Uncertainty perception by humans and LLMs in a multi-agent social network simulation

Tim Donkers[1], Jürgen Ziegler[1]

[1]University of Duisburg-Essen, Germany

Abstract

We investigate whether Large Language Models can serve as reliable proxies for human judgment when assessing perceived uncertainty in polarized social media discourse. Our approach uses calibrated LLM "mirror personas" that replicate individual human participants, enabling direct paired comparison between human and LLM assessments of the same experimental conditions. Results show that both humans and the LLM reliably distinguish polarized from moderate discourse, but the LLM amplifies condition differences, most strongly for uncertainty (5.0×). Mediation analysis reveals that humans use uncertainty as an integrative lens through which polarization colors other perceptions, while the LLM processes constructs independently.

Keywords

uncertainty quantification, large language models, human-AI alignment, validation study, social media discourse

1. Introduction

Although uncertainty has been a long-standing subject of research, uncertainty is not a well-defined concept. Its interpretation depends on the method and context for which uncertainty is assessed. The term uncertainty is closely related, but not equivalent, to confidence [1]. Yet, the terms are often used interchangeably. One important distinction in uncertainty assessment in AI lies in the type of task the system is meant to fulfill. Uncertainty in classical machine learning tasks such as classification or regression has been mostly analyzed with respect to the expected variability of some accuracy metrics [2]. The challenge of measuring uncertainty becomes much more complex, however, when generative tasks are concerned, such as when prompting an LLM for a textual response. These tasks are typically open-ended, there is not one correct or best target answer the system should output. Instead, a request may be answered by a large variety of textual responses, some of which may be semantically equivalent despite different surface forms.

A large body of work has emerged in recent years to quantify uncertainty in LLM-based text-generation [3, 4]. The goal of these uncertainty quantification methods is mostly to provide an estimate of the factual correctness of the system's response to a given prompt. Several approaches have been reported in which agents verbalize statements (often in numerical form) about the uncertainty of the response [5, 6]. For communicating accuracy-related uncertainty, most works present an uncertainty metric to users [4]. Some approaches have applied this form of uncertainty communication for creating uncertainty-aware explanations [7, 8]. As an alternative to presenting metrics, [9] have studied counterfactuals but found them still unreliable.

Yet, uncertainty may also refer to concepts beyond mere factuality and correctness, involving linguistic characteristics of generated texts such as the stance expressed in a statement or the level of doubt or self-reflection expressed by an agent about its own output [10]. Our research addresses this multi-faceted type of uncertainty which arises in a multi-agent setting where agents autonomously create messages that are communicated to and interpreted by other agents. Uncertainty assessment in this context goes beyond the prompt-response pattern and needs to draw upon linguistic markers of uncertainty embedded in the generated text itself. We study uncertainty and related factors in context of a self-developed platform for simulating polarization dynamics in a synthetic social network [11]. The

Joint Proceedings of the ACM Intelligent User Interfaces (IUI) Workshops 2026, March 23-26, 2026, Paphos, Cyprus

tim.donkers@uni-due.de (T. Donkers); juergen.ziegler@uni-due.de (J. Ziegler)

0000-0002-9230-1243 (T. Donkers); 0000-0001-9603-5272 (J. Ziegler)

platform is a fully-functional copy of real-world social media (such as X) in which synthetic LLM-based agents interact with each other and also with human users to provide an experimental testbed for studying the temporal development of opinion in humans and agents. Agents in this network generate and post messages and react to other agent's messages by liking, forwarding or commenting them. Their behavior is fully probabilistic, governed by injected personas and probability distributions of agent stance (towards a given topic of discourse) and engagement likelihood. Agents are capable of interpreting incoming messages with respect to a number of factors expressed in the message, including opinion, group salience, and uncertainty. With this testbed, we have been studying factors accelerating or mitigating the development of extreme positions that may result in network polarization.

The effectiveness of this simulation approach for conducting user studies critically depends on the human-likeness of the posts created by the agents and their behavior. We therefore performed a study comparing the perception of generated posts by humans and by virtual agents. With this study we aim to answer the following research questions: (1) How well is human and agent perception of posts aligned with respect to expressed opinion, emotionality, sense of group belonging, and uncertainty? (2) What is the perception of these factors in polarized and unpolarized settings and how are they related to each other?

Our results reveal a dissociation between detection and calibration. Both humans and LLMs reliably distinguish polarized from moderate discourse across all four constructs, establishing convergent validity. However, significant rater-by-condition interactions show that LLMs amplify condition differences, most strongly for uncertainty itself (5.0× effect size amplification). Emotionality emerges as a striking exception: human and LLM perception align without significant interaction, suggesting that emotional register represents a perceptually stable dimension where explicit linguistic markers suffice for comparable assessment. Most notably, mediation analysis uncovers a fundamental architectural difference: for humans, uncertainty mediates 22–32% of polarization's effects on other constructs, functioning as a lens through which polarization colors broader discourse impressions. For LLMs, no mediation pathways reach significance, as constructs are processed independently rather than integratively. These findings validate LLM-based assessment for comparative research while highlighting processing differences with implications for AI systems that must communicate uncertainty to human users.

2. Background

Classical uncertainty research in the field of machine learning has mainly addressed closed tasks such as classification, regression, or ranking which have well-defined goals, mostly with categorical or ordinal outputs. In these contexts, uncertainty is usually interpreted as the variability of the accuracy value of a prediction, where accuracy refers to the distance of the predicted value from some ground truth [12]. This interpretation is hard to uphold, however, for open-ended tasks which are typical for generative AI where there is mostly not a single correct outcome but rather a set of generated artifacts that are more or less adequate for the intended goal. Determining uncertainty for such systems therefore poses a major challenge, which is further exacerbated when the size of the generated artifact increases, e. g., when moving from single generated statements to complete text products. In our research, we investigate uncertainty expressed in LLM-generated medium-sized messages which are not requested in a typical prompt-response cycle but are generated in a multi-agent network where agents autonomously decide whether and when to post messages or to react to other agents' posts.

2.1. Uncertainty quantification in large language models

In line with the sharp rise in interest in LLMs, a large body of work has emerged in recent years to quantify uncertainty (UQ) in LLM-based text-generation [13, 14]. A main distinction exists between uncertainty estimates for black-box and white-box models [15]. The majority of UQ methods for LLMs applies to white-box models which allow complete or partial access to the internal parameters of the model [16]. Token-level UQ is a typical white-box approach that exploits the probability distributions of output tokens. Black-box UQ methods typically require evaluating the similarity between multiple

responses generated by an LLM on the same or similar prompts. Because differently formulated responses that can convey the same semantic content, methods have been proposed that perform a semantic clustering of a set of responses, from which, for instance, semantic entropy is calculated [17].

Recently, several methods have been proposed that train black-box models to self-verbalize an estimate of their confidence [18] in the output they create. Most of these methods map self-assessed uncertainty to a confidence value that is presented to the user either in numerical form or as a textual statement, similar to the way humans convey their confidence to a communication partner. In a related line of research, approaches have been proposed that use an LLM to judge the output of another model with respect to different criteria [19]. To assess uncertainty, some approaches draw upon linguistic markers in the generated text [10]. Almost all LLM-based assessment methods aim at evaluating the correctness or factuality of an LLM's response to a given prompt. Yet, uncertainty can also refer to concepts beyond mere factuality, involving a range of factors such as such as vagueness of the stance towards a topic, epistemic uncertainty acknowledging knowledge limitations, or uncertain argumentation. All these factors, indicated through linguistic markers in the generated text, can express different levels of doubt or self-reflection expressed by an agent about its own output. This wider perspective on uncertainty is in the focus of our work on multi-agent social network simulation, yet no specific research has, to our knowledge, addressed these questions so far. Furthermore, only limited work has compared LLM-based uncertainty assessment of agent-generated messages to human perception [20] which is a central question for our human-agent communication testbed.

2.2. LLM-Based Social Network Simulation

LLM-powered agents enable social simulations with unprecedented behavioral realism. Park et al.'s generative agents demonstrated believable social behaviors through memory and reflection mechanisms [21]. This paradigm has extended to networked opinion dynamics: the S3 system populates synthetic networks with content-generating agents [22]; Chuang et al. showed LLM networks reproduce opinion fragmentation, though with bias toward accurate information [23]; recent work demonstrates spontaneous emergence of human-like polarization and echo chambers among thousands of LLM agents [24]; and Wang et al. validated simulations against bounded confidence and Friedkin-Johnsen models [25].

These studies validate emergent network-level phenomena but pay less attention to construct-level alignment, specifically whether features like uncertainty or emotionality are perceived similarly by humans and LLMs. Our study addresses this by directly comparing matched human-LLM assessments of identical messages.

3. Method

We investigate whether LLMs can serve as reliable proxies for human judgment when assessing perceived uncertainty in social media discourse, using calibrated "mirror personas" that replicate individual participants for direct paired comparison.

3.1. User Study

Participants ($N = 122$) evaluated AI-generated social media posts discussing Universal Basic Income (UBI). Messages were generated by an agent-based simulation [11] varying the *opinion distribution*: in the *polarized* condition ($n = 58$), agents held bimodal distributions at opposing extremes (± 0.7 on a $[-1, +1]$ scale); in the *moderate* condition ($n = 64$), agents held unimodal distributions centered at neutral positions ($\mu = 0, \sigma = 0.1$). Each participant evaluated approximately 20 messages from their assigned condition. Table 1 presents representative messages illustrating the stylistic differences that emerged.

Table 1
Example Messages by Condition

Polarized Condition
Pro-UBI: "Enough is enough! It's time to EMBRACE Universal Basic Income and reject the shackles of outdated welfare systems! UBI can finally dismantle the poverty cycle, offering true financial SECURITY and dignity to ALL. Opponents? Just scared guardians of a broken status quo!" *Contra-UBI:* "The UBI madness is spreading like a toxic weed! Cash handouts won't save Economica; they'll destroy our work ethic! Those peddling this insanity are enemies of progress!"
Moderate Condition
Pro-UBI: "Feeling a bit hopeful about Universal Basic Income! It could provide financial security for all, but how might it impact motivation and innovation? Let's keep the dialogue going!" *Contra-UBI:* "While the idea of UBI sounds enticing, we must consider the potential downsides. Could it reduce motivation to work or drive inflation? It's vital to examine how we can support community contributions without losing individual drive."

Note: Messages generated by LLM agents with calibrated opinion distributions.

3.2. Constructs

Participants rated messages on four constructs using multi-item scales with two positively-worded and two reverse-coded items each, developed for our simulation framework [11]. Three constructs used 5-point Likert scales; perceived polarization used a 7-point scale. After examining internal consistency, reverse-coded items for group salience and perceived polarization were excluded due to low item-total correlations, yielding acceptable reliability across rater types (α_H/α_L denote Cronbach's α for human/LLM raters). Full scale items and additional psychometric detail are reported in [11].

Perceived uncertainty ($\alpha_H = .74$, $\alpha_L = .90$) captures the degree to which authors signal openness to being wrong, assessing whether authors acknowledge knowledge limitations, express doubt, or make absolute statements (reverse-coded).

Emotionality ($\alpha_H = .84$, $\alpha_L = .76$) measures the emotional versus logical register of discourse, assessing whether messages are charged with emotional content or maintain a calm, reasoned tone.

Group salience ($\alpha_H = .64$, $\alpha_L = .75$) captures the prominence of us-versus-them rhetoric, including emphasis on group distinctions and reference to group memberships in arguments.

Perceived polarization ($\alpha_H = .60$, $\alpha_L = .88$) assesses the degree to which discourse appears entrenched, measuring expression of polarized viewpoints and extreme or entrenched positions.

These constructs were chosen to capture both epistemic openness (uncertainty) and the core dimensions of affective polarization, emotional reactivity and group-based identity [26, 27], that jointly determine the downstream impression of dialogic closure (perceived polarization).

3.3. LLM Simulation

To enable direct human-LLM comparison, we developed a "mirror persona" methodology in which each human participant is matched with a calibrated LLM counterpart. Each persona received calibration context derived from the full human sample ($N = 122$), including population-level rating distributions ($M \pm SD$ per item) and guidance on typical rating variance. Crucially, the LLM never received the matched participant's actual ratings for the items being assessed; instead, calibration provided aggregate human patterns to anchor responses within realistic ranges. Each persona additionally inherited the matched participant's pre-interaction UBI stance (5-point scale), ensuring comparable prior attitudes.

Rather than eliciting point estimates, we instructed the LLM to predict a rating distribution (mean and standard deviation) for each item. Final ratings were sampled from these predicted distributions, better capturing the inherent uncertainty in human judgment. Each construct was evaluated in a separate LLM call to prevent cross-construct contamination. We used DeepSeek Chat as our model with temperature set to 1.5.

Table 2
Descriptive Statistics by Condition and Rater Type

	Polarized ($n = 58$)		Moderate ($n = 64$)	
Construct	Human	LLM	Human	LLM
Uncertainty	2.08 (0.66)	1.64 (0.30)	2.90 (0.64)	3.86 (0.39)
Emotionality	3.60 (0.78)	4.29 (0.36)	2.43 (0.74)	2.95 (0.53)
Group Salience	3.12 (0.64)	3.98 (0.50)	2.43 (0.66)	2.19 (0.56)
Polarization†	4.68 (1.20)	6.29 (0.41)	3.25 (0.86)	2.98 (0.64)

Note: M (SD). Scales: 1–5 except †1–7.

3.4. Analysis

We employed mixed-design ANOVAs (polarization × rater type) to test detection validity and interaction effects, computed Cohen's d amplification ratios, and conducted mediation analysis testing whether uncertainty mediates polarization's effects on other constructs differently for humans versus LLMs.

4. Results

Table 2 presents means and standard deviations by condition and rater type. Both humans and the LLM showed expected patterns: polarized conditions elicited lower uncertainty, higher emotionality, stronger group salience, and more perceived polarization. The LLM produced more extreme ratings with reduced variance, as visualized in Figure 1.

Mixed-design ANOVAs confirmed that both raters reliably distinguished conditions, with polarization explaining 66–81% of variance across all constructs ($p < .001$). However, significant rater × condition interactions emerged for uncertainty ($F(1, 120) = 115.98$, $p < .001$, $\eta_p^2 = .49$), group salience ($F(1, 120) = 46.81$, $p < .001$, $\eta_p^2 = .28$), and perceived polarization ($F(1, 120) = 76.18$, $p < .001$, $\eta_p^2 = .39$), indicating systematic LLM amplification. Emotionality stands as exception: while both raters detected differences ($F(1, 120) = 229.46$, $p < .001$, $\eta_p^2 = .66$), the interaction was non-significant ($F(1, 120) = 1.21$, $p = .27$, $\eta_p^2 = .01$), suggesting emotional register is perceptually stable across rater types.

The magnitude of amplification varied by construct (Table 3). Uncertainty emerged as most amplified: humans showed $d = 1.27$, but the LLM produced $d = 6.34$ (5.0× ratio). Emotionality showed minimal amplification (1.9×), consistent with its non-significant interaction.

Beyond detection validity, we examined whether uncertainty mediates the relationship between polarization and other discourse perceptions, and whether this mediating role differs between rater types. For human raters, uncertainty served as a significant partial mediator. Polarization's effect on perceived polarization was mediated by uncertainty (indirect effect = 0.32, 95% CI [0.06, 0.63], $z = 2.55$, $p = .011$), accounting for 22% of the total effect. Similarly, the effect on group salience was mediated (indirect effect = 0.22, 95% CI [0.06, 0.42], $z = 2.76$, $p = .006$), with uncertainty explaining 32% of the variance. The path to emotionality was not significantly mediated ($p = .55$).

For LLM raters, no mediation pathways reached significance ($p > .68$ for all outcomes). Despite the LLM's strong sensitivity to both polarization and uncertainty, these operated as parallel, independent effects rather than a causal chain. This asymmetry reveals a fundamental difference in processing architecture: humans appear to use uncertainty as a lens through which polarization colors other perceptions, whereas the LLM registers each dimension independently.

5. Discussion

Our findings reveal that LLMs reliably detect uncertainty signals in discourse, but process them through fundamentally different architectures than humans. We organize our discussion around three aspects of this dissociation.

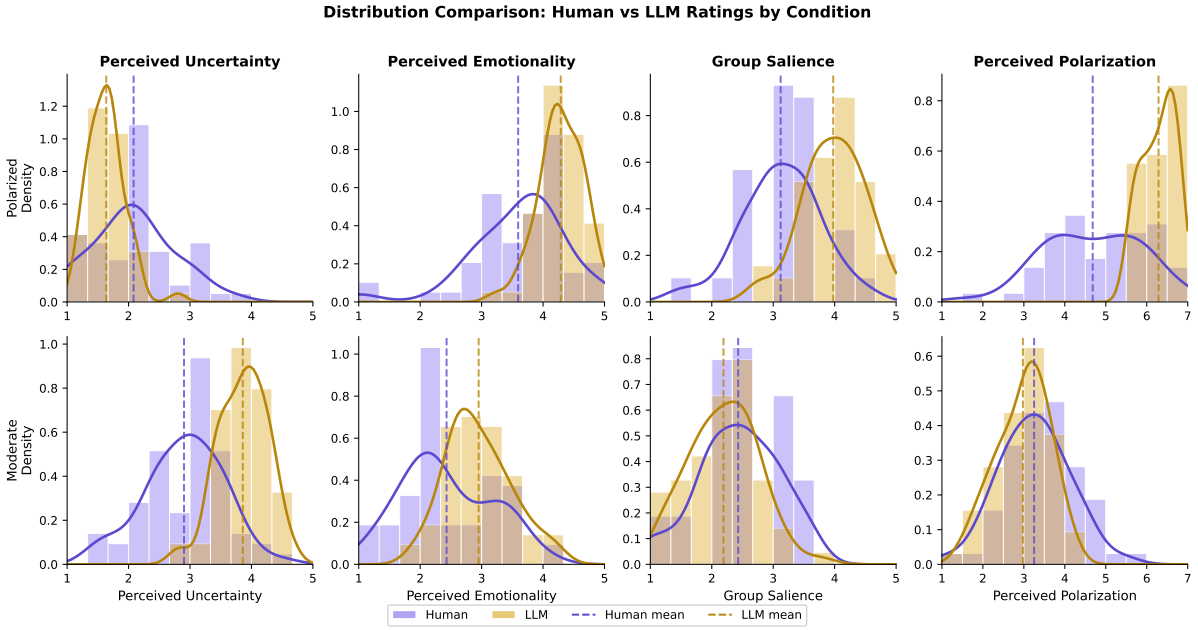

Figure 1: Rating distributions by condition and rater type. LLM ratings show reduced variance and greater condition separation.

Uncertainty detection is valid but amplified. Both humans and the LLM distinguished polarized from moderate discourse on perceived uncertainty, with polarization explaining 81% of variance. This confirms that LLMs can serve as valid instruments for comparative uncertainty assessment, since conditions humans perceive as more epistemically open are judged similarly by LLMs. However, the LLM amplified uncertainty differences by a factor of 5.0×, the largest amplification of any construct. This pattern held for group salience (3.1×) and perceived polarization (4.4×) as well, suggesting that LLMs function as contrast-maximizing systems that are valid for ranking but require calibration for absolute measurement.

Why uncertainty is amplified: the role of context-dependence. Emotionality provides a revealing contrast. It alone showed no significant interaction ($\eta_p^2 = .01$) and minimal amplification (1.9×). We hypothesize this reflects differences in context-dependence. Emotional expressions such as exclamation points, charged vocabulary, and affective intensifiers tend to be interpreted at face value. An exclamation point signals emphasis regardless of context; words like "outrageous" carry consistent affective valence.

Uncertainty markers, however, are more context-dependent. A hedge like "I think" might signal genuine epistemic doubt in one context but function as a politeness convention in another. "Perhaps" can indicate real uncertainty or rhetorical softening of a firmly held position. Humans, drawing on social and pragmatic knowledge, may discount hedges they recognize as formulaic. LLMs, lacking access to these contextual cues, weight all uncertainty markers consistently, which produces amplification precisely where human interpretation is most nuanced. This interpretation is speculative, but the data clearly establish context-independent constructs like emotionality as boundary conditions where human-LLM perception aligns.

Uncertainty as lens versus feature. Beyond detection, we find that uncertainty plays different *functional* roles in human versus LLM perception. For humans, uncertainty mediates the relationship between polarization and other discourse perceptions: perceiving epistemic closure leads to perceiving stronger group boundaries (32% mediated) and greater entrenchment (22% mediated). Uncertainty functions as a lens through which polarization colors interpretation.

For LLMs, no mediation pathways reached significance. Despite strong sensitivity to both polarization and uncertainty, these operated as parallel, independent assessments. Human perception appears *integrative*: constructs bleed into each other, with uncertainty shaping how other dimensions are

Table 3
Effect Size Comparison Between Rater Types

Construct	d_{Human}	d_{LLM}	Ratio
Uncertainty	1.27	6.34	5.0×
Emotionality	1.53	2.93	1.9×
Group Salience	1.06	3.35	3.1×
Polarization	1.39	6.06	4.4×

Note: Ratio = $|d_{\text{LLM}}|$ / $|d_{\text{Human}}|$.

interpreted. LLM perception appears *modular*: each construct is assessed through independent feature detection, missing the relational dynamics through which uncertainty shapes human interpretation.

Implications for uncertainty communication. These findings carry practical implications. LLMs provide valid comparative measurement of uncertainty, making them useful for screening or ranking discourse by epistemic openness. However, if LLMs assess uncertainty independently while humans experience it as interconnected with other perceptions, LLM-based analysis may miss relational dynamics. An intervention that increases expressed uncertainty might cascade to reduce perceived group salience and polarization for human readers; this effect would not be detected or predicted by LLM analysis alone.

The integrative nature of human uncertainty perception has direct consequences for the design of uncertainty communication in user-facing AI systems. Because humans use uncertainty as a lens that reshapes broader discourse impressions, the way an AI system expresses uncertainty does not merely set user expectations about factual reliability; it colors how users perceive the system's social positioning, group affiliation, and openness to dialogue. Designing calibrated uncertainty cues therefore requires attending to these downstream perceptual cascades rather than treating uncertainty as an isolated signal. Our findings suggest that when systems aim to foster realistic user expectations, they should consider that hedges, qualifiers, and epistemic markers will be interpreted holistically: a system that expresses appropriate doubt may simultaneously appear less partisan and more dialogically inviting. Conversely, the 5.0× amplification we observed warns that LLM-generated uncertainty cues risk overshooting, as what an LLM calibrates as moderate hedging may register to human readers as excessive equivocation, potentially undermining rather than fostering trust. Multi-agent settings, where AI-generated content circulates alongside human contributions, amplify these stakes: uncertainty signals shape not only individual user expectations but also the collective epistemic climate of the discourse environment.

Limitations. We used a single LLM (DeepSeek Chat); calibration parameters likely differ across models and should be validated. Messages addressed one topic (UBI) from one simulation framework; generalization requires testing across domains. The recursive design, in which LLMs assess LLM-generated content, may yield different patterns than assessment of human-authored text, a question future work should address.

Acknowledgements

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 551054271.

Declaration on Generative AI

During the preparation of this work, the authors used Claude Opus in order to: Grammar and spelling check, Paraphrase and reword, Improve writing style. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] A. Pouget, J. Drugowitsch, A. Kepecs, Confidence and certainty: distinct probabilistic quantities for different goals (2016). doi:10.1038/nn.4240.

[2] R. Ghanem, D. Higdon, H. Owhadi, Handbook of Uncertainty Quantification (2017).

[3] O. Shorinwa, Z. Mei, J. Lidard, A. Z. Ren, A. Majumdar, A Survey on Uncertainty Quantification of Large Language Models: Taxonomy, Open Research Challenges, and Future Directions (2025). doi:10.1145/3744238.

[4] X. Liu, T. Chen, L. Da, C. Chen, Z. Lin, H. Wei, Uncertainty Quantification and Confidence Calibration in Large Language Models: A Survey (2025). doi:10.1145/3711896.3736569.

[5] M. Xiong, Z. Hu, X. Lu, Y. Li, J. Fu, J. He, B. Hooi, Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs (2024). URL: https://arxiv.org/abs/2306.13063.

[6] S. Liu, Z. Li, X. Liu, R. Zhan, D. F. Wong, L. S. Chao, M. Zhang, Can LLMs Learn Uncertainty on Their Own? Expressing Uncertainty Effectively in A Self-Training Manner (2024). doi:10.18653/v1/2024.emnlp-main.1205.

[7] M. Förster, M. Hagn, N. Hambauer, P. Jaki, T. Darmstadt, A. Obermeier, M. Pinski, T. Darmstadt, A. Schauer, A. Schiller, A. Benlian, T. Darmstadt, B. Heinrich, E. Jussupow, T. Darmstadt, M. Klier, M. Kraus, D. Schnurr, A taxonomy for uncertainty-aware explanable ai, Amman, Jordan, 2025.

[8] L. Sheng, F. Chang, Q. Sun, D. Wangzha, Z. Gu, Uncertainty reports as explainable AI: A cognitive-adaptive framework for human-AI decision systems in context tasks, Advanced Engineering Informatics 68 (2025) 103634. URL: https://www.sciencedirect.com/science/article/pii/S1474034625005270. doi:10.1016/j.aei.2025.103634.

[9] H. Mayne, R. O. Kearns, Y. Yang, A. M. Bean, E. Delaney, C. Russell, A. Mahdi, LLMs Don't Know Their Own Decision Boundaries: The Unreliability of Self-Generated Counterfactual Explanations, in: EMNL 2025 - Conference on Empirical Methods in Natural Language Processing, 2025, pp. 24161–24186.

[10] J. Liu, Q. Zong, W. Wang, Y. Song, Revisiting epistemic markers in confidence estimation: Can markers accurately reflect large language models' uncertainty?, arXiv preprint arXiv:2505.24778 (2025).

[11] T. Donkers, J. Ziegler, Human-Agent Interaction in Synthetic Social Networks: A Framework for Studying Online Polarization (2025). URL: https://arxiv.org/abs/2502.01340.

[12] R. Ghanem, D. Higdon, H. Owhadi, et al., Handbook of uncertainty quantification, volume 6, Springer New York, 2017.

[13] Y. Xiao, W. Y. Wang, Quantifying Uncertainties in Natural Language Processing Tasks, Proceedings of the AAAI Conference on Artificial Intelligence 33 (2019) 7322–7329. URL: https://ojs.aaai.org/index.php/AAAI/article/view/4719. doi:10.1609/aaai.v33i01.33017322.

[14] J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher, M. Shahzad, W. Yang, R. Bamler, X. X. Zhu, A survey of uncertainty in deep neural networks, Artificial Intelligence Review 56 (2023) 1513–1589. URL: https://doi.org/10.1007/s10462-023-10562-9. doi:10.1007/s10462-023-10562-9.

[15] O. Shorinwa, Z. Mei, J. Lidard, A. Z. Ren, A. Majumdar, A Survey on Uncertainty Quantification of Large Language Models: Taxonomy, Open Research Challenges, and Future Directions, ACM Comput. Surv. 58 (2025) 63:1–63:38. URL: https://dl.acm.org/doi/10.1145/3744238. doi:10.1145/3744238.

[16] X. Liu, T. Chen, L. Da, C. Chen, Z. Lin, H. Wei, Uncertainty Quantification and Confidence Calibration in Large Language Models: A Survey, in: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2, KDD '25, Association for Computing Machinery, New York, NY, USA, 2025, pp. 6107–6117. URL: https://dl.acm.org/doi/10.1145/3711896.3736569. doi:10.1145/3711896.3736569.

[17] L. Kuhn, Y. Gal, S. Farquhar, Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation, 2023. URL: http://arxiv.org/abs/2302.09664. doi:10.48550/

arXiv.2302.09664, arXiv:2302.09664 [cs].
[18] S. Liu, Z. Li, X. Liu, R. Zhan, D. F. Wong, L. S. Chao, M. Zhang, Can LLMs Learn Uncertainty on Their Own? Expressing Uncertainty Effectively in A Self-Training Manner, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 21635–21645. URL: https://aclanthology.org/2024.emnlp-main.1205. doi:10.18653/v1/2024.emnlp-main.1205.
[19] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al., Judging llm-as-a-judge with mt-bench and chatbot arena, Advances in neural information processing systems 36 (2023) 46595–46623.
[20] C. G. Belem, M. Kelly, M. Steyvers, S. Singh, P. Smyth, Perceptions of Linguistic Uncertainty by Language Models and Humans, 2024. URL: http://arxiv.org/abs/2407.15814. doi:10.48550/arXiv.2407.15814, arXiv:2407.15814 [cs].
[21] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, M. S. Bernstein, Generative agents: Interactive simulacra of human behavior (2023).
[22] C. Gao, X. Lan, Z. Lu, J. Mao, J. Piao, H. Wang, D. Jin, Y. Li, S3: Social-network Simulation System with Large Language Model-Empowered Agents (2023). URL: https://arxiv.org/abs/2307.14984.
[23] Y.-S. Chuang, A. Goyal, N. Harlalka, S. Suresh, R. Hawkins, S. Yang, D. Shah, J. Hu, T. Rogers, Simulating Opinion Dynamics with Networks of LLM-based Agents (2024). URL: https://aclanthology.org/2024.findings-naacl.211. doi:10.18653/v1/2024.findings-naacl.211.
[24] J. Piao, Z. Lu, C. Gao, F. Xu, Q. Hu, F. P. Santos, Y. Li, J. Evans, Emergence of human-like polarization among large language model agents (2025). URL: https://arxiv.org/abs/2501.05171.
[25] C. Wang, Z. Liu, D. Yang, X. Chen, Decoding Echo Chambers: LLM-Powered Simulations Revealing Polarization in Social Networks (2024). URL: https://arxiv.org/abs/2409.19338.
[26] S. Iyengar, G. Sood, Y. Lelkes, Affect, Not Ideology: A Social Identity Perspective on Polarization (2012). doi:10.1093/poq/nfs038.
[27] H. Tajfel, J. C. Turner, The Social Identity Theory of Intergroup Behavior. (2004). doi:10.4324/9780203505984-16.