Pragmatic and Logical Inferences in NLI Systems: The Case of Conjunction Buttressing

Paolo Pedinotti
University of Pisa
pedinotti.paolo@gmail.com

Emmanuele Chersoni
The Hong Kong Polytechnic University
emmanuelechersoni@gmail.com

Enrico Santus
Bayer Pharmaceuticals
esantus@gmail.com

Alessandro Lenci
University of Pisa
alessandro.lenci@unipi.it

Abstract

An intelligent system is expected to perform reasonable inferences, accounting for both the literal meaning of a word and the meanings a word can acquire in different contexts. A specific kind of inference concerns the connective *and*, which in some cases gives rise to a temporal succession or causal interpretation in contrast with the logic, commutative one (Levinson, 2000). In this work, we investigate the phenomenon by creating a new dataset for evaluating the interpretation of *and* by NLI systems, which we use to test three Transformer-based models. Our results show that all systems generalize patterns that are consistent with both the logical and the pragmatic interpretation, perform inferences that are inconsistent with each other, and show clear divergences with both theoretical accounts and humans' behavior.

1 Introduction

Implicature is the term used in semantics and pragmatics to describe an inference that goes beyond the literal sense of what is said. Implicatures have received relatively limited attention in computational linguistics, since they are highly dependent on the communication context and on commonsense knowledge. However, the notion of **Generalized Conversational Implicature (GCI)** (Grice, 1975) captures the fact that some of these meaning enrichments are more general than others: They are still dependent on context, but they are also strongly conventionalized and they act as *default inferences*, which are carried out *unless* canceled by additional contextual information.

With this study, we aim at contributing to the research on GCIs in NLP systems by focusing on a specific type of GCI, namely Levinson's **i-implicatures** associated with the conjunction *and* (Levinson, 2000). Studies have noted that *and* is regularly interpreted as a temporal succession or causal connective (from *John repaired the engine and the car started* we understand that the car started as a result of John repairing the engine) (Carston, 1988). This implicature, which is referred to as **conjunction buttressing** by Levinson (2000), contradicts the *commutative* interpretation of *and* traditionally assumed in formal logic and semantics: If *A and B* entails *B after A*, *A and B* is not equivalent to *B and A*. Moreover, the implicature takes place only when the conjuncts express dynamic events, while with static ones *and* preserves the commutative property (e.g., *John was awake and the dog slept* entails *The dog slept and John was awake*).

To address the problem of the scarcity of data for the study of GCIs and conjunction buttressing in particular, we created a dataset for the study of the interpretation of *and* by NLI systems, using manual annotation to obtain quality data and control for features relevant for the implicature according to theoretical accounts. We assigned two different label sets based on a *pragmatic hypothesis* (*and* triggers the implicature) and a *logic* one (*and* is commutative), to distinguish logical vs. pragmatic behavior of the systems.

We tested three Transformer-based NLI systems fine-tuned on MNLI (Williams et al., 2018) on our dataset. We identified systematic inference patterns involving the interpretation of *and* that are common to all three systems. Some of these patterns are in accordance with the pragmatic hypothesis and others with the logic one. We found that the systems make inferences that are inconsistent with each other, and in many cases their interpretation of *and* is different from both the human interpretation and theoretical accounts. To see whether the results are due to biases in the systems' training set, we ran an analysis of MNLI aimed at identifying inference patterns involving *and* that are used by annotators, finding that the inferences generalized by systems are exemplified to varying degrees.

After describing related work in Section 2, in Section 3 we describe how we collected data to

Proceedings of the Second Workshop on Understanding Implicit and Underspecified Language, pages 8 - 16

assess logical and pragmatic interpretations in NLI systems.[1] Results of the experiments with NLI systems are illustrated in Section 4, along with the analysis on MNLI and the results of a human behavioral study. Conclusions are devoted to suggestions for future work and to the discussion of the limitations of the present work. By highlighting limitations of current systems on our dataset, we argue for a stronger convergence of neural systems for inference and cognitive models of GCIs.

2 Related work

Previous NLP studies on implicatures mostly focused on *scalar implicatures*, inferences involving sets of words that together form a lexical scale (e.g., <*all*, *some*>). The use of an alternate excludes the other from the interpretation (e.g., *Some of the boys came* +> (implicates) *Not all of the boys came*).

Jeretic et al. (2020) created a large scale dataset of automatically generated sentences following the NLI format, where a premise-hypothesis pair is labeled according to a *logical annotation* (following the logical, literal meaning) and a *pragmatic annotation* (following scalar implicature). The authors measured the accuracy of a BERT model (Devlin et al., 2019) fine-tuned on MNLI according to the logical and the pragmatic annotation. The authors showed that BERT reasoning is more pragmatic than logical for the sentences involving *all* and *some*, even if the results vary depending on how the premise and the hypothesis are built.

Scalar implicatures are not the only type of generalized implicatures. Levinson (2000) proposed a categorization of GCIs based on underlying inferential heuristics related to Grice's maxims of conversation (Grice, 1975). He considered scalar implicatures as an instance of Q-implicatures, a category of GCIs motivated by the principle *Select the informationally strongest paradigmatic alternate that is consistent with the facts*. They are distinguished from **I-implicatures**, motivated by the principle *Assume the richest temporal, causal and referential connections between described situations or events, consistent with what is taken for granted*. A phenomenon in the latter group involves the enrichment of the meaning of *and* (the so-called **conjunction buttressing**): *John repaired the engine* ***and*** *the car starts* implicates ***After*** *John repaired the engine, the car started* (from logical conjunction to temporal succession) and *The car started* ***because*** *John repaired the engine* (from logical conjunction to cause). The inferred meaning of *and* contrasts with the commutative meaning attributed to it in logic and formal semantics.

To our knowledge, Pandia et al. (2021) is the only NLP study dealing with conjuction buttressing: the authors tested if Transformer-based masked language models can predict the temporal connective corresponding to the correct interpretation of the enriched *and*, using the stimuli by Politzer-Ahles et al. (2017). Unlike their study, we created and used labeled data for the evaluation of NLI systems, testing a pragmatic hypothesis (enriched interpretation of *and*) vs. a logical one (commutative interpretation).

3 Data

Given the scarcity of existing resources for GCIs, we collected and annotated new data in NLI format, focusing on different interpretations of the connective *and*. We assigned two different sets of labels, one in accordance with the pragmatic hypothesis (i.e., the implicature is labeled as an entailment) and the other with the logic hypothesis (i.e., only logical inferences are treated as entailments).

Methodology. To obtain data to test the **temporal succession** and the **causal** interpretation of *and*, we first used a multigenre English corpus (UkWac, Ferraresi et al. (2008)) to extract sentences where a main and a subordinate clause are explicitly encoded in a temporal succession or causal relation by a connective (e.g., *Frazier quit* ***before*** *I did*).[2] Then, we replaced the original connective with *and* (*Frazier quit* ***and*** *I did*). The generated and the original sentences are, respectively, the premises and the hypotheses of our experiment (see the first two rows of Table 1). Because the implicature only takes place when two clauses describe events that are presented as a *dynamic process* (Levinson, 2000) (i.e., an event is described as a dynamic situation when it is a process with subparts, such as in *Frazier quit and I did* which implicates succession while *I have two sons and Mary has three* does not), we further manually refined the set to include only those instances. According to the pragmatic hypothesis, the systems should assign the entailment label to these pairs. According to the logical hypothesis, the label is neutral since a literal interpretation of *and* does not entail a temporal

[1] The dataset can be found in the supplementary materials and we will make it available for free use.

[2] See Appendix A for more details about data collection.

Interpretation of *and*	Premise	Hypothesis	Logical label	Pragmatic label
Temporal succession	A and B	B after A	N	E
Causal	A and B	B because A	N	E
Temporal precedence	A and B	B before A	N	C
Temporal synchronous	A and B	B while A	N	C
Commutative (dynamic)	A (dynamic) and B (dynamic)	B (dynamic) and A (dynamic)	E	C
Commutative (static)	A (static) and B (static)	B (static) and A (static)	E	C

Table 1: Dataset structure.

succession or causal relation between events.

From the premises used to test causal interpretation (e.g., *He refused to sign and he lost his job*) we produced new hypotheses where the clauses are linked by other temporal relations contradicting succession, namely **precedence** (***Before*** *he refused to sign, he lost his job*) and **synchronous** (***While*** *he refused to sign, he lost his job*). This is to ensure that systems do not perform an enriched interpretation of *and* that goes in the wrong direction (either temporal or causal). Since the pragmatic interpretation of *and* is temporal succession and this excludes a precedence or synchronous one, we assigned the gold pragmatic label contradiction to these pairs.

We also wanted to test whether NLI systems assign a logical interpretation to the connective *and*, namely **commutativity**. Here we studied the influence of the semantics of the conjuncts: While commutativity is a more natural inference with conjuncts describing static situations (*The rooms are comfortable and the food is super* entails *The food is super and the rooms are comfortable*), with conjuncts describing dynamic situations it is less natural, since it is overridden by the inference stemming from pragmatic enrichment (*He fell off a ladder and he had concussion* contradicts *He had concussion and he fell off a ladder*). To obtain instances of inferences involving the commutativity of *and* with **dynamic conjuncts**, we used the sentences with a causal relation from our dataset. For instance, from the sentence *He had concussion because he fell off a ladder* we generated the premise *He fell off a ladder and he had concussion* and the hypothesis *He had concussion and he fell off a ladder*. For **static conjuncts**, we manually annotated clause pairs linked by *and* in UkWac, and selected only pairs where the main verb of both clauses is stative (*The food is super and the rooms are comfortable*) or has an habitual reading (*Platypus builds nest, and echidna develops pouch*). While commutativity is entailed from the logic perspective, a contradiction would be produced if a pragmatic interpretation of *and* was selected, since temporal succession is not a commutative relation.

	Logical label	Pragmatic label
Temporal succession	0.02	0.94
Causal	0.02	0.98
Temporal precedence	0.07	0.51
Temporal synchronous	0	0
Commutative (dynamic)	1	0
Commutative (static)	1	0

Table 2: Accuracy of the DeBERTa-based system (He et al., 2021) according to the logic and pragmatic label.

Statistics. We collected 653 premise-hypotheses pairs for testing temporal succession interpretation, 270 for testing commutativity (static conjuncts) and 623 for each of causal, precedence, synchronous and commutativity (dynamic), ending up with a total of 3,470 instances.

4 Experiments

Systems. We used our data to evaluate a BERT (Devlin et al., 2019), a RoBERTa (Liu et al., 2019) and a DeBERTa (He et al., 2021) language model fine-tuned on MNLI. For BERT and RoBERTa, we adopted the fine-tuned versions by Poth et al. (2021).[3] We did not perform additional training, as our goal is to test existing systems and our dataset has been built only for evaluation purposes.

Results. We report in Table 2 the results for DeBERTa only, as they are the best ones and there are just slight variations across systems.[4] With *logical* and *pragmatic accuracy*, we refer to accuracy on labels following from the logical and the pragmatic hypotheses respectively.

Results show: a) Pragmatic accuracy close to 1 for Temporal succession and Causal (systems generalize the pattern *A and B* entails *B after A* and *B because A*), but logical accuracy 1 for commutative (systems generalize *A and B* entails *B and A* inde-

[3] See Appendix C for more details about the systems.

[4] Results for all systems can be found in Appendix D

pendent of the semantics of conjuncts); b) Accuracies 0 for temporal synchronous (systems generalize *A and B* entails *B while A*), c) divergent behavior of systems on examples involving a temporal precedence interpretation of *and* (RoBERTa-based: *and* nearly always entails a temporal precedence interpretation; BERT-based: *and* entails a temporal precedence interpretation in 74% of the cases and contradicts it in only 7%; DeBERTa-based: *and* entails temporal precedence in 42% of the cases, and contradicts it in 51%).

Results analysis. We first observe that the inferences drawn by the systems show inconsistent patterns. In many cases the systems assign a succession, precedence and synchronous interpretation to the same pair of conjuncts, which is an overt contradiction. Second, the systems' behavior is not aligned with theoretical accounts of implicatures. Linguistic theory predicts that only a limited set of relations between conjuncts can be inferred (among which succession and cause), while systems consider all the relations we tested as valid inferences. Moreover, while the dynamic event type of the conjuncts is expected to lead to the rejection of the commutative interpretation in favor of an enriched one, systems prefer the commutative pattern irrespective of the context.

MNLI analysis. To see whether results can be explained by biases in the dataset used for training of the systems, we performed an analysis of the MNLI training set aimed at identifying and quantifying inference patterns involving the connective *and* that are used by annotators. To identify examples of pragmatic inference patterns involving the connective *and*, we selected instances where the premise or the hypothesis contains two main clauses linked by *and* using the SpaCy dependency parser (Honnibal and Montani, 2017). We manually inspected 500 out of the 11,208 obtained pairs for cases where the gold label can be explained by assuming the triggering of a pragmatic inference. We found those patterns to be used by MNLI annotators: 26 cases can be explained by assuming an enriched interpretation of *and*. Temporal succession is the most frequent interpretation with 20 cases. Synchronous, causal and inclusion are less present with 3, 1 and 1 cases respectively (see the Appendix B for examples). We found the logic, commutative interpretation of *and* to be much less used for inference by MNLI annotators than the pragmatic one. Out of the 500 examples we analyzed, only 2 can be explained by assuming a commutative interpretation of *and* by annotators (see Appendix B). This analysis shows that inference patterns generalized by systems are exemplified to varying degrees in the training set.

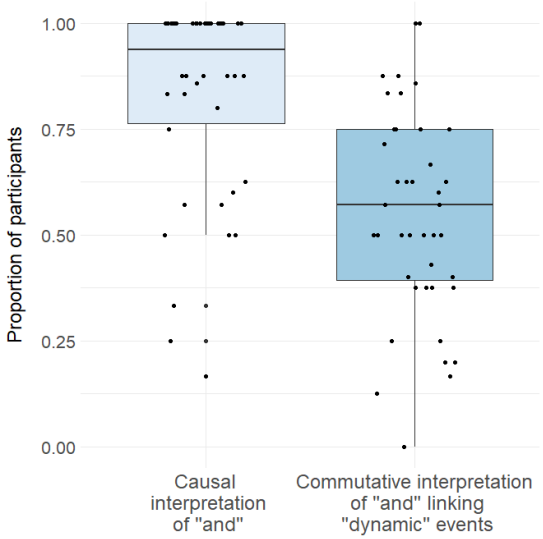


Figure 1: Human behavioral study. The y-axis reports, for each pair, the proportion of participants performing the interpretation on the x-axis.

Human behavioral study. The dataset annotation is based on linguistic theory and expert annotation. To compare it with actual intuitions people have about the meaning of the sentences, we performed a behavioral study using a small subset of premise-hypothesis pairs from the dataset.

Details of the study are given in E. For each of 40 pairs of type Causal, we asked 8 participants to judge if a speaker is implying "B because A" by saying "A and B" or not (that is, we tested if they assign a causal interpretation to *and*). For each of 43 pairs of type Commutativity (dynamic) we asked to judge whether, given a situation where a speaker uses the sentence of form "A and B" and another speaker uses the form "B and A" to describe the same fact, it is possible that both sentences are true at the same time (that is, we tested if a logical, commutative interpretation is assigned to *and*).

The left box in Figure 1 involves pairs of type Causal and shows, for each pair, the proportion of participants assigning a causal interpretation to *and*.[5] If judgments were in perfect agreement with our pragmatic labels, proportion should be 1 for all pairs (0 for logical). In the majority of cases (31 out

[5]The sentences used for the experimental study along with the proportion of participants choosing each answer are provided as a separate file in the supplementary material.

of 40) the proportion is equal to or higher than 0.8. This shows that, in most cases, the responses of almost all participants are in line with our previous annotations. In other cases, there is less support for the causal interpretation, and in a few cases the majority of participants reject it (e.g., *I went to a mass meeting one night and that happened* +> *That happened because I went to a mass meeting one night*, proportion of "Yes": 0.166). We attribute this result to a) Our expert annotation being open to challenge, and b) Limitations of Levinson's theory (possibly there are other factors affecting the pragmatic inference in addition to the situation type of the conjuncts, for example more stereotypical event sequences).

The right box involves pairs of type Commutativity (dynamic) and shows, for each pair, the proportion of participants considering the forms "A and B" and "B and A" true at the same time. If judgments were in perfect agreement with our pragmatic labels, proportion should be 0 for all pairs (1 for logical). Generally, questions receive more variable answers than in the previous group, which can be due to the survey questions being less clear than in the previous case (see E for the form of questions). In some cases, the majority of participants converge on the "Yes" (e.g., *People found them practical and they came into use* and *They came into use and people found them practical* are both true of the same situation according to 85.7% of participants) or the "No" (e.g., *I won an award at 16 for my poetry and I went to Russia* and *I went to Russia and I won an award at 16 for my poetry* are both true of the same situation according to 0% of participants) answer. We argue that answers are determined by the triggering of pragmatic inference (if the inference takes place, the two sentences are not considered true at the same time). The inference takes place differently across our set of pairs, possibly for the reasons we outlined in the paragraph above.

With this experiment, we have explored the distance between our dataset annotation and actual human intuitions about the interpretation of *and*, along with identifying interpretation tendencies.

Confidence scores. To get a more accurate evaluation of the systems and compare their output with human behavioral data, we analyzed the confidence scores of the label entailment for the pairs used for the behavioral study. We found that all systems' scores are concentrated in a small interval near 1 (BERT: [.945, .994], RoBERTa: [.936, .996], DeBERTa: [.950, .999], except for an outlier with score 0.558). The tendency to consistently assign high scores to the entailment label is confirmed by the mean $\overline{x}_n$ and the variance s_n^2 of the samples containing confidence scores of entailment in the whole dataset (BERT: $\overline{x}_n$=.775, s_n^2=.106; RoBERTa: $\overline{x}_n$=.851, s_n^2=.086; DeBERTa: $\overline{x}_n$=.814, s_n^2=.105).

The visualization of the relation between the systems' confidence score of entailment for a given pair and the frequency with which participants consider that pair an example of entailment (given in F) shows no positive correlation. We take the results of this analysis as evidence of a divergence between systems (who consistently choose the entailment label) and humans (who choose entailment label with different frequency across the dataset, showing a variability that does not correlate with the limited variability in the systems' output).

5 Conclusion

We found that NLI systems generalize "pragmatic" and "logical" inference patterns involving the connective *and*. This gives rise to unsatisfactory predictions, since in many cases inferences are not consistent with each other and are not aligned with human ones and theoretical accounts of implicatures. It should be noted that alternative accounts of implicatures exist: For scalar implicatures it has been shown that inference takes place with different strength depending on the context (Degen, 2015). A better assessment of the systems' abilities could be obtained by using implicature strength data. Finally, at this stage we cannot draw general conclusions about whether our results also extend to systems trained on other NLI datasets.

Based on the highlighted limitations of the tested systems, we argue for the need of a stronger convergence of neural systems with theories of GCIs to improve systems' interpretation of *and*. Levinson (2000) proposed that I-implicatures can be explained by assuming that the hearer knows that the speaker tried to achieve her communicative goals by maximizing economy, and thus enriches the interpretation in *stereotypical* ways (since it assumes that the speaker has left stereotypical information unsaid). Stereotypical relations between events in the form of event chains could be automatically collected from texts (Chambers and Jurafsky, 2008) and provided as additional information to systems.

References

Robyn Carston. 1988. Implicature, Explicature, and Truth-Theoretic Semantics. *Mental Representations: The Interface between Language and Reality*, pages 155–181.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised Learning of Narrative Event Chains. In *Proceedings of ACL-HLT*.

Judith Degen. 2015. Investigating the Distribution of Some (But Not All) Implicatures Using Corpora and Web-based Methods. *Semantics & Pragmatics*, 8(11):1–55.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*.

Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. 2008. Introducing and Evaluating UkWaC, a Very Large Web-derived Corpus of English. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4)*, page 47–54.

Herbert Paul Grice. 1975. Logic and Conversation. In *Speech Acts*, pages 41–58. Brill.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. In *International Conference on Learning Representations*.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing. *To appear*, 7(1):411–420.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are Natural Language Inference Models IMPPRESsive? Learning IMPlicature and PRESupposition. In *Proceedings of ACL*.

Stephen C. Levinson. 2000. *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. The MIT Press, Cambridge, MA.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Lalchand Pandia, Yan Cong, and Allyson Ettinger. 2021. Pragmatic Competence of Pre-trained Language Models through the Lens of Discourse Connectives. In *Proceedings of CONLL*.

Stephen Politzer-Ahles, Ming Xiang, and Diogo Almeida. 2017. " Before" and" After": Investigating the Relationship between Temporal Connectives and Chronological Ordering Using Event-related Potentials. *PloS One*, 12(4).

Clifton Poth, Jonas Pfeiffer, Andreas Rücklé, and Iryna Gurevych. 2021. What to Pre-Train on? Efficient Intermediate Task Selection. In *Proceedings of EMNLP*.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of LREC*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of NAACL-HLT*.

A Details about Data Collection

Using the SpaCy dependency parser (`https://spacy.io/`), we extracted sentences from UkWac (Ferraresi et al., 2008) matching the dependency pattern CONNECTIVE-*mark*-V-*advcl*-V-*ROOT*, where CONNECTIVE is a connective that unambiguously signal the discourse relation of interest (*before*, *after* and *once* for temporal succession, *because* for causal) and V is a verb according to the SpaCy POS tagger. We selected clauses linked by connectives that are unambiguous in term of their discourse function according to the English Penn Discourse Treebank (Prasad et al., 2008).

For our experiments, we used the SpaCy pipeline `en_core_web_s` from the most recent version 3.2. SpaCy is licensed under the MIT license.

UkWac is a large-scale corpus (>2 billion words) created with texts from URLs in the `.uk` web domain. URLs were selected based on the presence of a pair of words, where pairs are from a list created by choosing random medium-frequency words from BNC (written and spoken version) and a vocabulary list for foreign learner of English. This strategy ensures variety of content. As a result, the corpus covers various domains and demographic groups. The prevailing language is British English, but the presence of other variety of English cannot be excluded. The corpus is freely downloadable at `https://wacky.sslmit.unibo.it/`.

B Analysis of MNLI

Examples of pragmatic interpretation of *and*. Temporal succession interpretation: *Thorn turned and left* entails *Thorn left after he turned* (pairID: 17201c). Temporal synchronous interpretation: *The man roared out and cleaved off the demon's other arm* entails *The man made a loud noise as he injured the demon* (pairID: 35017e). Causal interpretation: *After 37 years of rule, Solomon died*

and the kingdom was split between the northern and southern tribes entails *Solomon was the ruler for 37 years and his death resulted in the divide of the kingdom between north and south* (pairID: 56084e). Temporal inclusion interpretation: *we came here and they had parking lots in the schools and i couldn't understand it you know all the kids had cars* entails *I was surprised to see that all the kids had cars when we came here* (pairID: 2744e).

Examples of commutative interpretation of *and*. *Several years ago a radio broke in my car and i never i got out of the habit of listening to the radio* entails *Several years ago a radio broke in my car and i never i got out of the habit of listening to the radio and I always stuck to the habit of listening to the radio, and mine broke* (pairID: 24186e).

C Systems details

The three systems we used for our experiments are Transformer models fine-tuned on MultiNLI (Williams et al., 2018). MNLI was built based on the following procedure. First, text sources of ten different genres (including written and spoken speech) are used to select sentences that are used as premises. Sources are from the Open American National Corpus and a selection of works of contemporary fiction. Then, a crowdworker is asked to produce a hypothesis for each NLI label (entailment, neutral, contradiction). Finally, other crowdworkers are asked to assign a label to each premise-hypothesis pair and a gold label is assigned based on the majority of labels. The corpus comes with a training/test/development split (392,702/ 20,000/ 20,000 examples respectively). MNLI can be freely used and may be modified and redistributed. The corpus is released under several licenses (cf. Williams et al. (2018) for details).

The three systems can be downloaded freely from `https://huggingface.co/` and are `bert-base-uncased-pf-mnli` (Poth et al., 2021), `roberta-base-pf-mnli` (Poth et al., 2021) and `deberta-v2-xlarge-mnli` (He et al., 2021). `deberta-v2-xlarge-mnli` is licensed under the MIT license. Details about the tested systems are provided in Table 3. We refer the reader to the original paper for further details.

D Results for all Systems

E Survey details

Platform. We launched the survey on Prolific Academic (`https://www.prolific.co/`).

	bert-base-uncased-pf-mnli	roberta-base-pf-mnli	deberta-v2-xlarge-mnli
Paper	Poth et al. 2021	Poth et al. 2021	He et al. 2021
Number of parameters (Language Model)	110M	125M	900M
Computational budget		8 × 32GB V100 GPUs	6 × 96GB V100 GPUs
Fine-tuning strategy	Adapter-based	Adapter-based	Scale-invariant
MNLI accuracy	84.2	87.5	91.7

Table 3: Details about the tested systems.

	Logical label	Pragmatic label
Temporal succession	0.02	0.81
Causal	0.03	0.97
Temporal precedence	0.19	0.07
Temporal synchronous	0	0.02
Commutative (dynamic)	1	0
Commutative (static)	1	0

Table 4: Results for the BERT-based system (Poth et al., 2021).

Participation requirements. Participants were required to a) Be born in the U.S., b) Be a U.S. citizen, c) Be in the U.S. at the time of the test, d) Have English as their first language, e) Have an approval rate of previous studies on Prolific between 90% and 100%, f) Have completed at least 50 tests on Prolific. We used Prolific's internal screening system for excluding participants who did not meet the requirements.

Survey structure. Each test consisted of 20 questions. Possible answers for each question in a survey were "Yes" and "No". 5 questions targeted the pragmatic interpretation of the *and* connective, 5 questions targeted the logical (commutative) interpretation, and the other 10 were comprehension question. Each question targeting the **pragmatic** interpretation of *and* has the following structure:

- Imagine that a speaker says PREMISE. In your opinion, is the speaker implying HY-

	Logical label	Pragmatic label
Temporal succession	0.01	0.91
Causal	0.01	0.98
Temporal precedence	0	0.07
Temporal synchronous	0	0
Commutative (dynamic)	0.98	0.02
Commutative (static)	0.98	0.02

Table 5: Results for the RoBERTa-based system (Poth et al., 2021).

POTHESIS?

PREMISE and HYPOTHESIS are examples of type "Causal" from the dataset presented in this article (an example of question is: *Imagine that a speaker says "I got bored in the first year and I dropped out of university". In your opinion, is the speaker implying "I dropped out of university because I got bored in the first year"?*).

Each question targeting the **logical (commutative)** interpretation of *and* has the following structure:

- Imagine that two speakers A and B know the same fact and are telling it. A says PREMISE, and we know she is telling things as they actually happened. Now imagine B says HYPOTHESIS. Is B also telling things as they happened?

PREMISE and HYPOTHESIS are examples of type "Commutativity (dynamic situation)" from the dataset presented in this article (an example of question is: *7. Imagine that two speakers A and B know the same fact and are telling it. A says "IBM used Intel and Intel became standard ", and we know she is telling things as they actually happened. Now imagine B says "Intel became standard and IBM used Intel". Is B also telling things as they happened?*).

Comprehension questions were added to a) Prevent participant from associating questions of a given form with a given answer, b) Mitigate the bias of questions of a given form towards a given answer type (given our previous annotation, we expected questions targeting pragmatic interpretation to have "Yes" as prevailing answer), c) Prompt participants to pay more attention to the meaning of the sentences in the survey, and d) Exclude from the final dataset the answer of participant who are suspected of not comprehending the task or not paying the right attention to the questions.

The comprehension questions have the same form of the other questions, but instead of targeting inference patterns involving the interpretation of *and*, they asked participants to make simple inferences based on other elements of sentences. Examples were inferences based on presuppositions (e.g., *Imagine that a speaker says "Europe tried to sweep itself clean of Jews and it came into existence". In your opinion, is the speaker implying that there were Jews in Europe?*), paraphrases (*Imagine that two speakers A and B know the same fact and are telling it. A says "Phillip adamantly and persistently refused to pay her a penny piece and she succeeded", and we know she is telling things as they actually happened. Now imagine B says "She was not given a penny by Philip and she succeeded". Is B also telling things as they happened?*), contradictions based on negation or antonyms (*Imagine that a speaker says "Christian voice intimidated 1/3 of the venues into dropping out and the tour became financially impossible". In your opinion, is the speaker implying "Christian voice intimidated 1/3 of the venues into dropping out and the tour became financially sustainable"?*).

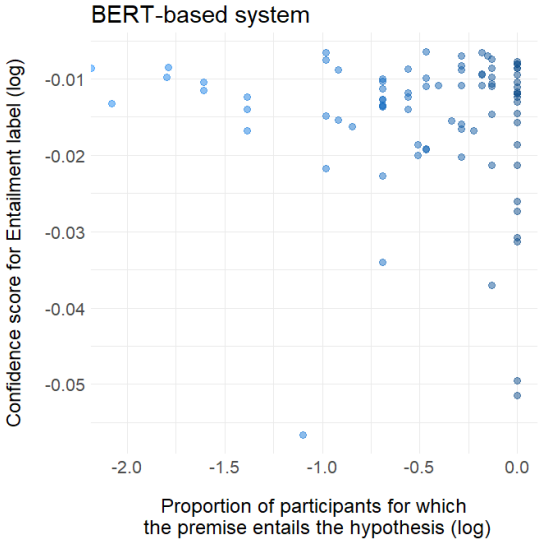
Since they involve straightforward inference patterns and they are not the focus of the experiment, we had gold standard answers for comprehension questions, which we used to exclude answers of participants from the dataset.

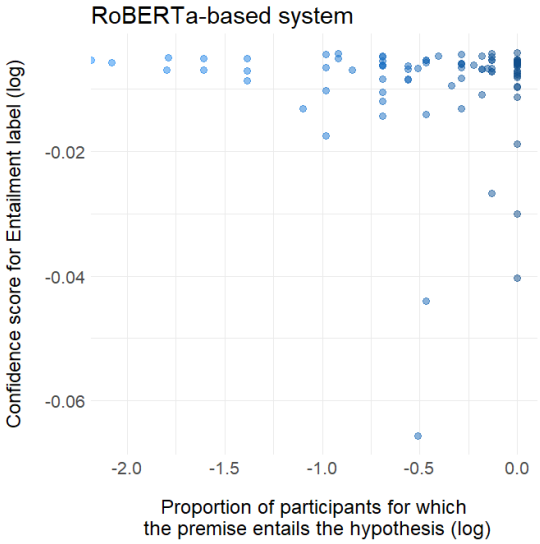
To ensure participants made choices based on their intuitions, no examples were provided in the instructions.

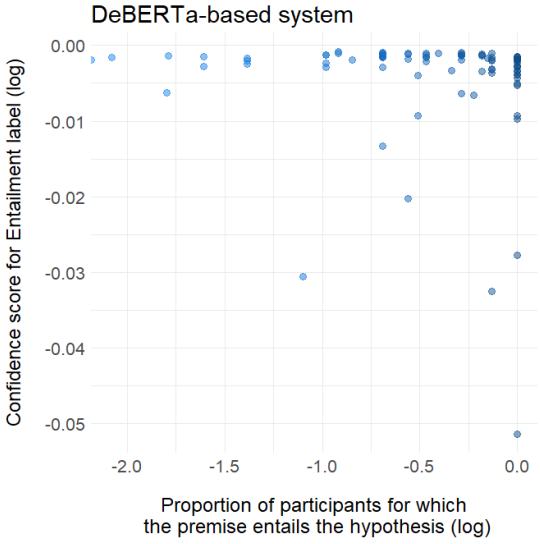
Number of participants and reward. Each survey was presented to 8 participants. 9 surveys were created in total, for a total of 72 participants taking part in the experiment. Participants were not allowed to take part in more than one survey. They received a reward of 0.55£ (0.65€, 0.67$).

Requirements for inclusion in the dataset. In order for a participant answers to be included in the final dataset, the participant must give the gold standard answer to at least 7 of the 10 comprehension question in the survey. This strategy led to the exclusion of the answers of 5 out of 72 participants.

F Systems' confidence scores for sentences from the experimental study


BERT-based system
Confidence score for Entailment label (log)
-0.01
-0.02
-0.03
-0.04
-0.05
-2.0
-1.5
-1.0
-0.5
0.0
Proportion of participants for which
the premise entails the hypothesis (log)


RoBERTa-based system
Confidence score for Entailment label (log)
-0.02
-0.04
-0.06
-2.0
-1.5
-1.0
-0.5
0.0
Proportion of participants for which
the premise entails the hypothesis (log)


DeBERTa-based system
Confidence score for Entailment label (log)
0.00
-0.01
-0.02
-0.03
-0.04
-0.05
-2.0
-1.5
-1.0
-0.5
0.0
Proportion of participants for which
the premise entails the hypothesis (log)