Transformers Learning Contrafactives: Investigation of a Negative Language Universal

Anonymous ACL submission

Abstract

No natural language is known to have contrafactive attitude verbs, yet factives are common across natural languages. Several experiments by Strohmaier and Wimmer (2022; 2023; 2025) try to explain this asymmetry via a 'learnability differential', using transformers as model learners. But they do not explore empirically-founded data distributions. We fill this gap, further improving the overall quality of training data distributions using linear programming. Our results confirm Strohmaier and Wimmer's 2025 conclusion that there is no learnability differential in production, while establishing the impact of differences in data distributions.

1 Introduction

To date, no natural language is known to have a 'contrafactive' (a morphologically atomic attitude verb that entails belief in the content of its embedded clause, but presupposes that clause's falsity), yet the contrafactive's 'factive' mirror image (a morphologically atomic attitude verb like *know* that entails belief in the content of its embedded clause and presupposes that clause's truth) is commonly, if not universally, attested (cf. Holton, 2017; Roberts and Özyıldız, 2025). As Strohmaier and Wimmer (2023) note, this raises the question of why these two verb types differ so much in their frequency.

Recent work on linguistic universals uses computational experiments to suggest that attested expressions are easier to learn than some unattested ones. E.g., Kallini et al. (2024) found that GPT-2 models learn English more easily than languages humans cannot learn. Steinert-Threlkeld and Szymanik (2019) explain the conservativity, monotonicity, and quantity universals for determiners by showing that LSTMs learn conservative, etc. determiners more easily than ones that are not.

Given these suggestive correlations between neural model and human performance, and given that a learnability difference between contrafactives and factives cannot be usefully tested with human subjects[1], Strohmaier and Wimmer (2022; 2023; 2025) also use computational experiments to (partially) explain the frequency difference between contrafactives and factives. They initially tested how easily transformers learn to comprehend attitude ascriptions fed into them. (Their transformer models effectively acted as classifiers for truth-values.) But this left open how easily transformers learn to produce attitude ascriptions. In reply, Strohmaier and Wimmer (2025) trained transformers to produce attitude ascriptions. Curiously, whilst they previously found a learnability difference between contrafactives and factives in comprehension, they did not find one in production. This leaves open whether contrafactives are overall harder to learn than factives, as they must be if a learnability difference is to explain the frequency difference between contrafactives and factives.

However, the data distribution they used for their experiment has two limitations, in light of which we might question their results. First, 50% of their training data required their transformer models to produce attitude ascriptions that are not true, effectively biasing the models towards producing lies.[2] But it is not the case that 50% of the attitude ascriptions human language users produce are lies. According to Serota et al. (2022), human language users only lie in 7% of total communication, and we see no reason to say that they lie far more frequently than that when they use attitude ascriptions.

Second, since contrafactives are unattested, how often would language learners need to produce them

[1] Since those subjects speak natural languages that have factives, but lack contrafactives, results from artificial language learning experiments would be biased by those subjects' previous knowledge of factives.

[2] For simplicity, we assume that utterances the model takes to be presupposition failures are lies. Philosophical work on lies, e.g. Stokke (2024), tends to classify such utterances as misleading instead. But Serota et al. (2010, 6), e.g., operationalise lies so as to include attempts to mislead.

if they were to learn them as part of a natural language? Our best evidence is that, as Sander (2025) argues, we would need to use contrafactives less frequently than factives, because we ascribe true beliefs to others by default. But the data distribution Strohmaier and Wimmer (2025) explore effectively assumes that factives and contrafactives would need to be used equally frequently.

Our experiment overcomes these two limitations by exploring data distributions that reflect our best evidence as to how frequently human language learners lie and how often they would need to use contrafactives if they were attested. Importantly, our results confirm Strohmaier and Wimmer (2025)'s conclusion. Our improved implementation of their paradigm also shows no relevant learnability difference.

Our main contributions are:

1. We explore a **range of empirically-informed data distributions**.
2. We provide a **deeper data analysis of our results, including significance testing** training trajectories.
3. We publicly **release our dataset and code**.[3]

2 Related Work

Holton (2017) introduced the frequency difference between contrafactives and factives. Cross-linguistic evidence relevant to the difference can be found in Rosenberg (1975); Hannon (2015); Kierstead (2015); Krifka (2016); Hsiao (2017); Anvari et al. (2019); Sander (2020); Hoeksema (2021); Bochnak and Hanink (2022); Bossi (2022); Strohmaier and Wimmer (2023); McGregor (2024); Sander (2025).

A learnability difference to explain linguistic universals is explored with comprehension-oriented experiments on human subjects in Maldonado et al. (2022) and on neural networks in Steinert-Threlkeld (2020); Steinert-Threlkeld and Szymanik (2019, 2020), and Strohmaier and Wimmer (2022; 2023). We can find production-oriented experiments on human subjects in Maldonado and Culbertson (2019) and on neural networks in Strohmaier and Wimmer (2025); Johnson et al. (2021) report experiments with both kinds of subject.

Work that encourages us to take transformers to approximate human language learning closely enough to draw conclusions about humans from results about transformers includes Ross and Pavlick (2019); Merkx and Frank (2021); Schrimpf et al. (2021); Caucheteux and King (2022); Paape (2023); Ziembicki et al. (2023); Kallini et al. (2024).

[3]LINK REDACTED FOR PEER-REVIEW

3 Data

Like Strohmaier and Wimmer (2025), we use a sequence-to-sequence task. Fig. 1 gives Strohmaier and Wimmer's function from an attitude content and one of 21 possible combinations of main value, sub value, and mind-world relation to an output ascription that consists of an attitude verb and an embedded clause (for details see Appendices A and B).

main value × sub value × mind-world relation × attitude content → attitude verb × embedded clause

Figure 1: Form of function from input to output.

We overcome the two limitations mentioned in Section 1 by generating 9 data sets that vary in how often we require the model to produce specific attitude verbs and attitude ascriptions of specific main values. Table 1 lists all 9 distributions.

	balanced	t-medium	t-heavy
equal	1:1:1 × 1:1:1	1:1:1 × 2:1:1	1:1:1 × 93:3.5:3.5
c-light	58:21:21 × 1:1:1	58:21:21 × 2:1:1	58:21:21 × 93:3.5:3.5
c-heavy	42:42:16 × 1:1:1	42:42:16 × 2:1:1	42:42:16 × 93:3.5:3.5

Table 1: Relative proportions of attitude verbs (factives : contrafactives : non-factives) and main values (true : false : p-failure) for our target distributions.

T-heavy distributions require 93% true attitude ascriptions, in line with Serota et al. (2022), with the rest evenly split between false and presupposition failure; t-medium distributions match Strohmaier and Wimmer (2025)'s distribution; finally, balanced distributions require a third each of attitude ascriptions to be true, false, and presupposition failures. C-heavy distributions require as many contrafactives to be produced as factives; equal distributions require the same number of each; and c-light distributions require many more factives than contrafactives, in line with (Sander, 2025).[4] The most plausible distribution, by our current evidence, is thus c-light/t-heavy.

Strohmaier and Wimmer (2025) sample their data sets by considering two dimensions: Distributions of required attitude verbs and distributions of required main values. This approach has two

[4]The number of contrafactives in c-light distributions is inspired by the proportion of non-factives to factives in Bartsch and Wellman (1995)'s study of over 200,000 spontaneous utterances by English-speaking children up to age 6.

downsides. First, it does not ensure sufficient data points for each of the 21 possible combinations of main value, sub value, and mind-world relation. Second, it double-samples combinations that allow the model to produce factives or contrafactives.

We address both downsides by using linear programming to determine the number of instances for the 21 possible combinations.[5] We restrict our datasets in two ways: First, they must reflect the relative proportions of one of the distributions in Table 1. Second, the total number of data points must be 160,000.[6] The optimization goal is to maximize the minimum number of data points observed across each of the combinations minus the maximum number observed. That is, we keep these combinations in as narrow a band as possible.

4 Experimental Design

For the experiment, we split the dataset into a train, dev, and test split. Each of the latter two makes up 10% of the overall dataset.

Like Strohmaier and Wimmer (2025), we base our experiments on the standard transformer model included in the pyTorch library (Paszke et al., 2017). As in Vaswani et al. (2017), we use an encoder-decoder architecture (for details see Appendix C).

Like Strohmaier and Wimmer (2025), we do not consider word order or syntactic effects and so fix the vocabulary the model can produce at each token position by using separate heads. Consequently, the model always produces a token for each position.

4.1 Training

Strohmaier and Wimmer (2025) explored 41 hyperparameter settings using a randomised search. We undertook a more extensive search, exploring 60 settings for each of our 9 distributions, using Optuna with TPE sampling algorithm (Akiba et al., 2019; Bergstra et al., 2011; Watanabe, 2023)

We trained the final model on the highest performing setting selected by Optuna on the dev split. We then evaluated the final model on the test split.

4.2 Evaluation

Our evaluation metric was the correctness of the output ascription given the input sequence. If an input allows a factive or contrafactive, either is correct. To assess the robustness of our evaluation results, we use an MC dropout-based method (Gal and Ghahramani, 2016) and run the model 250-times on the test set with active dropout.

[5]We use the PuLP library (Mitchell et al., 2011).

[6]For simplicity, we use floats, which we round. This leads to a negligible number of deviations.

5 Results

To compare the ease of learning of our attitude verbs, we compare performance every 20 training steps. We ignore training steps where performance is equal, which primarily occur when the model has achieved 100% performance for both verbs. Table 2 gives the results of these comparisons and whether the difference in performance is significant. Appendix G describes the significance tests in more detail and gives finer-grained results.

verb	semantic	f>c	sig	nf>c	f>nf
equal	balanced	41.7%		83.3%	9.7%
equal	t-medium	48.3%		96.4%	3.7%
equal	t-heavy	40.9%	✓	79.2%	17.9%
c-light	balanced	95.2%	✓	60.0%	95.2%
c-light	t-medium	100.0%	✓	27.8%	72.2%
c-light	t-heavy	26.7%		40.0%	46.7%
c-heavy	balanced	27.3%		15.0%	81.0%
c-heavy	t-medium	40.0%		4.5%	100.0%
c-heavy	t-heavy	53.7%		15.4%	77.5%

Table 2: Comparative performance advantage over percentage of time steps (factive=f, contrafactive=c, non-factive=nf). Comparison occurs every 20th training step.

For 6 of 9 distributions, our model learns contrafactives faster than factives. Of the remaining 3 models that learn factives faster, 2 are trained on c-light distributions, which include fewer contrafactives than factives. Notably, the only c-light distribution where contrafactives are learned faster than factives, though not significantly so, is the most plausible distribution: c-light/t-heavy.

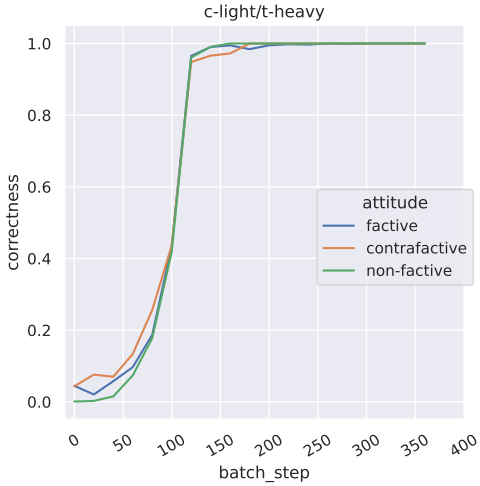


Figure 2: Performance of c-light/t-heavy.

Looking at completely trained models, they learn factives and contrafactives across all target distributions. Performance deviates only slightly from 100% correct. Even using dropout, the lowest performance found is a correctness of 99.87% for con-

trafactives in the equal/balanced distribution. Appendix I says more about remaining errors.

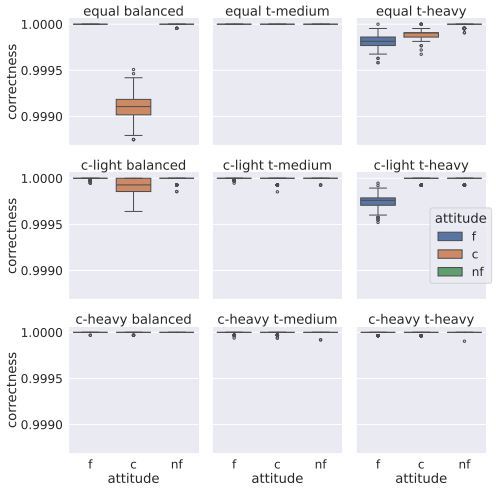


Figure 3: Performance for attitude verbs (factive=f, contrafactive=c, non-factive=nf) across distributions.

Since some inputs permit the production of a factive or contrafactive, the model can prefer one verb (i.e. given those inputs, produce it more frequently than the other) and still produce 100% correct output (Strohmaier and Wimmer, 2025). For 6 distributions the model came to consistently prefer one verb across the 250 runs with dropout: in 3 cases, it prefers factives; in 3 cases, contrafactives. We observe no systematic pattern in how preferences change over time. For details see Appendix H.

6 Discussion of Results

Although some c-light models learn contrafactives more slowly than factives, there are two reasons why these results do not support the general conclusion that contrafactives are harder to learn. First, the c-light distributions involved are not the most plausible, since they are not t-heavy. Second, c-light distributions contain far fewer contrafactives than factives (less than half); it is then no surprise that c-light distributions generally make learning contrafactives slower than factives.

Also, **for the few distributions where contrafactives are learned more slowly, performance rapidly catches up**: the penalty for contrafactives occurs mostly in the first 300–400 training steps (see Appendix G) and is then quickly overcome. E.g., for c-light/balanced, correctness for contrafactives rose from 47.5% to 99.3% in eighty training steps, reaching almost the same percentage as for factives (99.6%). A learning dynamic that allows for such rapid catch-up is unlikely to explain the frequency difference between factives and contrafactives. Many existing lexical items are rapidly learned later, leaving open why contrafactives with the same dynamic do not exist.[7]

Trained on c-light/t-heavy, the most plausible distribution, our model does not prefer factives over contrafactives. Across all 3 c-light distributions, which contain far fewer contrafactives than factives, we find 2 distributions where contrafactives are as likely or even more likely to be produced than factives if both are allowed.

Looking at fully trained models, those trained on t-heavy distributions, which require 93% true ascriptions, perform worse. Given our evidential support for this distribution, this suggests that **previous research understates the challenge of learning our attitude verbs**. In fact, the model trained on equal/t-medium, which corresponds most closely to the distribution of Strohmaier and Wimmer (2025), performs better than any other (see Appendix E).[8]

7 Conclusion

In line with Strohmaier and Wimmer (2025), our results do not support the hypothesis that contrafactives are harder to learn than factives. However, like Strohmaier and Wimmer (2025)'s results, our results need to be interpreted with care. They are consistent with contrafactives being harder to learn to comprehend, as Strohmaier and Wimmer (2022; 2023) argued. So, they do not obviously entail that contrafactives are not harder to learn overall. Still, our results do raise a key question: are contrafactives hard enough to learn overall for this to explain, even partly, why contrafactives are so much less frequent than factives?

Future research might try to find features of human language learning relevant to learning contrafactives other than those we got from Sander (2025); Serota et al. (2022). By implementing such features in neural models, we might well recover a learnability-based explanation of why contrafactives are so much rarer than factives.[9]

[7] For accidental learning by reading a text, it has been shown that two exposures can be sufficient, see Hulme et al. (2019).

[8] The distributions with equally many factives, contrafactives, and non-factives had long training times, i.e. over 2500 training steps, during the hyperparameter search. But, tracking performance over time shows that almost all additional steps are unnecessary, except for the equal/t-heavy distribution (see Appendix G).

[9] Strohmaier and Wimmer (2025, 406) note that pragmatic-syntactic bootstrapping (see Hacquard and Lidz 2022) might be one such feature.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Amir Anvari, Mora Maldonado, and Andrés Soria Ruiz. 2019. The puzzle of Reflexive Belief Construction in Spanish. *Proceedings of Sinn und Bedeutung*, 23(1):57–74.

Karen Bartsch and Henry M. Wellman. 1995. *Children talk about the mind*. Oxford University Press, New York.

James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. 2011. Algorithms for hyper-parameter optimization. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, NIPS'11, pages 2546–2554, Red Hook, NY, USA. Curran Associates Inc.

M. Ryan Bochnak and Emily A. Hanink. 2022. Clausal embedding in Washo: Complementation vs. modification. *Natural Language & Linguistic Theory*, 40(4):979–1022.

Madeline Bossi. 2022. Unifying negative bias and reminding functions: The case of Kipsigis *par*. In Özge Bakay, Breanna Pratley, Evan Neu, and Peyton Deal, editors, *Proceedings of the Fifty-Second Annual Meeting of the North East Linguistic Society*, pages 95–104. Amherst.

Charlotte Caucheteux and Jean-Rémi King. 2022. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1).

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 1050–1059. PMLR.

Valentine Hacquard and Jeffrey Lidz. 2022. On the Acquisition of Attitude Verbs. *Annual Review of Linguistics*, 8(1):193–212.

Michael Hannon. 2015. The universal core of knowledge. *Synthese*, 192(3):769–786.

Jack Hoeksema. 2021. Verbs of deception, point of view and polarity. *Proceedings of the International Conference on Head-Driven Phrase Structure Grammar*, pages 26–46.

Richard Holton. 2017. I—Facts, Factives, and Contrafactives. *Aristotelian Society Supplementary Volume*, 91(1):245–266.

Pei-Yi Katherine Hsiao. 2017. On counterfactual attitudes: a case study of Taiwanese Southern Min. *Lingua Sinica*, 3(1):4.

Rachael C. Hulme, Daria Barsky, and Jennifer M. Rodd. 2019. Incidental Learning and Long-Term Retention of New Word Meanings From Stories: The Effect of Number of Exposures. *Language Learning*, 69(1):18–43.

Tamar Johnson, Kexin Gao, Kenny Smith, Hugh Rabagliati, and Jennifer Culbertson. 2021. Investigating the effects of i-complexity and e-complexity on the learnability of morphological systems. *Journal of Language Modelling*, 9(1):97–150. Number: 1.

Julie Kallini, Isabel Papadimitriou, Richard Futrell, Kyle Mahowald, and Christopher Potts. 2024. Mission: Impossible language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, page 14691–14714, Bangkok, Thailand. Association for Computational Linguistics.

Gregory Weiss Kierstead. 2015. Projectivity and the Tagalog Reportative Evidential. Master's thesis, The Ohio State University.

Manfred Krifka. 2016. Realis and Non-Realis Modalities in Daakie (Ambrym, Vanuatu). *Semantics and Linguistic Theory*, pages 566–583.

Mora Maldonado and Jennifer Culbertson. 2019. Learnability as a window into universal constraints on person systems. *Proceedings of the Amsterdam Colloquium*, 22:484–493.

Mora Maldonado, Jennifer Culbertson, and Wataru Uegaki. 2022. Learnability and constraints on the semantics of clause-embedding predicates.

William B. McGregor. 2024. On the expression of mistaken beliefs in Australian languages. *Linguistic Typology*, 28(1):101–145.

Danny Merkx and Stefan L. Frank. 2021. Human Sentence Processing: Recurrence or Attention? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 12–22. Association for Computational Linguistics.

Stuart Mitchell, Michael O'Sullivan, and Iain Dunning. 2011. PuLP: A Linear Programming Toolkit for Python. *Preprint*, Optimization Online:11731.

Dario Paape. 2023. When Transformer models are more compositional than humans: The case of the depth charge illusion. *Experiments in Linguistic Meaning*, 2:202–218.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch.

Tom Roberts and Deniz Özyıldız. 2025. A causal explanation for the contrafactive gap. LingBuzz Published In:.

Marc Stephen Rosenberg. 1975. *Counterfactives: A Pragmatic Analysis of Presupposition*. Ph.D. thesis, University of Illinois at Urbana-Champaign.

Alexis Ross and Ellie Pavlick. 2019. How well do NLI models capture verb veridicality? 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 2230–2240, Hong Kong, China. Association for Computational Linguistics.

Thorsten Sander. 2020. Fregean Side-Thoughts. *Australasian Journal of Philosophy*, 0(0).

Thorsten Sander. 2025. A Puzzle About Anti-Factives. *Journal of the American Philosophical Association*, pages 1–20.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45).

Kim B. Serota, Timothy R. Levine, and Franklin J. Boster. 2010. The Prevalence of Lying in America: Three Studies of Self-Reported Lies. *Human Communication Research*, 36(1):2–25.

Kim B. Serota, Timothy R. Levine, and Tony Docan-Morgan. 2022. Unpacking variation in lie prevalence: Prolific liars, bad lie days, or both? *Communication Monographs*, 89(3):307–331. Publisher: NCA Website _eprint: https://doi.org/10.1080/03637751.2021.1985153.

Shane Steinert-Threlkeld. 2020. An Explanation of the Veridical Uniformity Universal. *Journal of Semantics*, 37(1):129–144.

Shane Steinert-Threlkeld and Jakub Szymanik. 2019. Learnability and semantic universals. *Semantics and Pragmatics*, 12:4:1–39.

Shane Steinert-Threlkeld and Jakub Szymanik. 2020. Ease of learning explains semantic universals. *Cognition*, 195.

Andreas Stokke. 2024. Lies are assertions and presuppositions are not. *Inquiry*, 0(0):1–24. Publisher: Routledge _eprint: https://doi.org/10.1080/0020174X.2024.2389582.

David Strohmaier and Simon Wimmer. 2023. Contrafactives and Learnability: An Experiment with Propositional Constants. *Post-Proceedings of Logic and Engineering of Natural Language Semantics 19*.

David Strohmaier and Simon Wimmer. 2025. Contrafactives, Learnability, and Production. *Experiments in Linguistic Meaning*, 3:395–410.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. *31st Conference on Neural Information Processing Systems*, pages 1–11.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, and 15 others. 2020. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272.

Shuhei Watanabe. 2023. Tree-Structured Parzen Estimator: Understanding Its Algorithm Components and Their Roles for Better Empirical Performance. *Preprint*, arXiv:2304.11127.

Simon Wimmer and David Strohmaier. 2022. Contrafactives and Learnability. In Marco Degano, Tom Roberts, Giorgio Sbardolini, and Marieke Schouwstra, editors, *Proceedings of the 23rd Amsterdam Colloquium*, pages 298–305.

Daniel Ziembicki, Karolina Seweryn, and Anna Wróblewska. 2023. Polish natural language inference and factivity: An expert-based dataset and benchmarks. *Natural Language Engineering*, pages 1–32.

A Input-Output Function

	MW	Main Value	Sub Value	Attitude Verb	Embedded Clause
CTT	C	True	True	Factive	Matching
CTF	C	True	False	IMPOSSIBLE	—
CTU	C	True	Unknown	IMPOSSIBLE	—
CFT	C	False	True	Factive	Non-Matching
CFF	C	False	False	Contrafactive	Non-Matching
CFU	C	False	Unknown	Non-Factive	Non-Matching
CPT	C	P-Failure	True	Contrafactive	Matching
CPF	C	P-Failure	False	Factive	Non-Matching
CPU	C	P-Failure	Unknown	Factive or Contrafactive	Non-Matching
ITT	I	True	True	IMPOSSIBLE	—
ITF	I	True	False	Contrafactive	Matching
ITU	I	True	Unknown	IMPOSSIBLE	—
IFT	I	False	True	Factive	Non-Matching
IFF	I	False	False	Contrafactive	Non-Matching
IFU	I	False	Unknown	Non-Factive	Non-Matching
IPT	I	P-Failure	True	Contrafactive	Non-Matching
IPF	I	P-Failure	False	Factive	Matching
IPU	I	P-Failure	Unknown	Factive or Contrafactive	Non-Matching
UTT	U	True	True	IMPOSSIBLE	—
UTF	U	True	False	IMPOSSIBLE	—
UTU	U	True	Unknown	Non-Factive	Matching
UFT	U	False	True	Factive	Non-Matching
UFF	U	False	False	Contrafactive	Non-Matching
UFU	U	False	Unknown	Non-Factive	Non-Matching
UPT	U	P-Failure	True	Contrafactive	Non-Matching
UPF	U	P-Failure	False	Factive	Non-Matching
UPU	U	P-Failure	Unknown	Factive or Contrafactive	Matching

Table 3: Possible combinations of semantic-pragmatic conditions and output tokens.

Table 3 specifies possible combinations of semantic-pragmatic conditions and output tokens. 6 rows in the table give impossible combinations, leaving 21 that are possible.

Main value is the required value of the output attitude ascription. Sub value the required value of the embedded clause used in the output attitude ascription. The mind-world relation, MW, has three possible values: mind and world are compatible (C), incompatible (I), or the world state is unknown (U). The only input not listed in table 3 is the attitude content, which tells the model what the subject of the ascription believes.

Output tokens partly consist of one of three attitude verbs: factive, contrafactive, and non-factive. (A non-factive is a verb like *believe* that entails belief in the content of its embedded clause, but presupposes neither the truth nor the falsity of that clause.) Rows with main value P-failure and sub value Unknown are notable because they allow for factives and contrafactives in the output.

The embedded clause column gives a relation between the embedded clause required in the output attitude ascription and the attitude content given to the model. The embedded clause can either match the attitude content or fail to match it. E.g., "eat rory tomato basil soup lunch tomorrow" matches "Rory will-eat tomato-basil soup for lunch tomorrow". To get a matching clause, the model must output a specific embedded clause. To get a non-matching clause, the model can output any embedded clause other than the matching one.

B Vocabulary

The vocabulary for the input attitude content is slightly larger than in Strohmaier and Wimmer (2025). This allows us to generate sufficient data. Fig. 4 gives the function from attitude content to embedded clause. The only added vocabulary in the embedded clause are prepositions such as 'for' and tense markers such as 'will.'

Category	Lexical Items
Verb	eat, cook, order, buy, season
Subject	rory, lorelai, lane, paris, timon, ahab
Ingredient	tomato, pumpkin, mushroom, carrot, potato
Spice	basil, oregano, pepper, chili, coconut
Dish	soup, pie, rice, stew, curry
Meal	lunch, dinner, breakfast, brunch
Day	day-before-yesterday, yesterday, now, today, tomorrow, day-after-tomorrow

Table 4: Attitude content vocabulary. Content has one token of each category.

$$\text{verb} \times \text{agent} \times \text{ingredient} \times \text{spice} \times \text{dish} \times \text{meal} \times \text{day} \rightarrow \times \text{agent} \times \text{verb (with tense)} \times \text{main ingredient} + \text{spice} \times \text{preposition} \times \text{dish} \times \text{meal} \times \text{day}$$

Figure 4: Form of function from attitude content to embedded clause.

C Architectural Details

We implemented the models using pyTorch and parallelised using the lightning-Fabric library. For training and evaluation, we used two A100 GPUs.

Like Strohmaier and Wimmer (2025), we use the Binary Cross Entropy (BCE) loss function. Since this is not a typical language modelling task, we have special cases, e.g. with more then one allowed embedded clause, where we set the target value for all allowed embedded clauses to 0.5. Also, if two attitude verbs are allowed, we set the target value for both to 1.

D Hyperparameter Search Details

The target of the hyperparameter search was to maximize the overall correctness of output created on

the dev set. All 9×60 settings were fully explored (no pruning). The 60 settings were explored by 4 processes, which communicated via a remote mySQL database.

Name	Lower	Upper	Step Size	Log-Space
Embedding dim.	48	480	24	
Hidden dim.	48	480	24	
# Encoder layers	5	25	5	
# Decoder layer	5	25	5	
Dropout prob.	0.1	0.3	0.1	
Learning rate	1e-07	1e-03	–	✓
Epochs	5	50	1	
Batch size	8	3200	8	

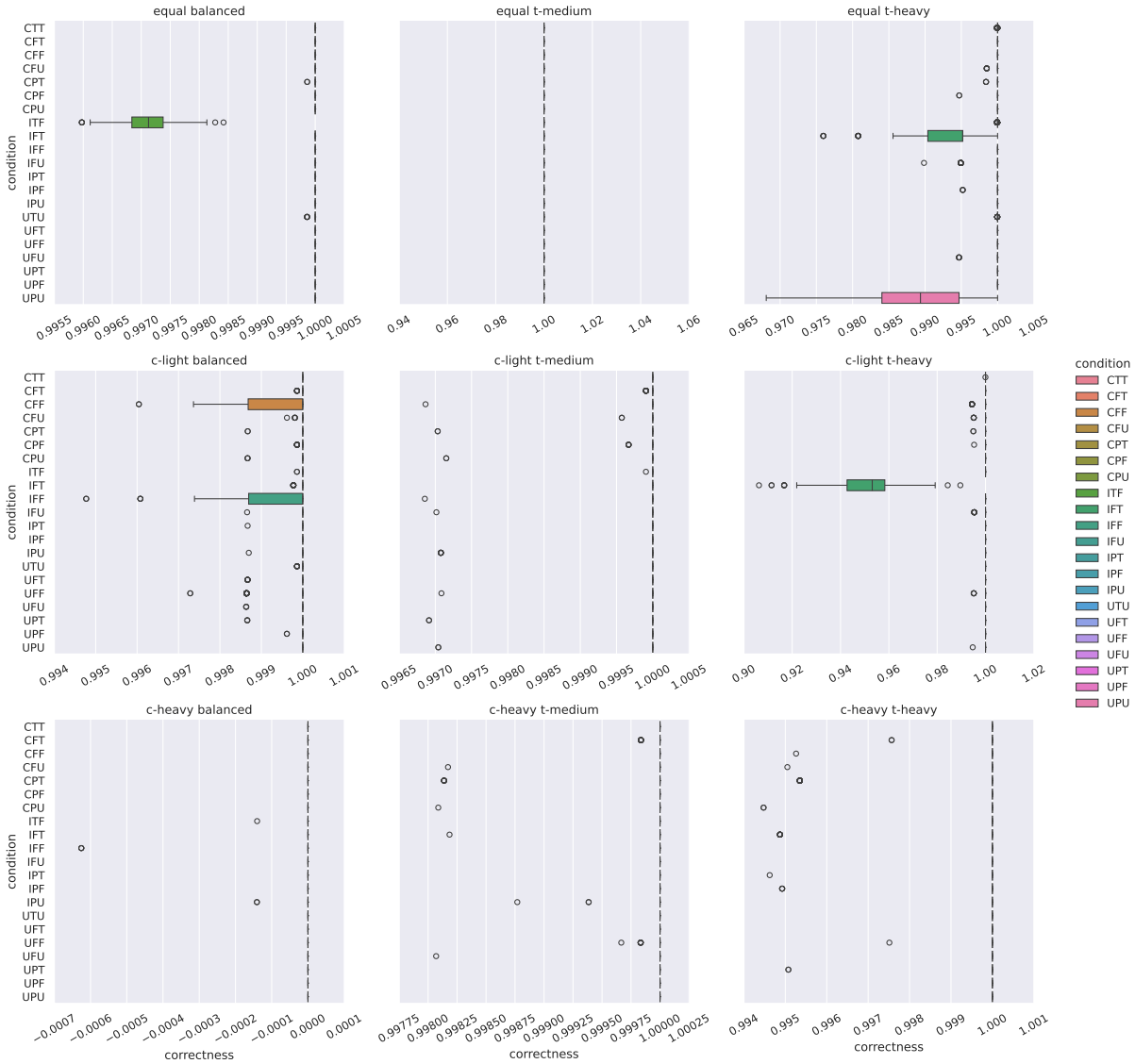
Table 5: Hyperparameter settings used for search by Optuna.

E Final Performance

verb	semantic	att	min	mean	std
equal	balanced	f	100%	100%	0
		c	99.8749%	99.9088%	0.00013
		nf	99.9955%	99.9999%	6.3e-06
	t-medium	f	100%	100%	0
		c	100%	100%	0
		nf	100%	100%	0
	t-heavy	f	99.9583%	99.9825%	7.1e-05
		c	99.9674%	99.9894%	5e-05
		nf	99.9907%	99.9992%	1.8e-05
c-light	balanced	f	99.9948%	99.9995%	1.1e-05
		c	99.964%	99.9918%	7.6e-05
		nf	99.9856%	99.9995%	2e-05
	t-medium	f	99.9947%	99.9997%	8.9e-06
		c	99.9853%	99.9994%	2.1e-05
		nf	99.9927%	99.9999%	8e-06
	t-heavy	f	99.9521%	99.9743%	7.7e-05
		c	99.9926%	99.9992%	2.3e-05
		nf	99.9926%	99.9995%	1.8e-05
c-heavy	balanced	f	99.9969%	100%	2.8e-06
		c	99.9968%	99.9999%	4.9e-06
		nf	100%	100%	0
	t-medium	f	99.9938%	99.9998%	8.3e-06
		c	99.9938%	99.9998%	9.3e-06
		nf	99.9919%	99.9999%	7.3e-06
	t-heavy	f	99.9963%	99.9997%	9.5e-06
		c	99.9963%	99.9997%	1.1e-05
		nf	99.9904%	100%	6.1e-06

Table 6: Correct output by required attitude verb (f=factive, c=contrafactive, nf=non-factive) across target distributions. Minimum, mean, and standard deviation are calculated using the 250 runs with active dropout.

F Detailed Results

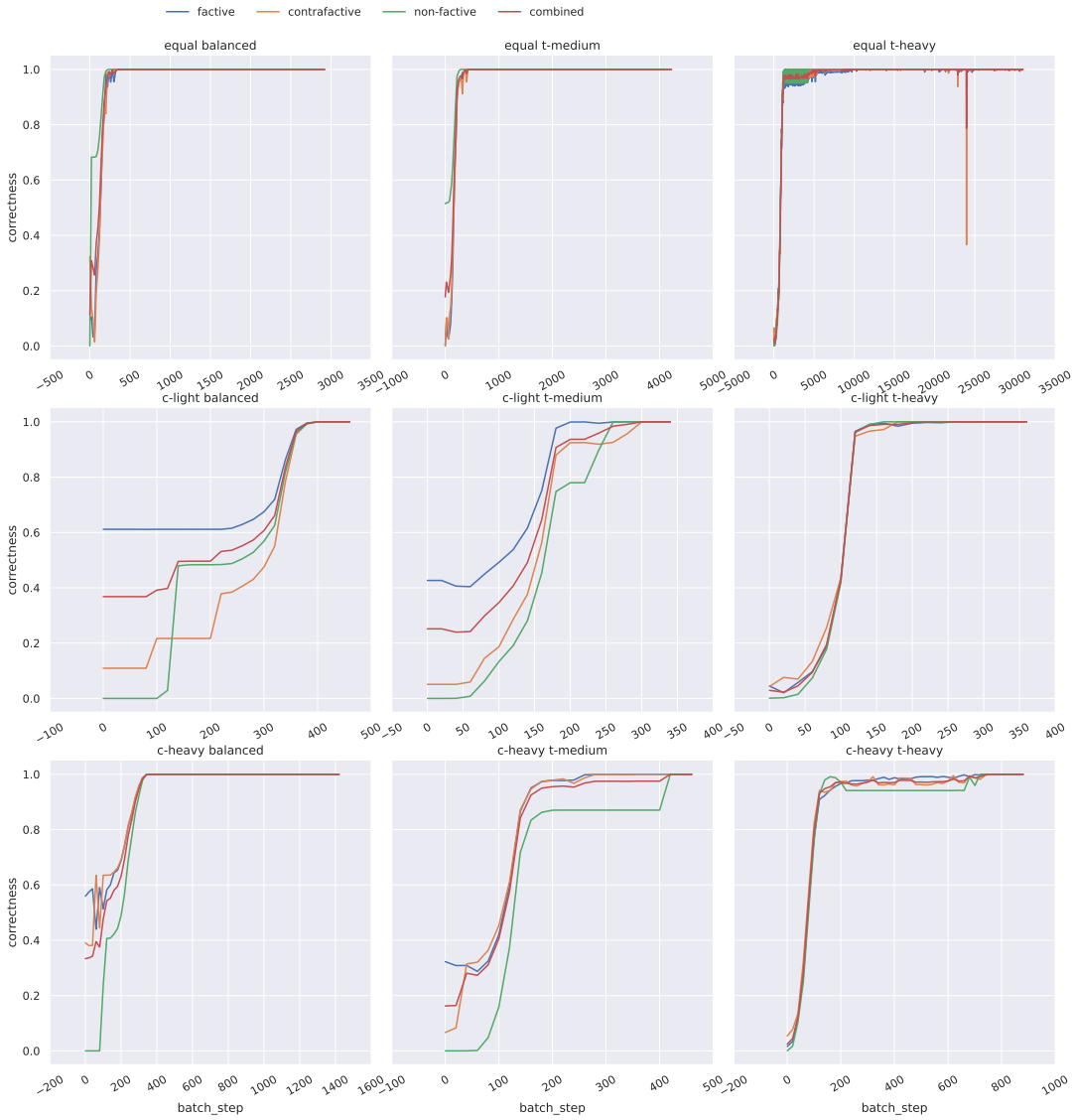


Performance on the 21 possible conditions given in Appendix A.

G Learning Trajectory

Significance Test We use the two-sided permutation-test implementation of SciPy (Virtanen et al., 2020) with 9999 resamples and permutation type "samples". The test statistic was the difference between the sum of comparisons in which the model performed better on factives than on contrafactives and the sum of comparisons where the reverse held.

Further Results A closer look at Table 2 suggests that non-factives differ across distributions. E.g., non-factives are learned fastest with distributions with as many non-factives as contrafactives and factives. But with distributions with fewer non-factives than factives (c-heavy, c-light), non-factives tend to be learned more slowly than contrafactives.



Changes in performance over the training for 3 attitude verbs and 9 data distributions.

step	f	c	nf	comb
0	9.546%	32.237%	0.027%	11.279%
200	98.280%	83.964%	99.362%	93.685%
420	99.996%	100.000%	100.000%	99.998%
640	100.000%	100.000%	100.000%	100.000%
840	100.000%	100.000%	100.000%	100.000%
1040	100.000%	100.000%	100.000%	100.000%
1260	100.000%	99.991%	100.000%	99.997%
1480	100.000%	100.000%	100.000%	100.000%
1680	100.000%	100.000%	100.000%	100.000%
1880	100.000%	100.000%	100.000%	100.000%
2100	100.000%	100.000%	100.000%	100.000%
2320	99.982%	100.000%	100.000%	99.994%
2520	100.000%	100.000%	100.000%	100.000%
2720	100.000%	100.000%	100.000%	100.000%
2920	100.000%	100.000%	100.000%	100.000%

(a) Trajectory for distribution equal/balanced

step	f	c	nf	comb
0	0.118%	0.005%	51.458%	17.745%
300	96.863%	96.503%	99.995%	97.763%
600	100.000%	100.000%	100.000%	100.000%
900	100.000%	100.000%	100.000%	100.000%
1220	100.000%	100.000%	100.000%	100.000%
1520	100.000%	100.000%	100.000%	100.000%
1820	100.000%	100.000%	100.000%	100.000%
2120	100.000%	100.000%	100.000%	100.000%
2420	100.000%	100.000%	100.000%	100.000%
2720	100.000%	100.000%	100.000%	100.000%
3020	100.000%	100.000%	100.000%	100.000%
3340	100.000%	100.000%	100.000%	100.000%
3640	100.000%	100.000%	100.000%	100.000%
3940	100.000%	100.000%	100.000%	100.000%
4220	100.000%	100.000%	100.000%	100.000%

(b) Trajectory for distribution equal/t-medium

step	f	c	nf	comb
0	6.470%	1.939%	0.000%	2.183%
2220	96.841%	97.175%	99.671%	97.962%
4420	96.786%	99.860%	99.750%	98.812%
6640	98.921%	99.888%	100.000%	99.636%
8860	99.931%	99.986%	100.000%	99.972%
11080	99.986%	100.000%	100.000%	99.995%
13280	99.995%	99.995%	100.000%	99.998%
15500	99.977%	99.986%	100.000%	99.992%
17720	99.977%	99.977%	100.000%	99.992%
19920	100.000%	100.000%	100.000%	100.000%
22140	99.977%	99.921%	99.995%	99.972%
24360	100.000%	100.000%	100.000%	100.000%
26580	100.000%	100.000%	100.000%	100.000%
28780	99.935%	99.935%	99.884%	99.939%
30980	100.000%	100.000%	100.000%	100.000%

(c) Trajectory for distribution equal/t-heavy

step	f	c	nf	comb
0	61.151%	10.916%	0.000%	36.807%
40	61.151%	10.916%	0.000%	36.807%
60	61.151%	10.916%	0.000%	36.807%
100	61.151%	21.688%	0.000%	39.146%
140	61.146%	21.688%	47.956%	49.534%
180	61.146%	21.681%	48.338%	49.618%
200	61.138%	21.695%	48.338%	49.620%
240	61.538%	38.426%	48.763%	53.573%
280	64.731%	43.045%	52.809%	57.279%
300	67.472%	47.514%	56.854%	60.692%
340	86.343%	78.463%	81.806%	83.643%
380	99.569%	99.280%	99.264%	99.448%
420	99.997%	100.000%	100.000%	99.998%
440	100.000%	100.000%	100.000%	100.000%
460	100.000%	100.000%	100.000%	100.000%

(d) Trajectory for distribution c-light/balanced

step	f	c	nf	comb
0	42.654%	5.093%	0.000%	25.158%
20	42.657%	5.093%	0.000%	25.160%
60	40.413%	5.945%	0.783%	24.185%
80	44.969%	14.543%	6.297%	29.850%
100	49.181%	18.629%	13.273%	34.640%
120	53.754%	28.557%	19.080%	40.651%
160	75.072%	56.298%	45.298%	64.562%
180	97.706%	88.000%	74.814%	90.744%
200	99.915%	92.460%	77.980%	93.664%
240	99.481%	91.894%	89.871%	95.934%
260	99.950%	92.585%	100.000%	98.420%
280	99.992%	95.760%	100.000%	99.098%
300	99.992%	99.978%	100.000%	99.995%
340	99.992%	99.978%	100.000%	99.995%

(e) Trajectory for distribution c-light/t-medium

step	f	c	nf	comb
0	4.480%	4.310%	0.089%	2.920%
20	2.050%	7.598%	0.251%	2.256%
60	9.673%	13.353%	7.377%	9.419%
80	18.697%	25.617%	17.778%	19.503%
100	43.512%	43.650%	41.817%	42.950%
140	99.023%	96.623%	99.076%	98.625%
160	99.465%	97.223%	100.000%	99.100%
200	99.502%	99.993%	100.000%	99.706%
220	99.782%	100.000%	100.000%	99.872%
240	99.691%	100.000%	100.000%	99.819%
280	99.920%	100.000%	100.000%	99.953%
300	100.000%	100.000%	100.000%	100.000%
320	100.000%	100.000%	100.000%	100.000%
360	100.000%	100.000%	100.000%	100.000%

(f) Trajectory for distribution c-light/t-heavy

step	f	c	nf	comb
0	55.907%	39.030%	0.000%	33.341%
100	51.340%	63.507%	23.683%	47.625%
200	68.717%	69.152%	48.880%	63.139%
300	95.509%	95.849%	92.343%	94.797%
420	100.000%	100.000%	100.000%	100.000%
520	100.000%	100.000%	100.000%	100.000%
620	100.000%	100.000%	100.000%	100.000%
720	100.000%	100.000%	100.000%	100.000%
820	100.000%	100.000%	100.000%	100.000%
920	100.000%	100.000%	100.000%	100.000%
1020	100.000%	100.000%	100.000%	100.000%
1140	100.000%	100.000%	100.000%	100.000%
1240	100.000%	100.000%	100.000%	100.000%
1340	100.000%	100.000%	100.000%	100.000%
1420	100.000%	100.000%	100.000%	100.000%

(g) Trajectory for distribution c-heavy/balanced

step	f	c	nf	comb
0	32.243%	6.649%	0.000%	16.257%
40	30.873%	31.429%	0.016%	28.068%
60	28.732%	32.060%	0.147%	27.316%
100	41.997%	45.298%	15.857%	40.489%
140	86.567%	87.152%	71.716%	84.168%
180	97.307%	97.474%	86.237%	94.972%
200	97.797%	97.843%	87.003%	95.516%
240	97.902%	96.719%	87.027%	95.344%
280	99.923%	99.944%	87.027%	97.469%
300	99.944%	99.978%	87.027%	97.478%
340	99.916%	99.858%	87.027%	97.441%
380	100.000%	100.000%	87.027%	97.513%
420	100.000%	99.997%	100.000%	99.998%
440	100.000%	100.000%	100.000%	100.000%
460	100.000%	100.000%	100.000%	100.000%

(h) Trajectory for distribution c-heavy/t-medium

step	f	c	nf	comb
0	1.607%	5.404%	0.048%	2.438%
60	27.054%	32.388%	25.079%	28.802%
120	90.888%	94.114%	92.854%	93.088%
200	96.647%	97.351%	97.101%	97.162%
260	97.690%	95.760%	94.115%	96.433%
320	97.869%	99.111%	94.115%	97.878%
380	98.148%	96.499%	94.115%	96.941%
440	98.316%	98.588%	94.115%	97.811%
520	99.114%	96.122%	94.115%	97.095%
580	99.187%	96.653%	94.115%	97.305%
640	99.253%	97.092%	94.173%	97.541%
700	99.865%	98.639%	95.955%	98.705%
780	100.000%	100.000%	100.000%	100.000%
840	99.993%	100.000%	100.000%	99.997%
880	99.993%	100.000%	100.000%	99.997%

(i) Trajectory for distribution c-heavy/t-heavy

Table 7: Learning Trajectory for 9 distributions (factive=f, contrafactive=c, non-factive=nf, comb=combined). "step" refers to the number of training steps taken, which is equivalent to the number of batches seen.

H Preference where Factives and Contrafactives Both Permitted

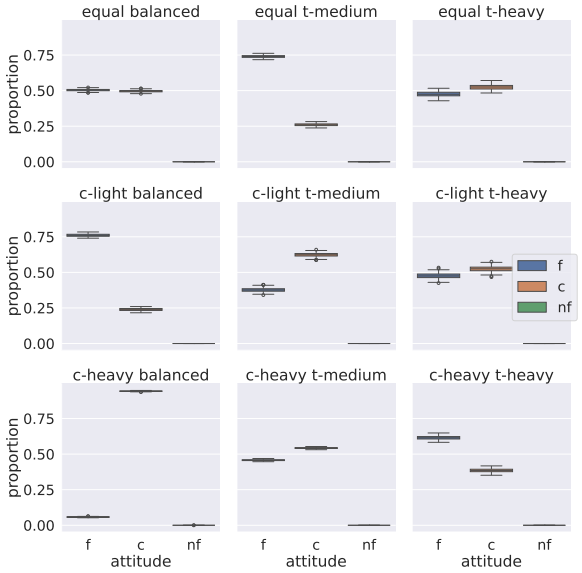


Figure 5: Attitude verb preference of final trained models across target distributions.

verb	semantic	att	min	mean	max	std
equal	balanced	f	48.5%	50.4%	52.2%	0.7
		c	47.8%	49.6%	51.5%	0.7
		nf	0.0%	0.0%	0.0%	0.0
	t-medium	f	71.8%	74.0%	76.2%	0.8
		c	23.8%	26.0%	28.2%	0.8
		nf	0.0%	0.0%	0.0%	0.0
	t-heavy	f	42.9%	47.6%	51.7%	1.8
		c	48.3%	52.4%	57.1%	1.8
		nf	0.0%	0.0%	0.0%	0.0
c-light	balanced	f	74.0%	76.1%	78.4%	0.9
		c	21.6%	23.9%	26.0%	0.9
		nf	0.0%	0.0%	0.0%	0.0
	t-medium	f	34.0%	37.7%	41.3%	1.3
		c	58.7%	62.3%	66.0%	1.3
		nf	0.0%	0.0%	0.0%	0.0
	t-heavy	f	42.5%	47.6%	53.3%	1.9
		c	46.7%	52.4%	57.5%	1.9
		nf	0.0%	0.0%	0.0%	0.0
c-heavy	balanced	f	5.3%	5.8%	6.4%	0.2
		c	93.6%	94.2%	94.7%	0.2
		nf	0.0%	0.0%	0.0%	0.0
	t-medium	f	44.7%	45.7%	46.9%	0.4
		c	53.1%	54.3%	55.3%	0.4
		nf	0.0%	0.0%	0.0%	0.0
	t-heavy	f	58.3%	61.5%	64.9%	1.3
		c	35.1%	38.5%	41.7%	1.3
		nf	0.0%	0.0%	0.0%	0.0

Table 8: Proportion of attitude verb chosen where factive and contrafactive are allowed (f=factive, c=contrafactive, nf=non-factive). Numbers concern final trained models. Minimum, mean, max, and standard deviation are calculated using the 250 runs with dropout. Standard deviation given in percentage points.

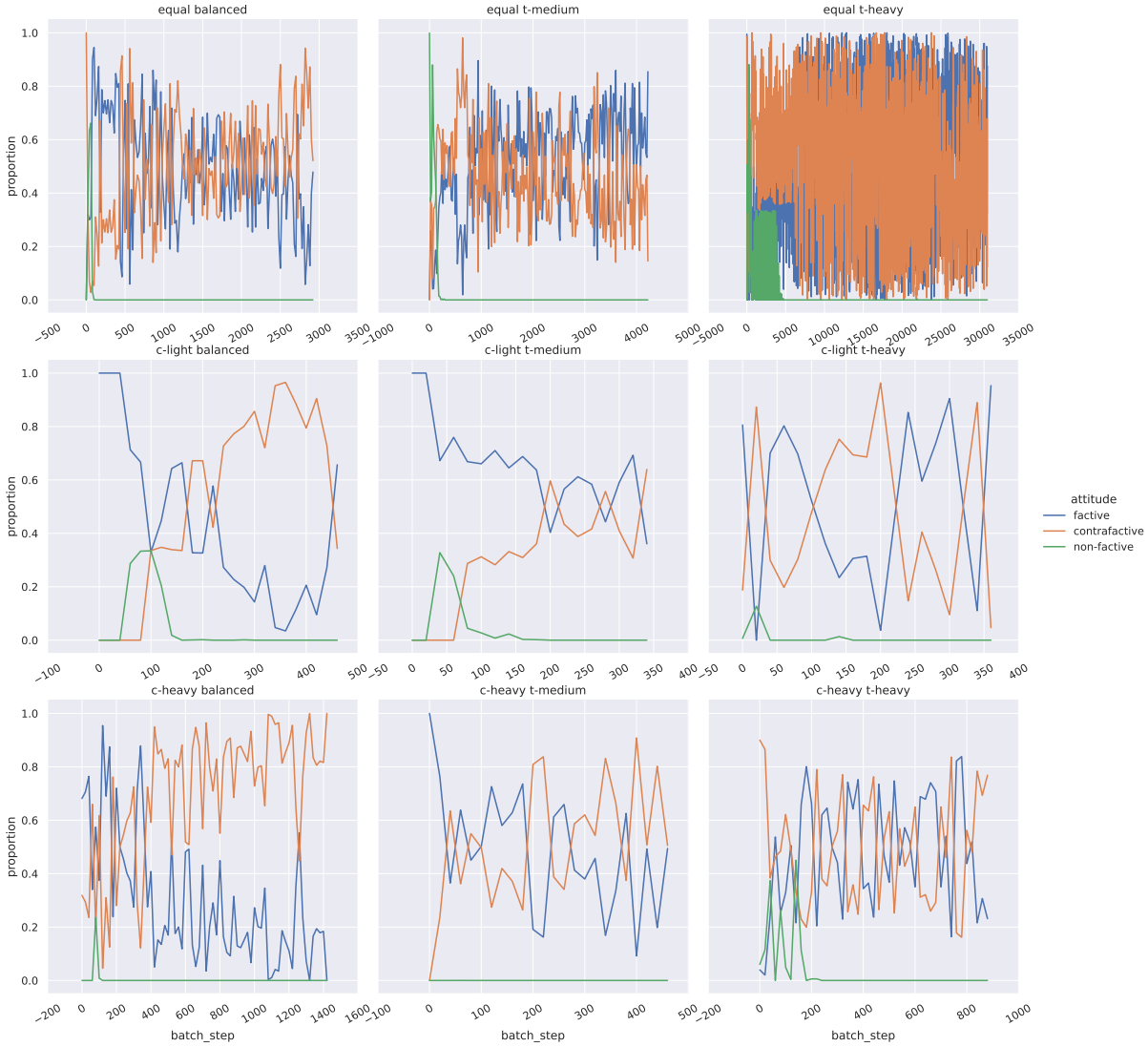
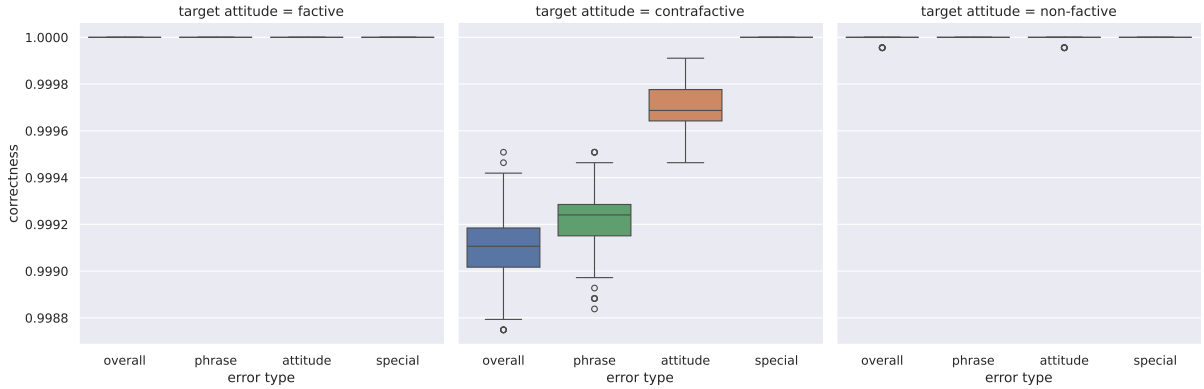


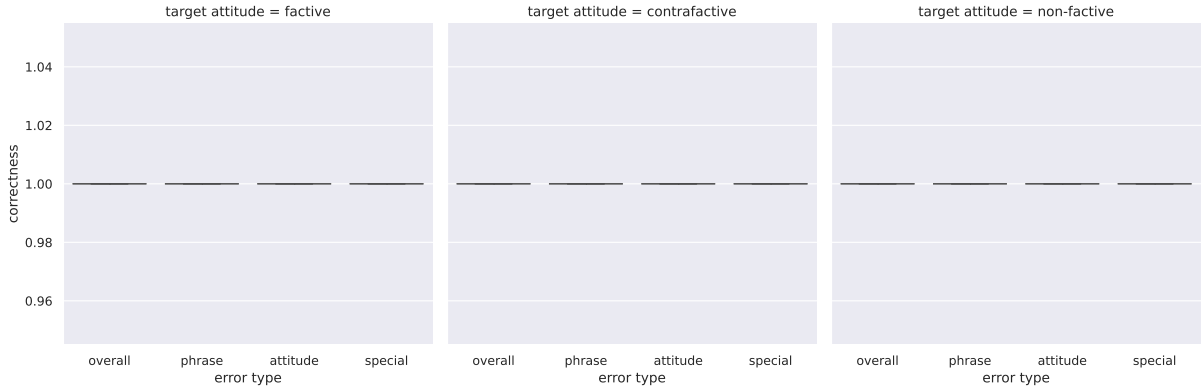
Figure 6: Changes over time in attitude verb preference across target distributions.

I Error Types

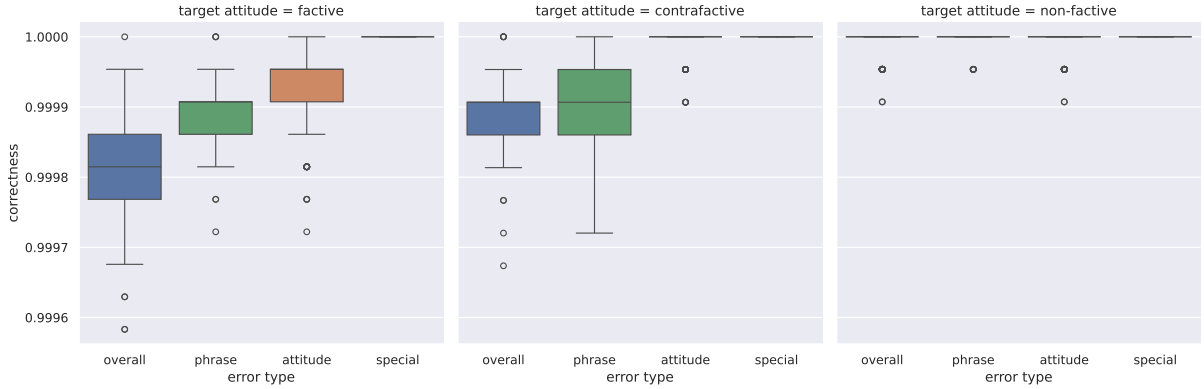
We can distinguish three parts of an output that can be responsible for an error: first, the tokens of the embedded clause; second, the token for the attitude verb; or, third, the special tokens. Because the vocabulary is fixed for each token position special token errors are architecturally impossible. That we find no such errors merely reflects the absence of coding errors. Notably, the worst performance on input that allows contrafactives, found in distribution equal/balanced, is primarily due to errors in the production of embedded clause tokens.



(a) Error types for allowed attitude verbs given the distribution equal/balanced.

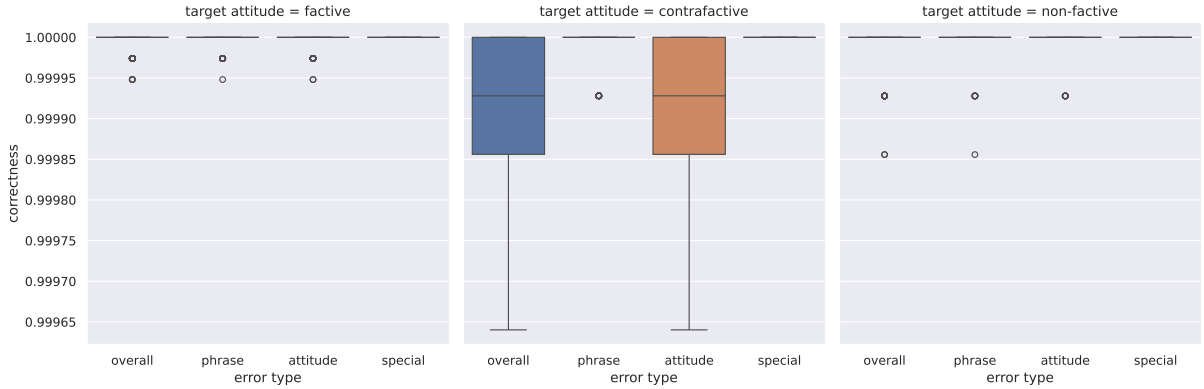


(b) Error types for allowed attitude verbs given the distribution equal/t-medium.

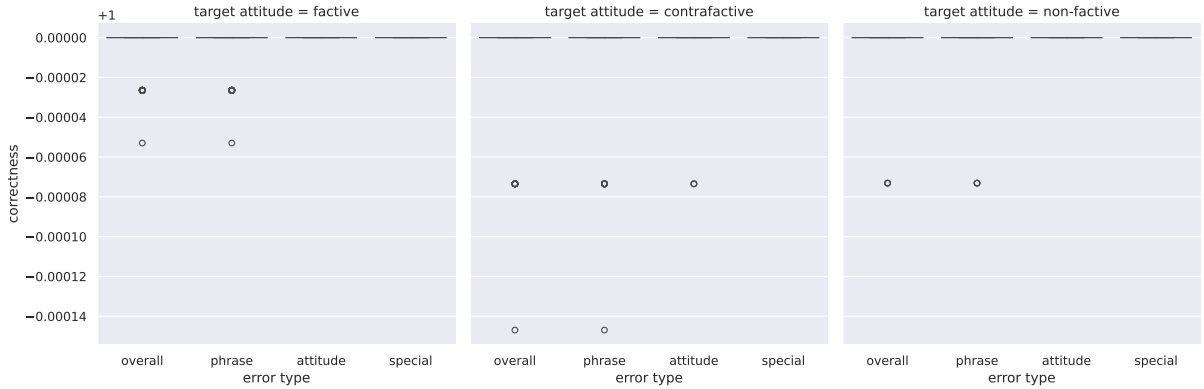


(c) Error types for allowed attitude verbs given the distribution equal/t-heavy.

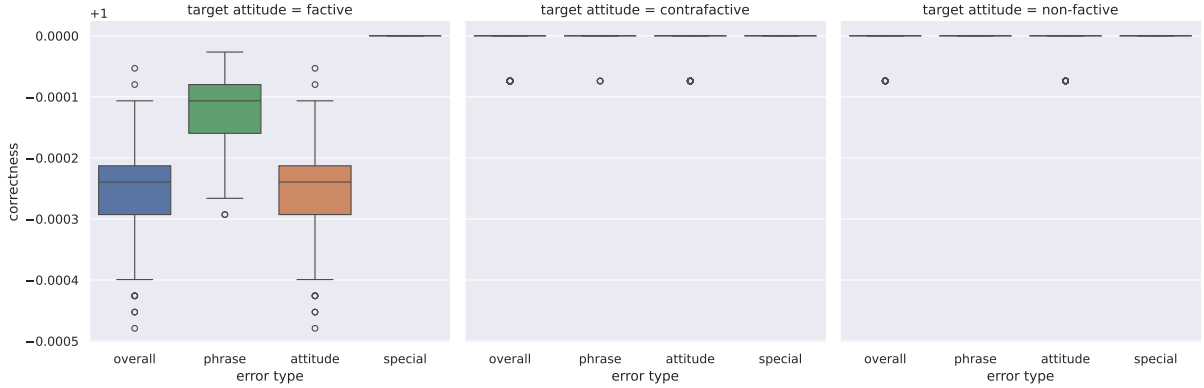
Figure 7: Error types by attitude verb given distributions that are equal.



(a) Error types for allowed attitude verbs given the distribution c-light/balanced.

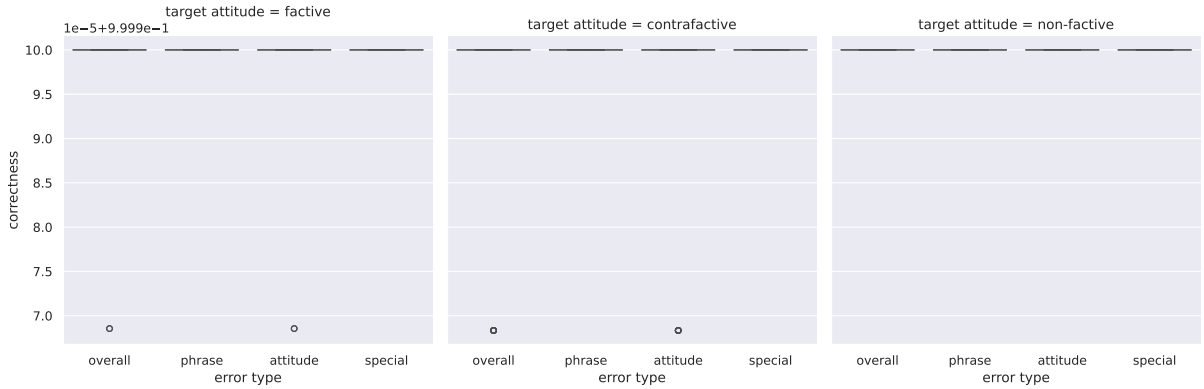


(b) Error types for allowed attitude verbs given the distribution c-light/t-medium.

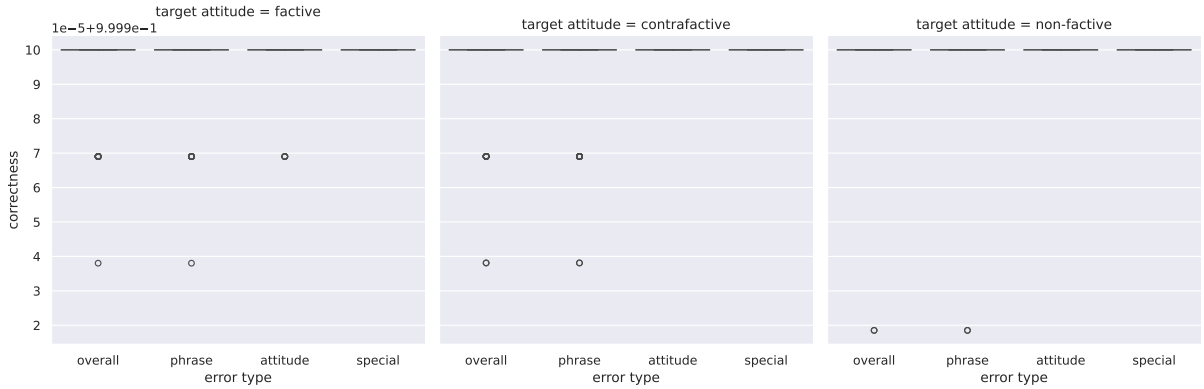


(c) Error types for allowed attitude verbs given the distribution c-light/t-heavy.

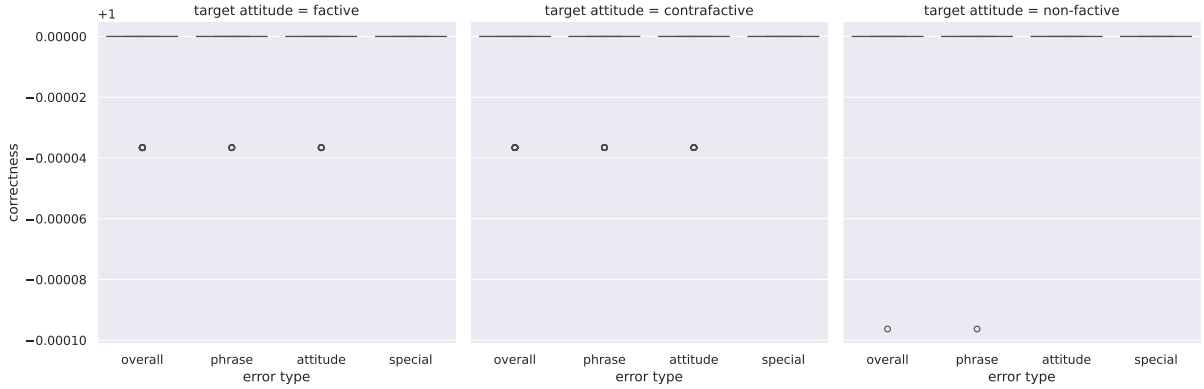
Figure 8: Error types by attitude verb given distributions that are c-light.



(a) Error types for allowed attitude verbs given the distribution c-heavy/balanced.



(b) Error types for allowed attitude verbs given the distribution c-heavy/t-medium.

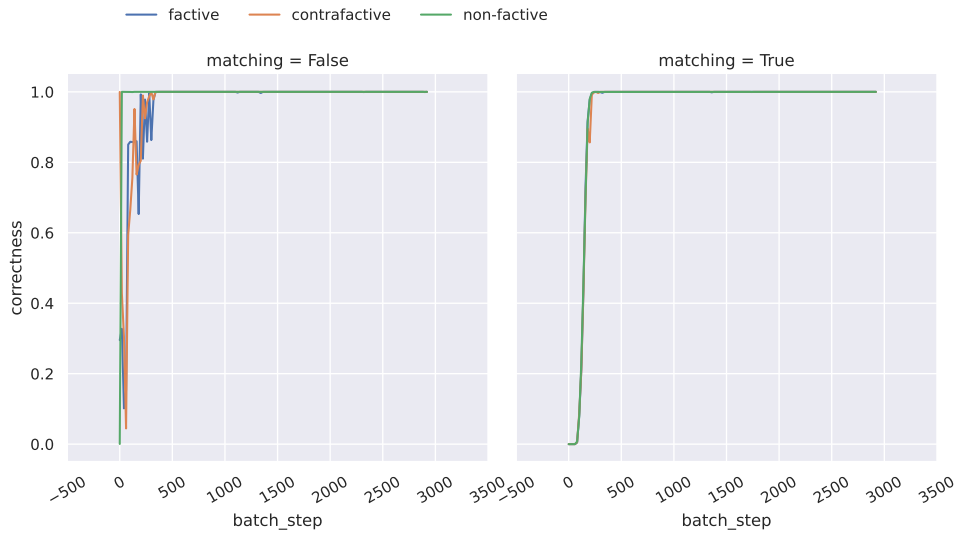


(c) Error types for allowed attitude verbs given the distribution c-heavy/t-heavy.

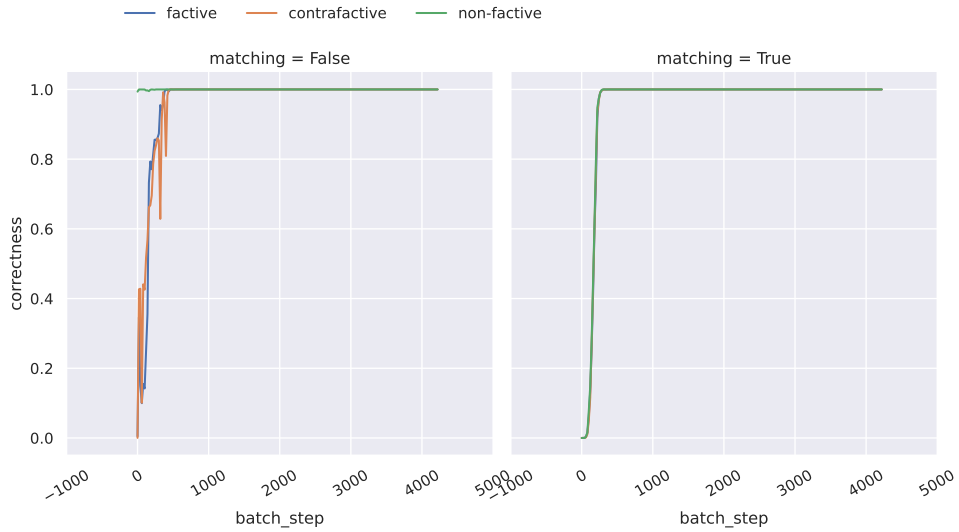
Figure 9: Error types by attitude verb given distributions that are c-heavy.

J Matching vs. Non-Matching

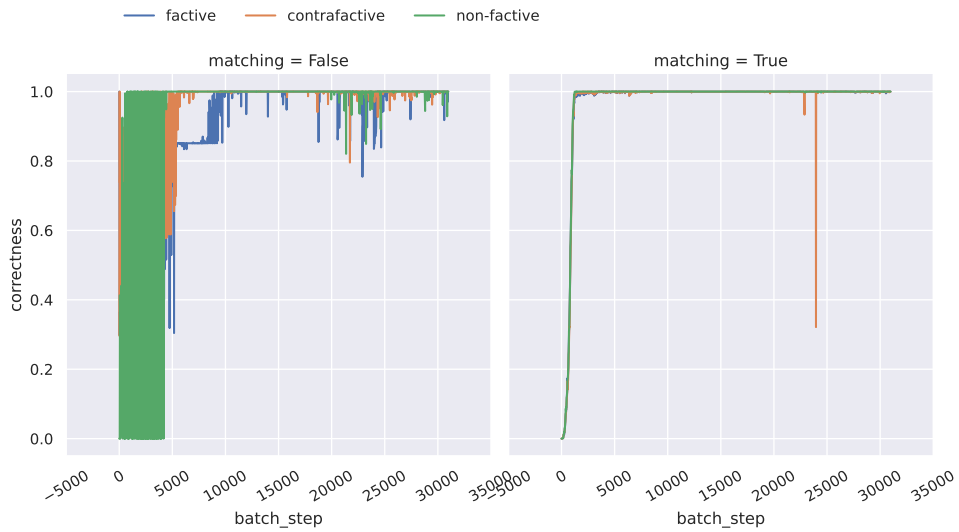
Strohmaier and Wimmer (2025) reported strong differences in performance between conditions that require matching embedded clauses and conditions that require non-matching embedded clauses (see Table 3 to see into which group each of the 21 conditions falls). We provide here the learning graphs split up by what kind of embedded clause is required for all 9 distributions. Our results replicate those of Strohmaier and Wimmer (2025) across distributions.



(a) Matching vs. non-matching performance for allowed attitude verbs given the distribution equal/balanced.

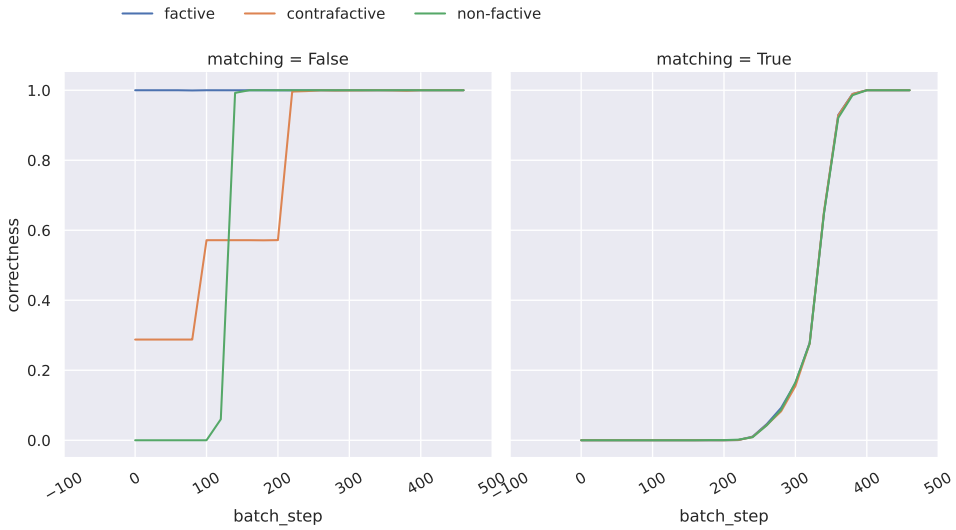


(b) Matching vs. non-matching performance for allowed attitude verbs given the distribution equal/t-medium.

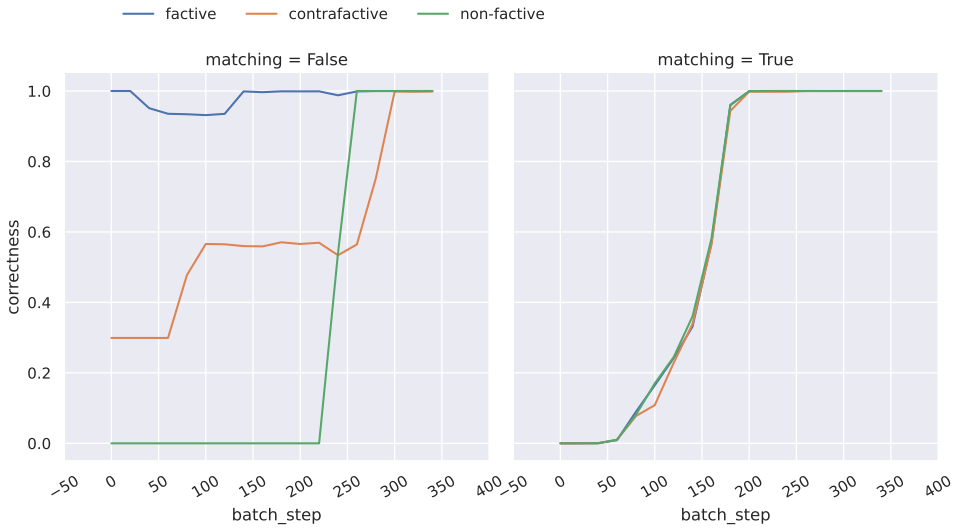


(c) Matching vs. non-matching performance for allowed attitude verbs given the distribution equal/t-heavy.

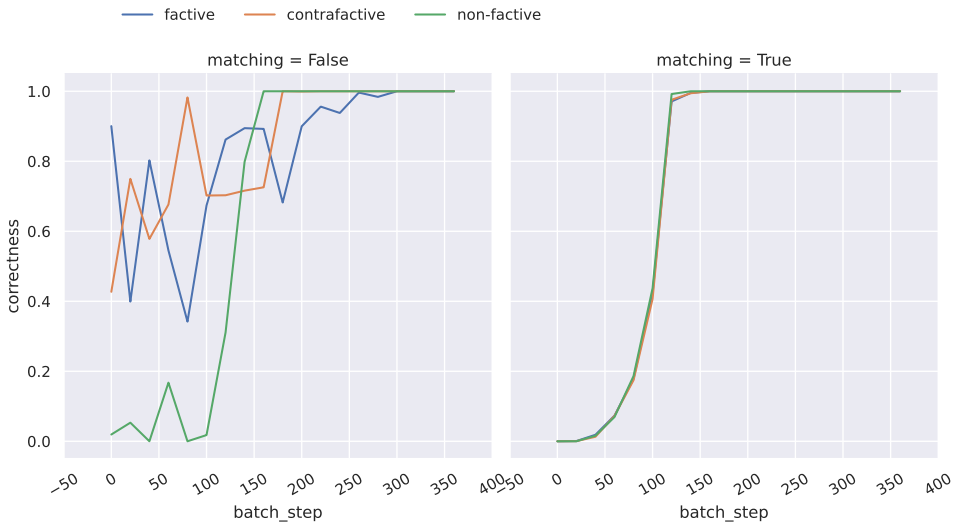
Figure 10: Matching vs. non-matching performance by attitude verb given distributions that are equal.



(a) Matching vs. non-matching performance for allowed attitude verbs given the distribution c-light/balanced.

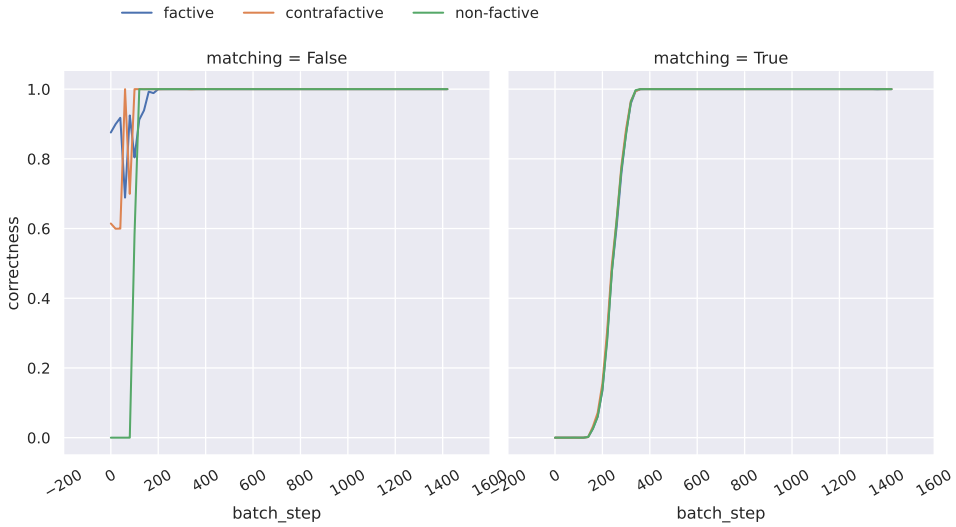


(b) Matching vs. non-matching performance for allowed attitude verbs given the distribution c-light/t-medium.

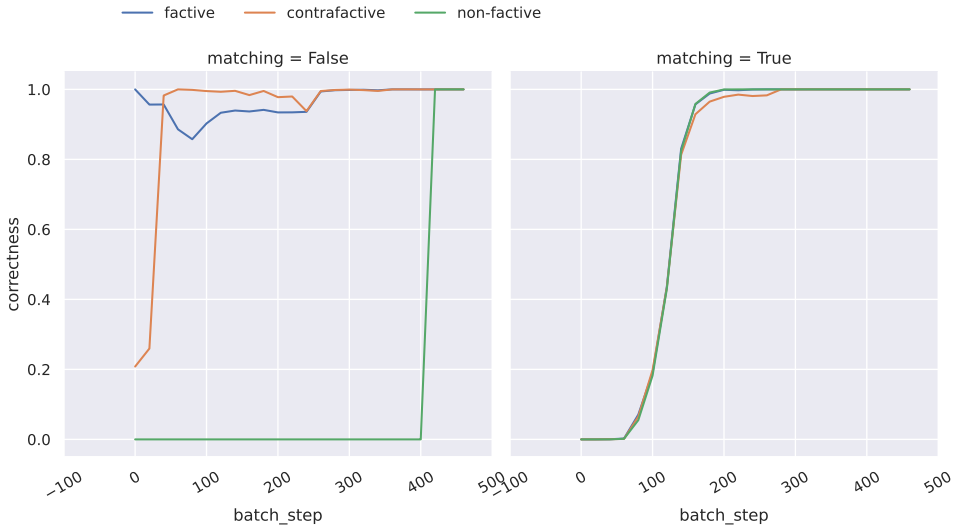


(c) Matching vs. non-matching performance for allowed attitude verbs given the distribution c-light/t-heavy.

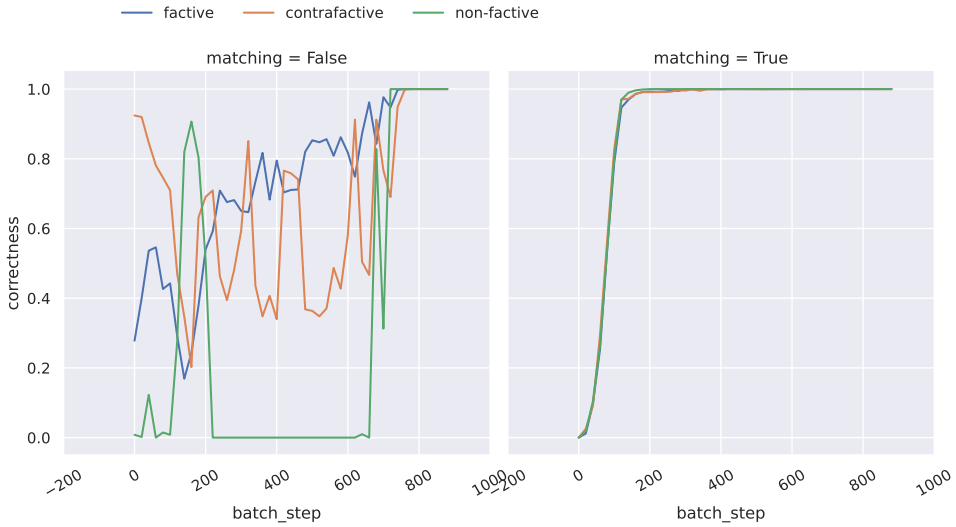
Figure 11: Matching vs. non-matching performance by attitude verb given distributions that are c-light.



(a) Matching vs. non-matching performance for allowed attitude verbs given the distribution c-heavy/balanced.



(b) Matching vs. non-matching performance for allowed attitude verbs given the distribution c-heavy/t-medium.



(c) Matching vs. non-matching performance for allowed attitude verbs given the distribution c-heavy/t-heavy.

Figure 12: Matching vs. non-matching performance by attitude verb given distributions that are c-heavy.