POSITION: LOGICAL SOUNDNESS IS NOT A RELIABLE CRITERION FOR NEUROSYMBOLIC FACT-CHECKING WITH LLMS

Jason Chan, Robert Gaizauskas & Zhixue Zhao
School of Computer Science
University of Sheffield
United Kingdom
{jlychan1,r.gaizauskas,zhixue.zhao}@sheffield.ac.uk

ABSTRACT

As large language models (LLMs) are increasing integrated into fact-checking pipelines, formal logic is often proposed as a rigorous means by which to mitigate bias, errors and hallucinations in these models' outputs. For example, some neurosymbolic systems verify claims by using LLMs to translate natural language into logical formulae and then checking whether the proposed claims are logically sound, i.e. whether they can be validly derived from premises that are verified to be true. **We argue that such approaches structurally fail to detect misleading claims due to systematic divergences between conclusions that are logically sound and inferences that humans typically make and accept.** Drawing on studies in cognitive science and pragmatics, we present a typology of cases in which logically sound conclusions systematically elicit human inferences that are unsupported by the underlying premises. Consequently, we advocate for a complementary approach: leveraging human-like reasoning tendencies of LLMs as a feature rather than a bug, and using these models to validate the outputs of formal components in neurosymbolic systems against potentially misleading conclusions.

1 INTRODUCTION

Large language models (LLMs) have enabled substantial progress in automated fact-checking and claim verification (Guo et al., 2022; Augenstein et al., 2024; Huang et al., 2025). However, their outputs remain vulnerable to hallucinations, bias, and logical inconsistency (Ji et al., 2023; Quelle & Bovet, 2024; Pelrine et al., 2023). To mitigate these limitations, a growing body of work has proposed neurosymbolic approaches that use formal logic to constrain or validate LLM outputs. In some architectures, for example, LLMs act as semantic parsers that translate natural-language statements, e.g. from user input or model-generated text into logical formulae such as in first-order logic (FOL) that are then evaluated against a verified knowledge base or external source (Wang & Shu, 2023; Asghari et al., 2026). A common underlying assumption is that if a claim is logically sound, i.e. validly derivable from true premises in a verified knowledge base under formal inference rules such as those of FOL, then the claim is supported and acceptable.

We argue that existing work relying on this approach has a critical blind spot: **logical soundness is not a reliable criterion for detecting misleading information, because human inferences often diverge from formal logical operations such that logically sound conclusions can be *systematically* misleading**. This yields three implications that challenge the current paradigm: (a) as opposed to using formal logic to validate LLM outputs, LLMs also have the potential of modeling human inference and detecting misleading conclusions otherwise unrecognized by systems relying on formal logic; (b) research on LLM reasoning should investigate these models' human-like tendencies not only as bugs (i.e., models committing human-like genuine errors and biases in logical tasks) but also as features (i.e., models predicting inferences that humans would typically accept despite being unsound in formal logic); and (c) if LLMs are to model human reasoning effectively, they should not be trained or optimized to conform strictly to formal logic in their reasoning processes.

2 MISLEADING INFORMATION AND LOGIC-BASED APPROACHES TO FACT-CHECKING

In the context of recognizing and mitigating misinformation, fact-checking aims to identify not only explicitly false statements but also those that are technically true yet misleading (Chen & Shu, 2024; Khraisat et al., 2026; Guo et al., 2022; Akhtar et al., 2023). In line with existing literature, we use "misleading" to mean factually correct content that nonetheless conveys an *implied meaning* that is false or unsupported (Tang et al., 2025). Within this paradigm, the starting point of our position is that "implied meaning" includes inferences that human readers would typically draw or accept. Consider the sentence "John had a fight with Marilyn and decided to break up with her." As predicted by the theory of conjunction buttressing, readers would typically interpret this compound statement as (i) a temporal sequence (fight first, breakup decision second) and often (ii) a causal relation (the fight contributed to the decision) (Levinson, 2000). Accordingly, if John had already decided to break up with Marilyn before the fight, the sentence should be considered misleading despite being literally compatible with the facts.

In parallel, existing research combines LLMs with explicit logic-based verification to improve faithfulness and robustness in fact-checking. Example methods include using LLMs to decompose complex natural language claims into formulae composed of FOL predicates whose truth values are individually verified by external knowledge sources (Wang & Shu, 2023; Vladika et al., 2025; Asghari et al., 2026). Other methods rely on natural logic, using LLMs to segment claims and evidence into corresponding text spans, and then relying on automated theorem-provers in natural logic to determine whether the claims are logically entailed by the evidence (Krishna et al., 2022; Strong et al., 2024). More generally, this approach of using LLMs to translate natural language into logical formulae which are then checked with a symbolic solver has also been proposed to ensure that LLMs' own reasoning processes and resulting outputs are logically sound, i.e. based on true premises and derived in a logically valid manner (Pei et al., 2025; Zheng et al., 2025). Likewise, formal logic underpins other verification paradigms, including knowledge-graph-based methods (Opsahl, 2024; Hao & Wu, 2025) (e.g. to infer unstated relationships between entities in graph) and program-based approaches (Pan et al., 2023). The underlying assumption across these logic-based approaches is that if a claim can be derived in a logically valid manner from premises that are trusted or assumed to be true, then the claim itself should be treated as supported and verified by that system.

3 THE DIVERGENCE BETWEEN HUMAN REASONING AND FORMAL LOGIC

Existing work in cognitive science has shown that humans accept and make inferences in ways that diverge from what is considered valid under formal logical systems (Johnson-Laird, 2010; Khemlani & Johnson-Laird, 2019; Ragni et al., 2019). For example, consider a premise in disjunctive form: "*A or B*". According to the theory of mental models, humans represent a disjunctive statement as a conjunction of possibilities that hold in default to the contrary. This account explains why, on hearing that "*A or B*", humans tend to accept the inferences "*it is possible that A*" and "*it is possible that B*" as intuitive and correct, even though both these inferences are invalid (*and therefore unsound regardless of whether the premises are true*) in standard modal logic that is introduced to reason with statements about possibilities and necessities (Johnson-Laird & Ragni, 2025).

We now illustrate why such divergences between human inference and formal logic are problematic for logic-based approaches in fact-checking. While we present an example statement in disjunctive form ("*A or B*") commonly found in official announcements[1] and news reports[2] (thus making them realistic targets for fact-checking pipelines), we also cover a broader range of other linguistically natural phenomena and statements e.g. involving conjunctions, conditionals and modals in Table 1.

Given a knowledge base with only one verified true premise S1:

[1] See e.g. "*[i]f you look at the Summary of Economic Projections, things are moving by just a tick* ***or*** *even a semi tick between now and March*" (https://www.federalreserve.gov/mediacenter/files/FOMCpresconf20180613.pdf)

[2] See e.g. "*fire [...] may have spread through the plane and caused the explosion,* ***or*** *the jet could have caught fire after colliding with an object on the ground*" (https://www.bbc.co.uk/news/articles/ckgky7djx5eo); "*Sometimes that's human error, someone misconfiguring something somewhere,* ***or*** *in extreme cases a cyber attack*" (https://www.bbc.co.uk/news/articles/cev1en9077ro)

Logically Valid Operation Applied to S1	**Conclusion Derived by Applying Operation (Misleading)**	**What Humans Infer from Conclusion**	**Cognitive/Pragmatic Basis of Human Inference**
Disjunction Introduction ($A \vdash A \vee B$)	**S2**: "Tariffs for either France or some other European countries will go up 10%."	**S2-I**: "It is possible that tariffs for some other European countries will go up 10%."	Humans interpret disjunctions as a conjunction of default possibilities: i.e. given "A or B", inferring the possibility that "B" (Johnson-Laird & Ragni, 2025).
Conjunction Introduction ($A \vdash A \wedge A$)	**S3**: "Tariffs for France will go up 10%, and tariffs for France will go up 10%."	**S3-I**: "Tariffs will go up 10% twice in a row."	Humans infer that the repetition refers to two separate events (by the Gricean Maxim of Quantity (Grice, 1975)), happening sequentially in time (by conjunction buttressing (Levinson, 2000)).
Material Implication (Positive Paradox) ($A \vdash B \rightarrow A$)	**S4**: "If interest rates do not decrease, tariffs for France will go up 10%."	**S4-I**: "If interest rates **do** decrease, tariffs will **not** go up 10%."	Conditional perfection. Humans routinely interpret natural-language conditionals as biconditionals (Geis & Zwicky, 1971; Horn, 2000).
Material Implication (Vacuous Truth) ($A \vdash \neg A \rightarrow B$)	**S5**: "If tariffs for France will not go up 10%, our domestic market will be flooded with French goods."	**S5-I**: "If tariffs for France **do** go up 10%, our domestic market will **not** be flooded"	As above
By Axioms B and T in Modal Logic ($A \vdash \diamond A$)	**S6**: "It is possible that tariffs for France will go up 10%."	**S6-I**: "It is possible that tariffs for France will **not** go up 10%."	Humans infer uncertainty from a statement about possibility, by Gricean Maxim of Quality (Grice, 1975; Johnson-Laird & Ragni, 2019)

Table 1: Examples of logically sound derivations that can induce ungrounded human inferences.

S1: Tariffs for France will go up 10%.

Suppose a logic-based fact-checking system is tasked with verifying the claim that:

S2: Tariffs for either France or some other European countries will go up 10%.

Denoting S1 in propositional logic as A, the system can validly derive S2 from S1 by applying the inference rule of disjunction introduction ($A \vdash A \vee B$), hence verifying that S2 is a logically sound conclusion given S1.[3]

As explained above however, humans typically accept the inference that "*it is possible that B*" as following from a statement that "*A or B*". On this basis, given S2, humans typically accept S2-I:

S2-I: It is possible that tariffs for some other European countries will go up 10%.

even though S2-I is not explicitly supported by the premise S1.

S2 should therefore be considered misleading: it implies S2-I, a claim with no basis in S1. Taken altogether, the example illustrates how a conclusion verified by a system to be logically sound can still be misleading because human inference does not track formal logical operations. Assuming the same verified premise S1, Table 1 now presents examples (including S2 and S2-I as discussed) in which logically valid operations yield sound conclusions that invite ungrounded human inferences. Together, these examples support our central claim that, **despite a common assumption underlying neurosymbolic approaches using LLMs, logical soundness is not a reliable criterion for detecting misleading statements when fact-checking for misinformation**.

[3]To be clear, our main concern is not that a system might spontaneously use disjunction introduction to generate new facts or statements wholly unrelated to its input context. Rather, the problem arises when a fact-checking system applies this rule to formally derive and verify an *existing* claim that is in disjunctive form.

4 IMPLICATIONS: RETHINKING THE ROLE OF LLMS IN NEUROSYMBOLIC FACT-CHECKING SYSTEMS

Given the above, we argue that the current paradigm of constraining and validating LLM outputs through formal logic captures only part of what reliability in fact-checking and misinformation detection requires. Rather, **future work should also explore the potential for LLMs to evaluate whether formally sound conclusions (as verified or generated by logic-based systems) are likely to mislead human readers.**

This direction is plausible as prior work has shown that LLMs can be trained to recognize certain pragmatic inferences routinely made by humans (Yue et al., 2024; Ma et al., 2025; Sravanthi et al., 2025). For example, LLMs improved through reinforcement learning (Somayajula et al., 2025) can recognize scalar implicatures with certain gradable adjectives (e.g. the statement that "*the coffee is warm*" implying that the coffee is not "*hot*" (Nizamani et al., 2024)) and encode such implicatures in their hidden representations (Lin et al., 2024). In parallel, existing work has also found that LLMs display various human-like reasoning tendencies (Eisape et al., 2024; Mondorf & Plank, 2024), even though studies in this category commonly investigate these tendencies only as systematic error patterns and cognitive biases in logical reasoning tasks (i.e., mistakes that people themselves would recognize once corrected) (Lampinen et al., 2024; Parmar et al., 2024). For example, Lampinen et al. (2024) and Eisape et al. (2024) respectively found that LLMs exhibit *content effect*, misjudging logical validity based on factual plausibility of the conclusion, and *figural effect*, being unduly affected by the ordering of words and terms in the premises.

One possible approach to capitalize on these human-like pragmatic capabilities and reasoning tendencies is to use LLMs for what we term "*claim expansion*". Returning to the earlier example, when a system is tasked with fact-checking S2, an LLM can be prompted to generate a set of *n follow-on* inferences that human readers are likely to make based on S2, especially including those which are logically unsound but intuitively acceptable such as S2-I.[4] The system would then verify this expanded set of inferences using formal logic, and assign the original claim S2 a soft score based on the proportion of follow-on inferences that are logically verifiable. For example, because S2-I cannot be logically derived from S1, S2's score would decrease accordingly and reflect its potentially misleading nature. In this framework, our proposed method effectively utilizes LLMs' human-like reasoning tendencies and cognitive biases as a valuable feature rather than a bug to be eliminated.

That said, we recognize that LLMs used in such a capacity still inherit certain challenges in terms of model hallucination and social biases (Ji et al., 2023; Gallegos et al., 2024; Kalai et al., 2025). For example, similar to issues observed when using LLMs to decompose claims into atomic subclaims for verification (see e.g. Hu et al., 2025), LLMs may hallucinate or generate irrelevant follow-on claims. Furthermore, while human-like reasoning tendencies are desirable for our use case, models may still exhibit specific social biases when expanding claims to be fact-checked in a way that prejudices minority groups (Lee et al., 2024; Shrawgi et al., 2024). In both these respects, our approach remains dependent on existing mitigation methods, such as more rigorous data curation (Gunasekar et al., 2024) and test-time interventions (Li et al., 2023; Siddique et al., 2026), to improve models' fairness and factuality. Moreover, existing work shows that LLMs can be prompted to explicitly express their confidence levels in their own generated outputs, in a way that provides well-calibrated estimates of how likely these outputs are correct (Lin et al., 2022; Tian et al., 2023; Xia et al., 2025). Our proposed approach could incorporate these uncertainty scores either by rejecting outputs (i.e. model-generated follow-on inferences) that fall below a certain confidence threshold, or proportionately discounting the impact of low-confidence outputs when computing the overall soft score for a particular claim being verified.

Taken altogether, our proposed approach shows that LLMs' human-like reasoning tendencies and pragmatic capabilities can serve as valuable features, complementing formal logic-based verification by modelling inferences that are logically unsound but typically considered intuitive and acceptable by humans. Importantly, for LLMs to serve effectively in such a capacity, we argue that these models' training should not enforce or encourage conformity to formal logic. Methods that reward formal logical consistency (Calanzone et al., 2025) or bias training data toward strictly logic-conforming

[4]The size (n) of this set of follow-on inferences can be flexible and dependent e.g. on the importance of the claim being fact-checked and computational budget limits.

patterns (Morishita et al., 2024; Tan et al., 2025) risk suppressing the very pragmatic sensitivities needed for LLMs to be useful in detecting logically sound but misleading outputs.

5 Alternative Views

AV: We can simply augment logical soundness with an auxiliary criterion X

Requiring that claims be both logically sound whilst meeting another criterion could systematically result in unintended false-positives. This is because, as we have demonstrated in the previous section, acceptable human inferences are often logically unsound. If a knowledge base contains the statement that "*it is possible that the Fed will raise interest rates by 0.5%*", we would not want our claim verification system to reject, e.g., "*it is possible that the Fed will not raise interest rates by 0.5%*", which is otherwise an intuitive and acceptable inference (Johnson-Laird & Ragni, 2019). However, rejecting claims only if they are logically unsound **and** do not meet a criterion X is equally problematic. As shown above, such a system would systematically fail to detect misleading statements such as the ones in Table 1.

AV: There is no need to focus on compound statements involving logical connectives because malicious actors are unlikely to use them. *Such actors do not care about logical soundness and simple assertive claims (corresponding to atomic propositions in logic) are often more persuasive.*

As LLMs have the potential of producing and spreading misinformation at scale (Ferrara, 2024), we should expect that these models could be optimized to bypass safeguards. In other words, even if human actors prefer direct assertions, automated systems can learn to generate formally sound compound claims that evade logic-based filters while still inducing misleading inferences.

AV: We can simply expose logic derivation traces to users and let humans decide which conclusions to accept.

This view assumes, first, that users will consistently inspect formal traces, and second, that they can reliably detect pragmatic misleadingness from those traces. In practice, many users may not be capable of performing this review especially at scale.

Moreover, when users delegate trace interpretation to LLMs, the problem reappears: once again models are depended upon to flag mismatches between formal logic and typical human inferences. As such, trace transparency does not eliminate the reliability issue raised here.

6 Conclusion

In this work, we argue that logical soundness is not a reliable criterion for assessing claims in the context of fact-checking and misinformation detection, as it *systematically* fails to detect certain statements that are misleading due to the divergence between human reasoning and formal logic. We discuss and list examples of logically sound conclusions that are nonetheless misleading since as they tend to elicit ungrounded inferences from humans, as predicted by various cognitive science and linguistics studies. On this basis, we present an alternative to the current paradigm of constraining and validating LLM outputs with formal logic, and call instead for future work to explore and maximize the potential of LLMs for detecting such misleading statements that would otherwise be licensed in strictly formal systems.

Acknowledgment

This work was supported by the UKRI AI Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1]. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

REFERENCES

Mubashara Akhtar, Michael Schlichtkrull, Zhijiang Guo, Oana Cocarascu, Elena Simperl, and Andreas Vlachos. Multimodal automated fact-checking: A survey. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5430–5448, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.361. URL https://aclanthology.org/2023.findings-emnlp.361/.

Sara Asghari, Laks V. Lakshmanan, Venkatesh Srinivasan, and Alex Thomo. Fact-checking with large language models via cost-effective first-order logic reformulation. In *Social Networks Analysis and Mining: 17th International Conference, ASONAM 2025, Niagara Falls, ON, Canada, August 25–28, 2025, Proceedings, Part III*, pp. 66–78, Berlin, Heidelberg, 2026. Springer-Verlag. ISBN 978-3-032-14106-4. doi: 10.1007/978-3-032-14107-1_6. URL https://doi.org/10.1007/978-3-032-14107-1_6.

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, Eduard Hovy, Heng Ji, Filippo Menczer, Ruben Miguez, Preslav Nakov, Dietram Scheufele, Shivam Sharma, and Giovanni Zagni. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence*, 6(8):852–863, Aug 2024. ISSN 2522-5839. doi: 10.1038/s42256-024-00881-z. URL https://doi.org/10.1038/s42256-024-00881-z.

Diego Calanzone, Stefano Teso, and Antonio Vergari. Logically consistent language models via neuro-symbolic integration. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=7PGluppo4k.

Canyu Chen and Kai Shu. Can LLM-generated misinformation be detected? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ccxD4mtkTU.

Tiwalayo Eisape, Michael Tessler, Ishita Dasgupta, Fei Sha, Sjoerd Steenkiste, and Tal Linzen. A systematic comparison of syllogistic reasoning in humans and language models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8425–8444, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.466. URL https://aclanthology.org/2024.naacl-long.466/.

Emilio Ferrara. Genai against humanity: nefarious applications of generative artificial intelligence and large language models. *Journal of Computational Social Science*, 7(1):549–569, February 2024. ISSN 2432-2725. doi: 10.1007/s42001-024-00250-1. URL http://dx.doi.org/10.1007/s42001-024-00250-1.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179, September 2024. doi: 10.1162/coli_a_00524. URL https://aclanthology.org/2024.cl-3.8/.

Michael L. Geis and Arnold M. Zwicky. On invited inferences. *Linguistic Inquiry*, 2(4):561–566, 1971. ISSN 00243892, 15309150. URL http://www.jstor.org/stable/4177664.

H. P. Grice. *Logic and Conversation*, pp. 41 – 58. Brill, Leiden, The Netherlands, 1975. ISBN 9789004368811. doi: 10.1163/9789004368811_003. URL https://brill.com/view/book/edcoll/9789004368811/BP000003.xml.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio Cesar Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Conti Kauffmann, Gustavo Henrique de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Behl, Xin Wang, Sebastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. Textbooks are all you need, 2024. URL https://openreview.net/forum?id=Fq8tKtjACC.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206, 02 2022. ISSN 2307-387X. doi: 10.1162/tacl_a_00454. URL https://doi.org/10.1162/tacl_a_00454.

Yuanzhen Hao and Desheng Wu. Fact verification on knowledge graph via programmatic graph reasoning. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 5480–5495, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.293. URL https://aclanthology.org/2025.findings-emnlp.293/.

Laurence R. Horn. From if to iff: Conditional perfection as pragmatic strengthening. *Journal of Pragmatics*, 32(3):289–326, 2000. ISSN 0378-2166. doi: https://doi.org/10.1016/S0378-2166(99)00053-3. URL https://www.sciencedirect.com/science/article/pii/S0378216699000533.

Qisheng Hu, Quanyu Long, and Wenya Wang. Decomposition dilemmas: Does claim decomposition boost or burden fact-checking performance? In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6313–6336, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.320. URL https://aclanthology.org/2025.naacl-long.320/.

Tianyi Huang, Jingyuan Yi, Peiyang Yu, and Xiaochuan Xu. Unmasking digital falsehoods: A comparative analysis of llm-based misinformation detection strategies, 2025. URL https://arxiv.org/abs/2503.00724.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12), March 2023. ISSN 0360-0300. doi: 10.1145/3571730. URL https://doi.org/10.1145/3571730.

P. N. Johnson-Laird and Marco Ragni. Reasoning about possibilities: Modal logics, possible worlds, and mental models. *Psychonomic Bulletin & Review*, 32(1):52–79, Feb 2025. ISSN 1531-5320. doi: 10.3758/s13423-024-02518-z. URL https://doi.org/10.3758/s13423-024-02518-z.

Philip N. Johnson-Laird. Against logical form. *Psychologica Belgica*, 50(3–4):193–221, 2010. URL https://psychologicabelgica.com/articles/pb-50-3-4-193.

P.N. Johnson-Laird and Marco Ragni. Possibilities as the foundation of reasoning. *Cognition*, 193:103950, 2019. ISSN 0010-0277. doi: https://doi.org/10.1016/j.cognition.2019.04.019. URL https://www.sciencedirect.com/science/article/pii/S0010027719301039.

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. Why language models hallucinate, 2025. URL https://arxiv.org/abs/2509.04664.

Sangeet Khemlani and P. N. Johnson-Laird. Why machines don't (yet) reason like people. *KI - Künstliche Intelligenz*, 33(3):219–228, Sep 2019. ISSN 1610-1987. doi: 10.1007/s13218-019-00599-w. URL https://doi.org/10.1007/s13218-019-00599-w.

Ansam Khraisat, Manisha, Lennon Chang, and Jemal Abawajy. Survey on deep learning for misinformation detection: Adapting to recent events, multilingual challenges, and future visions. *Social Science Computer Review*, 44(2):209–230, 2026. doi: 10.1177/08944393251315910. URL https://doi.org/10.1177/08944393251315910.

Amrith Krishna, Sebastian Riedel, and Andreas Vlachos. ProoFVer: Natural logic theorem proving for fact verification. *Transactions of the Association for Computational Linguistics*, 10:1013–1030, 2022. doi: 10.1162/tacl_a_00503. URL https://aclanthology.org/2022.tacl-1.59/.

Andrew K Lampinen, Ishita Dasgupta, Stephanie C Y Chan, Hannah R Sheahan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. Language models, like humans, show content effects on reasoning tasks. *PNAS Nexus*, 3(7):pgae233, 07 2024. ISSN 2752-6542. doi: 10.1093/pnasnexus/pgae233. URL https://doi.org/10.1093/pnasnexus/pgae233.

Messi H.J. Lee, Jacob M. Montgomery, and Calvin K. Lai. Large language models portray socially subordinate groups as more homogeneous, consistent with a bias observed in humans. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pp. 1321–1340, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3658975. URL https://doi.org/10.1145/3630106.3658975.

Stephen C. Levinson. *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. The MIT Press, 04 2000. ISBN 9780262278256. doi: 10.7551/mitpress/5526.001.0001. URL https://doi.org/10.7551/mitpress/5526.001.0001.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=aLLuYpn83y.

Fangru Lin, Daniel Altshuler, and Janet B. Pierrehumbert. Probing large language models for scalar adjective lexical semantics and scalar diversity pragmatics. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 13033–13049, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.1141/.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words, 2022. URL https://arxiv.org/abs/2205.14334.

Bolei Ma, Yuting Li, Wei Zhou, Ziwei Gong, Yang Janet Liu, Katja Jasinskaja, Annemarie Friedrich, Julia Hirschberg, Frauke Kreuter, and Barbara Plank. Pragmatics in the era of large language models: A survey on datasets, evaluation, opportunities and challenges. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8679–8696, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.425. URL https://aclanthology.org/2025.acl-long.425/.

Philipp Mondorf and Barbara Plank. Comparing inferential strategies of humans and large language models in deductive reasoning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9370–9402, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.508. URL https://aclanthology.org/2024.acl-long.508/.

Terufumi Morishita, Gaku Morio, Atsuki Yamaguchi, and Yasuhiro Sogawa. Enhancing reasoning capabilities of llms via principled synthetic logic corpus. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 73572–73604. Curran Associates, Inc., 2024. doi: 10.52202/079017-2340. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/8678da90126aa58326b2fc0254b33a8c-Paper-Conference.pdf.

Rashid Nizamani, Sebastian Schuster, and Vera Demberg. SIGA: A naturalistic NLI dataset of English scalar implicatures with gradable adjectives. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 14784–14795, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.1288/.

Tobias Aanderaa Opsahl. Fact or fiction? improving fact verification with knowledge graphs through simplified subgraph retrievals. In Michael Schlichtkrull, Yulong Chen, Chenxi Whitehouse, Zhenyun Deng, Mubashara Akhtar, Rami Aly, Zhijiang Guo, Christos Christodoulopoulos, Oana Cocarascu, Arpit Mittal, James Thorne, and Andreas Vlachos (eds.), *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pp. 307–316, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.fever-1.32. URL https://aclanthology.org/2024.fever-1.32/.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. Fact-checking complex claims with program-guided reasoning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6981–7004, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.386. URL https://aclanthology.org/2023.acl-long.386/.

Mihir Parmar, Nisarg Patel, Neeraj Varshney, Mutsumi Nakamura, Man Luo, Santosh Mashetty, Arindam Mitra, and Chitta Baral. LogicBench: Towards systematic evaluation of logical reasoning ability of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13679–13707, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.739. URL https://aclanthology.org/2024.acl-long.739/.

Yu Pei, Yongping Du, and Xingnan Jin. FoVer: First-order logic verification for natural language reasoning. *Transactions of the Association for Computational Linguistics*, 13:1340–1359, 2025. doi: 10.1162/tacl.a.41. URL https://aclanthology.org/2025.tacl-1.61/.

Kellin Pelrine, Anne Imouza, Camille Thibault, Meilina Reksoprodjo, Caleb Gupta, Joel Christoph, Jean-François Godbout, and Reihaneh Rabbany. Towards reliable misinformation mitigation: Generalization, uncertainty, and GPT-4. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6399–6429, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.395. URL https://aclanthology.org/2023.emnlp-main.395/.

Dorian Quelle and Alexandre Bovet. The perils and promises of fact-checking with large language models. *Frontiers in Artificial Intelligence*, Volume 7 - 2024, 2024. ISSN 2624-8212. doi: 10.3389/frai.2024.1341697. URL https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2024.1341697.

Marco Ragni, Hannah Dames, and Philip N. Johnson-Laird. A meta-analysis of conditional reasoning. In *Proceedings of the 17th International Conference on Cognitive Modeling*, pp. 151–156. Applied Cognitive Science Lab, Penn State, 2019. URL https://modeltheory.org/papers/2019meta-conditionals.pdf.

Hari Shrawgi, Prasanjit Rath, Tushar Singhal, and Sandipan Dandapat. Uncovering stereotypes in large language models: A task complexity-based approach. In Yvette Graham and Matthew Purver (eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1841–1857, St. Julian's, Malta, March 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.eacl-long.111. URL https://aclanthology.org/2024.eacl-long.111/.

Zara Siddique, Irtaza Khalid, Liam Turner, and Luis Espinosa-Anke. Shifting perspectives: Steering vectors for robust bias mitigation in LLMs. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 809–820, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-386-9. doi: 10.18653/v1/2026.findings-eacl.41. URL https://aclanthology.org/2026.findings-eacl.41/.

Sai Ashish Somayajula, Bokai Hu, Qi Cao, Xin Pan, and Pengtao Xie. Improving the language understanding capabilities of large language models using reinforcement learning. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings*

of the Association for Computational Linguistics: EMNLP 2025, pp. 25552–25567, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.1392. URL https://aclanthology.org/2025.findings-emnlp.1392/.

Settaluri Lakshmi Sravanthi, Kishan Maharaj, Sravani Gunnu, Abhijit Mishra, and Pushpak Bhattacharyya. Understand the implication: Learning to think for pragmatic understanding. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 23778–23790, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1218. URL https://aclanthology.org/2025.findings-acl.1218/.

Marek Strong, Rami Aly, and Andreas Vlachos. Zero-shot fact verification via natural logic and large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 17021–17035, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.991. URL https://aclanthology.org/2024.findings-emnlp.991/.

Xingwei Tan, Marco Valentino, Mahmud Elahi Akhter, Maria Liakata, and Nikolaos Aletras. Enhancing logical reasoning in language models via symbolically-guided Monte Carlo process supervision. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 31886–31900, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1624. URL https://aclanthology.org/2025.emnlp-main.1624/.

Yixuan Tang, Jincheng Wang, and Anthony Kum Hoe Tung. The missing parts: Augmenting fact verification with half truth detection. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 33979–33996, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1724. URL https://aclanthology.org/2025.emnlp-main.1724/.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5433–5442, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.330. URL https://aclanthology.org/2023.emnlp-main.330/.

Juraj Vladika, Ivana Hacajova, and Florian Matthes. Step-by-step fact verification system for medical claims with explainable reasoning. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 805–816, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-190-2. doi: 10.18653/v1/2025.naacl-short.68. URL https://aclanthology.org/2025.naacl-short.68/.

Haoran Wang and Kai Shu. Explainable claim verification via knowledge-grounded reasoning with large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 6288–6304, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.416. URL https://aclanthology.org/2023.findings-emnlp.416/.

Zhiqiu Xia, Jinxuan Xu, Yuqian Zhang, and Hang Liu. A survey of uncertainty estimation methods on large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics:*

ACL 2025, pp. 21381–21396, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1101. URL https://aclanthology.org/2025.findings-acl.1101/.

Shisen Yue, Siyuan Song, Xinyuan Cheng, and Hai Hu. Do large language models understand conversational implicature- a case study with a Chinese sitcom. In Sun Maosong, Liang Jiye, Han Xianpei, Liu Zhiyuan, and He Yulan (eds.), *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 1: Main Conference)*, pp. 1270–1285, Taiyuan, China, July 2024. Chinese Information Processing Society of China. URL https://aclanthology.org/2024.ccl-1.98/.

Xinyi Zheng, Ningke Li, Xiaokun Luan, Kailong Wang, Ling Shi, Meng Sun, and Haoyu Wang. Beyond correctness: Exposing llm-generated logical flaws in reasoning via multi-step automated theorem proving, 2025. URL https://arxiv.org/abs/2512.23511.