Can Large Language Models *Transform* Computational Social Science?

Caleb Ziems[1*], **William Held**[2*], **Omar Shaikh**[1*], **Jiaao Chen**[2*], **Zhehao Zhang**[3*], **Diyi Yang**[1]
[1]Stanford University, [2]Georgia Institute of Technology, [3]Dartmouth

* All substantially contributed to the implementation of this work

Overview

We evaluate **13 LLMs** on **25 CSS tasks** zero-shot and draw a road map 🚗🗺️ for social scientists who want to use them.

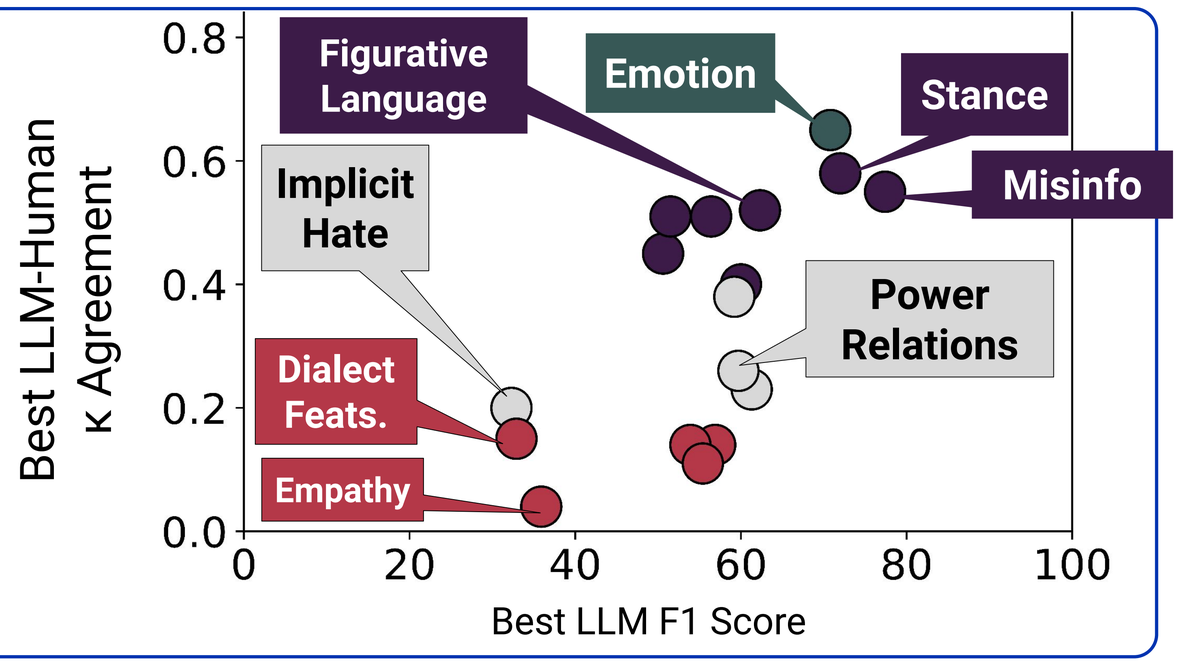
White House Ousts Top Climate Change Official

Which of the following describes the above news headline?
A: Misinformation
B: Trustworthy
Constraint: Answer with only the option above that is most accurate and nothing else.

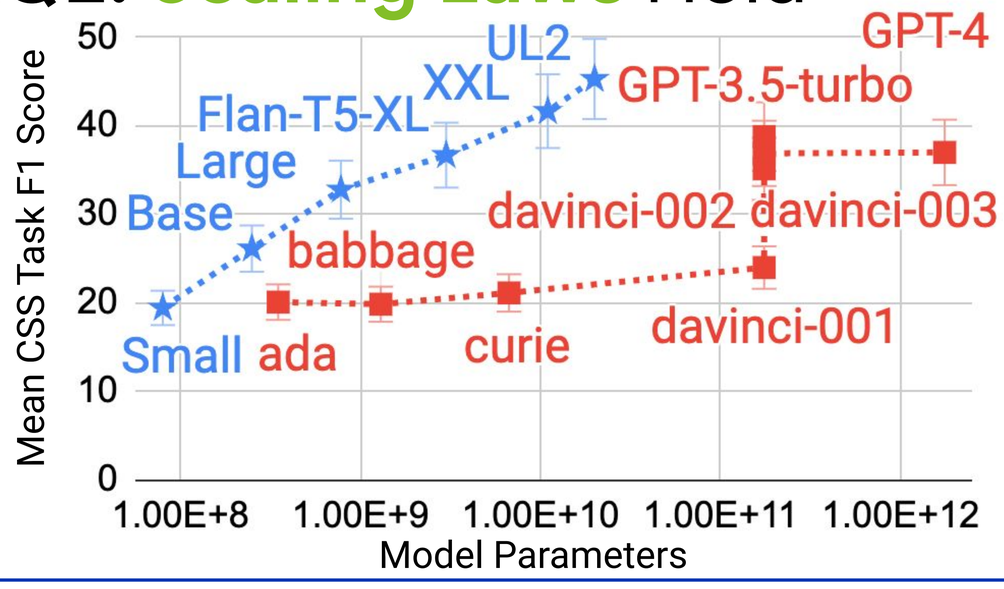
- **Sociology**
  - **Social Dynamics**
    - Persuasiveness
    - Power
  - **Anti-Social Behavior**
    - Toxicity Prediction
    - Hate Speech
  - **Cultural Analysis**
    - Social Bias Inference
    - Figurative Language Explanation
- **Psychology**
  - **Social Psych**
    - Emotion
    - Humor
    - Politeness
  - **Mental Health**
    - Empathy
    - Positive Reframing
    - Emotion Summarization
- **Literature**
  - **Literary Themes**
  - **Narrative Analysis**
    - Character Tropes
    - Relationship Dynamics
- **History**
  - **Historical Events**
    - Event Extraction
  - **Cultural Evolution**
    - Semantic Change
- **Linguistics**
  - **Sociolinguistic Variation**
    - Dialect Feature Identification
  - **Social Language Use**
    - Figurative Language
    - Persuasion Strategies
    - Discourse Acts
- **Pol. Sci**
  - **Framing**
    - Misinformation
    - Event Framing
  - **Ideology**
    - Stance
    - Statement Ideology
    - Media Slant

RQ1: LLMs can Augment but not Replace Human Annotation

- **Augment?** Yes ✅ for ~half of tasks with κ>0.4 (fair or good agreement with humans)
- **Replace?** No ❌ especially not for expert taxonomies (Dialect Feats.) or parsing tasks



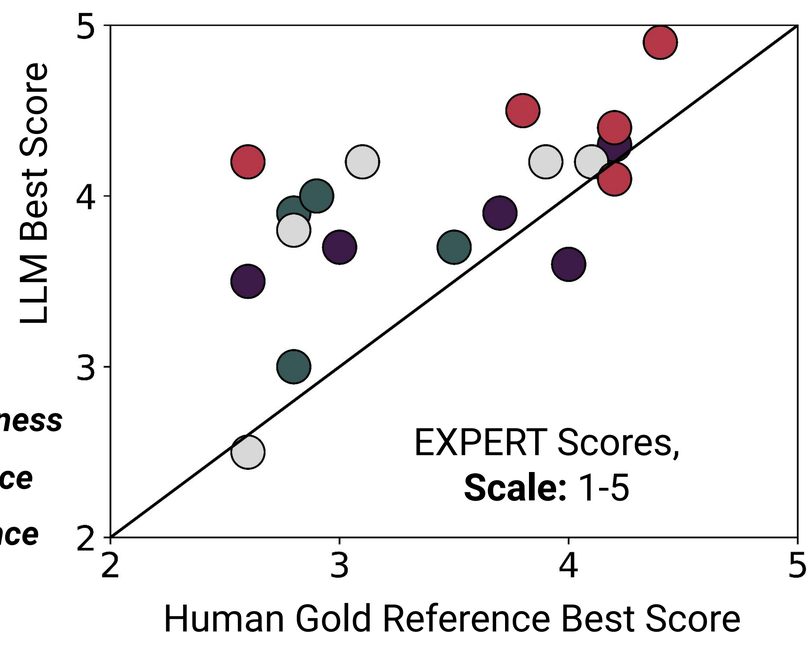
RQ2: Scaling Laws Hold



RQ4:

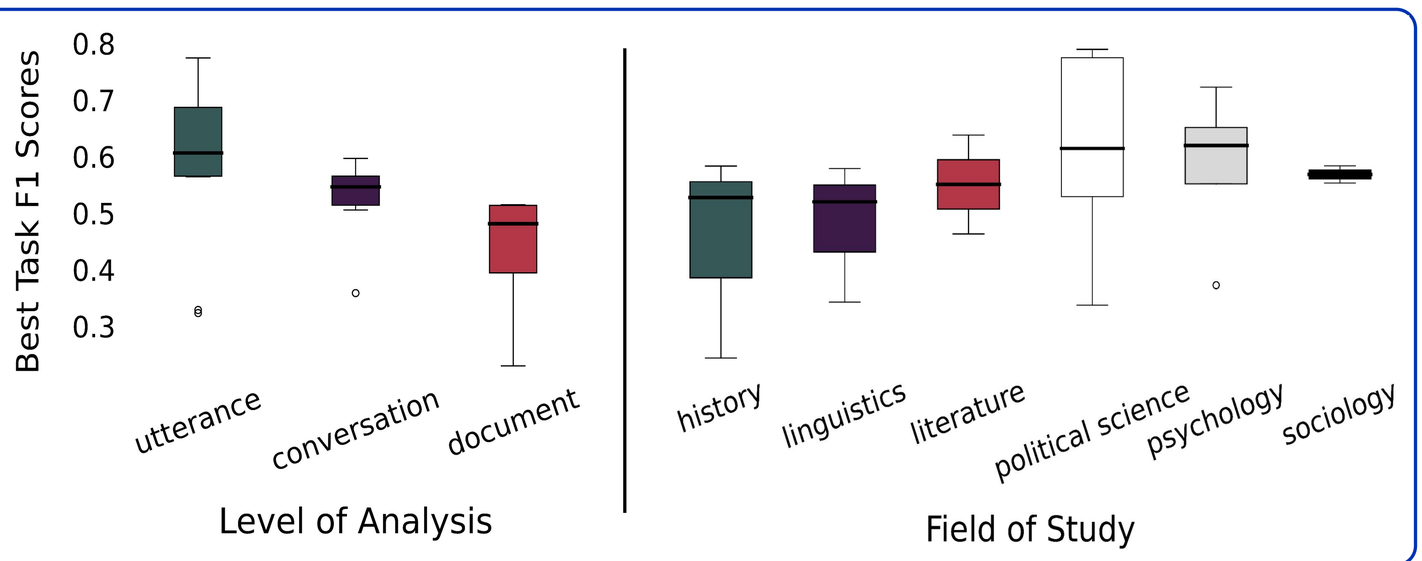
LLMs Can Also **Generate New Labels** that ***beat*** reference levels of:

- ***Faithfulness***
- ***Relevance***
- ***Coherence***
- ***Fluency***



RQ3:

Performance Varies by **Complexity**, **not Field of Study**



Recommends:

1. LLM + Human co-annotation
2. Open Source models for classification
3. Reinvest in experts for reproducible CSS