

Researching Responsible and Trustworthy Natural Language Processing

Replication, Reproducibility, and Generalisation

Emily Allaway

20 October 2025

School of Informatics
University of Edinburgh

eallaway@ed.ac.uk

Replication and Reproducibility

Generalisation

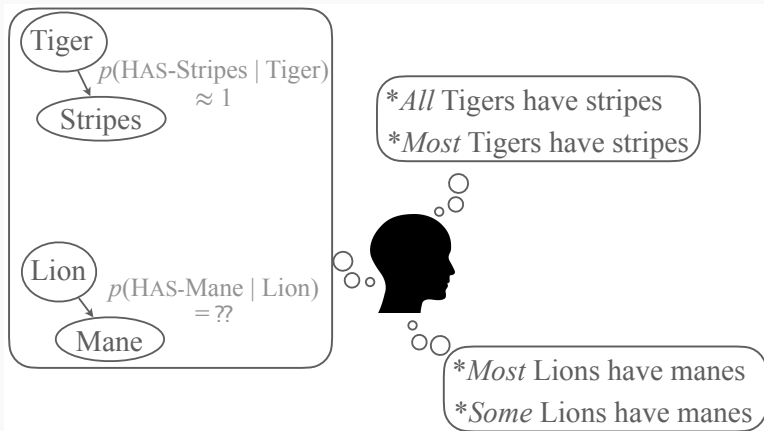
Replication and Reproducibility

Generalisation

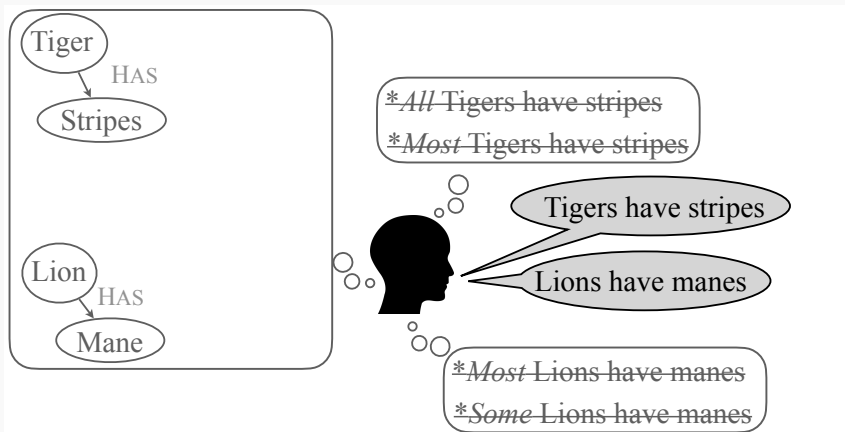
Many ways to express generalisations

- Quantifiers
 - *Most* models perform well on the MT tasks
 - We find that *generally* our method produces fewer hallucinations than the baselines.
- Scoping
 - *When using beam search*, models produce more fluent outputs.
- Generics
 - Our model outperforms or performs comparably to the baselines.

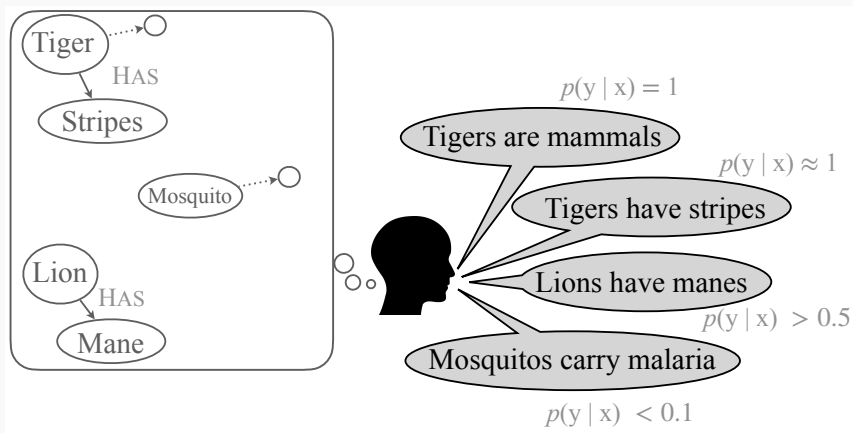
Generalisations and frequency



Generalisations and frequency



Generalisations and frequency





Lions have manes

This lion has a mane





Lions have manes

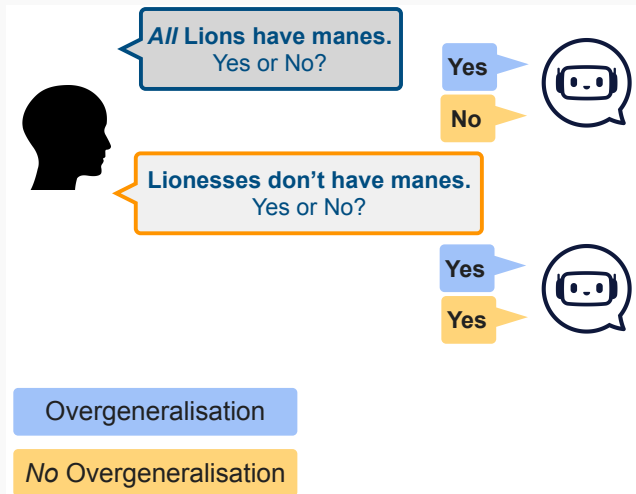
This lion has a mane



Exception

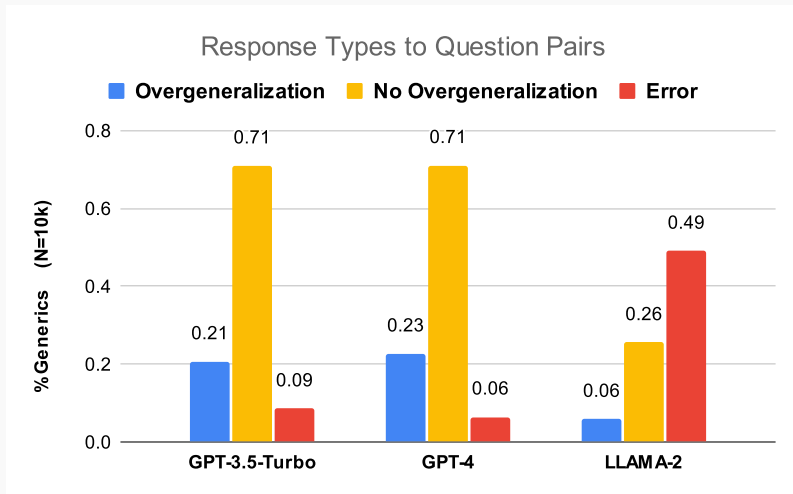
Lionesses don't
have manes

Probing for overgeneralization

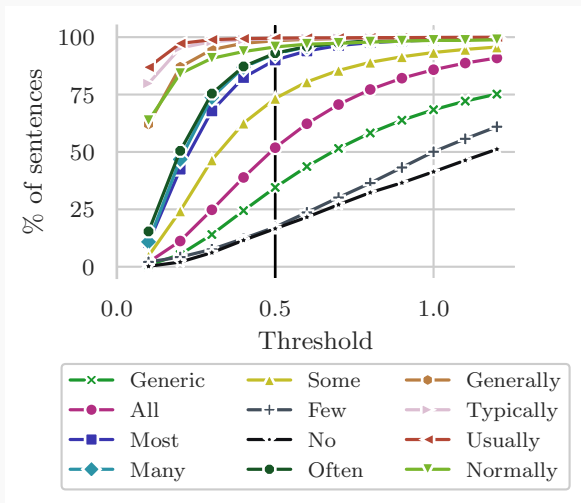


From [Allaway et al. \(2024\)](#).

Consistent overgeneralization in LLMs



Generic generalisations



Generics are not like quantified statements ([Calderón et al., 2025](#)).

Hasty generalisations

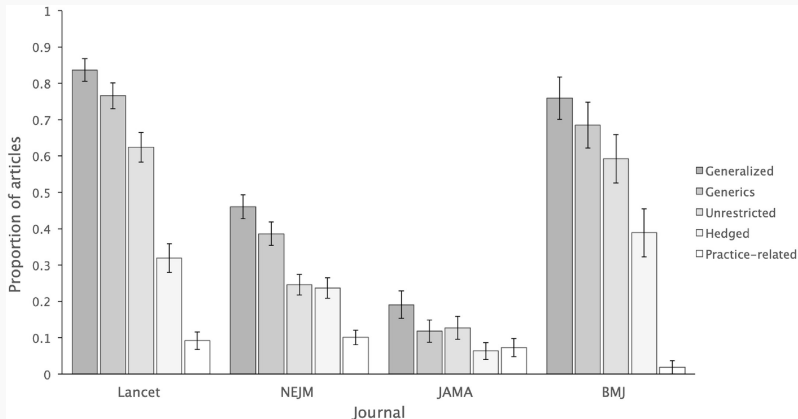


Fig 2. The proportion of articles with generalized, generic, unrestricted, hedged, or practice-related result claims (derived by dividing the number of articles with these claims with the total number of articles of each journal). Error bars indicate standard error for the variability in proportion estimates.

From [Peters et al. \(2024\)](#)

Hasty generalisations

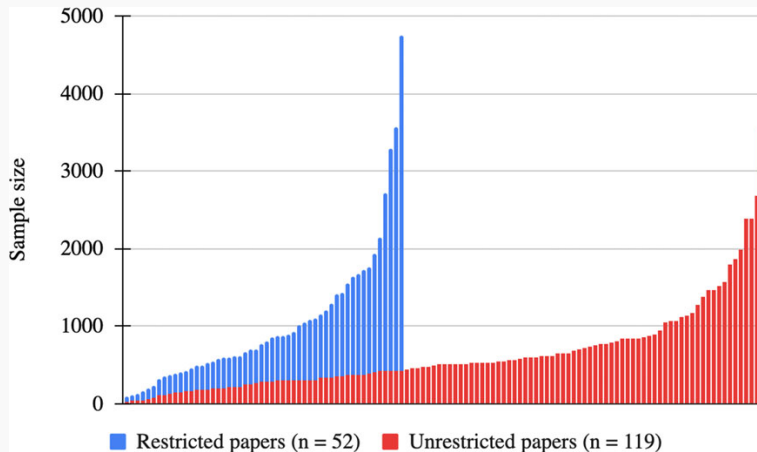


Figure 2. Full distribution of sample sizes within each group of x-phi articles.

From [Peters and Lemeire \(2024\)](#)

Exercise - evaluating claim soundness

1. Rate each claim on a scale of 1-5 (least to most) in terms of *robustness*, *replicability*, and *generalisability*.
2. Find and identify where the evidence for each claim is located.
 - Highlight table or figure labels (e.g., Figure 1) or sentences.
 - Use a separate color for the evidence corresponding to each claim.
3. Consider whether there is any *non-supporting* evidence.
 - If so, highlight or mark this evidence (e.g., cells in a table).
4. Rerate the claims, taking into account the evidence you have seen.
 - If your rating is different from initially, please describe why briefly.
5. Discuss the claims and your ratings with your partner and come to a consensus for each aspect.

Exercise - evaluating claim soundness

Observations?

- How did evidence impact your perception?
- Did non-supporting evidence influence your perception?
- Did either of these things surprise you?

References i

- Emily Allaway, Chandra Bhagavatula, Jena D. Hwang, Kathleen McKeown, and Sarah-Jane Leslie. 2024. [Exceptions, instantiations, and overgeneralization: Insights into how language models process generics](#). *Computational Linguistics*, 50:1211–1275.
- Gustavo Cilleruelo Calderón, Mahrad Almotahari, Emily Allaway, Barry Haddow, and Alexandra Birch. 2025. A new path from language models to semantic theory. In *In Submission to ARR*.
- K. Bretonnel Cohen, Jingbo Xia, Pierre Zweigenbaum, Tiffany Callahan, Orin Hargraves, Foster Goss, Nancy Ide, Aurélie Névéol, Cyril Grouin, and Lawrence E. Hunter. 2018. [Three dimensions of reproducibility in natural language processing](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Scientific Data, Kirstie Whitaker, and The Ludic Group LLP. 2017. [Kirstie Whitaker - Better Science through Better Data 2017 keynote scribe images](#).
- Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. [Show your work: Improved reporting of experimental results](#). 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 2185–2194, Hong Kong, China. Association for Computational Linguistics.

- Sarah-Jane Leslie, Sangeet Khemlani, and Sam Glucksberg. 2011. [Do all ducks lay eggs? The generic overgeneralization effect](#). *Journal of Memory and Language*, 65(1):15–31.
- Ian Magnusson, Noah A. Smith, and Jesse Dodge. 2023. [Reproducibility in NLP: What have we learned from the checklist?](#) In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12789–12811, Toronto, Canada. Association for Computational Linguistics.
- Uwe Peters and Olivier Lemeire. 2024. [Hasty generalizations are pervasive in experimental philosophy: A systematic analysis](#). *Philosophy of Science*, 91(3):661–681.
- Uwe Peters, Henrik Røed Sherling, and Benjamin Chin-Yee. 2024. Hasty generalizations and generics in medical research: A systematic review. *Plos one*, 19(7):e0306749.
- Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d’Alché Buc, Emily Fox, and Hugo Larochelle. 2021. Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). *Journal of machine learning research*, 22(164):1–20.
- Anna Rogers, Timothy Baldwin, and Kobi Leins. 2021. [‘just what do you think you’re doing, dave?’ a checklist for responsible data use in NLP](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4821–4833, Punta Cana, Dominican Republic. Association for Computational Linguistics.

- Kristin Yvonne Rozier and Eric WD Rozier. 2014. Reproducibility, correctness, and buildability: The three principles for ethical public dissemination of computer science and engineering research. In *2014 IEEE International Symposium on Ethics in Science, Technology and Engineering*, pages 1–13. IEEE.
- Shane Storks, Keunwoo Yu, Ziqiao Ma, and Joyce Chai. 2023. [NLP reproducibility for all: Understanding experiences of beginners](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10199–10219, Toronto, Canada. Association for Computational Linguistics.
- Rachael Tatman, Jake VanderPlas, and Sohier Dane. 2018. A practical taxonomy of reproducibility for machine learning research.
- Martijn Wieling, Josine Rawee, and Gertjan van Noord. 2018. Reproducibility in computational linguistics: Are we willing to share? *Computational Linguistics*, 44(4):641–649.