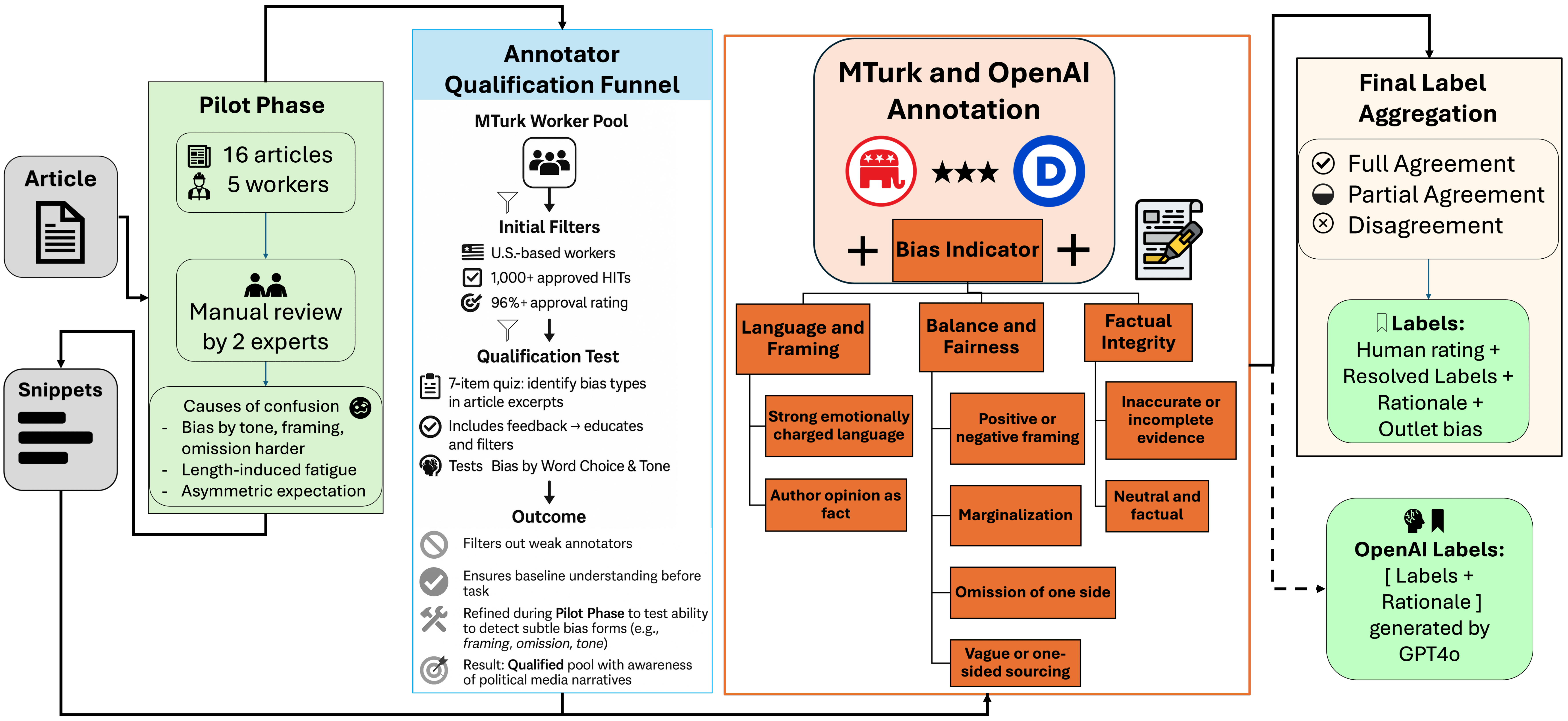
BiasLab: Explainable Political Bias Detection via Dual-Axis Human Annotations and Rationale Indicators

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

Annotation Pipeline



Article

Snippets

Pilot Phase
16 articles
5 workers
Manual review by 2 experts
Causes of confusion
- Bias by tone, framing, omission harder
- Length-induced fatigue
- Asymmetric expectation

Annotator Qualification Funnel
MTurk Worker Pool
Initial Filters
U.S.-based workers
1,000+ approved HITs
96%+ approval rating
Qualification Test
7-item quiz: identify bias types in article excerpts
Includes feedback → educates and filters
Tests Bias by Word Choice & Tone
Outcome
Filters out weak annotators
Ensures baseline understanding before task
Refined during **Pilot Phase** to test ability to detect subtle bias forms (e.g., *framing, omission, tone*)
Result: **Qualified** pool with awareness of political media narratives

MTurk and OpenAI Annotation
★★★
D
+ Bias Indicator +
Language and Framing
Strong emotionally charged language
Author opinion as fact
Balance and Fairness
Positive or negative framing
Marginalization
Omission of one side
Vague or one-sided sourcing
Factual Integrity
Inaccurate or incomplete evidence
Neutral and factual

Final Label Aggregation
Full Agreement
Partial Agreement
Disagreement
Labels: Human rating + Resolved Labels + Rationale + Outlet bias

OpenAI Labels: [Labels + Rationale] generated by GPT4o

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

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

Dual-Axis Human Ratings

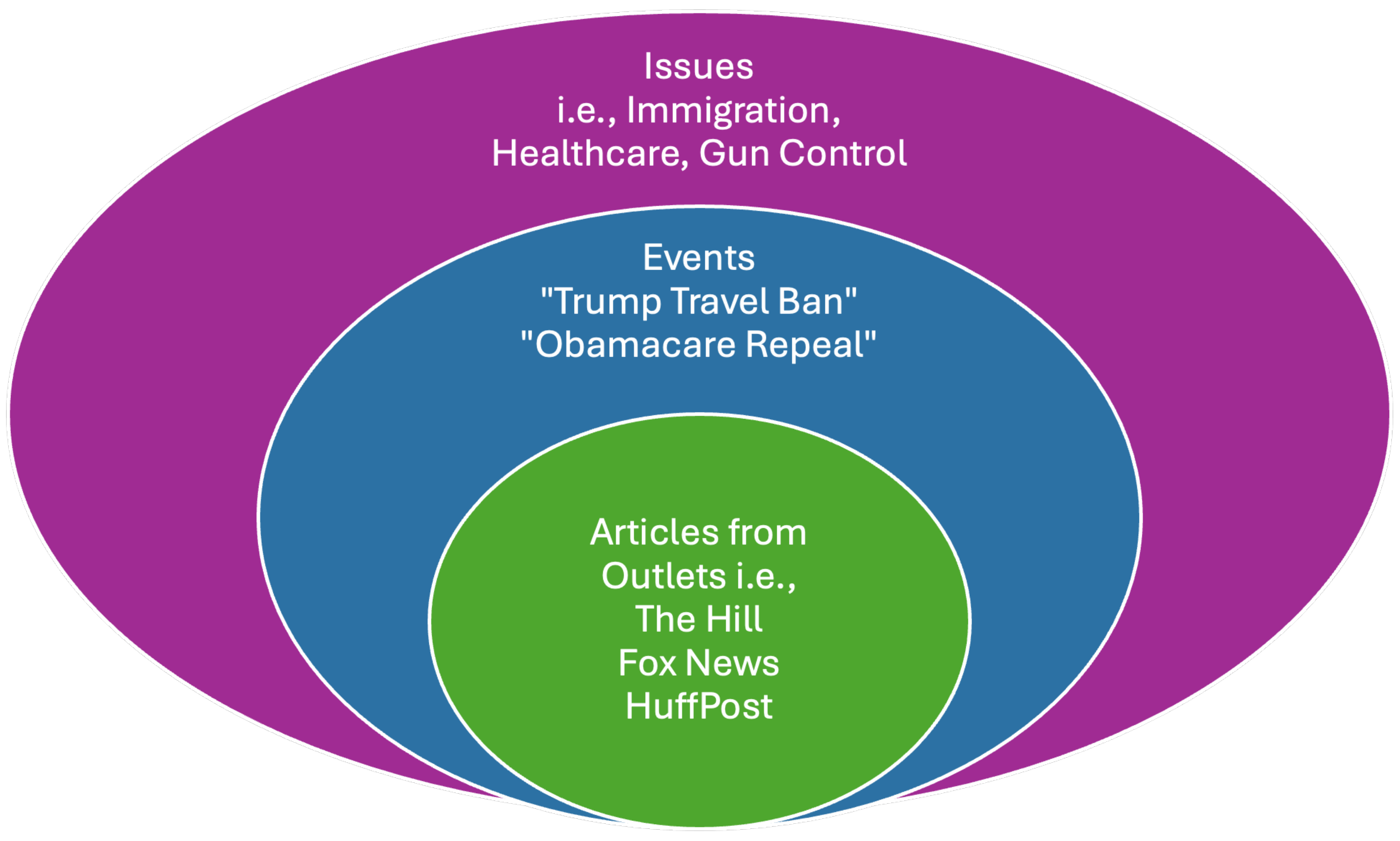
Democratic

Negative	Somewhat Negative	Neutral	Somewhat Positive	Positive

Republican Party

Negative	Somewhat Negative	Neutral	Somewhat Positive	Positive

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



Issues
i.e., Immigration, Healthcare, Gun Control
Events
"Trump Travel Ban"
"Obamacare Repeal"
Articles from Outlets i.e.,
The Hill
Fox News
HuffPost

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

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.

Dataset, Paper and Code

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