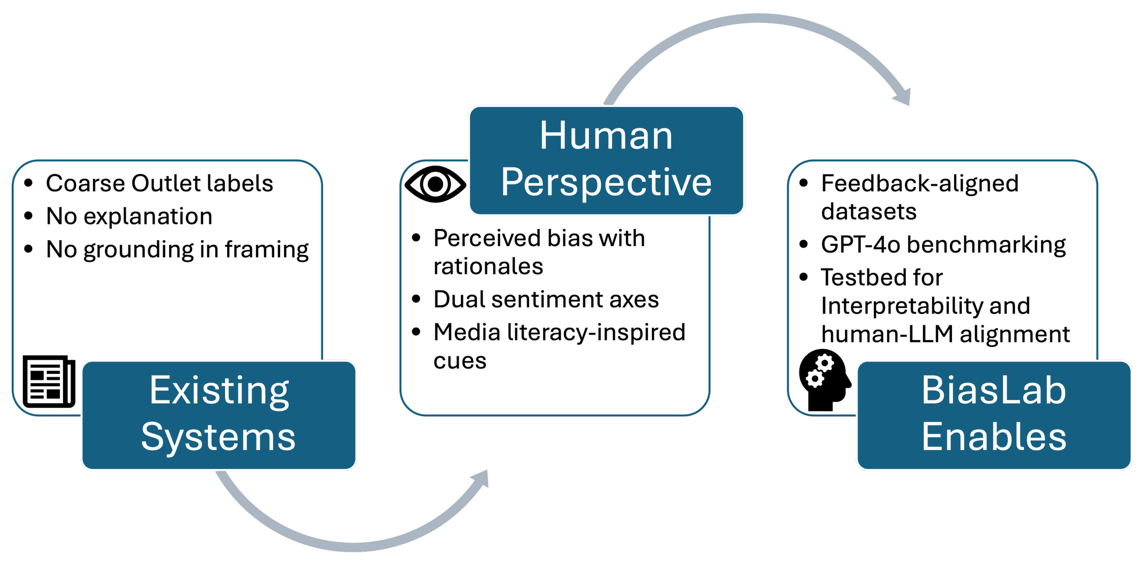
BiasLab: Explainable Political Bias Detection via Dual-Axis Human Annotations and Rationale Indicators

KMA Solaiman
Department of Computer Science and Electrical Engineering,
University of Maryland, Baltimore County (UMBC), Baltimore, Maryland, USA

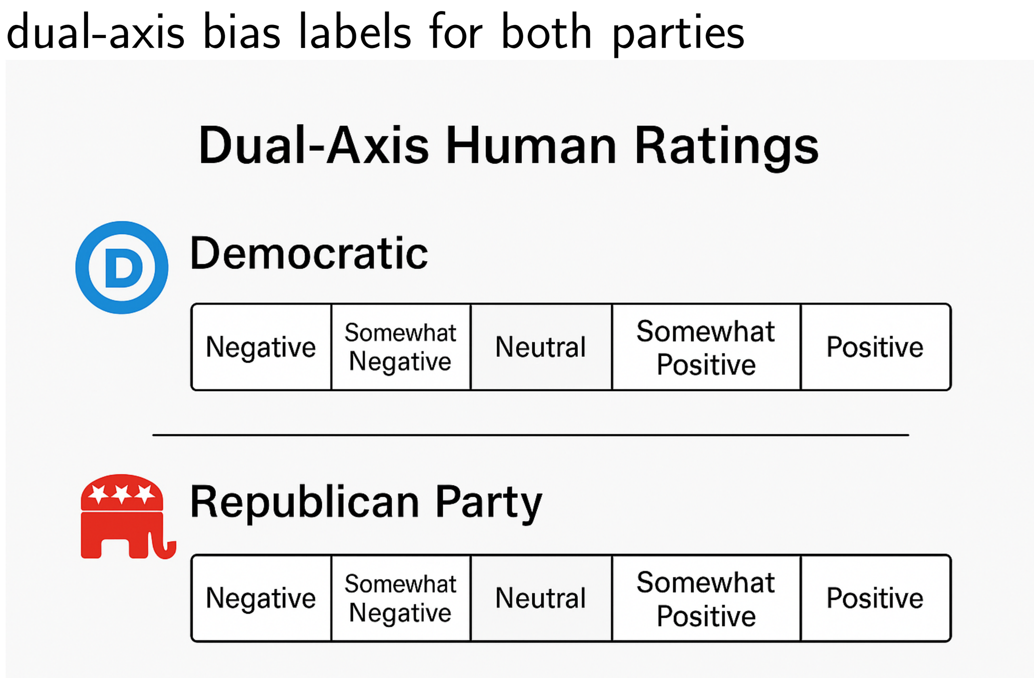
Motivation: From Coarse Labels to Perception Alignment



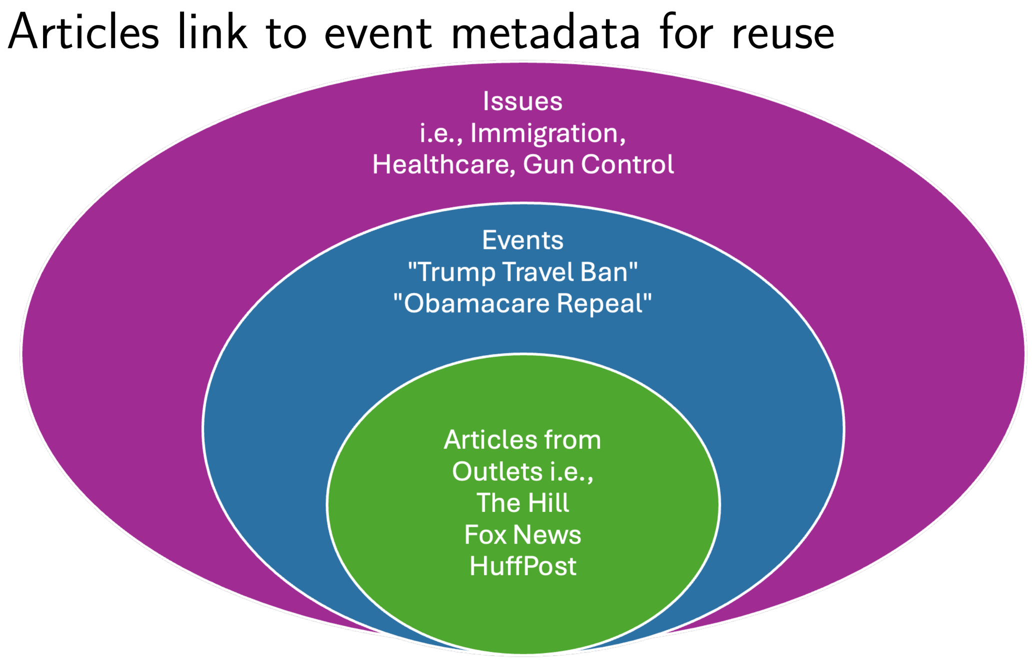
BiasLab captures *what* readers perceive and *why* to support human-LLM alignment.

Dataset Overview

- **900 partisan political articles** curated across major U.S. events (2016–2018)
- **300 articles annotated** via MTurk with dual-axis bias labels for both parties



- Each annotation also includes **bias rationale indicators** (e.g., *labeling, omission, framing*)
- Articles link to event metadata for reuse



Dataset structure: Articles are nested within events and issue categories.

- **Designed for** alignment, disagreement, and rationale modeling

How BiasLab Captures Perceived Bias

Example Annotation Entry

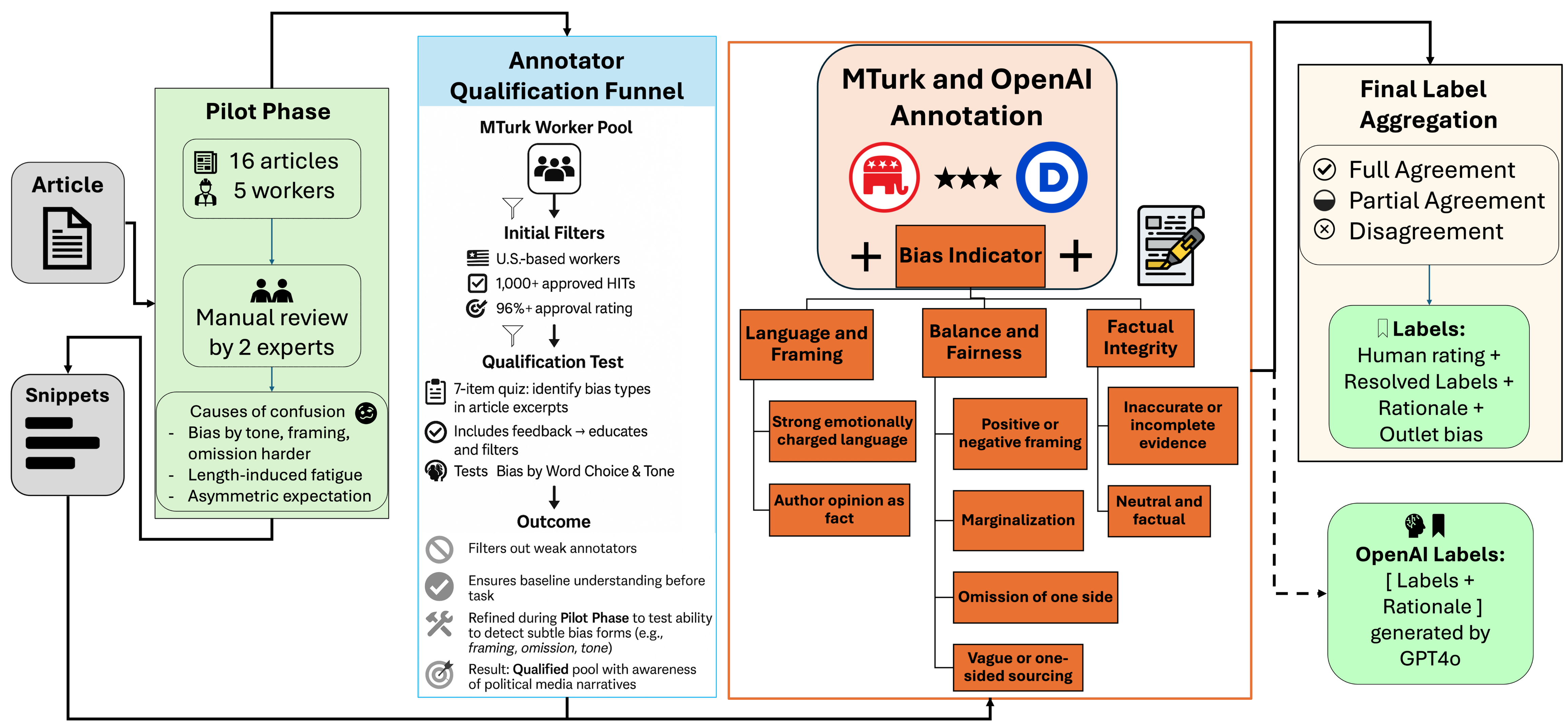
Title: *Anti-Trump celebs plan 'People's State of the Union'*
Event: President Trump will deliver his first State of the Union

Article Snippet (excerpt):
A group of **Hollywood elites***, progressive groups, and other Trump opponents are planning a "People's State of the Union" to counter the president's first official address. The event, coordinated by unions, organizers of the Women's March and Planned Parenthood, is being marketed as a celebration of the "resistance," closer to "the people's point of view," USA Today reported.*

Marked Bias Indicators:
- **Marginalization of one side** (Indicator 4): *"A group of Hollywood elites . . . celebration of the resistance"*
- **Emotionally charged language** (Indicator 0): *"Hollywood elites," "social activists," "public alternative"*

Worker Labels: Right, Right
Final Human Label: Right
Outlet Bias: Right

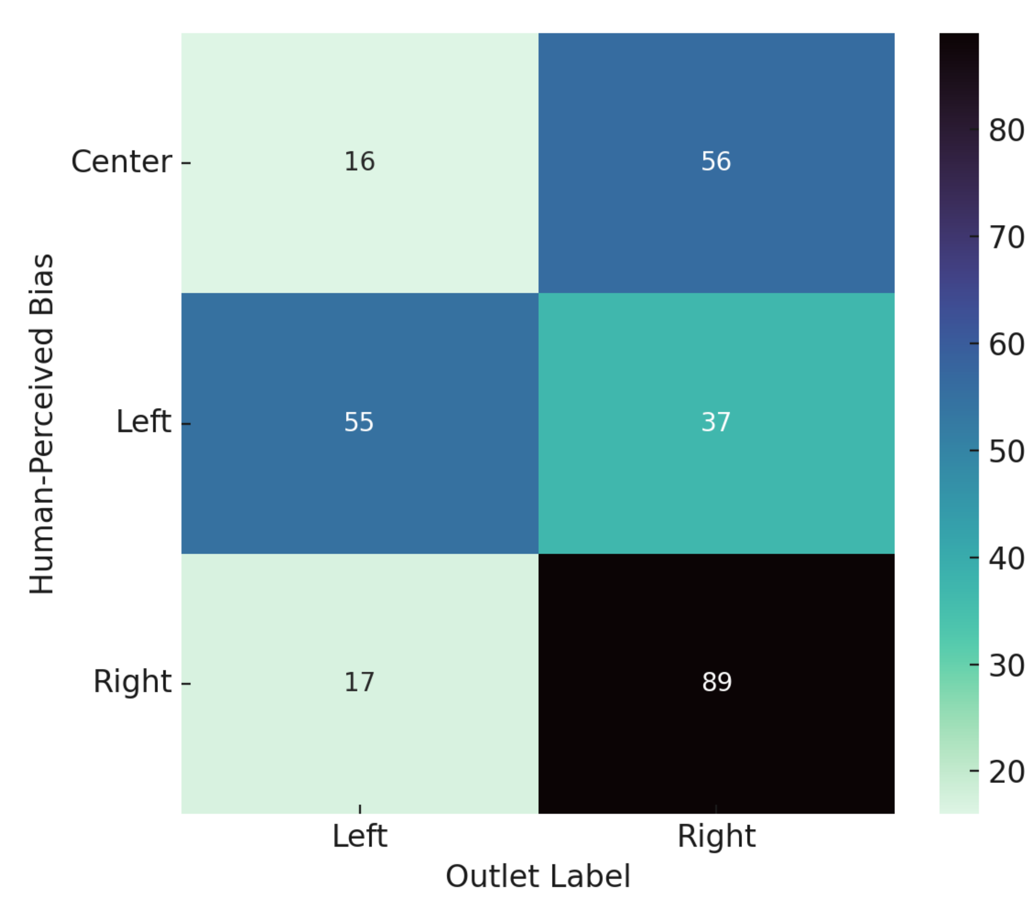
Annotation Pipeline



Pipeline overview: *Each article is split into snippets. Annotators rate tone toward both parties and select rationale indicators with highlighted text.*

Findings: Human Bias Perception

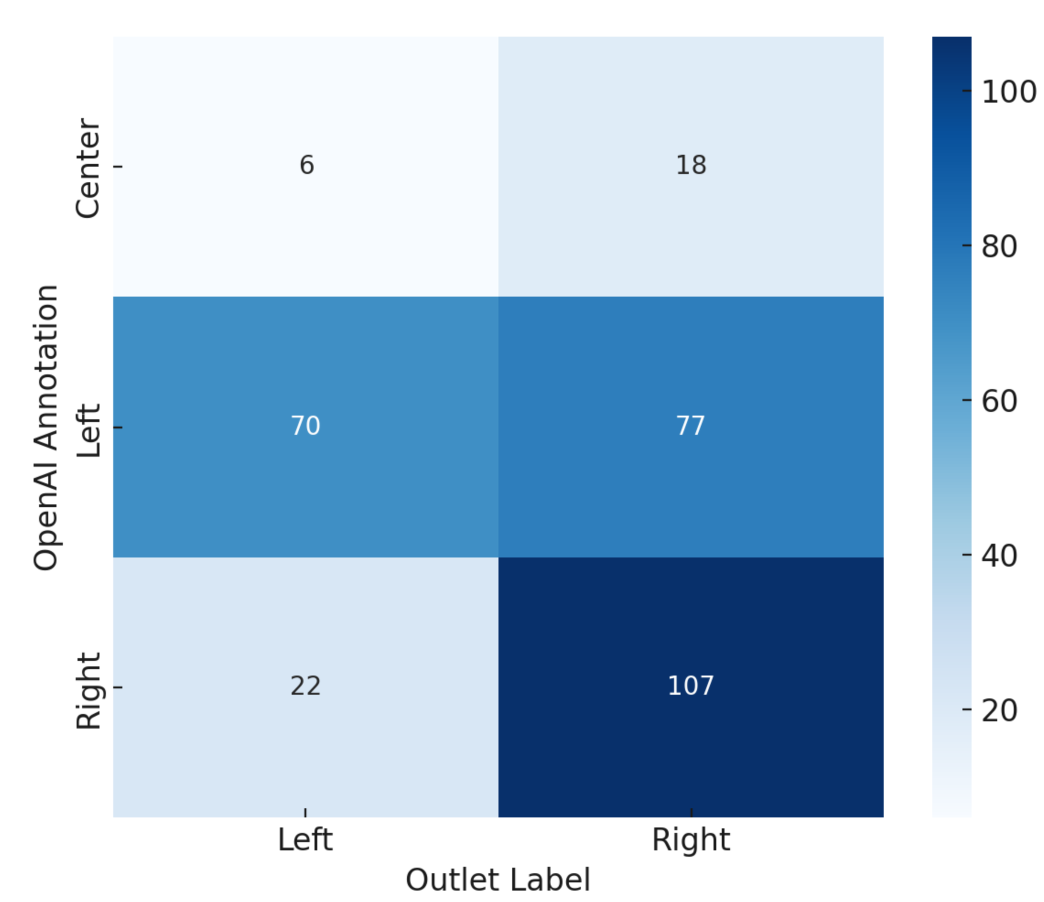
- Annotators underdetect right-leaning bias
- Agreement better on overt partisanship



Annotators often rate subtle right-leaning content as neutral - diverging from outlet bias.

Human vs GPT-4o Alignment

- GPT-4o achieved higher outlet-label agreement (59%) vs. human annotators (48%)



GPT-4o mirrors human bias misclassifications.

Feedback Alignment Tasks

Task 1: Perception Drift

- Can models detect when human-perceived bias **diverges** from outlet-level ideology?
- **Logistic Regression+TF-IDF**: 55.6% accuracy
- Perception drift is learnable, but very subtle

Task 2: Rationale Classification

- Can models learn to predict annotator-marked **rationale types** and relate to perceived bias?
- Human rationales as interpretable supervision
- Multi-label task over rationale types:
  1. **Directional** (Framing-dominated)
  2. **Structural** (Balance & Fairness and Factual)
  3. **Neutral**

Rationale Type	Precision	Recall	F1 Score
directional	0.62	0.54	0.51
structural	0.61	0.56	0.54
neutral_other	0.70	0.64	0.61

Structural and neutral rationales are more learnable than directional (e.g., emotionally charged language).

Key Takeaways

Perceived bias ≠ Outlet ideology
More prevalent for subtle right-leaning content

- **Snippet-level tone + Rationale** annotations help expose interpretive judgments
- GPT-4o mimics both strengths and blind spots in human bias judgment
- Structured annotations support **alignment** and **interpretability modeling**, not just classification

Usable for critique modeling, alignment feedback, explainability tasks, and temporal drift analysis.

Resources

- **Dataset:** DOI: 10.5281/zenodo.15571668
- **Paper**: `https://arxiv.org/abs/2505.16081`
- **Code:**

H.A.R.M.O.N.I. Lab
Human-Aligned, Resilient, Multimodal, Open-ended, Novelty-Informed Intelligence