Evaluating and Applying LLMs for Social Science Data

From Evaluation Pipelines to Deployment on Roar

Cassandra Tai

Center for Social Data Analytics (C-SoDA)

Social Science Research Institute Open House
Penn State University
December 1, 2025

Outline

- Evaluation Pipeline:*From annotation to oversight* (Ko, Tai, and Webb Williams, 2025)
  - Annotation (text and multimodal data) + Impersonating respondents
  - Text Annotation
- Application case: *GenAI vs. Human Fact-checker* (Tai et al., 2025)
  - zero-shot
- Open-source LLMs on Roar / Roar Collab
  - Environment Setup
  - Model Deployment: DeepSeek, Llama, quantized models,
  - OpenAI API: GPT-4o

Opportunities and Challenges in GenAI

- GenAI in Social Science: New data sources and tasks
  - Synthetic respondents, experiments, multi-agent simulations(Argyle et al. 2023; Bisbee et al.2024; Aher et al, 2023; Park et al. 2023)
  - Large-scale annotation of text, images, and multimodal data (Gilardi et al 2023; Davidson 2024)
  - Misinformation and conspiracy detection(Diab et al. 2024; Ziems et al. 2024)
  - ...
- Challenges of GenAI in Social Science:
  - Bias and hallucinations (Abid et al. 2021; Felkner et al. 2024; Haim et al. 2024; Augenstein et al. 2024)
  - Black-box reasoning and lack of transparency, explainability, and reproducibility (Bail 2024; Bisbee et al. 2024)
  - ...

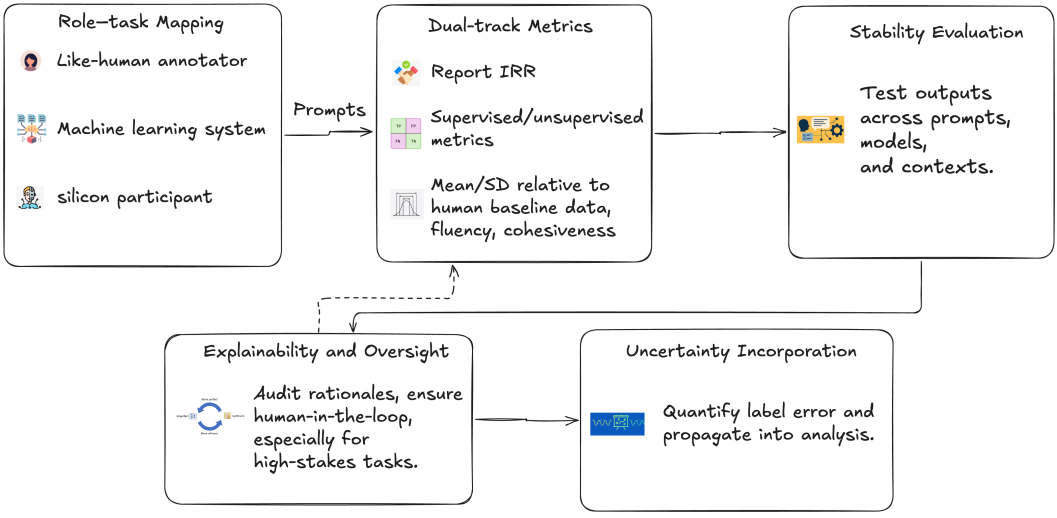
Why We Need a Systematic Framework

Can an LLM do this task?

V.S.

How can we validate and document LLM use so that our inferences remain credible?

5-Step Pipeline: Ko et al., 2025



Step 1-Role-Task Mapping: Clear role definition guides evaluation criteria

- Like-Human Annotator
  - Replicating human annotation patterns for labeling tasks
- Machine Learning System
  - Recover gold labels with either zero-shot or few-shot or fine-tuning with gold labels (Close to Supervised Learning)
  - Clustering tasks with zero-shot and without gold labels (Close to Unsupervised Learning)
- Silicon Participant
  - Simulating human responses to survey questions

Step 2-Dual-Track Metrics: Evaluation metrics must align with LLMs' role

- Reliability Track: Inter-rater reliability
  - Cohen's kappa
  - Krippendorff's alpha
- Validity Track
  - Supervised: Precision, Recall, F1, AUC, MCC, etc
  - Unsupervised: Silhouette coefficient, pair comparison, etc
  - Silicon Participant: Mean/SD relative to human baseline, fluency, coherence, etc

Roles are not exclusive. Reliability and validity can be tested **together**—GenAI may act as both a human-like annotator and a computational system in the same study.

Step 3-Stability Evaluation:Reduce cross-prompt/model variance

- Prompt Sensitivity
  - Test multiple prompt templates / phrasings
  - Assess variation in metrics across prompts
- Model sensitivity
  - Compare models (GPT, Llama, Deepseek, etc.)
  - Examine where models disagree with humans and with each other

Step 4-Explainability & Oversight: Automated systems require oversight

- Use model-generated rationales as auditable artifacts
- Audit for:
  - Logical coherence and conceptual validity
  - Biases, hallucinations, ethical red flags
- Two-way human–AI collaboration
  - LLM rationales expand human awareness
  - Human experts correct, constrain, and document model behavior

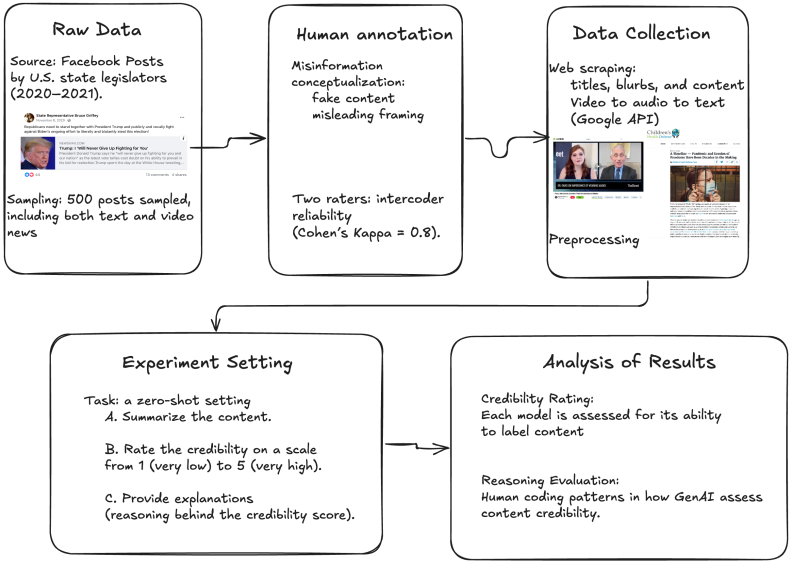
Step 5-Uncertainty & Error Correction:Acknowledging AI uncertainties

- Misclassification bias can distort regression and descriptive analyses
- Treat AI labels as noisy measurements, not ground truth
- Methods:
  - Design-based Supervised Learning (DSL): gold-standard subsample + inclusion probabilities (Egami et al. 2023)
  - Misclassification / maximum-likelihood adjustment (MLA) (Teblunthuis et al. 2024)

Applications- GenAI vs. Human Fact-checker (Tai et al., 2025)

- Research Question: Can GenAI effectively assess content credibility?
- Data:
  - An archive that systematically tracks online communications of federal, state, and local officials across multiple digital platforms.
  - 28,834 public officials
  - Daily online activity since 2020: over 6 million posts from X and Facebook
  - Public Accessible

Applications- GenAI vs. Human Fact-checker (Tai et al., 2025)



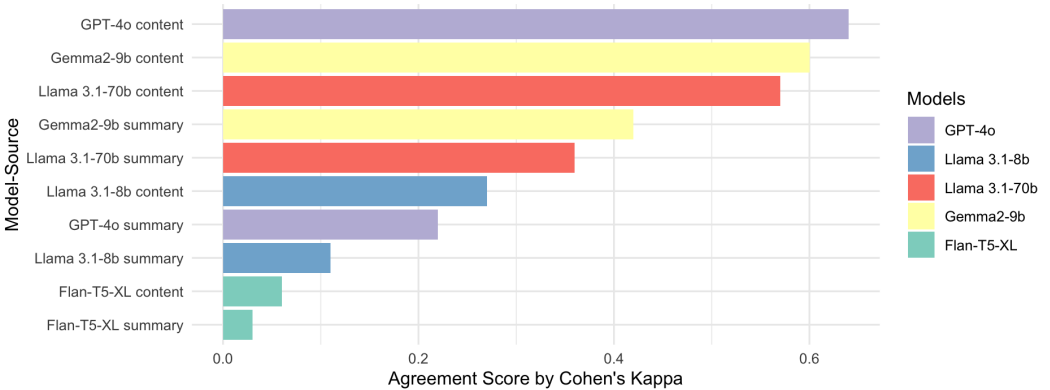
LLM Deployment

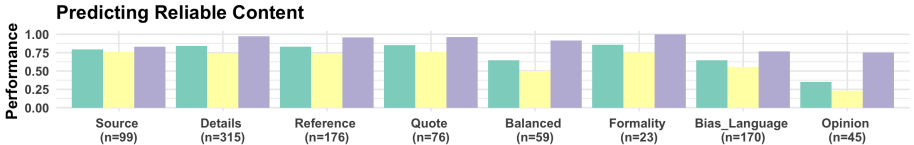
Steps 1-2 Summary

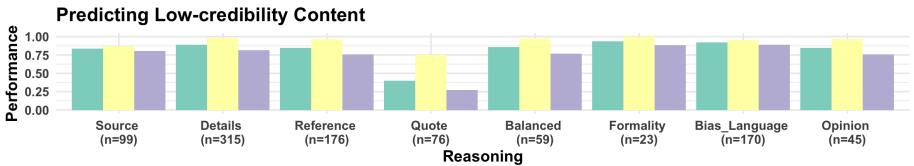
Conception	Task	Prompts	Gold Labels	Suggested Metrics
Like-human annotator	Create high quality labels/clusters as gold labels	Zero-shot or Few-shots	No	*Reliability:* Cohen's kappa, Krippendorff's alpha, etc.
			Yes	*Validity:* Precision, recall, F1, AUC, Matthews Correlation Coefficient (MCC), etc.
Machine learning system	Recover gold labels	Zero-shot	Yes	*Reliability + Validity:* IRR, Precision, recall, F1, AUC, etc.
		Few-shot; fine-tuned	Yes	*Validity:* Precision, recall, F1, etc;
	Create clusters of materials without prior schema	Zero-shot	No	*Validity:*Silhouette coefficient, pair comparison/adjusted normalized mutual information, etc.
Like-human subject (silicon participant)	Take a survey, play a game, or otherwise stand in for a human subject	–	Yes	*Validity:* Mean/SD relative to human baseline data, or and humanness such as fluency, cohesiveness, objectivity, readability, etc.

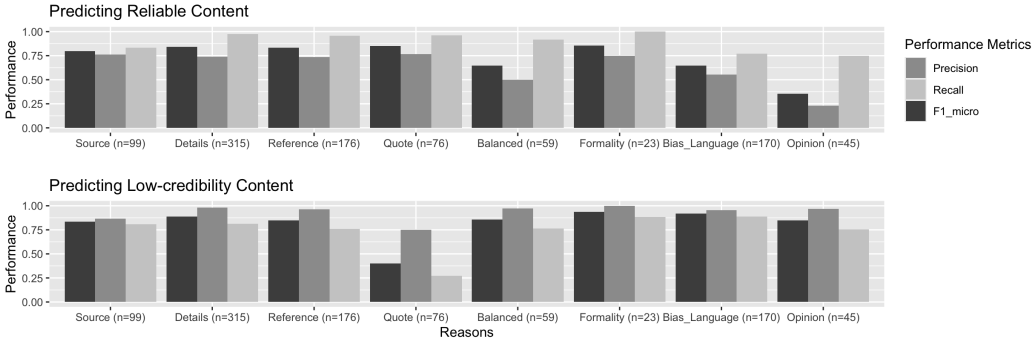
Applications- GenAI vs. Human Fact-checker (Tai et al., 2025)

- Models:
  - GPT-4o (OpenAI)
  - Llama 3.1 (Meta): 8B and 70B parameters
  - Gemma 2 (Google): 9B parameters
  - Flan-T5-XL (Google)
- Prompt Design:
  - 5 variations for rating and reasoning tasks, following standard evaluation scales.
  - A 5-point rating scale, ranging from very low credibility (1) to very high credibility (5), with a threshold of 3 for content reliability.
  - Evaluation based on the average performance across the five different prompt variants.
- Results: GenAI has potential but is fundamentally limited in its capacity to detect political content credibility. Human oversight remains critical


GPT-4o content
Gemma2-9b content
Llama 3.1-70b content
Gemma2-9b summary
Llama 3.1-70b summary
Llama 3.1-8b content
GPT-4o summary
Llama 3.1-8b summary
Flan-T5-XL content
Flan-T5-XL summary
Model-Source
0.0
0.2
0.4
0.6
Agreement Score by Cohen's Kappa
Models
GPT-4o
Llama 3.1-8b
Llama 3.1-70b
Gemma2-9b
Flan-T5-XL


Predicting Reliable Content
Performance
1.00
0.75
0.50
0.25
0.00
Source (n=99)
Details (n=315)
Reference (n=176)
Quote (n=76)
Balanced (n=59)
Formality (n=23)
Bias_Language (n=170)
Opinion (n=45)


Predicting Low-credibility Content
Performance
1.00
0.75
0.50
0.25
0.00
Source (n=99)
Details (n=315)
Reference (n=176)
Quote (n=76)
Balanced (n=59)
Formality (n=23)
Bias_Language (n=170)
Opinion (n=45)
Reasoning
Performance Metrics
Precision
Recall
F1_micro


Predicting Reliable Content
Performance
1.00
0.75
0.50
0.25
0.00
Source (n=99)
Details (n=315)
Reference (n=176)
Quote (n=76)
Balanced (n=59)
Formality (n=23)
Bias_Language (n=170)
Opinion (n=45)
Performance Metrics
Precision
Recall
F1_micro
Predicting Low-credibility Content
Performance
1.00
0.75
0.50
0.25
0.00
Source (n=99)
Details (n=315)
Reference (n=176)
Quote (n=76)
Balanced (n=59)
Formality (n=23)
Bias_Language (n=170)
Opinion (n=45)
Reasons