Frame Detection in High-Stakes Discourse: A Hybrid Human-AI Approach

David Farr [1], Stephen Proschaska [1], Iain Cruickshank [2], Jevin West [1], Kate Starbird[1]

[1]University of Washington, [2]Carnegie Mellon University

Correspondence: dtfarr@uw.edu

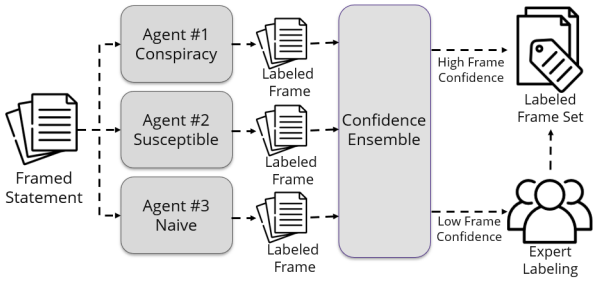

Figure 1: System diagram that demonstrates proposed frame detection methodology. Each statement is passed to all annotators to provide a data tag. The tagged statements are compared via a measurement of agent confidence prior to being assigned a final label. Labels with low confidence can be routed to human annotators for increased system performance.

Abstract

This paper explores the application of language models with human-in-the-loop routing for frame detection, with a focus on natural language learning techniques that capture both explicit and implicit frames. We benchmark two labeled framing datasets and propose a methodology that uses role-prompting across multiple language agents, enabling the detection of ambiguous or unstated frames. In addition, we incorporate uncertainty quantification strategies into the data labeling process, improving system performance in complex tasks where meaning is influenced by contextual and background knowledge. Our findings contribute to the advancement of scalable and efficient frame detection techniques, enhancing our understanding of how language reflects and shapes social and communicative structures.

1 Introduction

Recent advances in Natural Language Learning (NLL) have dramatically improved our ability to process large-scale textual data, yet key challenges remain in understanding how meaning emerges in interactive text, including online discourse. Computational approaches often assume that meaning is explicit within a given text, but insights from discourse semantics and sociolinguistics suggest that meaning is often constructed through social interaction and shaped by cognitive frames (Nguyen et al., 2016). This challenge is particularly pronounced in the context of Computational Social Science (CSS), where scholars study how social meaning is collectively negotiated in digital spaces.

A core issue in analyzing such discourse is context-dependent ambiguity — i.e. content that could be interpreted in multiple ways and that may mean different things to different readers. For example, in online discussions about election integrity different audience segments interpret the same textual event differently based on prior beliefs, cultural factors, and exposure to specific narratives. When tabulator errors occurred on election day in 2022, some social media users framed these errors as evidence of election fraud, while others framed them as incompetence but not intentional manipulation. These competing interpretations arise from underlying framing processes, a concept long studied in discourse analysis, sociolinguistics, and political communication (Goffman, 1974; Entman, 1993). Frames function as cognitive schemas that shape meaning construction, highlighting some elements of reality while obscuring others. Understanding such implicit frames is crucial for discourse analysis, yet traditional NLL models struggle with perspective-driven meaning, as frames often lack direct linguistic markers and reside in the interpretive process of the audience rather than the text itself (Ziems et al., 2024).

To address this challenge, we propose a multi-agent natural language learning system that models diverse interpretive perspectives. Our methodology extends prior frame detection efforts in NLP by simulating the varied ways individuals interpret the same text. We implement a multi-agent annotation system in which multiple LLM-based annotators, each representing a distinct social per-

spective, classify frames in text. These perspectives are reconciled through confidence-aware aggregation, ensuring robustness in cases of interpretive ambiguity. Low-confidence cases are escalated to human experts, facilitating a hybrid human-AI framework that improves both scalability and accuracy. Our work builds on recent advances in Computational Social Science and discourse semantics, operationalizing framing as a computational problem while maintaining theoretical fidelity to sociolinguistic models of meaning construction.

To evaluate our system, we benchmark three different frame detection architectures on two datasets, systematically assessing how different system components impact performance. Our results contribute to context-sensitive natural language learning by introducing an approach that accounts for interpretive variation in high-stakes, ambiguous discourse. By integrating insights from discourse semantics, sociolinguistics, and CSS, this research advances scalable methodologies for analyzing contested narratives, offering a bridge between theoretical models of meaning and practical natural language learning applications.

2 Related Works

2.1 Framing

In sociologist Erving Goffman's Frame Analysis (1974), he introduces the concept of a "frame" to describe how people make sense of social situations (Goffman, 1974). Frames are mental schemas that shape individuals' interpretations of the world, highlighting certain elements of a situation while downplaying others. Because people rely on different frames, even when exposed to the same event or data, their understanding of reality can diverge. Text frequently conveys sociolinguistic cues that signal a particular frame to readers, influencing interpretation and perception.

Online discourse is deeply shaped by framing, making frame detection a crucial task for researchers studying how communities interact with text. However, framing has been conceptualized across multiple disciplines—including sociology, communication, and political science—contributing to challenges in reliably operationalizing it for computational analysis. Entman (Entman, 1993) unifies several perspectives on framing, emphasizing that frames are not merely cognitive constructs residing in the minds of individuals. Instead, frames exist in both the communication itself and the interpretation of audiences, meaning that framing can be observed in linguistic patterns, rhetorical strategies, and the structure of discourse. In other words, frames are both "in the heads" of communicative speakers and receivers, and in the text of the speech itself.

Despite advances in natural language processing (NLP), computational models still struggle with frame detection due to the fluid, implicit, and dynamic nature of frames. Frames are numerous and context-dependent, their boundaries can be subjective, and they are often not explicitly signaled in text, making annotated data scarce. Traditional NLP methods rely on surface-level lexical cues or topic modeling, which often fail to capture deeper framing mechanisms. Large language models (LLMs) have shown promise in zero-shot and few-shot classification tasks, but they remain limited by data scarcity and difficulty in generalizing across framing contexts.

Klein et al. distinguish between data—the raw information in a situation—and frames, which guide what information is attended to and how it is interpreted. In this view, a perceiver selects a frame while simultaneously interpreting the data, with the potential to adjust or switch frames as new information emerges (Klein et al., 2007). This theoretical distinction is useful for computational modeling, as it allows NLP systems to disentangle the linguistic markers of framing from the factual content of discourse.

By applying data-frame theory to NLP, we propose an approach that moves beyond surface-level linguistic features to model how framing influences meaning construction in natural language. Our method incorporates both the explicit framing cues present in text and the implicit interpretive processes that shape how information is received. This perspective provides a foundation for improving frame detection in zero-shot and few-shot settings, offering a step toward more reliable and scalable computational models for analyzing framing in socially generated data.

2.2 Data Annotation using Large Language Models

Using LLMs for data annotation has become increasingly common in both industry and academia (Tan et al., 2024; Goel et al., 2023; Li, 2024; Cruickshank and Ng, 2024). These models often perform well on out-of-domain data, where traditional supervised models can struggle with generalization

(Ng and Carley, 2022). The capability to generalize on previously unseen data makes LLMs a valuable tool for rapid data annotation, especially when training data is limited or unavailable.

While LLMs are frequently applied to data annotation, limited research has addressed their use for frame detection. Weinzierl and Harabagiu (2024) demonstrated the efficacy of Chain-of-Thought prompting for identifying frames in communication about COVID-19 vaccines, while Pastorino et al. (2024) analyzed LLM performance on frame detection using zero-shot and few-shot prompting. Although important, these works rely on datasets that loosely represent topic framing with clear ground-truth classifications. Both studies primarily focus on single-agent prompt engineering for improved classification, leaving substantial room for exploring more complex prompt designs and testing on datasets that align with social science concepts of framing.

Recent studies such as Long et al. (2024); Lan et al. (2024) show that multi-expert prompting can enhance LLM reliability and performance as general-purpose problem solvers. However, this work does not address data annotation tasks within a broader system context. Similarly, Kong et al. (2023) demonstrates how role-prompting – in which the LLM is assigned a certain personality or role to play – can improve LLM performance in datasets with established ground truths, raising the question of whether role-prompting within a systems methodology could benefit context-dependent classification tasks.

For human-in-the-loop (HITL) data annotation, Wang et al. (2024) and Farr et al. (2024b) both present HITL methodologies on benchmark datasets. However, Farr et al. (2024b) focuses on the training of the downstream model with a single agent in ground-truth datasets rather than directly assessing HITL performance. Although Wang et al. (2024) explores different HITL sampling thresholds, it does not address frame detection, employs a different sampling approach, and also utilizes a single LLM agent.

In this work, we aim to combine various aspects of the related works, to propose a novel systems methodology for leveraging multi-agent role-prompting for improved HITL data annotation on difficult and ambiguous tasks. We specifically apply this methodology to frame detection due to the uniqueness of the task, but also demonstrate its performance on a benchmark ideology dataset to show it's generalization power and performance on a well-established benchmark.

3 Data

We use two manually curated datasets in our evaluation, derived from previous work examining collective sensemaking and rumoring about U.S. elections in 2020 and 2022 (Prochaska et al.; Starbird et al.). Rumors, which can be true or false, often emerge from collective sensemaking processes as communities come together to understand novel or ambiguous information (Arif et al., 2017; DiFonzo and Bordia, 2007; Maddock et al., 2015; Shibutani, 1966). The data we rely on here are drawn from larger Twitter collections of election-related discourse. The 2020 collection is described in Kennedy et al. (Kennedy et al., 2022) while the 2022 collection is discussed in Schafer et al. (Schafer et al., 2024). These datasets make ideal test sets because as online audiences collectively made sense of emergent events, they used different frames to interpret evidence as it spread online. These frames facilitated the emergence of online rumors, resulting in different interpretations of the same evidence depending on the political community from which the frames emerged (Starbird et al.). The two datasets we benchmark for the task of unsupervised frame detection and apply our methodology are discussed in detail below.

3.1 Deep Stories - 2020 Election Sensemaking

The first dataset is described in more detail in Prochaska et al. (Prochaska et al.) and compares sensemaking processes across 2020 and 2022. This dataset relies on what the researchers term "incidents" to group-related tweets. Generally speaking, an incident refers to a single rumor or collection of temporally adjacent, related rumors surrounding a specific event. For both the 2020 and 2022 collections, incidents were defined at the time of sampling as a larger research team attempted to track and analyze election rumoring in real time.

Prochaska et al.'s (Prochaska et al.) study focuses on five incidents from 2020 and five incidents from 2022 (including both liberal- and conservative-leaning rumoring), aiming to understand how specific narratives influence collective sensemaking over time. In the process of developing a qualitative coding scheme to measure collective sensemaking and storytelling online, the research team developed and applied codes that

measured specific frames that commonly appeared in the data. It is this frame category that we use to evaluate our single- and multi-agent systems for this dataset.

3.2 Maricopa Dataset

The second dataset is discussed in Starbird et al. (Starbird et al.) and focuses narrowly on collective sensemaking about Maricopa County, Arizona on and after election day in 2022. On election day in 2022, printer errors in Maricopa County, Arizona resulted in tabulator failures. The breakdown of expected election-day administration caused online audiences to come together to make sense of what was happening, discussing what the tabulator rejections meant, whether they could trust remedies, and what the best avenue was to ensure that their votes were counted. As part of this sensemaking process, rumors emerged using different frames to interpret the tabulator errors.

Starbird et al. analyzed this sensemaking process on Twitter by exploring related conversation using quantitative and qualitative methods that guided iterations of purposive sampling (Starbird et al.). During the process of exploration, the researchers noted that conversations were often driven by the introduction of perceived "evidence" of election fraud, which was framed in different ways as audiences shared the evidence and made sense of its meaning. In order to understand how framing processes influenced the perceived meaning of a piece of evidence, the researchers developed a coding scheme that measured different types of evidence and frames present in the conversation, including both implicit and explicit frames. In particular, Starbird et al. (Starbird et al.) identify multiple "meta-frames" that guided audience interpretations. We use four of the most common meta-frames that Starbird et al. applied as labels.

4 Methodology

All prompts are located in Appendix A. For all testing we use GPT-4o and Llama2-7b (Touvron et al., 2023) with temperature of zero and logit bias of 10 for OpenAI. We evaluated each frame dataset against the following system configurations:

4.1 Single-Agent Annotation

Our initial system methodology involves a single-agent setup, where we prompt the model to annotate statements using a concise definition of the task. This approach reflects standard prompting practices currently applied in frame and ideology detection tasks (Ziems et al., 2024; Pastorino et al., 2024).

To evaluate performance under different sampling strategies, we apply a human-in-the-loop (HITL) approach using both random sampling and confidence-informed sampling. For the latter, we employ the confidence metric introduced by Farr et al. (2024a), which calculates the distribution between the top two token log probabilities for a predefined set of constrained labels. This single-agent methodology provides a straightforward benchmark for comparison with our multi-agent setup, helping us assess whether adding multiple agents with varied role prompts enhances system performance. We refer to this system in the results as "1-Agent CI" when using confidence-informed sampling for HITL frame annotation and "1-Agent Rand" when using random sampling.

4.2 Multi-Agent Annotation with Confidence Ensemble

Our second system methodology employs a multi-agent setup, where each agent is designed to represent a distinct user perspective through role-prompting. For implicit frame detection, we define agents as a *conspiracy theorist*, a *user susceptible to conspiracy theories*, and *a naive web user*. This configuration enables the model to simulate the interpretative diversity that may arise from different user points of view, an approach that approximates frames derived from audience members' cultural groups. This is representative of the original instructions given to hand annotators for the dataset to labels frames as a susceptible user (Starbird et al.).

The annotations for each agent are aggregated in a confidence ensemble. If two or more agents agree on an annotation, this becomes the meta label, with a confidence score equal to the sum of confidence between the agreeing agents. When all agents provide different annotations, the meta-label is selected based on the annotation with the highest individual confidence score. A graphical representation of this system can be seen in Figure 1.

To evaluate the multi-agent system, we conduct HITL annotation testing with both random sampling and confidence-informed sampling. This comparison allows us to measure the effectiveness of confidence ensemble methods across dif-

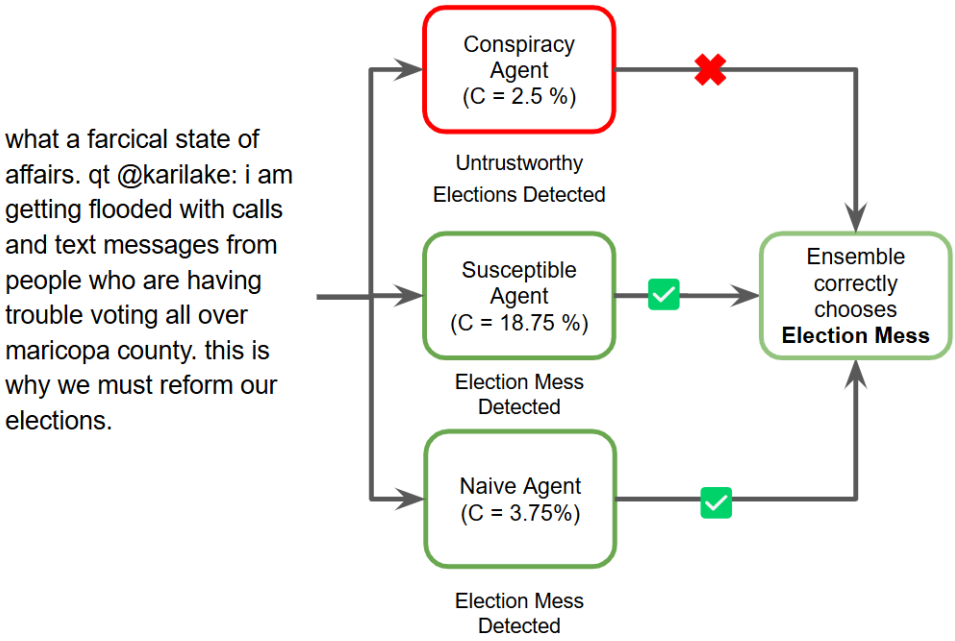

Figure 2: Actual framed statement passing through multi-agent frame detection system with associated confidence scores for each agent. Although the system correctly annotated this framed statement, the low to moderate confidence means it would likely go to a human annotator. Confidence values are the actual score over the max possible confidence score and not divided among the different classification possibilities.

ferent sampling strategies, providing insight into how multi-agent role-prompting impacts system performance relative to the single-agent baseline. We refer to this system in the results as “MA Conf CI" when using confidence-informed sampling and “MA Conf Rand" when using random sampling.

4.3 Multi-Agent Annotation with Judge

Our final system methodology retains the same role-prompted agents as in the confidence ensemble system but introduces a meta-agent, referred to as the “judge” agent, to arbitrate between the annotations of the agents. Rather than relying on a confidence ensemble, this approach allows the judge agent to review and determine the final annotation based on the prior agents’ perspectives and associated confidence in their assessment.

We use the same sampling methodologies for HITL annotation—random sampling and confidence-informed sampling—as in the previous systems. This methodology is designed to assess the effectiveness of using a confidence ensemble versus assigning a dedicated agent to act as an arbitrator, providing insights into the comparative benefits of ensemble consensus versus a centralized judgment approach. We refer to this system in the results as “MA Judge CI" when using confidence-informed sampling for and “MA Judge Rand" when using random sampling.

4.4 Human Routing

Each annotated, framed statement is accompanied by a confidence score, which quantifies the language model’s certainty in its annotation. These confidence scores can be normalized to a scale of 0 to 100 or ranked to facilitate prioritization, as described in (Farr et al., 2024a). This scoring system allows users to make informed decisions about how much of the data to user-label for improved system performance.

By integrating confidence scores into the annotation workflow, users can implement a human-in-the-loop (HITL) strategy that prioritizes statements based on their ambiguity or uncertainty. For example, statements with low confidence scores may be flagged for expert review, ensuring that the most challenging or uncertain cases receive human oversight. Conversely, high-confidence annotations can be accepted automatically, reducing the overall burden of labeling.

This methodology empowers users to balance efficiency and accuracy by specifying the percentage of data they wish to self-label, depending on their resource constraints or quality requirements in an informed manner. We evaluated and compared the impact of these strategies on overall system

	Accuracy - GPT		
	Single Agent	MA Judge	MA Conf
maricopa	60.1	61.2	**61.9**
deep	**62.6**	61.9	62.2

Table 1: Initial accuracy scores with no human intervention for GPT. Accuracy shows efficacy in using language models for further exploration in frame detection.

	Accuracy - Llama		
	Single Agent	MA Judge	MA Conf
maricopa	45.1	**51.5**	**50.4**
deep	**55.6**	55.9	**61.5**

Table 2: Initial accuracy scores with no human intervention for Llama. Demonstrates relatively high-performance in lower-parameter, open-source, language models.

performance, particularly when paired with HITL interventions.

4.5 Performance Evaluation Metrics

We evaluate our system performance using two primary metrics. First, we measure the accuracy of the initial system without any human intervention, establishing a baseline for each system shown in Table 1 and Table 2. To comprehensively evaluate the performance of HITL, we calculated the area under the curve (AUC) for accuracy as a function of expert-labeled data as illustrated in Figure 4. An AUC score of 50 would be equivalent to random or normal sampling whereas a score of 100 would be a perfect sampling strategy. As can be seen in Figure 4, at 20 percent of the data evaluated for the highest performing metric, a difference of 10 points results in approximately twice the amount of human intervention needed for the same performance.

5 Results

5.1 Task Performance

Our initial task performance shows little variance in the accuracy of frame detection using GPT-4o across different system settings; however, much more variance is seen in Llama-7b, with the MA Conf method demonstrating the most robustness across tests. This could potentially be due to GPT being less sensitive to prompt changes than open-sourced models and multi-agent systems off setting some of the sensitivity. Results for initial GPT accuracy on the Maricopa and sense-making frame detection datasets are in Table 1. Results for Llama across both frame detection datasets are in Table 2. However, the multi-agent ensemble does a better job in representing ambiguity in the data when multiple agents have disagreement signaling the frame can be interpreted different ways or is not straightforward.

5.2 HITL Performance

HITL performance reflects the real-world applicability of our systems in assisting human annotators by prioritizing ambiguous or complex samples for labeling. The task of providing high-quality labels on context-dependent data in rapidly evolving environments is difficult. This makes context-dependent labeling a well-suited candidate for HITL annotation. If we can confidently label a subset of data, we can save human-annotators significant amounts of time. This makes the HITL performance of each system perhaps more important than initial accuracy metrics. In order to provide a comparison across all confidence thresholds in a succinct manner, we use the AUC of the accuracy to human-labeling curve for each specified system with both random sampling for annotation and confidence-informed annotation as represented in Figure 4.

Table 3 and Table 4 demonstrate that confidence-informed annotation consistently outperformed random sampling across both GPT and Llama models. Moreover, the MA Conf system configuration outperforms the single agent system across both datasets and both language models. It outperforms the judge system in three models and is comparable in the fourth model.

These results show that using multiple agents is a promising method for detecting user frames due to the ability of Language Models to interpret statements in a manner consistent with their prompted values. Moreover, using uncertainty quantification methods to triage the uniquely difficult frames to experts can save human labelers significant time and resources.

Given the AUC curves, the amount of data that should be routed to humans is dependent on the risk tolerance for the task at hand. For general sentiment and understanding of the population, 5-10 percent expert labels provides a fairly robust measure with accuracy measures around 70 percent.

	AUC - GPT					
	1-Agent Rand	1-Agent CI	MA Judge Rand.	MA Judge CI	MA Conf Rand	MA Conf CI
maricopa	78.9	85	78.8	85.3	80.4	**87.7**
deep	81.8	83.8	80	**84.6**	80.7	**84.4**

Table 3: AUC accuracy scores across conditions for GPT model. Multi-Agent models show more effective frame detection, especially when integrated with confidence-informed sampling for expert annotation. To build further intuition on score calculations, all graphs AUCs are based on are available in Appendix C.

	AUC - Llama					
	1-Agent Rand	1-Agent CI	MA Judge Rand.	MA Judge CI	MA Conf Rand	MA Conf CI
maricopa	71.9	77.1	75.7	80.7	74.3	**80.9**
deep	77.8	80.1	78.1	77.8	79.9	**82.3**

Table 4: AUC scores across conditions for Llama. Multi-agent systems with confidence-informed sampling continue robustness and high performance, with confidence ensemble showing unique efficacy across all annotation tasks and models. Graphs for AUC metrics are available in Appendix C.

MA Conf CI System Performance with HITL				
% Expert Labeled	**10%**	**20%**	**30%**	**40%**
GPT -Maricopa	69.8	76.9	82.8	88.8
GPT - Deep	67.1	73.1	78.3	82.1
LLAMA - Maricopa	57.8	65.7	73.1	80.2
LLAMA - Deep	66.4	72	74.8	78.7

Table 5: Multi-Agent Confidence Ensemble System with confidence informed sampling for expert labeling accuracy levels across varying levels of intervention by expert annotators.

For downstream model training, only taking labels with high confidence scores and no human labeling is sufficient (greater than 95 percent accuracy for 60 percentile data). For a more nuanced view of individual data points, 40-50 percent of expert-labeled data offers over 90 percent accuracy. The amount of human-in-the-loop assistance should be entirely dependent on the task, resources, and risk tolerance.

6 Discussion

Our findings contribute to advancing computational approaches for frame detection, human-in-the-loop (HITL) annotation, and the broader challenge of analyzing implicit meaning in large-scale social discourse. Specifically, the integration of a multi-agent system with confidence-ordered replacement demonstrated superior performance across multiple datasets, suggesting that modeling the perspectives of diverse online audiences can improve the detection of context-dependent, implicit frames that are often overlooked in communication artifacts. These results provide empirical support for sociolinguistic theories of framing as cognitive structures that shape interpretation rather than explicit textual elements (Entman, 1993; Goffman, 1974).

The data sets used in our evaluation, primarily composed of short-form social media discourse, illustrate the linguistic ambiguity and context dependence that make automated frame detection particularly challenging. Unlike traditional NLP tasks focused on explicit stance or sentiment, frame detection requires computational models to infer the underlying social, cultural, or ideological lenses through which statements are interpreted. This distinction highlights a core limitation in natural language learning: models must move beyond surface-level pattern recognition to engage in a form of structured interpretive reasoning. Our findings indicate that multi-agent simulation, when coupled with confidence scoring, can significantly enhance annotation of ambiguous language, thereby improving both large-scale frame analysis and human annotation workflows. Not only does this methodology most often find the correct classification, but importantly denotes when classifications may be uncertain. This underscores the need for natural language learning systems to incorporate mechanisms for modeling audience perspectives and contextual variation when developing computational methods for discourse analysis.

Our study offers a pathway toward disentangling factual claims from interpretive frames at scale, allowing researchers to more effectively analyze how audiences construct meaning from evidence rather than merely identifying "true" versus "false" statements. By refining our ability to computationally model these interpretive processes, we provide a stepping stone for future sociolinguistic and NLP research into how framing influences discourse dy-

namics in online spaces.

6.1 Limitations and Future Work

One of the primary challenges in frame detection—especially for natural language learning systems—is determining which type of frame a model should prioritize. Expert coders intuitively differentiate between specific frames (e.g., "voting machines are untrustworthy") and broader meta-frames (e.g., "elections are vulnerable"), but computational models struggle to replicate this hierarchical reasoning. In cases where frames are not explicitly stated, distinguishing between overlapping or competing frames remains difficult, making it challenging for models to generalize across different annotation schemes.

This issue highlights a key limitation in our approach: to avoid overfitting, we refrained from extensively tailoring prompts to specific datasets. While this ensured that our methodology remained generalizable, it likely introduced performance trade-offs. Future research could explore prompt engineering strategies that better balance dataset-specific framing nuances with a more standardized, transferable approach to frame detection. Additionally, refining the role descriptions within our multi-agent system may yield improvements in both interpretability and downstream accuracy.

Another challenge relates to platform affordances that influence textual framing. Our system struggled with quote tweets, hashtags, and cases where ambiguous responses (e.g., emoji reactions) altered the implied meaning of the original post. These affordances introduce structural complexities that natural language learning models often overlook. Future work should examine how discourse formatting—such as threaded conversations, hyperlinking, or multimodal elements—affects framing detection and model performance.

Overall, the significance of our study lies in its ability to enhance computational understanding of web-based, ambiguous social-communication processes that have historically been difficult to analyze. By identifying the interpretive lens through which social media users engage with information, our system could inform interventions that prompt perspective shifts, foster constructive narratives, and support healthier collective sensemaking. Such applications may reduce polarization and help maintain trust and community in online spaces.

7 Availability and Resources

All code, codebooks, prompts to produce these experiments can be found at https://anonymous.4open.science/r/frame_detection-D7C5/readme.md. Two NVIDIA A6000 GPUs were used over the course of 18 hours for local LLMs.

References

Ahmer Arif, John J. Robinson, Stephanie A. Stanek, Elodie S. Fichet, Paul Townsend, Zena Worku, and Kate Starbird. 2017. A Closer Look at the Self-Correcting Crowd: Examining Corrections in Online Rumors. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 155–168, Portland Oregon USA. ACM.

Iain J Cruickshank and Lynnette Hui Xian Ng. 2024. Diverse: Deciphering internet views on the us military through video comment stance analysis, a novel benchmark dataset for stance classification. *arXiv preprint arXiv:2403.03334*.

Nicholas DiFonzo and Prashant Bordia. 2007. *Rumor psychology: Social and organizational approaches*. Rumor psychology: Social and organizational approaches. American Psychological Association, Washington, DC, US. Pages: x, 292.

Robert M Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of communication*, 43(4):51–58.

David Farr, Iain Cruickshank, Nico Manzonelli, Nicholas Clark, Kate Starbird, and Jevin West. 2024a. Llm confidence evaluation measures in zero-shot css classification. *Preprint*, arXiv:2410.13047.

David Farr, Nico Manzonelli, Iain Cruickshank, and Jevin West. 2024b. Red-ct: A systems design methodology for using llm-labeled data to train and deploy edge classifiers for computational social science. *Preprint*, arXiv:2408.08217.

Akshay Goel, Almog Gueta, Omry Gilon, Chang Liu, Sofia Erell, Lan Huong Nguyen, Xiaohong Hao, Bolous Jaber, Shashir Reddy, Rupesh Kartha, et al. 2023. Llms accelerate annotation for medical information extraction. In *Machine Learning for Health (ML4H)*, pages 82–100. PMLR.

Erving Goffman. 1974. *Frame analysis: An essay on the organization of experience*. Frame analysis: An essay on the organization of experience. Harvard University Press, Cambridge, MA, US. Pages: ix, 586.

Ian Kennedy, Morgan Wack, Andrew Beers, Joseph S. Schafer, Isabella Garcia-Camargo, Emma S. Spiro, and Kate Starbird. 2022. Repeat Spreaders and Election Delegitimization: A Comprehensive Dataset of

Misinformation Tweets from the 2020 U.S. Election. *Journal of Quantitative Description: Digital Media*, 2.

Gary Klein, J.K. Phillips, E.L. Rall, and Deborah Peluso. 2007. A data-frame theory of sensemaking. *Expertise out of Context: Proceedings of the Sixth International Conference on Naturalistic Decision Making*, pages 113–155.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2023. Better zero-shot reasoning with role-play prompting. *arXiv preprint arXiv:2308.07702*.

Xiaochong Lan, Chen Gao, Depeng Jin, and Yong Li. 2024. Stance detection with collaborative role-infused llm-based agents. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 891–903.

Jiyi Li. 2024. A comparative study on annotation quality of crowdsourcing and llm via label aggregation. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6525–6529.

Do Xuan Long, Duong Ngoc Yen, Anh Tuan Luu, Kenji Kawaguchi, Min-Yen Kan, and Nancy F. Chen. 2024. Multi-expert prompting improves reliability, safety, and usefulness of large language models. *Preprint*, arXiv:2411.00492.

Jim Maddock, Kate Starbird, Haneen J. Al-Hassani, Daniel E. Sandoval, Mania Orand, and Robert M. Mason. 2015. Characterizing Online Rumoring Behavior Using Multi-Dimensional Signatures. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 228–241. Conference Name: CSCW '15: Computer Supported Cooperative Work and Social Computing ISBN: 9781450329224 Place: Vancouver BC Canada Publisher: ACM.

Lynnette Hui Xian Ng and Kathleen M Carley. 2022. Is my stance the same as your stance? a cross validation study of stance detection datasets. *Information Processing & Management*, 59(6):103070.

Dong Nguyen, A Seza Doğruöz, Carolyn P Rosé, and Franciska De Jong. 2016. Computational sociolinguistics: A survey. *Computational linguistics*, 42(3):537–593.

Valeria Pastorino, Jasivan A Sivakumar, and Nafise Sadat Moosavi. 2024. Decoding news narratives: A critical analysis of large language models in framing bias detection. *arXiv preprint arXiv:2402.11621*.

Stephen Prochaska, Julie Vera, Douglas Lew Tan, Ben Yamron, Sylvie Venuto, Amaya Kejriwal, Sarah Chu, and Kate Starbird. Deep Storytelling: Collective Sensemaking and Layers of Meaning in U.S. Elections. CSCW1. Submitted for publication to *Conference on Computer-Supported Cooperative Work & Social Computing* October 2025.

Joseph S. Schafer, Kayla Duskin, Stephen Prochaska, Morgan Wack, Anna Beers, Lia Bozarth, Taylor Agajanian, Mike Caulfield, Emma S. Spiro, and Kate Starbird. 2024. ElectionRumors2022: A Dataset of Election Rumors on Twitter During the 2022 US Midterms. *arXiv preprint*. ArXiv:2407.16051 [cs].

Tomotsu Shibutani. 1966. *Improvised news: A sociological study of rumor*. Improvised news: A sociological study of rumor. Bobbs-Merrill, Oxford, England. Pages: x, 262.

Kate Starbird, Stephen Prochaska, and Ben Yamron. What is going on? an evidence-frame framework for analyzing online rumors about election integrity. CSCW1. Submitted for publication to *Conference on Computer-Supported Cooperative Work & Social Computing* October 2025.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation: A survey. *Preprint*, arXiv:2402.13446.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024. Human-llm collaborative annotation through effective verification of llm labels. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Maxwell Weinzierl and Sanda Harabagiu. 2024. Discovering and articulating frames of communication from social media using chain-of-thought reasoning. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1617–1631.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Preprint*, arXiv:2305.03514.

8 Appendices

A Prompts

A.1 Maricopa Normal

Statement: STATEMENT

Framing is selecting aspects of data to communicate a point in text. Which of the following frames is the author most likely using to describe the democrat ran elections in Maricopa County, Arizona?

A: voter fraud - voting issues are part of deliberate cheating or manipulation.
B: election mess - organizers of the election are incompetent or disorganized, leading to issues.
C: untrustworthy elections - mistakes were unintentional, but they create doubt about the legitimacy of the election.
D: fake news - the news report itself is false or misleading.
E: trustworthy elections - the election results are legitimate, and the process was carried out fairly.

Pick the letter option 'A','B', 'C', 'D', or 'E' that makes most sense and nothing else.

A.2 Maricopa Susceptible

Framing is selecting aspects of data to communicate a point in text. You are playing the role of a social media user who could be susceptible to election controversies or conspiracies and see the following statement: STATEMENT
What frame do you believe the author of the statement is conveying?
Same multiple choice options and question statement as above.

A.3 Maricopa Conspiracy

Framing is selecting aspects of data to communicate a point in text. You are playing the role of a social media user who believes in many conspiracies related to election controversies and see the following statement: STATEMENT
What frame do you believe the author of the statement is conveying?
Same multiple choice options and questions as above.

A.4 Maricopa Susceptible

Framing is selecting aspects of data to communicate a point in text. You are playing the role of a social media user who could be susceptible to election controversies or conspiracies and see the following statement: STATEMENT
What frame do you believe the author of the statement is conveying?
Same multiple choice options and question statement as above.

A.5 Maricopa Naive

Framing is selecting aspects of data to communicate a point in text. You are playing the role of a social media user who has limited exposure to election controversies and see the following statement: STATEMENT
What frame do you believe the author of the statement is conveying?
Same multiple choice options and question statement as above.

A.6 Deep Stories Normal

Statement: STATEMENT
Framing is selecting aspects of data to communicate a point in text. Which of the following frames is the author most likely using in the above statement?
A: Explicit claim of fraud - voting issues are part of deliberate cheating or manipulation or officials should be jailed. In discussions of election or voting issues, if an issue is claimed to be intentional, that claim would qualify as a claim of fraud.
B: Explicit claim of suppression or disenfranchisement – Voting or election issues indicate intentional voter suppression or disenfranchisement. If a text claims voters are unable to vote due to procedural issues and fraud or doubt about election integrity is not insinuated, then this frame is present.
C: Sows doubt – Events and issues surrounding voting and elections might indicate fraud, but fraud is not explicitly stated or described. This frame is present in texts that ask leading questions, use provocative emojis, or otherwise provide a reason to doubt election integrity without outright stating that issues are intentional or that fraud is occurring.
D: Claim that fraud is not an issue – Widespread fraud is not an issue, and claims of fraud are misguided or intentionally false. More generally, election results are legitimate, and the election process was carried out fairly.
E: No insinuation of fraud or malfeasance – A neutral discussion of elections or election issues that doesn't claim or insinuate fraud, but also does not deny it or promote election integrity.
Pick the letter option 'A','B', 'C', 'D', or 'E' that makes most sense and nothing else.

A.7 Deep Stories Conspiracy

Framing is selecting aspects of data to communicate a point in text. You are playing the role of a social media user who believes in many conspiracies related to election controversies and see the following statement: STATEMENT
What frame do you believe the author of the statement is conveying?

Same multiple choice options and questions as above.

A.8 Deep Stories Susceptible

Framing is selecting aspects of data to communicate a point in text. You are playing the role of a social media user who could be susceptible to election controversies or conspiracies and see the following statement: STATEMENT
What frame do you believe the author of the statement is conveying?
Same multiple choice options and question statement as above.

A.9 Deep Stories Naive

Framing is selecting aspects of data to communicate a point in text. You are playing the role of a social media user who has limited exposure to election controversies and see the following statement: STATEMENT
What frame do you believe the author of the statement is conveying?
Same multiple choice options and question statement as above.

B Confidence Score Distributions

The distribution of confidence scores overlaid with correct and incorrect data annotation can be seen below.

C HITL Performance Graphs

Final graphic is shown in text as Figure 4.

D Model Parameters

Temperature was set to 0 for both Llama-7b and GPT-4o. Logit-bias for GPT-4o was set to 10 for each token ID being searched for. Additional information on parameter settings and all code can be found in referenced GitHub in the availability statement.

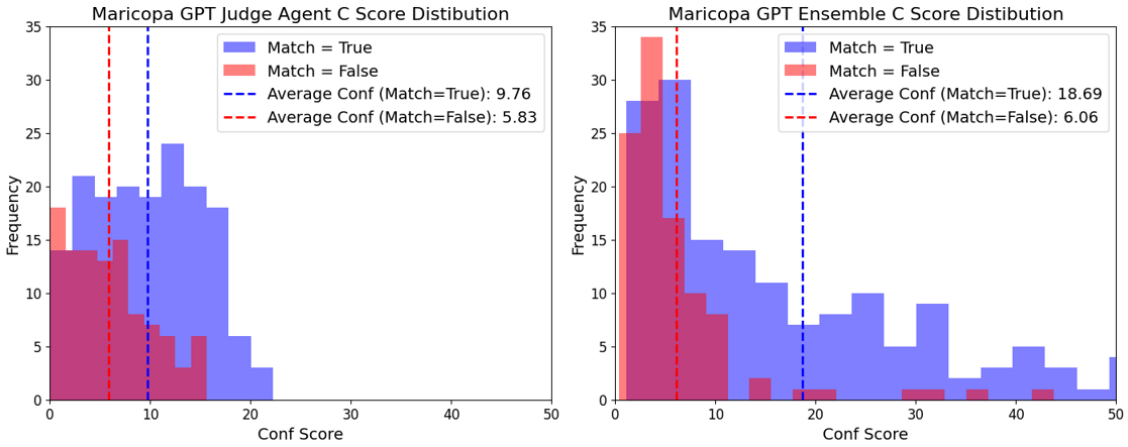

Figure 3: Confidence Score distributions of the Ensemble and Judge Agent methods. Change in the distribution of confidence scores to further separate the true annotations from the false annotations as shown in the graphic allow for better human sampling of incorrectly detected frames based on low confidence scores.

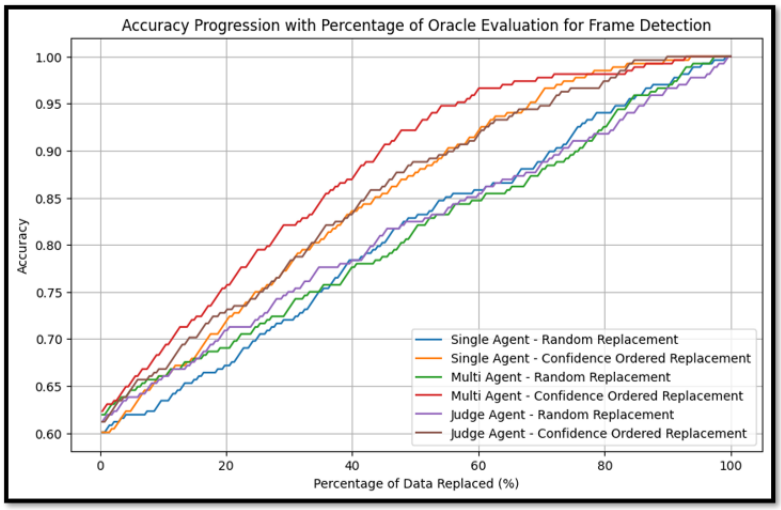

Figure 4: Accuracy progression as oracle/human-expert labels are applied across each system methodology for the Maricopa County Dataset using GPT-4o. Our multi-agent system with a confidence ensemble and informed annotation achieves a high performance with limited human intervention.

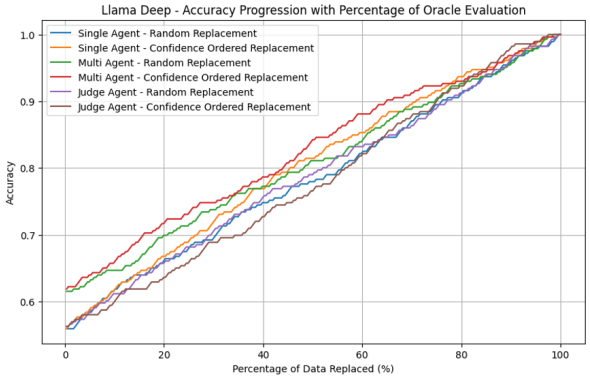

Figure 5: Accuracy progression as oracle/human-expert labels are applied across each system methodology for the Deep Stories Dataset using Llama-7b

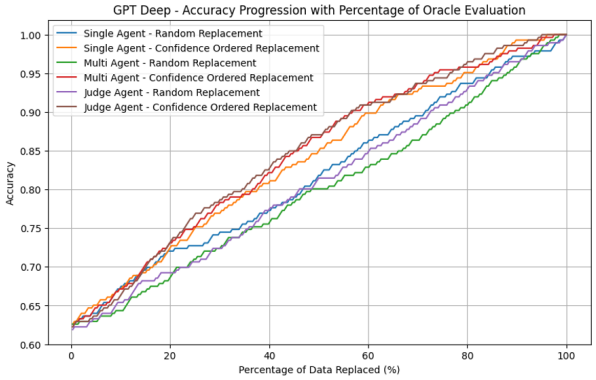

Figure 6: Accuracy progression as oracle/human-expert labels are applied across each system methodology for the Deep stories Dataset using GPT-4o.

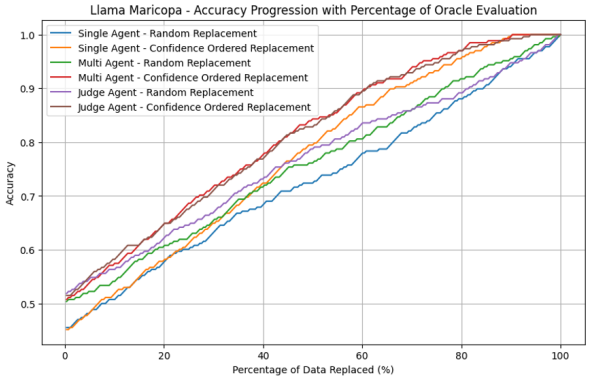

Figure 7: Accuracy progression as oracle/human-expert labels are applied across each system methodology for the Maricopa County Dataset using Llama-7b.