From Factuality to Meta-Factivity: A Cognitive Blueprint for Trustworthy LLMs

Daohuan Liu[1*], **Lun Xia**[1*], **Yuer Wang**[1], **Jiaoyang Su**[1], **Xuri Tang**[1†]

[1]School of Foreign Languages, Huazhong University of Science and Technology
{liudh, xiaxx, wyeer, sujiaoyang1, xrtang}@hust.edu.cn

Abstract

Current research on Event Factuality Prediction (EFP) predominantly treats LLMs as passive classifiers, where high aggregate metrics often mask shortcut learning and unreliable reasoning. In this position paper, we argue for a focus shift from event factuality to meta-factivity. We introduce the Meta-Factivity Framework (MFF), a theoretical roadmap that moves evaluation beyond surface recognition to belief trajectory reasoning and epistemic regulation. By framing hallucination as a failure of meta-cognitive control, we advocate for a transition from measuring black-box accuracy to evaluating white-box cognition, laying the groundwork for a more rigorous benchmark for explainable self-governance.

1 Introduction

Event Factuality Prediction (EFP) is the cornerstone of robust Natural Language Understanding, aiming to determine the veridical status of events in text (Saurí and Pustejovsky, 2009). While approaches have evolved from rule-based systems (Saurí and Pustejovsky, 2012) to deep neural baselines trained on large-scale datasets (e.g., Stanovsky et al., 2017; Rudinger et al., 2018; Veyseh et al., 2019; Liu et al., 2022), the evaluation paradigm remains largely focused on treating factuality as a static classification task, i.e., assigning a fixed label (e.g., *Factual*, *Counterfactual*, or *Uncertain*) to an event.

This position paper makes two key contributions: (1) diagnostically, we empirically expose how shortcut learning and fragile reasoning undermine current SOTA benchmarks; (2) prescriptively, based on linguistic insights, we introduce the Meta-Factivity Framework (MFF) as a roadmap to guide the transition from static labels to the agent's ability to regulate commitments. In this context, Meta-Factivity refers to a model's capacity to monitor and evaluate its own factuality reasoning process, and regulating commitments refers to how it dynamically adjusts its output certainty or factual stance to resolve logical conflicts.

For example, when a user inputs "The doctor did not deny that [the medicine caused the side effect]", the MFF not only requires the model to correctly label the bracketed event as *Factual*, but also assesses whether the model: (1) faithfully attributes this status to the interaction between the negation and the trigger verb, and (2) effectively regulates its commitments instead of succumbing to sycophancy upon receiving a contradictory update like "...but the patient reports no side effects". Hence, the MFF strictly prioritizes internal faithfulness over conversational plausibility.

2 Reflections on Event Factuality

2.1 Long-Tail Data and Shortcut Learning

Existing EFP benchmarks generally suffer from label skewness, and this long-tail distribution is widespread across languages. As shown in Table 1, the "Factual" category (the head class) consistently dominates across five representative benchmarks, while samples for "Uncertain" or "Counterfactual" categories (the tail classes) remain extremely sparse.

Although this imbalance reflects the natural distribution of texts, it does pose challenges for research. Lee et al. (2015) noted that sparse lexical cues may hinder learning efficiency: too few non-factual samples in the training set make it difficult for models to fully learn these tail categories. To alleviate this problem, recent work has attempted scaling and stratified sampling. MAVEN-FACT (Li et al., 2024) expanded the sample size to 11.5 times that of FactBank (Saurí and Pustejovsky,

[*] Equal Contribution
[†] Corresponding Author

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 62–69

Dataset	Language	Factual	Uncertain	Counterfactual
FactBank (Saurí and Pustejovsky, 2009)	English	58.1%	39.6%	2.3%
UW 2015 (Lee et al., 2015)	English	68.2%	29.0%	2.8%
ACE2005 (Cao et al., 2013)	Chinese	66.0%	32.0%	2.0%
MAVEN-FACT (Li et al., 2024)	English	93.7%	4.2%	2.1%
ModaFact (Rovera et al., 2025)	Italian	69.1%	26.5%	4.4%

Table 1: The long-tail distribution of factuality. Diverse dataset labels are mapped into three categories: "Factual" includes labels such as *CT+*, *Factual*, and *3*; "Uncertain" includes *PR*, *PS*, *Uu*, *Possible*, *Underspecified*, and *2,1,0,-1*; "Counterfactual" includes *CT-*, *Counterfactual*, and *-2,-3*.

2009) to increase the absolute volume of rare labels, whereas ModaFact (Rovera et al., 2025) adopted stratified sampling to ensure consistent proportions of rare categories across sets to guarantee evaluation fairness. While these strategies provide practical improvement schemes for model training and evaluation, the fundamental mechanistic vulnerability driven by the long-tail phenomenon remains unresolved.

As shown in Table 2, the F1 scores across categories exhibit distinct long-tail characteristics, showing a strong positive correlation with sample size. This suggests that high aggregate scores might be driven by shortcut learning, i.e., exploiting the prior probability to maximize expected accuracy on the majority class (e.g., Geirhos et al., 2020; McCoy et al., 2019). Notably, stronger models (like GPT-4) exhibit a larger performance drop, suggesting they are particularly prone to over-fitting head-class priors.

Dataset	Model	Head F1	Tail F1	ΔF1
MAVEN-FACT (Li et al., 2024)	*Sample %*	*93.7%*	*0.3%*	–
	GPT-4*	94.4	0.0	-94.4
	Mistral-7B	86.2	8.5	-77.7
	LLAMA 3	82.7	6.8	-75.9
ModaFact (Rovera et al., 2025)	*Sample %*	*69.1%*	*3.4%*	–
	mT5*	88.8	46.0	-42.8
	seqBIO	87.6	49.0	-38.6
	Aya	86.3	36.9	-49.4

Table 2: Performance degradation (ΔF1) across representative models (* denotes SOTA), illustrating the extreme correlation between head/tail data scarcity (Sample %) and F1 score drop.

Fine-grained analyses disclose deeper shortcut mechanisms. Take ModaFact (Rovera et al., 2025) as an example: although its *Counterfactual* and *Underspecified* cases share similar scarcity in the long-tail distribution, performance on the former significantly exceeds the latter. This disparity arises because *Counterfactual* samples typically contain explicit negation markers (e.g., not, never), whereas *Underspecified* samples require deep contextual inference. This aligns with broader Natural Language Inference (NLI) findings (e.g., White et al., 2018; Gururangan et al., 2018; Schuster et al., 2019), which confirm that neural models tend to exploit negation heuristics rather than engage in complex evidence reasoning. Li et al. (2024) also demonstrated that in approximately 30% of correct factuality predictions, models failed to extract the supporting evidence. This "right for the wrong reasons" phenomenon suggests that high performance may not stem from genuine semantic comprehension.

Crucially, high accuracy can obscure such reasoning failures. This may lead to blind affirmation in unseen contexts, thereby masking risks in high-stakes domains like medicine and law (Li et al., 2024). Therefore, future benchmarks must incorporate more adversarial probes to rigorously distinguish true reasoning from statistical shortcuts.

2.2 Unreliable Reasoning Trajectories

While Li et al. (2024) attempted to improve interpretability of current end-to-end EFP results via evidence word prediction, this approach struggles with analytic languages like Chinese. Unlike Indo-European languages marked by explicit morphology, Chinese factuality inference relies on implicit constructions, collocations, and pragmatic inference (Li and Yuan, 2023; Yuan, 2020). Thus, identifying discrete keywords is insufficient; understanding the compositionality of truth values is required.

Chain-of-Thought (CoT) offers a window into internal logic but introduces distinct risks such as post-hoc rationalization (Turpin et al., 2023) and hallucination snowballing (Zhang et al., 2025), where reasoning is fabricated to justify an initial guess. We examine this unreliability through two dimensions.

Internal stochasticity, or self-inconsistency (Elazar et al., 2021; Wang et al., 2022), refers to the inconsistency of LLMs across multiple runs for the same input. Meincke et al. (2025) report sig-

ID	Question Text	Multi-round Result	Self-consistency
Q1	Everyone knows that [the development of the West requires capital and technology].	9 *Uncertain*, 1 *Factual*	90%
Q2	People do not know that [the development of the West requires capital and technology].	10 *Uncertain*	100%
Q3	Everyone knows that [the development of the West requires capital and technology], but the person in charge pointed out that fundamentally, knowledge and talent are more needed.	10 *Factual*	100%
Q4	People do not know that [the development of the West requires capital and technology], because the person in charge pointed out that fundamentally, knowledge and talent are more needed.	6 *Uncertain*, 4 *Factual*	60%

Table 3: Multi-round reasoning results of Deepseek-V3.2 evaluating the factuality of the same bracketed event (10 runs per query). To test inferential robustness, we introduce linguistic perturbations including negation (Q2, Q4) and discourse extensions (Q3, Q4). Although the ground truth for all queries is *Factual*, the model exhibits severe judgment reversal and self-inconsistency. Self-consistency refers to the frequency of the majority answer.

nificant performance fluctuations in GPT-4o across 100 repeated runs on the GPQA dataset, even with the temperature fixed at 0.

External instability, or low robustness, refers to the model output being easily affected by non-critical perturbations. As shown in Table 3 from our probe experiment, the mere introduction of an additional discourse clause triggers severe volatility, including both judgment reversal (Q1 vs. Q3) and a drop in answer consistency (Q2 vs. Q4). This indicates that the model's factuality judgment is not anchored in the proposition itself, but is sensitive to surface-level noise (Jia and Liang, 2017).

Additionally, even when models perform well under explicit prompting, this underlying reasoning fragility highlights a broader limitation: they lack ecological validity in real-world scenarios requiring the implicit understanding of factual commitments.

2.3 The Ecological Validity Problem

Existing EFP research largely adheres to the NLI paradigm, treating models as passive classifiers in an idealized vacuum. However, LLMs function as dynamic interlocutors and generators. This role shift causes an ecological validity problem (Li et al., 2024), characterized by a "knowledge-action gap" in two dimensions:

First, the disconnect between discriminative ability and generative behavior. Traditional metrics assume that correct classification implies correct application. However, Kadavath et al. (2022) highlight that while models may possess accurate internal calibration, RLHF-driven "helpfulness" often leads to sycophancy (Sharma et al., 2023; Wei et al., 2023), i.e., prioritizing user intent over factual adherence. For instance, models frequently validate misleading premises to satisfy user requests, despite accurate internal factuality judgments. Thus, high discriminative scores mask the lack of a robust factual stance in generation.

Second, the misalignment between internal confidence and expression. Real-world interaction requires linguistic calibration—modulating output certainty through hedges (Eikema et al., 2025) such as *possibly* and *seems*. While Azaria and Mitchell (2023) show models can internally detect "lying," existing benchmarks fail to assess whether the surface realization matches this internal state. A possible event that is articulated with absolute certainty constitutes a severe safety risk (Liu et al., 2025).

In summary, trustworthy AI demands a shift from static classification to "full-link factuality evaluation," incorporating robustness, self-correction, and uncertainty management. Current research on factual faithfulness remains fragmented. This paper aims to systematize these efforts by establishing an evaluation framework grounded in human pragmatics.

3 Returning to Factivity: From Static Features to Dynamic Context

Factivity is classically defined as the linguistic property where a predicate presupposes the truth of its complement clause: "To know p entails the truth of p, whereas to believe p does not" (Kiparsky and Kiparsky, 1970; Hintikka, 2005). It is the natural language mechanism governing human factuality understanding. While factuality targets the final status (a static label of the event), factivity elucidates the generative mechanism (analyzing how triggers, negation, and modality dynamically derive truth). This distinction highlights a paradigmatic

divergence often overlooked in early EFP work.

Early linguistics also categorized factivity via discrete features, such as distinguishing hard/soft triggers (Abrusán, 2011) or mental/speech acts (Anand and Hacquard, 2014). However, recent experimental semantics reveals that factivity is actually a probabilistic gradient modulated by context and prior beliefs (Degen and Tonhauser, 2022; Degen, 2013). Static labels fail to capture this continuity. Therefore, valid evaluation must shift focus from recognizing atomic cues to assessing the dynamic pragmatic mechanisms that govern truth updates.

Gazdar's (1979) **Pragmatic Priority** model suggests that pragmatic implicatures can override semantic presuppositions in conflict, thus canceling them. This offers a theoretical explanation for sycophancy in LLMs: models may sacrifice factual accuracy (semantic concession) to maintain contextual coherence (pragmatic priority). Recent mechanistic studies reveal that hallucination and sycophancy share a common neural basis, where models are physically wired to prioritize conversational compliance over semantic truth (Gao et al., 2025) as a result of RLHF alignment.

The **Question Under Discussion (QUD)** model (Simons, 2013; Abrusán, 2011) posits that presupposition projection (Karttunen, 1973) depends on whether a proposition is At-issue (main point) or Non-at-issue (background). RLHF training often biases models toward explicitly answering the QUD, creating background blind spots where non-at-issue information goes unverified. This explains the Q1-Q3 answer shift in Table 3: as new information becomes the focus, the original target proposition recedes into the background.

Intersubjectivity represents the most important linguistic insight for our framework. Dynamic Semantics (Heim, 1982) defines meaning as a function updating the common ground (Stalnaker, 1978). Yuan (2020) highlights that factive predicates are inherently inter-subjective, inviting cognitive collaboration between agent and user. From this view, factivity is not a static label but a dynamic process of maintaining context consistency in multi-turn interactions.

Above linguistic insights indicate that a trustworthy agent requires not only "knowing that" (recognition) and "knowing why" (reasoning), but crucially, "prudence in knowing what one does not know" (regulation).

4 The Meta-Factivity Framework (MFF)

Grounded in meta-cognition (Flavell, 1979; Shea, 2018), we propose meta-factivity, the ability to conduct second-order monitoring, evaluation, and regulation of the factuality status of content it generates or understands. To operationalize this concept, we introduce the Meta-Factivity Framework (MFF), a hierarchical, progressive framework that shifts evaluation from "what the model recognizes as facts" to "how the model manages its commitments."

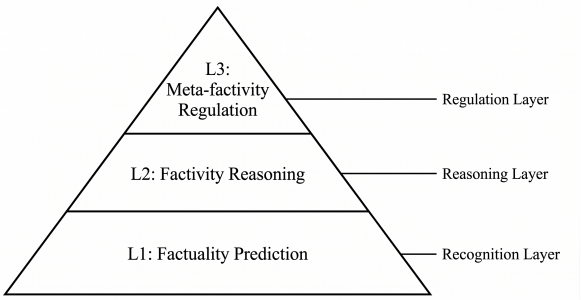

Figure 1: The Meta-Factivity Framework (MFF) and its three cognitive layers.

L1 Recognition Layer (Factuality Prediction): This foundational layer corresponds to traditional classification tasks.

L2 Reasoning Layer (Factivity Reasoning): Examines if the model can perform coherent logical deduction combining context and belief states, thereby distinguishing genuine semantic understanding from shortcut learning or reasoning unfaithfulness.

L3 Regulation Layer (Meta-Factivity Regulation): Evaluates reflexivity, i.e., the agent's capacity to monitor the generation process, serving as the ultimate defense against hallucination.

Dimension	Reasoning Layer	Regulation Layer
Output	Reliability	Calibration
Process	Faithfulness	Monitoring
Grounding	Robustness	Controllability

Table 4: The mapping dimensions between Reasoning Layer (L2) and Regulation Layer (L3).

4.1 L2: Factivity Reasoning

While factuality Prediction (L1) serves as the foundational layer, MFF prioritizes factivity reasoning (L2) to evaluate the belief trajectory (specifically, the dynamic evolution of truth values). We structure this layer around three key capabilities: reasoning output, process, and invariance (Table 4). To

operationalize these concepts, MFF calls for the development of comprehensive benchmarking suites, serving as the definitive architectural guide for designing the following computable, downstream metrics:

Reliability: Evaluates whether a model possesses firm knowledge rather than relying on lucky guesses. This dimension may decompose into Answer Accuracy, Answer Consistency (Wang et al., 2022), and the Calibration Gap (Lyu et al., 2025).

Faithfulness: Examines the causal link between evidence and conclusion (Turpin et al., 2023). This can be measured via established metrics such as step-by-step Reasoning Consistency (Wang et al., 2025; Lai et al., 2025), the RACE framework (Wang et al., 2026), and NLI-based CoT verification (Sadeddine and Suchanek, 2025).

Robustness: Tests the stability of factuality inferences against linguistic perturbations. This can be evaluated by measuring the retention of accurate predictions despite non-critical prompt variations, such as changes in context length and logical phrasing.

4.2 L3: Meta-Factivity Regulation

This layer marks the transition from understanding input (other-mind) to regulating output (self-mind). Current LLM pathologies (hallucination, self-contradiction, and failure in self-correction) can be viewed as a lack of factuality monitoring. MFF proposes three regulatory dimensions to assess if a model appears as a responsible agent:

Calibration: Evaluates whether the model's internal confidence aligns with its external accuracy, and whether it actively employs linguistic hedges to modulate factual commitment under uncertainty and avoid overconfidence (Lichtenstein et al., 1982; Nelson, 1990).

Monitoring: Demands the ability to detect internal contradictions during generation, and spontaneously rectify errors, e.g., via Constitutional AI paradigm (Bai et al., 2022). This moves beyond static generation to dynamic editing (Norman and Shallice, 1986), mirroring human meta-cognitive monitoring.

Controllability: Ensures the robustness of knowledge with a physical basis, grounded in recent findings that LLMs often maintain accurate internal states even when generating falsehoods (Azaria and Mitchell, 2023) and neurons for hallucinatory behaviors (Gao et al., 2025). Leveraging mechanistic interpretability techniques (e.g., Cunningham et al., 2023; Dunefsky et al., 2024), such white-box verification constitutes the ultimate test of whether an agent truly knows what it is saying.

5 Conclusion

The Meta-Factivity Framework (MFF) marks a transition in EFP research: from static accuracy metrics to a dynamic framework of cognitive control. It diagnoses structural deficits by evaluating how models establish, maintain, and regulate epistemic commitments. By viewing non-factual generation as a failure of meta-control rather than a simple knowledge deficit, MFF guides mitigation strategies toward second-order factivity awareness. Ultimately, MFF aims to move evaluation toward white-box cognition, as the ability to regulate factual commitments represents the definitive boundary between stochastic mimicry and true agentic intelligence.

Furthermore, because the vulnerabilities identified in this paper are systemic beyond EFP, MFF also functions as a universal diagnostic blueprint for general NLI tasks and provides a pathway to upgrade existing entailment benchmarks.

Limitations

This paper establishes a theoretical blueprint rather than a fully instantiated benchmarking suite. While we have outlined concrete pathways for operationalizing these dimensions, constructing the comprehensive, large-scale adversarial datasets required to empirically execute this framework across diverse natural language tasks remains necessary future work. Besides, the operationalization of MFF faces a key constraint in its dependence on white-box access. The neural controllability (L3) dimension relies on probing internal representations. This inherently limits the deepest level of evaluation to open-weights models, rendering it currently inaccessible for proprietary closed-source APIs where internal belief states are obfuscated.

Ethical Considerations

This work focuses on the theoretical diagnosis and framework construction for Event Factuality Prediction. All datasets discussed are established, publicly available benchmarks used strictly for academic analysis. Our work aims to improve the transparency and trustworthiness of LLMs, which is essential for the safe deployment of AI in high-stakes domains.

Acknowledgements

This work was supported by the Fundamental Research Funds for the Central Universities (HUST: No.YCJJ20252111).

References

Márta Abrusán. 2011. Predicting the presuppositions of soft triggers. *Linguistics and Philosophy*, 34(6):491–535.

Pranav Anand and Valentine Hacquard. 2014. Factivity, belief and discourse. *The Art and Craft of Semantics: A Festschrift for Irene Heim*, 1:69–90.

Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when it's lying. *arXiv preprint arXiv:2304.13734*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, and Cameron McKinnon. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Yuan Cao, Qiaoming Zhu, and Peifeng Li. 2013. Construction method of chinese event factuality corpus. *Journal of Chinese Information Processing*, 27(6):38–45.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*.

Judith Degen. 2013. Alternatives in pragmatic reasoning.

Judith Degen and Judith Tonhauser. 2022. Are there factive predicates? an empirical investigation. *Language*, 98(3):552–591.

Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. 2024. Transcoders find interpretable llm feature circuits. *Advances in Neural Information Processing Systems*, 37:24375–24410.

Bryan Eikema, Evgenia Ilia, José GC de Souza, Chrysoula Zerva, and Wilker Aziz. 2025. Teaching language models to faithfully express their uncertainty. *arXiv preprint arXiv:2510.12587*.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

John H Flavell. 1979. Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, 34(10):906.

Chuan Gao, Hao Chen, Chaojun Xiao, Zhipeng Chen, Zhiyuan Liu, and Maosong Sun. 2025. H-neurons: On the existence, impact, and origin of hallucination-associated neurons. *arXiv preprint arXiv:2512.01797*.

Gerald Gazdar. 1979. *Pragmatics: Implicature, Presupposition, and Logical Form*. Academic Press, New York.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112.

Irene Roswitha Heim. 1982. *The semantics of definite and indefinite noun phrases*. University of Massachusetts Amherst.

J. Hintikka. 2005. *Knowledge and Belief: An Introduction to the Logic of the Two Notions*. Texts in philosophy. King's College London Publications.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, and Eli Tran-Johnson. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Lauri Karttunen. 1973. Presuppositions of compound sentences. *Linguistic inquiry*, 4(2):169–193.

Paul Kiparsky and Carol Kiparsky. 1970. *Fact*, pages 143–173. The Hague: Mouton.

Huiyuan Lai, Xiao Zhang, and Malvina Nissim. 2025. Multidimensional consistency improves reasoning in language models. *arXiv preprint arXiv:2503.02670*.

Kenton Lee, Yoav Artzi, Yejin Choi, and Luke Zettlemoyer. 2015. Event detection and factuality assessment with non-expert supervision. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1643–1648, Lisbon, Portugal. Association for Computational Linguistics.

Chunyang Li, Hao Peng, Xiaozhi Wang, Yunjia Qi, Lei Hou, Bin Xu, and Juanzi Li. 2024. Maven-fact: A large-scale event factuality detection dataset. *arXiv preprint arXiv:2407.15352*.

Xinliang Li and Yulin Yuan. 2023. *Theory and Application of Narrative and Factuality*. Foreign Language Teaching and Research Press.

Sarah Lichtenstein, Baruch Fischhoff, and Lawrence D. Phillips. 1982. Calibration of probabilities: The state of the art to 1980. In Daniel Kahneman, Paul Slovic, and Amos Tversky, editors, *Judgment under Uncertainty: Heuristics and Biases*, pages 306–334. Cambridge University Press, Cambridge.

Gabrielle Kaili-May Liu, Gal Yona, Avi Caciularu, Idan Szpektor, Tim GJ Rudner, and Arman Cohan. 2025. Metafaith: Faithful natural language uncertainty expression in llms. *arXiv preprint arXiv:2505.24858*.

Xiao Liu, Heyan Huang, and Yue Zhang. 2022. End-to-end event factuality prediction using directional labeled graph recurrent network. *Information Processing & Management*, 59(2):102836.

Qing Lyu, Kumar Shridhar, Chaitanya Malaviya, Li Zhang, Yanai Elazar, Niket Tandon, Marianna Apidianaki, Mrinmaya Sachan, and Chris Callison-Burch. 2025. Calibrating large language models with sample consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 19260–19268.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*.

Lennart Meincke, Ethan Mollick, Lilach Mollick, and Dan Shapiro. 2025. Prompting science report 1: Prompt engineering is complicated and contingent. *Preprint*, arXiv:2503.04818.

Thomas O Nelson. 1990. Metamemory: A theoretical framework and new findings. In *The Psychology of Learning and Motivation*, volume 26, pages 125–173. Academic Press (Elsevier).

Donald A Norman and Tim Shallice. 1986. *Attention to action: Willed and automatic control of behavior*, pages 1–18. Springer.

Marco Rovera, Serena Cristoforetti, and Sara Tonelli. 2025. Modafact: Multi-paradigm evaluation for joint event modality and factuality detection. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6378–6396.

Rachel Rudinger, Aaron Steven White, and Benjamin Van Durme. 2018. Neural models of factuality. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, volume 1, page 731–744, New Orleans, Louisiana. Association for Computational Linguistics.

Zacchary Sadeddine and Fabian Suchanek. 2025. Verifying the steps of deductive reasoning chains. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 456–475.

Roser Saurí and James Pustejovsky. 2009. Factbank: A corpus annotated with event factuality. *Language Resources and Evaluation*, 43(3):227–268.

Roser Saurí and James Pustejovsky. 2012. Are you sure that this happened? assessing the factuality degree of events in text. *Computational Linguistics*, 38(2):261–299.

Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3419–3425.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, and Scott R Johnston. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

Nicholas Shea. 2018. *Representation in cognitive science*. Oxford University Press.

Mandy Simons. 2013. On the conversational basis of some presuppositions. In *Proceedings of Semantics and Linguistic Theory (SALT)*, volume 11, pages 431–448.

Robert Stalnaker. 1978. *Assertion*, pages 315–332. Academic Press, New York.

Gabriel Stanovsky, Judith Eckle-Kohler, Yevgeniy Puzikov, Ido Dagan, and Iryna Gurevych. 2017. Integrating deep linguistic features in factuality prediction over unified datasets. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 352–357.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965.

Amir Pouran Ben Veyseh, Thien Huu Nguyen, and Dejing Dou. 2019. Graph based neural networks for event factuality prediction using syntactic and semantic structures. *arXiv preprint arXiv:1907.03227*.

Boxuan Wang, Zhuoyun Li, Xinmiao Huang, Xiaowei Huang, and Yi Dong. 2025. Chasing consistency: Quantifying and optimizing human-model alignment in chain-of-thought reasoning. *arXiv preprint arXiv:2511.06168*.

Changyue Wang, Weihang Su, Qingyao Ai, and Yiqun Liu. 2026. Joint evaluation of answer and reasoning consistency for hallucination detection in large reasoning models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 33377–33385.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V Le. 2023. Simple synthetic data reduces sycophancy in large language models. *arXiv preprint arXiv:2308.03958*.

Aaron Steven White, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2018. Lexicosyntactic inference in neural models. *arXiv preprint arXiv:1808.06232*.

Yulin Yuan. 2020. Narrativity and factuality: Two navigational mechanisms for linguistic reasoning. *Studies in Chinese Language and Literature*, (1):1–9.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, and Yulong Chen. 2025. Siren's song in the ai ocean: A survey on hallucination in large language models. *Computational Linguistics*, pages 1–46.