The Magnitude of Truth: On Using Magnitude Estimation for Truthfulness Assessment

Michael Soprano, Denis Eduard Tapu, David La Barbera, Kevin Roitero, and Stefano Mizzaro

The 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2025)

July 16, 2025 - Padua, Italy

UNIVERSITÀ DEGLI STUDI DI UDINE

hic sunt futura

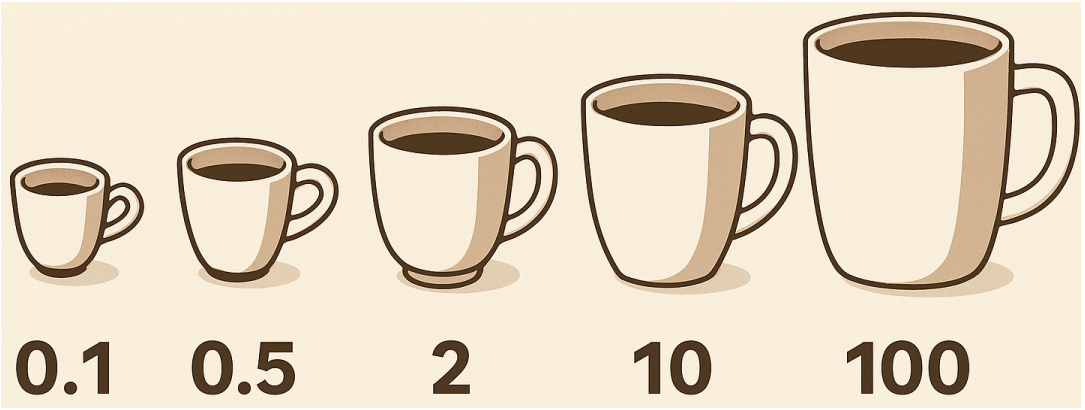
Assessing Truthfulness

- **Misinformation** grows faster than fact-checkers can respond
- Crowdsourcing scales assessment, but quality depends on the chosen **assessment scale**
- Traditional scales have intrinsic limits

Type	Levels	Limit
Binary	True / False	No nuances
Ordinal	3–6 (e.g., PolitiFact)	Subjective steps, equidistance?
Fine-grained	101-level (0–100)	Cognitive load, anchoring

Enter Magnitude Estimation (ME)

- **Continuous ratio scale** – captures nuance lost in fixed levels
- Workers assign **any positive number** $(0, +\infty)$
  - Scale never "runs out": always room for a larger or smaller value
- Already used in relevance assessment, linguistics, ...
  - Turpin et al. (SIGIR 2015), 50k relevance labels aligned with experts

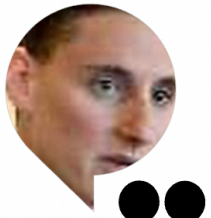


Research Questions

1. **RQ1: Alignment** – Do ME labels match expert ground truth?
2. **RQ2: Comparison** – How does ME perform compared to traditional scales (e.g., 6 levels)?
3. **RQ3: Insights** – Does ME reveal *additional* information?

Dataset & Ground Truth

- **120** PolitiFact statements (20 per level), reused from La Barbera et al. (IP&M 2024) study
- **1,200** crowdsourced assessments on a **six-level scale** (S6)
- Identical statements and assessments → **direct comparison**



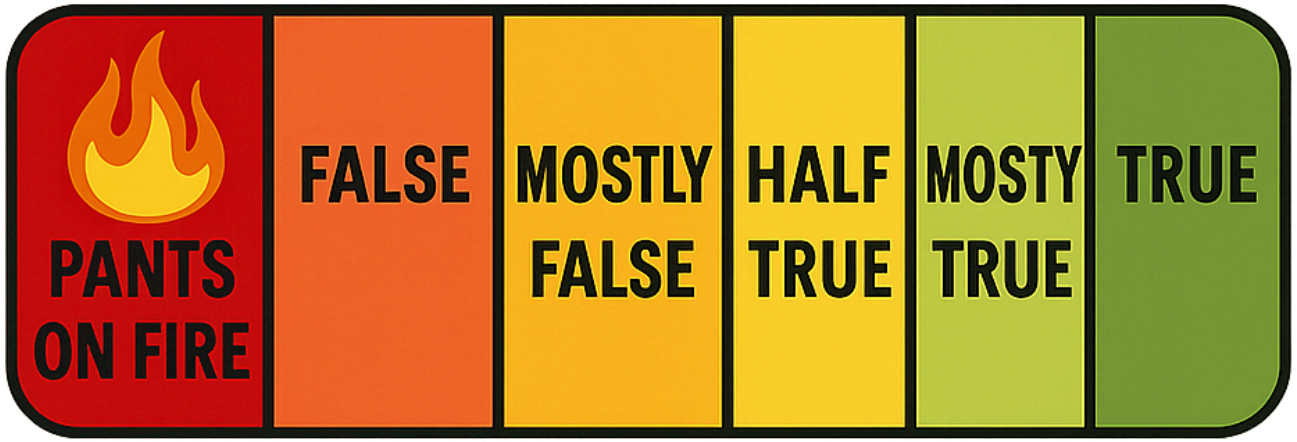
Josh Mandel

stated on February 25, 2022 in a speech at the 2022 CPAC meeting in Florida:

The 2020 election "was stolen from Donald J. Trump."

Processing ME Assessments

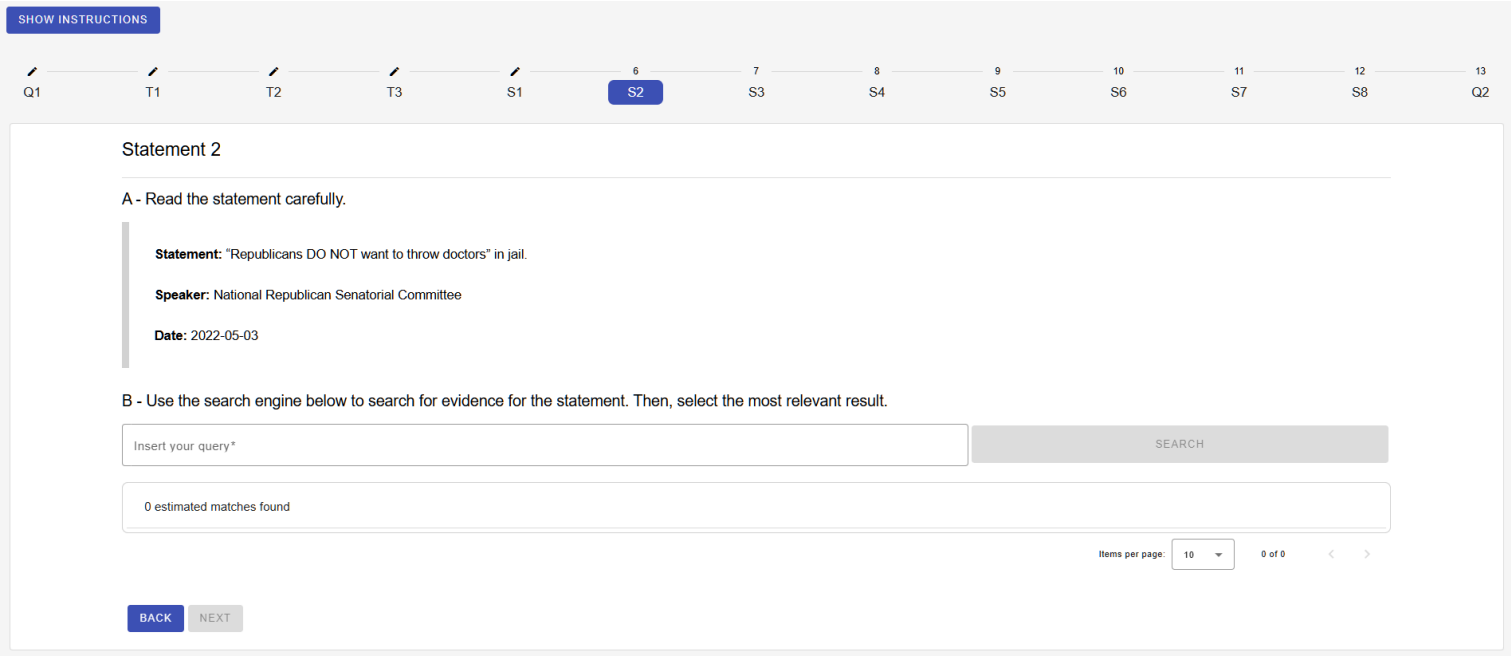
- **Normalise** each worker's ME range → 0 – 5
- **Aggregate & Group**: weighted mean → group **GT_6 → GT_3 → GT_2**
- **Metrics**: accuracy ($GT_{6/3/2}$), MAE/MSE, agreement (external, internal, pairwise)



PolitiFact's six-levels truthfulness assessment scale

Crowdsourcing Task

- **200** U.S. Prolific workers → **1,200** ME assessments after filtering
  - 10 workers per statement, each assessing 8; 2 gold statements
- ME warm-up → **evidence search** → **numeric assessment**
- Average completion **19 min**; median effective pay ≈ **£10/h**



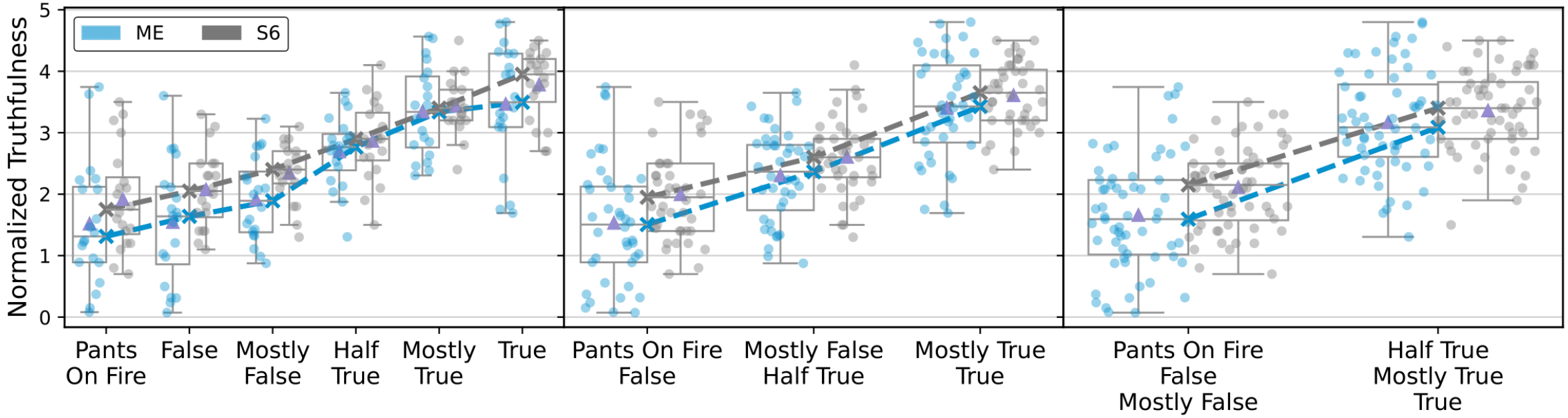
RQ1: Alignment with Experts

- **Binary accuracy 0.80**, close to automated systems
  - Prior results: 0.50–0.78
- **Medians rise** with ground-truth level
- Adjacent **false levels** are **difficult**

Aggregated, normalised ME vs. expert ground truth for GT_6, GT_3, GT_2

RQ2: ME Vs. S6

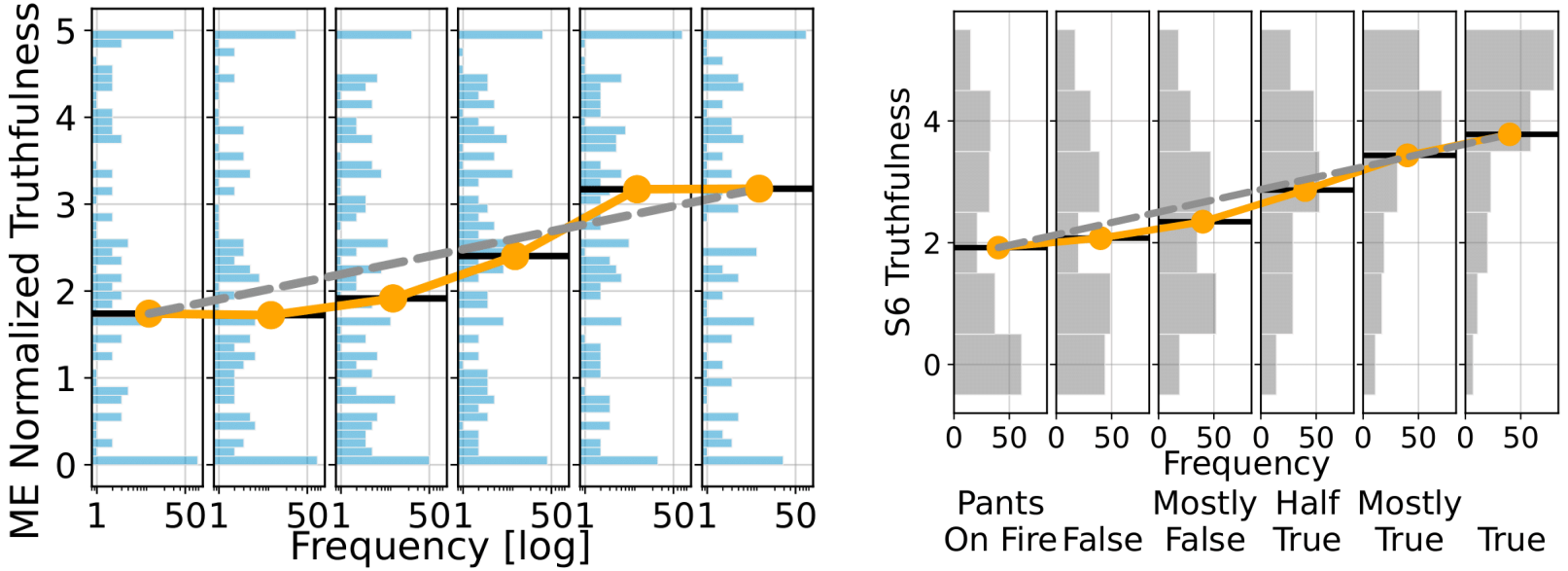
- Overall effectiveness **comparable** with S6
- **ME better** on **false** statements (lower norm. scores)
- **S6** slightly **better** on **true** statements (higher norm. scores)



S6 baseline (grey) vs. ME (blue) across GT_6, GT_3, GT_2

RQ3: Extra Insights from ME

- Perceived distances form a **sigmoid** not a linear progression
- *Half-True* is the **midpoint** of the scale
- Similar rankings, different views (pairwise agreement = 0.75)



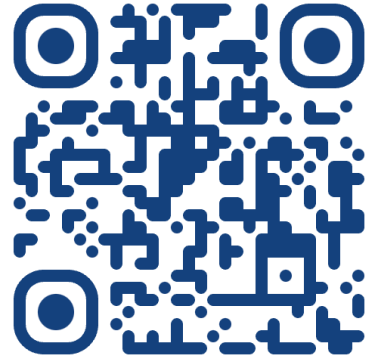
Gain profile curves: ME (left) vs. S6 (right)

Key Takeaways and Future Work

- **ME matches S6 accuracy** and scales easily with crowds
- **Uncovers nuances** invisible to fixed-level scales
- Next: expert-ME baselines, hybrid ME + S6 scales, benchmarks vs. LLM fact-checkers

Links and Acknowledgments

- **Paper** – doi.org/3726302.3730091
- **Repository** – osf.io/yux42
- **Contact** – michael.soprano@uniud.it



We welcome feedback and collaborations!

Partially supported by **PRIN 2022 — MoT: The Measure of Truth**

Finanziato dall'Unione europea NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA

UNIVERSITÀ DEGLI STUDI DI UDINE