Language comprehension reveals natural logical ability

Maksymilian Dąbkowski,[1*] Alyssa Loo[2,3], Ellie Pavlick[2,3], Roman Feiman[2,4]

[1]Department of Linguistics,
University of California, Berkeley, CA 94720, USA
[2]Program in Linguistics,
Brown University, Providence, RI 02912, USA
[3]Department of Computer Science,
Brown University, Providence, RI 02912, USA
[4]Department of Cognitive, Linguistic, and Psychological Sciences,
Brown University, Providence, RI 02912, USA

[*]To whom correspondence should be addressed;
E-mail: maksymilian.michal.dabkowski@gmail.com.

Classification: Social Sciences/Psychological and Cognitive Sciences;
secondary: Physical Sciences/Computer Sciences

Keywords: natural logic, dual process, reasoning, self-paced reading, next-word prediction, entailment, semantics, quantification, large language models

Abstract: Does logic only dictate how people ought to reason or can it also describe how they actually do? Research documenting pervasive reasoning errors has been taken to show that people are not naturally logical. However, previous studies did not examine the kinds of logic inherent in the formal structure of language. We hypothesized that, if ordinary language comprehension requires computing certain logical inferences, understanding a narrative would be disrupted just when those inferences are violated. Three self-paced reading experiments tested the signatures of these computations. A fourth tested GPT-2 and 3 as models of language based only on probabilistic prediction, without built-in logical structure. We find that humans, but not GPTs, spontaneously made fast, correct inferences, suggesting that natural logic constitutes part of human thought.

Significance statement: Logic characterizes how to draw reliable conclusions from existing beliefs, but people often fail to reason logically. According to the current majoritarian position, logical principles are only normative—they tell people how to reason, but do not successfully capture the way people naturally do. Using a novel experimental design, we demonstrate that people's understanding of a written story is disrupted when logical inferences are violated, even when no attention is drawn to the logical structure of the story. This suggests participants are making unprompted, logically valid inferences in the course of natural language comprehension. Our findings upend the majoritarian position, providing strong evidence that fast, accurate logical reasoning is a natural ability.

Main text: People form new beliefs not only from new evidence, but also by reasoning about what they already know. Logic characterizes how people should reason to guarantee that their conclusions are as good as their premises, but to the chagrin of logicians since Aristotle (*1*), people often fail to reason logically. The gap between how people should reason and how they actually do has fueled a debate about the role of logic in thought that has lasted from antiquity to

the present, dividing scholars in philosophy (*2–6*), psychology (*7–14*), and linguistics (*15–18*). One side argues that logical principles are only *normative laws*, akin to a legal code (*4, 14*): People can recognize that these principles are useful and learn to follow them, but they are at best only rough approximations of how people naturally think. The other side argues that at least some logical principles really are *natural laws* that characterize how thought works (*2, 8*): People deviate from them only in the way a falling object deviates from the gravitational constant—due to interference from other factors, forces, and laws.

Today, opposing views are dominant in different fields. In psychology, decades of evidence documenting pervasive errors have been taken to favor the normative view. Competing psychological theories of reasoning disagree about what causes particular errors, but they all assume that in those cases where people err, reasoning must not be naturally governed by logic (*9,11,13, 19–24*). Because different psychological theories predict different errors, their hypotheses are tested by generating and comparing error patterns on non-trivial reasoning tasks. These typically use puzzles (*23,25*), exam-style questions (given some information, "what follows?") (*9,10,22*), or more ordinary scenarios that are then accompanied by framing or subtle cues deliberately designed by experimenters to increase the appeal of wrong answers (*14, 19*). For instance, consider tests of whether people naturally reason according to Aristotle's *dictum de omni et nullo*. This is a valid logical schema, which states that if a property can be affirmed or denied of a kind, it can be correspondingly affirmed or denied of any subkind. Both children (*26–28*) and adults (*12, 29–31*) systematically diverge from this logical standard. For example, when adults are told that "all birds have sesamoid bones," they do not follow the *dictum* evenly across all subkinds. They judge more typical birds, such as robins, as more likely to have sesamoid bones than less typical birds, such as penguins, even as they agree that both penguins and robins are indeed birds (*12*). Although this kind of evidence leaves open whether people would reason logically when factors that lead them to a different conclusion (in this case, typicality) are un-

available or seem to them uninformative, these mistakes have generally been taken to indicate that the relevant logical competence is absent from ordinary intuitive reasoning. Even "dual process" theories, which hold that people can reason either by logic or by heuristics like typicality, still uniformly subscribe to the view that logic is merely normative (*14, 24, 32–35*). Dual process theorists maintain that logical reasoning must be taught explicitly, that persistent errors reflect a lack of logical competence, and that, at best, logic might become naturally intuitive only for people who have enough aptitude or training to avoid heuristic-driven error (*35, 36*).

However, on the view that logic provides natural laws of thought, there is an alternative explanation for these systematic errors: If they are due to non-logical cognitive processes, these processes might exist *alongside* natural logic. On this alternative, forming new beliefs is a multifactorial process, made up of both rational and non-rational components (*37*). In that case, errors might simply mask a more sophisticated reasoning competence that can naturally govern how people reason when no other process interferes. Indeed, *interference* with reasoning is already the leading explanation of the role that many factors play, including not only the various biases and heuristics (*14, 19*), but also prior content knowledge that conflicts with a logically valid conclusion (*38*), and the perceived relevance of task instructions (*39*).

Consistent with this alternative, decades of research in linguistics have shown how a natural logic can explain otherwise unexpected phenomena in people's ordinary language use (*18, 40–44*). For instance, English speakers intuitively know that it is possible to say either, "all birds ever discovered have sesamoid bones" or "no birds ever discovered have sesamoid bones," but unacceptable to say, "some birds ever discovered have sesamoid bones." This intuitive judgment has been supported by many psycholinguistic (*45–47*) and neuropsychological (*46–50*) studies. The leading explanation is that ordinary language processing is sensitive to Aristotle's very same *dictum*: the quantifiers *all* and *no* are **downward-entailing**, licensing inferences from properties of kinds to properties of subkinds. Adding "ever" specifies that the predicate applies

without exception, which only makes sense where exceptions are conceivable – when talking about a property of a kind that should otherwise apply (or not apply) to its subkinds. In contrast, because a statement about "some birds" entails nothing about any subkinds of birds, specifying that the predicate applies without exception makes—quite literally—no sense. Consistent with this explanation, the same pattern holds for other words that specify *without exception* (*51*) and for translation equivalents across languages (*52*). Moreover, this pattern does not depend on formal education or any instruction in logic; even preschool-aged children only say words that specify *without exception* in downward-entailing contexts, just like adults (*53, 54*). All this evidence suggests that a natural logic characterizes aspects of language knowledge.

Nevertheless, people's logical competence could be limited to grasping the meanings of sentences. After all, not all logical inferences leave linguistic fingerprints; many incoherent or contradictory statements are nevertheless otherwise unobjectionable grammatical sentences (*43*). Even if natural logic underlies language comprehension, it may not govern how people evaluate whether a comprehensible sentence is actually true, let alone how they reason about the truth of one sentence given another. That is, even if logic provides some of the natural laws of language, it is an open question whether it also provides any natural laws of thought: whether it characterizes belief formation, or how people make inferences from one thought to another.

We hypothesized that reasoning logically might be intuitive and automatic for people with no special training when it involves the logical relations that are inherent in the structure of language, such as Aristotle's *dictum*. To increase the chance of revealing this logical competence, we created a task aimed at minimizing interference from other cognitive processes. Participants were never instructed to reason, and the task was presented only as a test of reading comprehension. We hypothesized that if logic characterizes how people naturally think, participants would slow down whenever they expect to encounter information that logically follows from the preceding context, but instead encounter information that does not. Instead of asking participants

to evaluate claims about kinds and their properties (such as whether *all birds have sesamoid bones*), where they might be influenced by what they consider to be plausible given their knowledge about the world, we had them read about specific fictional characters and events where different conclusions are plausible. Instead of asking them to consider subkind relations that rely on world knowledge (such as the relation between birds and penguins), we introduced subkinds through the compositional rules of language (such as the relation between birds and small birds).

In each of the three pre-registered experiments, participants read short narratives line-by-line. After reading each line, they pressed [SPACE] to hide that line and reveal the next, and their reading time was recorded. The critical narratives contained a *premise* in line 4 and a *conclusion* in line 5, which differed by one word, creating different logical relations between the lines (see Figure 1). A key feature of our design is that across trials and participants, the same conclusion was paired with different premises. This enabled us to test whether the exact same information took longer to read when it did not logically follow from the preceding premise than when it did. The conclusion line always started with a *presupposition trigger* (*55*), a phrase which conveys that the information following it is old (i.e. already known to the reader). Prior research has shown that when information following a presupposition trigger is new rather than old, a *presupposition failure* occurs and people take longer to read the sentence (*56, 57*). We measured reading time of the conclusion line as a proxy for participants' processing cost; if participants process logically entailed information as old, they should take less time to read it than unentailed, new information. Finally, each story was followed by a comprehension question, which did not target the logical inference, but required paying attention to the narrative.

Experiment 1 tested whether participants ($N = 383$ after exclusions) would spontaneously detect logical contradictions. We manipulated the quantifiers (*some, all, none, not all*) in the premise and the conclusion to create six trial types: two where the premise was identical to the conclusion, two where the premise differed from but logically entailed the conclusion, and two

(1) *A group of scientists wanted to know whether spotted rats,*
(2) *who are pickier eaters than other rats, liked a new kind of food.*
(3) *They tested white, black, and spotted rats of both sexes.*
(4) *The scientists discovered that* QUANT1 *of the rats loved the food.*
(5) *Now that they knew that* QUANT2 *of the rats loved the food,*
(6) *they decided to issue a recommendation based on their findings.*

Figure 1: An example Experiment 1 item. QUANT1 was replaced by *some*, *all*, *not all*, or *none*; QUANT2 was replaced by *some* or *not all*. The box indicates the conclusion line. On each trial, the dependent measure was the time between participants pressing space to reveal this line, and pressing space again to hide it and reveal the next. Line numbers and the box around line 5 were not shown to participants.

where the premise contradicted the conclusion (Table 1). For example, given the premise that *all of the rats loved the food*, the conclusion that *some of the rats loved the food* is entailed. However, given the premise that *none of the rats loved the food*, the same conclusion would instead be a contradiction. We found that how long participants took to read the conclusion depended on its logical relation to the preceding premise (Figure 2; $\chi^2(2) = 401.88$, $p < 0.001$; model comparisons from linear mixed-effects regressions). On average, participants took $434\,\mathrm{ms}$ longer to read a conclusion that contradicted the preceding premise than one that was logically entailed by it ($t = 11.82$, $p < 0.001$, $d = 0.34$), and $694\,\mathrm{ms}$ longer than a conclusion that simply repeated the premise ($t = 19.95$, $p < 0.001$, $d = 0.57$).

TRIAL TYPE	QUANT1	QUANT2
IDENTITY	**some** *of the rats loved . . . . now that they knew that*	**some** *of the rats loved . . .*
IDENTITY	**not all** *of the rats loved . . . now that they knew that*	**not all** *of the rats loved . . .*
ENTAILMENT	**all** *of the rats loved . . . . . now that they knew that*	**some** *of the rats loved . . .*
ENTAILMENT	**none** *of the rats loved . . . . now that they knew that*	**not all** *of the rats loved . . .*
CONTRADICTION	**none** *of the rats loved . . . . now that they knew that*	**some** *of the rats loved . . .*
CONTRADICTION	**all** *of the rats loved . . . . . now that they knew that*	**not all** *of the rats loved . . .*

Table 1: Experiment 1 trial types. QUANT1 indicates the quantifier used in the Premise line. QUANT2 indicates the quantifier used in the conclusion line of the same trial.

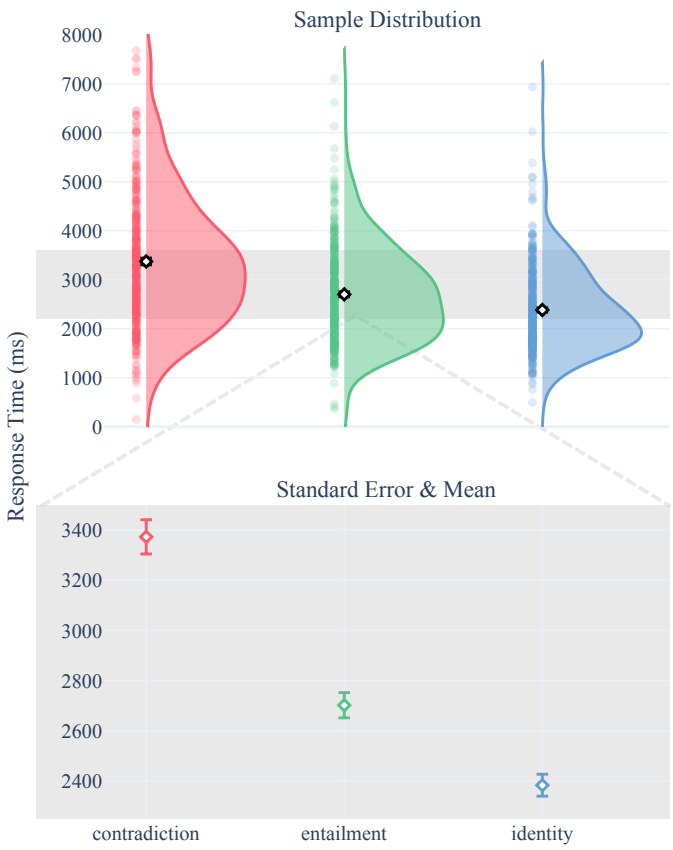


Figure 2: Results from Experiment 1, broken down by trial type. Each scatter point represents one participant's average response time over all vignettes they read within that trial type [$n = 384$ participants]. Contradiction trials in red, Entailment trials in green, Identity trials in blue. The lower portion zooms in on the relevant part of the distribution. Error bars show +/-1 S.E.

Experiments 2 and 3 investigated the ability to detect subtler unlicensed inferences in the absence of strict contradictions. They extended the same paradigm to test reasoning by Aristotle's *dictum de omni et nullo* (*1*) and related extensions in first-order logic. Per the *dictum*, if either *all* or *none of the rats loved the food*, it follows that any subset of the rats (e.g. *the spotted rats*, *the male spotted rats*, etc.) felt the same way. More generally, as mentioned above, the quantifiers *all* and *none* are **downward-entailing** with respect to the subject of the sentence, licensing inferences from properties of a set "down" to any subset. Inferences in the opposite direction are unlicensed; if *all* or *none of the male spotted rats loved the food*, the same might

not be true of *all* or *none* of the *spotted rats*. Conversely, the quantifiers *not all* and *some* are **upward-entailing** with respect to the subject of the sentence, licensing just the opposite pattern of inference, from sets "up" to their supersets. If *not all* or *some of the male spotted rats loved the food*, the same must be true of any superset (e.g. *the spotted rats*, *the rats*), but need not be true of any subset. Finally, two of these four quantifiers – *all* and *not all* – flip which inferences they license with respect to the predicate of the sentence as compared to its subject. For example, *all* is downward-entailing with respect to the subject, but upward-entailing with respect to the predicate: *all of the rats ate leafy vegetables* licenses the inference that *all of the rats ate vegetables*. Figure 3 summarizes the directions of inferences licensed by particular quantifiers in different parts of a sentence. More detailed tables, providing a breakdown of the licit and illicit inferences by set-subset relations, are given in the Supplementary Materials.

$$\underbrace{\textit{some of}}_{\text{quantifier}}\ \underbrace{\textit{the spotted rats}}_{\text{subject}}\ \underbrace{\textit{ate leafy vegetables}}_{\text{predicate}}$$

QUANTIFIER:	*some*	*not all*	*all*	*none*
SUBJECT	upward	upward	downward	downward
PREDICATE	upward	downward	upward	downward

Figure 3: The structure of a quantified sentence and the entailment profiles of the four quantifiers used in Experiments 2 and 3 with respect to the subject and the predicate of a sentence. Experiment 2 manipulates the subject (e.g. *the spotted rats*), whereas Experiment 3 manipulates the noun phrase contained in the predicate (e.g. *ate leafy vegetables*).

Experiment 2 ($N = 384$) and Experiment 3 ($N = 393$) tested whether participants would spontaneously detect inferences that are unlicensed with respect to this logical pattern. Unlike in Experiment 1, the quantifier was held constant between the premise and the conclusion lines. Instead, we manipulated the noun phrases in the *subject* (Experiment 2) and the *predicate* (Experiment 3). In Experiment 2, the subject in the premise appeared with two modifiers (*male*

EXPERIMENT 2

(1) *A group of scientists wanted to know whether spotted rats,*
(2) *who are pickier eaters than other rats, liked a new kind of food.*
(3) *They tested white, black, and spotted rats of both sexes.*
(4) *The scientists discovered that* QUANT *of the ((male) spotted) rats loved the food.*
(5) *Now that they knew that* QUANT *of the spotted rats loved the food,*
(6) *they decided to issue a recommendation based on their findings.*

EXPERIMENT 3

(1) *A group of scientists wanted to know what rats liked to eat.*
(2) *They gave rats a choice of different meats,*
(3) *as well as leafy and root vegetables, both fresh and frozen.*
(4) *They discovered that* QUANT *of the rats ate ((frozen) leafy) vegetables.*
(5) *Now that they knew that* QUANT *of the rats ate leafy vegetables,*
(6) *they decided to issue a recommendation based on their findings.*

Figure 4: Example items from Experiments 2 (top) and 3 (bottom). Underlined elements varied between trial types. QUANT was replaced by *some*, *all*, *not all*, or *none*, with the same quantifier used in line 4 as in line 5. Line 5 always contained one modifier on the noun in the subject (Exp 2: *spotted rats*) or the predicate (Exp 3: *leafy vegetables*). Line 4 varied to create different containment relations (subset, identity, superset) relative to line 5: two modifiers (e.g. *male spotted rats* $\subset$ *spotted rats*), one modifier (*spotted rats* = *spotted rats*), or no modifiers (*rats* $\supset$ *spotted rats*). On each trial, the dependent measure was the time between participants pressing space to reveal line 5 (highlighted by the box) and pressing space again to hide it and reveal the next line. None of the line numbers, the underlining, or the box around line 5 were shown to participants.

spotted rats), one modifier (*spotted rats*), or no modifiers (*rats*). Likewise, in Experiment 3, the predicate in the premise appeared with two modifiers (*frozen leafy vegetables*), one modifier (*leafy vegetables*), or no modifiers (*vegetables*). The conclusion noun phrase always appeared with one modifier (Experiment 2: *spotted rats*; Experiment 3: *leafy vegetables*). Thus, in both experiments, the noun phrase in the premise described a subset, an identical set, or a superset of the conclusion noun phrase. These three containment relations, combined with the same four quantifiers in both experiments (*some*, *not all*, *all*, and *none*) yielded twelve trial types in each

experiment: four trial types where the premise was identical to the conclusion, four where the premise differed from but entailed the conclusion, and four where the premise did not entail the conclusion.

In each experiment, we found that how long participants took to read a conclusion line depended on whether it was entailed by the preceding premise (Figure 5; interaction of Entailment Direction (*upward* vs. *downward*) $\times$ Containment Relation (*subset* vs. *superset*) in Experiment 2: $\chi^2(1) = 11.0$, $p < 0.001$; Experiment 3: $\chi^2(1) = 7.18$, $p = 0.007$; LMER model comparison). This reflects participants taking longer to read conclusions that were not entailed compared to ones that were entailed (Exp 2: $M_{entailed} = 2708\,\mathrm{ms}$, $M_{unentailed} = 2842\,\mathrm{ms}$; Exp 3: $M_{entailed}$=$2653\,\mathrm{ms}$, $M_{unentailed}$=$2738\,\mathrm{ms}$). The effect on individual quantifiers was sometimes exacerbated and sometimes suppressed by an independent main effect of Containment in both experiments (Exp 2: $\chi^2(2) = 24.92$, $p < 0.001$; Exp 3: $\chi^2(1) = 3.98$, $p = 0.05$). This reflects lexical priming of conclusions sharing more words with their preceding premises on subset (e.g. *male spotted rats* $\prec$ *spotted rats*) than superset (e.g. *rats* $\prec$ *spotted rats*) trials, with the lexical repetition producing a priming effect that is orthogonal to the logical inferences (see Supplement). Figure 5 shows the residuals from this main effect to visualize the signature of logical inference: participants consistently took longer to read unentailed conclusions than entailed ones. In an even finer-grained test of logical sensitivity, we found that the effect of Containment on participants' reading times changed between Experiments 2 and 3 only for those quantifiers whose entailment profiles differ between the subject and the predicate. That is, Containment (Subset vs. Superset) interacted with Experiment (2 vs. 3) for *all* and *not all*, which license inferences in the opposite directions in each experiment ($\chi^2(1) = 9.17$, $p = 0.002$; linear hypothesis tests) but not for *some* and *none*, where the licensed inference does not change ($\chi^2(1) = 2.53$, $p = 0.112$). In sum, participants' reading times followed exactly the intricate pattern of logically licensed and unlicensed inferences: when they encountered statements that should have followed from

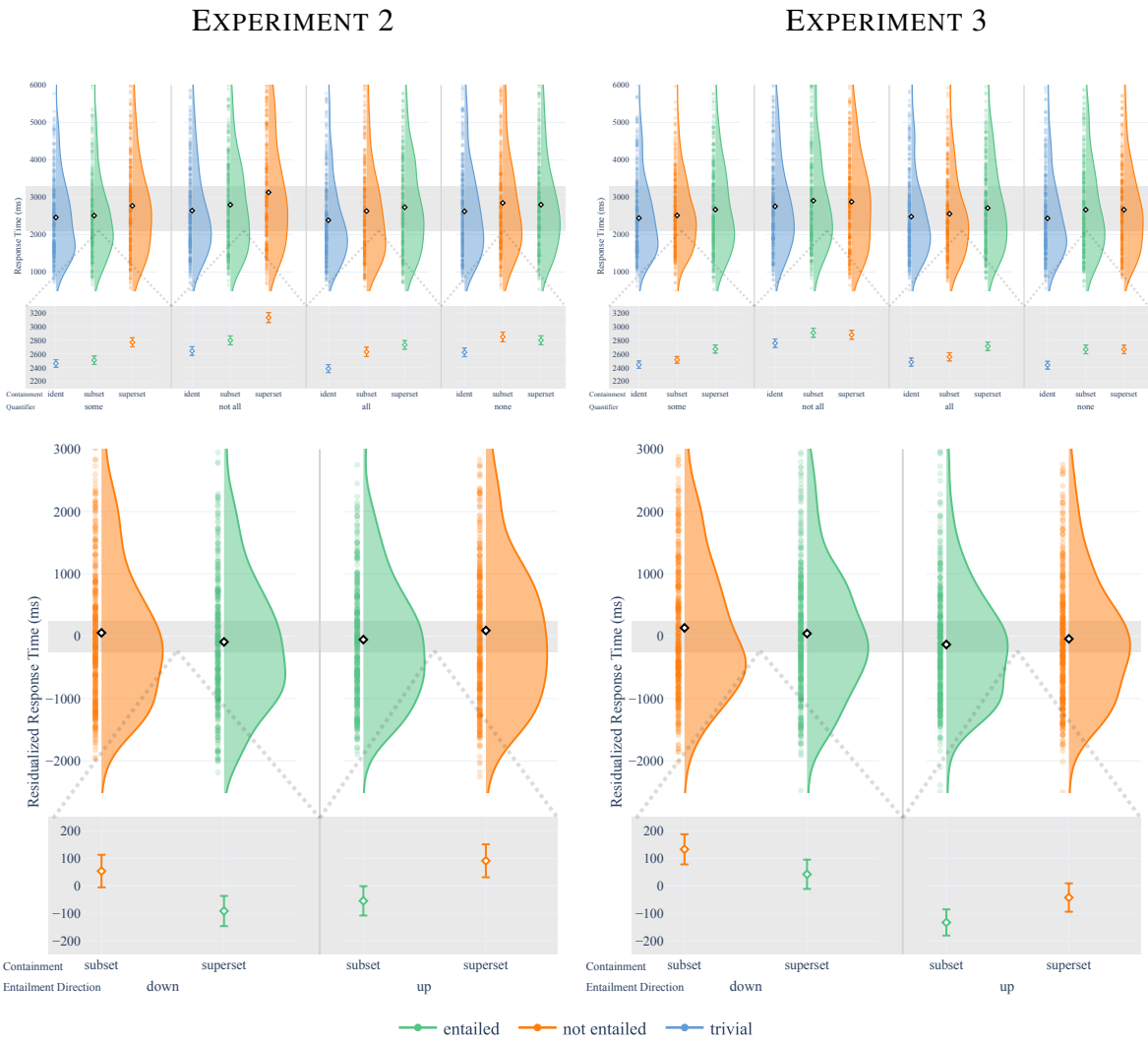


Figure 5: Results of Experiments 2 (left) and 3 (right) [subset: e.g. *male spotted rats* $\prec$ *spotted rats*, superset: e.g. *rats* $\prec$ *spotted rats*]. Trials, where the conclusion is entailed by the premise, are in green, trials where it is not entailed are in orange, and trials where the premise and conclusion are identical are in blue. Top: The full distribution of response times within each trial type. Each scatter point represents one participant's average response time across all vignettes with that quantifier-containment combination [Exp. 2: $N = 384$, Exp. 3: $N = 393$]. Bottom: Results of Experiments 2 and 3, with quantifiers grouped by entailment direction. The y-axis shows residual response times, subtracting the mean of the corresponding level of Containment (Subset or Superset) from each scatter point. This visualizes the interaction that reflects whether each conclusion was entailed by the preceding premise or not, independent of the main effect of Containment.

the preceding contexts but did not, they slowed down, suggesting that they registered the invalid inference.

In a last experiment, we use computational models to test an alternative explanation for these results: maybe people are not sensitive to invalid logical inferences, but only to unusual distributions of phrases in their language. It is well-established that reading times generally depend on the fast, automatic use of latent distributional knowledge to predict upcoming words during online processing, with unexpected words taking longer to read (*58,59*). It is possible, for example, that people represent that phrases of the form "all of the SUBJECT PREDICATE" are rarely followed by "not all of the SUBJECT PREDICATE". Perhaps these kinds of representations are all that determine their reading times in the preceding experiments.

Large language models (LLMs) effectively capture both the existence of this kind of distributional information, and people's mental representation of it. LLMs are trained by iteratively performing next-word prediction over extremely large corpora of text, and they learn to assign probabilities to each word based on the distributional information in their training data. When shown a novel naturalistic text, the probability that these models assign to each upcoming word maps both to human reading time of the same text (*60,61*), and to patterns of brain activity that correlate selectively with next-word prediction (*62–64*).

In Experiment 4, we compare humans' reading times on the conclusion sentences of Experiment 1-3 to the surprisal values that models assign to the same text. If the patterns of reading times can be explained by distributional knowledge, language models should likewise assign higher surprisals to conclusion sentences that contradict their premises than those that are entailed by their premises in Experiment 1, and higher surprisal to unentailed than entailed conclusions in Experiments 2 and 3. We use probabilities extracted from the largest GPT-3 (`text-davinci-002`) (*65*) and smallest GPT-2 (124M) (*66*) which have been shown to pattern well with human reading times (*60*). We do not use more recent models that are trained to

do more than next-word prediction (such as GPT-4), because they have been shown to be inferior predictors of human reading times (*67*) (details on model choice and inference procedure are given in the Supplementary Materials).

Figures 6 and 7 show the distributions of models' surprisal estimates for each trial type in each experiment. As an important check on the ability of the models to account for human data, we observe that the models show the same lexical priming effects that humans do. First, just like people were the fastest to read conclusions that were exactly identical to their premises, the models assigned the lowest surprisals to identical conclusions in each of the three experiments (see Supplement). Second, in Experiments 2 and 3, people read conclusions in which there was more lexical overlap with the preceding premise (i.e. Subset trials) faster, independently of whether this conclusion was entailed or not, reflected by a main effect of Containment. Models showed the same strong main effect (Figure 7a), assigning lower surprisal to conclusions following Subset than Superset trials in both experiments (Exp 2: GPT-2 $\chi^2(1) = 113.84$, $p < 0.0001$; GPT-3 $\chi^2(1) = 22.95$, $p =< 0.0001$. Exp 3: GPT-2 $\chi^2(1) = 151.74$, $p < 0.0001$; GPT-3 $\chi^2(1) = 16.04$, $p =< 0.0001$).

In contrast, across all three experiments, we find no evidence that either GPT 2 or 3 registered whether a conclusion was entailed by its premise or not. Evaluating both models on Experiment 1, we do not even find any evidence that the models registered contradictory conclusions (effect of Trial Type, Entailed vs. Contradiction; GPT-2: $\chi^2(1) = 0.57$, $p = 0.45$; GPT-3: $\chi^2(1) = 1.38$, $p = 0.24$). Evaluating both models on Experiments 2 and 3, we also find no evidence of an interaction between Entailment Direction $\times$ Containment, which reflects sensitivity to whether a conclusion is entailed from its premise (Exp 2. GPT-2: $\chi^2(1) = 0.88$, $p = 0.34$; GPT-3: $\chi^2(1) = 0.01$, $p = 0.90$; Exp 3. GPT-2: $\chi^2(1) = 0.53$, $p = 0.46$; GPT-3: $\chi^2(1) = 0.88$, $p = 0.35$). Figure 7b shows that residualizing over the effect of Containment results in indistinguishable surprisals for entailed and unentailed conclusions. This pattern of

results suggests that even as humans and these two large language models were both sensitive to lexical overlap between preceding and upcoming content, only humans were sensitive to the difference between logical and illogical inferences.

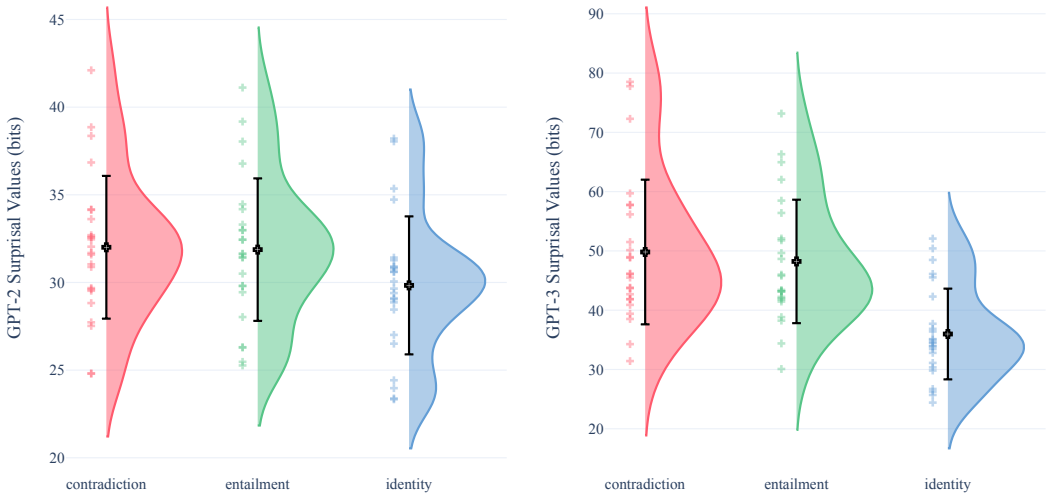
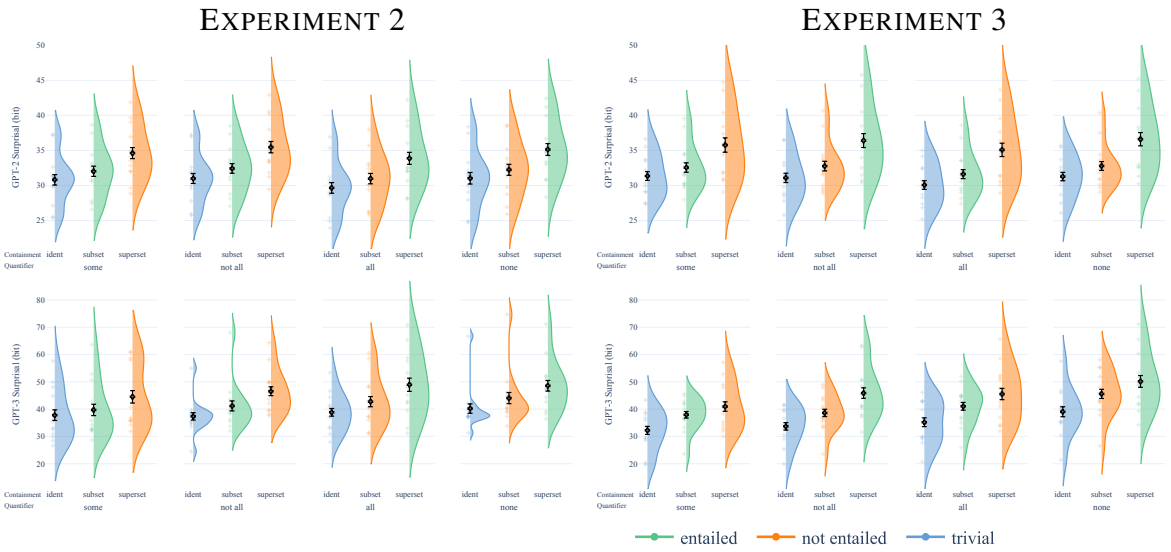
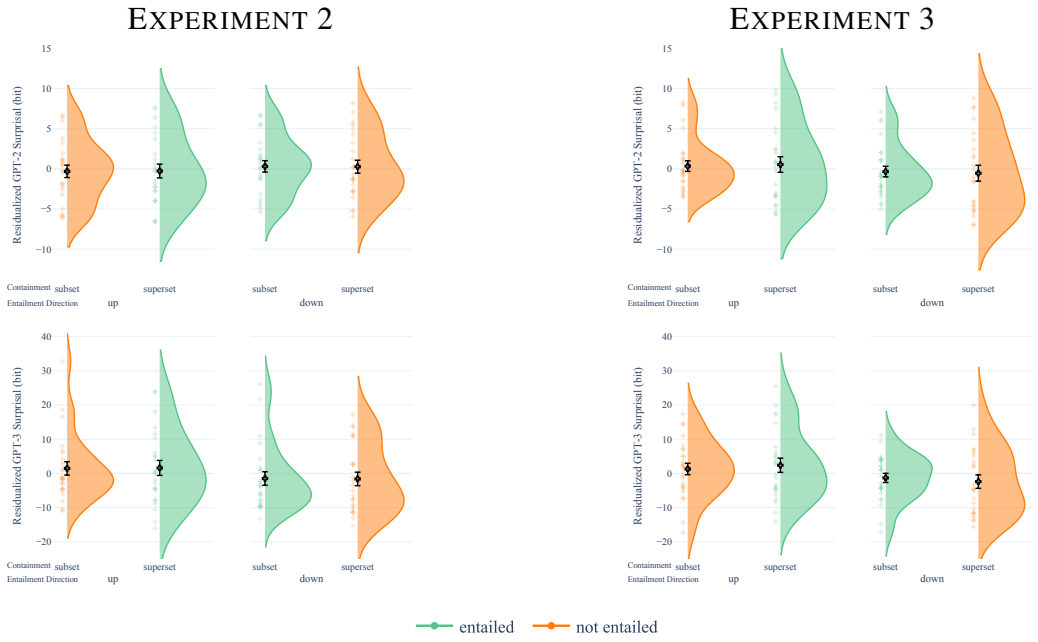


Figure 6: Results of Experiment 4. Distribution of surprisal estimates from GPT-2 and GPT-3 on the conclusion lines from Experiment 1, by entailment condition. Unlike the scatter points in the graphs of Experiments 1-3, each scatter point here represents one item: a vignette within a particular trial type [$N = 24$ per trial type].

In sum, we found that humans have an automatic, fast, and accurate logical ability. This ability is revealed when the influence of other cognitive processes that lead to other conclusions is minimized. This finding challenges theories on which people are only capable of logical reasoning when thinking slowly (*14*), when deliberately following a sequence of rules (*33*), when reasoning about specific kinds of content (*20, 21, 68*), or when they have had extensive instruction (*35*). Our participants were selected only for their knowledge of English and the inferences we tested are on the one hand about ordinary events, and on the other hand not typically taught prior to logic courses at the university level. Indeed, our tasks never instructed participants to reason at all, only to read short narratives. Our results therefore suggest that people engage in spontaneous reasoning that is well-characterized by logical rules. At least some of the laws of



(a) Distribution of surprisal estimates from GPT-2 (top) and GPT-3 (bottom). One scatter point represents the conclusion line in one premise/conclusion pair of each Quantifier × Containment Direction condition ($N = 12$ per condition).



(b) Distribution of residualized surprisal estimates from GPT-2 (top) and GPT-3 (bottom). Residualized surprisal values are computed by taking the surprisal values within each combination of Entailment Direction and Containment and subtracting the mean of the corresponding level of the Containment variable. One scatter point represents one premise/conclusion pair in each Containment × Entailment Direction condition ($N = 24$ per condition).

Figure 7: Results of Experiment 4. Distribution of surprisal estimates from GPT-2 and GPT-3 on conclusion lines from Experiment 2 (left) and 3 (right).

logic do appear to be not only normative, but natural laws of thought. Because forming and endorsing beliefs is governed by the interaction of many different cognitive processes, revealing the natural logical component requires carefully stripping away the rest.

At the same time, it is obvious that not all logical reasoning is spontaneous and automatic. A vast number of interesting mathematical truths are logically entailed by a small set of premises, but figuring out which conclusions follow and which do not takes deliberation by generations of mathematicians. If some kinds of logical inferences are slow and deliberate, and others are fast, automatic, and natural, our results raise the question of where the joint in nature lies. They also suggest an answer: the logical analysis of linguistic meaning proposes candidate components that might make up natural laws not only of language, but also of thought.

References

1. Aristotle, W. R. Roberts, I. Bywater, F. Solmsen, *Rhetoric* (Modern Library, New York, 1954), 7th edn.

2. G. Boole, *An investigation of the laws of thought: on which are founded the mathematical theories of logic and probabilities*, vol. 2 (Walton and Maberly, 1854).

3. J. S. Mill, *A System of Logic, Ratiocinative and Inductive: 1*, vol. 1 (Parker, 1856).

4. G. Frege, *Mind* **65**, 289 (1918/1956).

5. P. F. Strawson, *Introduction to logical theory* (Routledge, 1952).

6. M. Henle, *Psychological review* **69**, 366 (1962).

7. J. Bruner, J. Goodnow, G. Austin, *A Study of Thinking* (John Wiley and Sons, 1956).

8. J. T. Macnamara, *A Border Dispute: The Place of Logic in Psychology* (MIT Press, 1986).

9. P. N. Johnson-Laird, *Foundations of cognitive science*, M. I. Posner, ed. (MIT Press, 1989), pp. 469–499.

10. L. J. Rips, *The psychology of proof: Deductive reasoning in human thinking* (Mit Press, 1994).

11. M. D. S. Braine, D. P. O'Brien, *Mental Logic* (Psychology Press, 1998).

12. S. A. Sloman, *Cognitive Psychology* **35**, 1 (1998).

13. M. Oaksford, N. Chater, *Bayesian Rationality: The Probabilistic Approach to Human Reasoning* (Oxford University Press, 2007).

14. D. Kahneman, *Thinking, Fast and Slow* (Farrar, Straus and Giroux, 2011).

15. R. Montague, *Linguaggi nella societa e nella tecnica*, B. Visentini, ed. (Edizioni di Communita, 1970), pp. 188–221.

16. G. Lakoff, *Synthese* **22**, 151 (1970).

17. B. Hall-Partee, *Semantics from different points of view* (Springer, 1979), pp. 1–14.

18. G. Chierchia, *Logic in grammar: Polarity, free choice, and intervention* (OUP Oxford, 2013).

19. D. Kahneman, A. Tversky, *Cognitive Psychology* **3**, 430 (1972).

20. P. W. Cheng, K. J. Holyoak, *Cognitive Psychology* **17**, 391 (1985).

21. L. Cosmides, *Cognition* **31**, 187 (1989).

22. P. Koralus, S. Mascarenhas, *Philosophical Perspectives* **27**, 312 (2013).

23. M. Ragni, I. Kola, P. N. Johnson-Laird, *Psychological Bulletin* **144**, 779 (2018).

24. F. Lieder, T. L. Griffiths, *Behavioral and Brain Sciences* **43** (2020).

25. P. C. Wason, *Quarterly Journal of Experimental Psychology* **20**, 273 (1968).

26. B. Inhelder, J. Piaget, *London: Routledge & Kegan Paul. Jordan, WJ, & Nettles, SM (2000). How students invest their time outside of school: Effects on school-related outcomes. Social Psychology of Education* **3**, 217 (1964).

27. E. M. Markman, *Categorization and naming in children: Problems of induction* (mit Press, 1989).

28. V. M. Sloutsky, A. V. Fisher, *Journal of Experimental Psychology: General* **133**, 166 (2004).

29. S. A. Sloman, *Cognitive psychology* **25**, 231 (1993).

30. D. P. Calvillo, R. Revlin, *Psychonomic Bulletin & Review* **12**, 938 (2005).

31. A. K. Bright, A. Feeney, *Journal of Experimental Psychology: General* **143**, 2082 (2014).

32. J. S. B. T. Evans, K. E. Stanovich, *Perspectives on Psychological Science* **8**, 223 (2013).

33. S. A. Sloman, *Psychological Bulletin* **119**, 3 (1996).

34. S. S. Khemlani, R. M. J. Byrne, P. N. Johnson-Laird, *Cognitive Science* **42**, 1887 (2018).

35. W. De Neys, *The Behavioral and Brain Sciences* **46**, e146 (2023).

36. W. De Neys, G. Pennycook, *Current Directions in Psychological Science* **28**, 503 (2019).

37. N. Porot, E. Mandelbaum, *Wiley Interdisciplinary Reviews: Cognitive Science* **12**, e1539 (2021).

38. P. Pollard, J. S. B. Evans, *The American journal of psychology* pp. 41–60 (1987).

39. D. Sperber, F. Cara, V. Girotto, *Cognition* **57**, 31 (1995).

40. G. Fauconnier, *Proceedings of Chicago* (1975).

41. W. A. Ladusaw, Polarity sensitivity as inherent scope relations, Ph.D. thesis, University of Texas at Austin (1980).

42. K. Von Fintel, *Natural language semantics* **1**, 123 (1993).

43. J. Gajewski, *Manuscript, MIT* **3** (2002).

44. D. Fox, M. Hackl, *Linguistics and philosophy* **29**, 537 (2006).

45. C. Clifton Jr, L. Frazier, *Linguistic Inquiry* **41**, 681 (2010).

46. M. Xiang, J. Grove, A. Giannakidou, *Frontiers in psychology* **4**, 708 (2013).

47. D. Parker, C. Phillips, *Cognition* **157**, 321 (2016).

48. H. Drenhaus, S. Frisch, D. Saddy, *Linguistic Evidence* (De Gruyter Mouton, 2008), pp. 145–164.

49. S. Vasishth, S. Brüssow, R. L. Lewis, H. Drenhaus, *Cognitive Science* **32**, 685 (2008).

50. M. Xiang, B. Dillon, C. Phillips, *Brain and Language* **108**, 40 (2009).

51. N. Kadmon, F. Landman, *Linguistics and philosophy* **16**, 353 (1993).

52. J. Hoeksema, *Linguistic Analysis* **38**, 3 (2012).

53. L. S. Tieu, *Semantics and Linguistic Theory* (2010), vol. 20, pp. 19–37.

54. A. GUALMINI, *Linguistics* **42**, 957 (2004).

55. L. Karttunen, *Theoretical Linguistics* **1**, 181 (1974).

56. C. Clifton Jr, *Journal of Experimental Psychology: Learning, Memory, and Cognition* **39**, 487 (2013).

57. F. Schwarz, S. Tiemann, *Journal of Semantics* **34**, 61 (2017).

58. J. Hale, *Language and Linguistics Compass* **10**, 397 (2016).

59. R. Levy, *Cognition* **106**, 1126 (2008).

60. C. Shain, C. Meister, T. Pimentel, R. Cotterell, R. P. Levy, Large-Scale Evidence for Logarithmic Effects of Word Predictability on Reading Time (2022).

61. N. J. Smith, R. Levy, *Cognition* **128**, 302 (2013).

62. M. Schrimpf, *et al.*, *Proceedings of the National Academy of Sciences* **118**, e2105646118 (2021).

63. C. Caucheteux, J.-R. King, *Communications biology* **5**, 134 (2022).

64. A. Goldstein, *et al.*, *Nature neuroscience* **25**, 369 (2022).

65. T. B. Brown, *et al.*, Language Models are Few-Shot Learners (2020).

66. A. Radford, *et al.*, *OpenAI blog* **1**, 9 (2019).

67. T. Kuribayashi, Y. Oseki, T. Baldwin, Psychometric Predictive Power of Large Language Models (2023).

68. I. Dasgupta, *et al.*, *arXiv preprint arXiv:2207.07051* (2022).

69. N. Mostafazadeh, *et al.*, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Association for Computational Linguistics, San Diego, California, 2016), pp. 839–849.

70. R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria (2022).

71. D. Bates, M. Mächler, B. Bolker, S. Walker, *Journal of Statistical Software* **67**, 1 (2015).

72. H. Matuschek, R. Kliegl, S. Vasishth, H. Baayen, D. Bates, *Journal of memory and language* **94**, 305 (2017).

73. S. C. Levinson, *Cambridge UK* (1983).

74. A. W. Inhoff, *Journal of Psycholinguistic Research* **14**, 45 (1985).

75. F. Schwarz, *Annual review of linguistics* **2**, 273 (2016).

76. G. B. Forbach, R. F. Stanners, L. Hochhaus, *Memory & Cognition* **2**, 337 (1974).

77. K. I. Forster, C. Davis, *Journal of experimental psychology: Learning, Memory, and Cognition* **10**, 680 (1984).

78. Llama 2: Open Foundation and Fine-Tuned Chat Models | Meta AI Research, https://ai.meta.com/research/publications/llama-2-open-foundation-and-fine-tuned-chat-models/.

79. Introducing Claude, https://www.anthropic.com/index/introducing-claude.

80. R. Taori, *et al.*, Stanford alpaca: An instruction-following llama model, https://github.com/tatsu-lab/stanford_alpaca (2023).

81. OpenAI, GPT-4 Technical Report (2023).

82. M. van Schijndel, T. Linzen, *Neural Network Models* (2018).

Acknowledgments

For help with data collection, we are grateful to Heather Yu. For generous and helpful feedback, we thank Roger Levy, Scott AnderBois, Joshua Schechter, Thomas Icard, Susan Carey, and the audiences of the 2021 CUNY Sentence Processing Conference, Society of Philosophy and Psychology 2022, and a workshop on Monotonicity and Entailment funded by a network grant on the "Ontogenetic Origins of Abstract Combinatorial Thought" from the McDonnell Foundation. We especially thank Jason Ritt for invaluable assistance with the statistical analyses, particularly with the comparison of results between Experiments 2 and 3.

Funding

Experiments 1-3 were funded by startup funds from Brown University to RF.

Author contributions

Conceptualization: Exps 1-3: MD, RF. Exp 4: AL, EP, RF

Methodology: Exps 1-3: MD, RF. Exp 4: AL, EP, RF

Formal Analysis: Exps 1-3: MD, AL, RF. Exp 4: AL

Investigation: Exps 1-3: MD. Exp 4: AL

Visualization: AL, RF, MD

Funding acquisition: RF, EP

Project administration: RF

Supervision: RF, EP

Writing – original draft: MD, RF

Writing – review & editing: MD, AL, EP, RF

Competing interests

Authors declare that they have no competing interests.

Ethical compliance

All the participants were informed about the nature and possible consequences of the experiments and consented to participate in the study. The experiments were conducted with approval from the IRB at Brown University.

Data, code and materials availability

All data, pre-registrations, and an annotated code base are available at https://osf.io/gmbs8/?view_only=f2f5374ebcb14a5083985b3c6549afd3. Additionally, all the vignettes used in the study (including both target and filler items) are included in Supplementary Materials. There are no restrictions on data availability.

Supplementary materials

Experiments 1 to 4: Methods and Results

Appendix A: Target items

Appendix B: Filler items

Figs. 8 to 12

Tables 3 to 10

References (*68-81*)

Supplementary materials

Experiment 1

Experiments 1-3 were each preregistered separately. All aspects of the methods and results followed the preregistered plan, unless otherwise noted. All analysis scripts and details can be found on OSF, accompanying the preregistration for each experiment. Experiment 4 was not pre-registered. The preregistration for Experiment 1 can be found on OSF (link anonymized for peer review).

Methods

Participants

We tested 400 self-reported native English speakers. Participants were recruited through Amazon Mechanical Turk, using TurkPrime to recruit only CloudResearch Approved Participants, who had previously passed a battery of engagement measures. One participant did not finish the task, and 15 participants were excluded for failing to answer the attention checks (the T/F questions at the end of each item) correctly more often than chance (at least 21 out of 30 trials, $p < 0.05$ on a binomial test). The excluded participants were not replaced. The rest were compensated $2 for their participation. Study protocols were approved by Brown University's Institutional Review Board.

Design

This study used a line-by-line self-paced reading paradigm. Each participant was asked to read 12 target items and 18 filler items. Each item was displayed to them one line at a time, with line breaks at clausal boundaries. The reading time of each clause was measured as the length of time before the participant pressed a key to reveal the next clause and hide the present one. An example of a target item is given in Figure 8.

TRIAL TYPE	QUANT1	QUANT2
IDENTITY	**some** *of the rats loved . . . . now that they knew that*	**some** *of the rats loved . . .*
IDENTITY	**not all** *of the rats loved . . . now that they knew that*	**not all** *of the rats loved . . .*
ENTAILMENT	**all** *of the rats loved . . . . . now that they knew that*	**some** *of the rats loved . . .*
ENTAILMENT	**none** *of the rats loved . . . . now that they knew that*	**not all** *of the rats loved . . .*
CONTRADICTION	**none** *of the rats loved . . . . now that they knew that*	**some** *of the rats loved . . .*
CONTRADICTION	**all** *of the rats loved . . . . . now that they knew that*	**not all** *of the rats loved . . .*

Table 2: Experiment 1 trial types. QUANT1 indicates the quantifier used in the Premise line. QUANT2 indicates the quantifier used in the conclusion line of the same trial.

The first three lines of each target item set up background information. Each target item contained two quantified clauses: Line 4 contained the *premise* of the logical inference, and line 5 contained its *conclusion*. Line 6 provided a resolution to the narrative, designed to be compatible with any conclusion line. Lines 4 and 5 were identical except for the quantifiers they used, which varied to create different trial types. The quantifier in the conclusion line was either *some* or *not all*. The quantifier in the corresponding premise line was *some*, *all*, *not all*, or *none*, chosen in such a way as to create two trial types where the conclusion clause was identical to the premise clause, two trial types where the conclusion contradicted the premise, and two trial

(1) *A group of scientists wanted to know whether spotted rats,*
(2) *who are pickier eaters than other rats, liked a new kind of food.*
(3) *They tested white, black, and spotted rats of both sexes.*
(4) *The scientists discovered that* QUANT1 *of the rats loved the food.*
(5) *Now that they knew that* QUANT2 *of the rats loved the food,*
(6) *they decided to issue a recommendation based on their findings.*

Figure 8: An example item from Experiment 1. QUANT1 was replaced by *some*, *all*, *not all*, or *none*; QUANT2 was replaced by *some* or *not all*. The box indicates the conclusion line. On each trial, the dependent measure was the time between participants pressing space to reveal this line, and pressing space again to hide it and reveal the next. Line numbers and the box around line 5 were not shown to participants.

types where the conclusion was entailed by the premise. In total, there were six different trial types, including two Identity, two Contradiction, and two Entailment types (Table 2).

To create the target items, we wrote 12 vignettes (Appendix A), designed to tell coherent narratives in combination with any trial type (except Contradiction). Each vignette was fully crossed with each trial type to create $12 \times 6 = 72$ total target items. These were divided among 12 lists of 12 items each, such that each of the 6 trial types appeared in two different vignettes within a list. Each participant saw all of the target items from one list. Thus, within a participant, each vignette appeared once, and each trial type (Identity, Contradiction, Entailment) appeared four times. Between participants, every trial type appeared in every vignette. Each participant was assigned one of the 12 lists of target items at random.

Filler items were adapted from the Story Cloze Test and ROCStories Corpora (*69*). Like the test items, fillers were narrative paragraphs that the participant revealed line-by-line. Unlike test items, fillers included no premise or conclusion sentences, and no explicit violations of entailment or contradictions. The purpose of the fillers was to obscure the manipulation in the target items. The same fillers were used in Experiments 1-3 for all participants.

All items, filler and test, were followed by a T/F question about the content of the narrative. These questions served as an attention check. They did not ask about any content relevant to the logical inference.

Procedure

The experiment was administered using the Ibex platform (http://spellout.net/ibexfarm/). After consenting to participate, participants were told that they would read 30 short stories, but would not be able to see the entire story all at once. They were instructed to press the space bar to reveal the story line by line. They were also told that each story would be followed by a statement about it, which they would have to evaluate as true or false by pressing [T] or [F] (or clicking

true [*T*] or *false* [*F*] on the screen), and were warned that they would need to pay attention and get at least 21/30 questions right.

These instructions were followed by three practice trials of two lines each (*This is a practice trial ...*), meant to help participants get used to pressing space to reveal the next line. Each item was displayed one line at a time, and pressing the space bar simultaneously revealed the next line and hid the present one. The second practice trial was followed by a T/F question (*Is this a practice question?*)

After this practice phase, the 30 items (12 target + 18 filler) were administered in a dynamically generated pseudo-random order, constrained so that target items never appeared back to back, and were always separated by either one or two filler items. All items were separated by a forced one-second delay to prevent participants from rapidly advancing through the study. Following each item, the T/F question about that item appeared on a new screen.

Results

Data exclusion. Following our pre-registered plan, we excluded 25 individual trials from analysis where the RT on the Conclusion line was faster than 100 ms (assumed to be too fast to have read the clause, suggesting either a bot or a participant speeding through), and 1530 trials where the RT on the Conclusion line was slower than 10 seconds (assumed to be a distracted participant). We then excluded 1240 additional trials where the mean RT on the Conclusion line was more than +/- 3 standard deviations away from the mean for that line in that vignette. The remaining data included 70,873 trials from 384 participants.

Analysis Strategy. In all of the experiments we report, our dependent variable is the log-transformed reading time (RT) on the conclusion (i.e. penultimate) line of each target item. We used a series of mixed effects linear regressions to test whether participants' time to read these

conclusion lines depended on the logical relation between the conclusions and the immediately preceding premise lines.

We analyzed the data in R (v4.2.1) (*70*) using the lme4 package (v1.1.30) (*71*) to build a series of logistic mixed-effects models. We followed the model selection strategy recommended by Matuschek and colleagues (*72*). We start from a maximal model and use log-likelihood ratio tests to compare it with a sequence of models with reduced random parameter structures, stopping upon finding a significant decrease in model fit. Rather than choosing the most reduced of these models in each case, which would result in inconsistent and incomparable models across otherwise similar analyses, we select the 'lowest common' reduced model across such analyses. In each case we used log likelihood ratio tests to verify that there was no significant difference in model fit between these and the most maximal converging model. For estimates of the significance of main effects, we derived p-values from Type II Wald χ^2 tests comparing minimally different models with and without that effect. Estimates of simple effects were obtained from Wald z significance tests on coefficients of dummy coded contrast variables. The contrast coding scheme is explained further below. All data, reproducible analysis code, and exact model specifications are available at https://osf.io/gmbs8/?view_only=f2f5374ebcb14a5083985b3c6549afd3.

Analyses. The Identity trials require no inference, with the conclusion line just repeating the premise. This repetition means that participants may have been primed by the premise to process the exact lexical and semantic content of the conclusion, and so may take less time to read the conclusions of Identity trials compared to the other trial types. Both of these factors predict faster RTs for the Identity trials, meaning that a comparison between these and either the superset or the subset conditions would not be directly attributable to the presence or absence of an inference. We treat this trial type as a manipulation check for our design, and sought to confirm that participants read its conclusion faster. We built a regression including

an Identity variable, coded to compare the Identity trials to the average of the Entailment and Contradiction trials. As expected, we find significantly faster reading times in the Identity trials ($\chi^2(2) = 284.97, p < 0.001$).

Our main question was whether participants read the conclusion faster when it was entailed by the premise than when it contradicted the premise. An auxiliary question is whether this effect varied by the quantifier used in the conclusion. To answer both questions, we leave out Identity trials, and model the binary effect of Trial Type (Contradiction vs. Entailment), Conclusion Quantifier (Some vs. Not All), and their interaction. This revealed a highly significant main effect of Trial Type, with participants taking over half a second longer to read a conclusion when it had been contradicted by the preceding premise compared to when it had been entailed by that premise (Mean difference: 660 msec, $\chi^2(2) = 123.27$, $p < 0.001$). There was also a highly significant effect of Quantifier ($\chi^2(2) = 47.22$, $p < 0.001$), indicating that participants read conclusions with *not all* slower than with *some*. However, there was no significant interaction between the quantifier and whether the conclusions was contradicted or entailed by the premise ($\chi^2(2) = 2.49$, $p = 0.116$). Treatment coding further revealed that participants took longer to read the contradicting conclusions with each of these quantifiers separately (simple effect of Trial Type within *some*: $t = 7.47$, $p < 0.001$; within *not all*: $t = 9.49$, $p < 0.001$). Recall that conclusions containing *some* were entailed by premises containing *all* and contradicted by premises containing *none* and that the pattern is exactly reversed for conclusions containing *not all*. This means that the finding that contradicting conclusions took longer to read than entailed conclusions cannot be due to the particular lexical material in either the conclusions or the premises, but must instead be due to the logical relation between the two lines.

Finally, note that it is possible to compute an implicature with respect to the material in all of the conclusion lines. In the *all* $\rightarrow$ *some* condition, *some of the rats loved the food* implicates that *not all* did, which contradicts the information that *all of the rats loved the food*, given in the

premise. Similarly, in the *none* → *not all* condition, *not all* implicates *some*, which contradicts *none*. This could have led participants to interpret the trials that we classified as entailments as contradictions instead. This means that, in comparing the trials we classified as contradictions to those we classified as entailments, the analyses above actually compare literal contradictions to possible contradictions-by-implicature. This makes the present task a more conservative test of sensitivity to logical contradiction. The fact that we nevertheless find a large and highly significant difference between the Entailment and Contradiction trials suggests either that participants did not actually compute the implicatures on the conclusion material (maybe because the very alternative that the implicature would negate had just been affirmed in the immediately preceding premise line), or that detecting literal contradictions is a faster or more reliable process than detecting implicated contradictions.

Experiment 2

The preregistration for Experiment 2 can be found on OSF (link anonymized for peer review). As in the other experiments, we follow the preregistration exactly unless otherwise noted.

Methods

Participants

We tested a new sample of 400 self-reported native English speakers, none of whom had participated in Experiment 1. Recruitment and compensation were identical to Experiment 1. One participant did not finish the task, and 15 participants were excluded for failing to answer at least 21 out of 30 T/F attention check questions correctly. These participants were not replaced.

(1) *A group of scientists wanted to know whether spotted rats,*

(2) *who are pickier eaters than other rats, liked a new kind of food.*

(3) *They tested white, black, and spotted rats of both sexes.*

(4) *The scientists discovered that* QUANT *of the ((male) spotted) rats loved the food.*

(5) *Now that they knew that* QUANT *of the spotted rats loved the food,*

(6) *they decided to issue a recommendation based on their findings.*

Figure 9: An example item in Experiment 2. QUANT was replaced by *some*, *all*, *not all*, or *none*, with the same quantifier used in line 4 as in line 5. On each trial, the dependent measure was the time between participants pressing space to reveal line 5 (highlighted by the box) and pressing space again to hide it and reveal the next line. Neither the line numbers nor the box around line 5 were shown to participants.

	some	*not all*	*all*	*none*
SUBSET → of spotted rats →	... **some** *of the* **male spotted rats** *loved the food. Now that they knew that* **some** *of the* **spotted rats** ...	... **not all** *of the* **male spotted rats** *loved the food. Now that they knew that* **not all** *of the* **spotted rats** ...	... **all** *of the* **male spotted rats** *loved the food. Now that they knew that* **all** *of the* **spotted rats** ...	... **none** *of the* **male spotted rats** *loved the food. Now that they knew that* **none** *of the* **spotted rats** ...
IDENTICAL → to spotted rats →	... **some** *of the* **spotted rats** *loved the food. Now that they knew that* **some** *of the* **spotted rats** ...	... **not all** *of the* **spotted rats** *loved the food. Now that they knew that* **not all** *of the* **spotted rats** ...	... **all** *of the* **spotted rats** *loved the food. Now that they knew that* **all** *of the* **spotted rats** ...	... **none** *of the* **spotted rats** *loved the food. Now that they knew that* **none** *of the* **spotted rats** ...
SUPERSET → of spotted rats →	... **some** *of the* **rats** *loved the food. Now that they knew that* **some** *of the* **spotted rats** ...	... **not all** *of the* **rats** *loved the food. Now that they knew that* **not all** *of the* **spotted rats** ...	... **all** *of the* **rats** *loved the food. Now that they knew that* **all** *of the* **spotted rats** ...	... **none** *of the* **rats** *loved the food. Now that they knew that* **none** *of the* **spotted rats** ...

- Trivially entailed
- Entailed
- Not Entailed

Table 3: Trial types in Experiment 2. Depending on the combination of the quantifier and the containment, there were four conditions where the premise was identical to the conclusion, and so it Trivially Entailed it (blue), four conditions where the premise differed from but Entailed the conclusion (green), and four conditions where the conclusion was Not Entailed by the premise (orange).

Study design

Experiment 2 used a line-by-line self-paced reading task similar to Experiment 1. However, in Experiment 2, the task was modified to test for the capacity to detect unlicensed inferences in the absence of strict contradictions (Figure 9). In each item, the first three lines (the *background*) and line 6 (the *resolution*) were identical to Experiment 1. Lines 4 and 5 (the premise and the conclusion) contained one of four quantifiers—*some, all, none,* or *not all*. Unlike in Experiment 1, the quantifier was kept constant between the premise and the conclusion. We manipulated whether the noun phrase in the subject of the *premise* appeared with two modifiers, one modifier, or no modifiers. The subject noun in the conclusion always appeared with one modifier. Thus, the subject noun phrase in the premise was a subset (*male spotted rats* $\subset$ *spotted rats*), identical to (*spotted rats* = *spotted rats*), or a superset (*rats* $\supset$ *spotted rats*) of the subject noun phrase in the conclusion. The quantifiers *some* and *not all* are upward-entailing [$\uparrow$] on the subjects of these sentences, licensing inferences from a subset to any superset that includes it, while the quantifiers *all* and *none* are downward-entailing [$\downarrow$], licensing inferences in the opposite direction, from a set to any subset. The combination of four quantifiers and three containment relations yielded four instances of three trial types (Table 3): four identity trials, where the premise was identical to the conclusion, and so Trivially Entailed it; four trials where the premise differed from, but still Entailed the conclusion; and four trials where the conclusion was Not Entailed by the premise.

This design allows us to compare the reading times of conclusion lines that have exactly the same lexical content, but vary in whether that content is logically entailed. For example, the same conclusion line, "Now that they knew that some of the spotted rats loved the food," can be Trivially Entailed, non-trivially Entailed, or Not Entailed, depending on whether the premise that preceded it mentioned *spotted rats*, *male spotted rats*, or *rats*, respectively. Comparing RTs on the conclusion lines, any differences between trial types cannot be due to differences in the

conclusion lines themselves, but must instead be due to the logical relation of that conclusion to the preceding premise.

Of course, narratives regularly introduce new content that does not deductively follow from information given earlier. After all, if people do draw logical inferences while reading, any story that did not introduce unentailed information would risk being very boring indeed. To create a narrative in which new unentailed information would specifically impede comprehension, we aimed to ensure that the conclusion in line 5 would be understood as previously given information, having been entailed by the premise in line 4. We designed the beginning of line 5 ("Now that they knew that...") with this goal in mind. This phrase is a *presupposition trigger*, which conveys that the information that follows it is presupposed to be true given information in the preceding discourse (*73*). Studies of presupposition processing show that when a presupposition trigger is followed by content that has not been previously introduced, people are slower to read it compared to information that has already been given (*57, 74, 75*). Here, we leverage these findings to test whether unentailed content similarly clashes with the presupposition trigger and leads readers to slow down when they encounter it. Note that, while the same presupposition trigger was also present in Experiment 1, its theoretical role there in that experiment is less important. Unlike information that is only not entailed, information that directly contradicts what preceded it would be expected to impede comprehension even without a presupposition trigger.

For target items, we used the same 12 vignettes as in Experiment 1 (Appendix A), designed to tell coherent narratives in combination with any trial type (except Not Entailed). Each vignette was fully crossed with each combination of 4 Quantifiers and 3 Containment relations to create $12 \times 12 = 144$ target items. We created 12 lists of 12 target items each, such that each of the 12 combinations of Quantifier and Containment appeared in a different vignette within a list. Each participant saw the target items from one list. Thus, within a participant, each vignette appeared once, and each trial type (Trivially Entailed, Entailed, Not Entailed) appeared

four times. Between participants, every trial type appeared in every vignette. All participants also saw the same 18 fillers as in Experiment 1, again presented to participants in a dynamically generated pseudo-random order.

Results

Data Exclusion Following our preregistration, we excluded 15 individual trials with RTs on conclusion lines faster than 100 ms and 1469 trials with RTs slower than 10 seconds. We excluded an additional 1337 trials with RTs +/- 3 S.D. away from the mean for that line in that vignette. The remaining data included 26,592 trials on conclusion lines from 384 participants.

Analyses Except where otherwise noted, our analysis strategy was identical to Experiment 1 and follows our preregistration. A priori, we expected a maximal random effects structure that includes estimates of the full covariance matrix to lead to an overfit model. We therefore started with a model that excluded correlations between random effects (maximal-zero-correlation model), and reduced it further as guided by the rePCA() function in lme4. We then extend the final reduced model with a random correlation parameter to check for improvement in relative model fit.

As in Experiment 1, the Trivially Entailed trials in which the conclusion and the premise were identical should be faster than all other trials, both because they require no inference and because they might be subject to lexical priming from the preceding premise line. Checking whether participants read these conclusions faster than other trial types serves as a validation of the present method. As expected, we find that RTs on the Trivially Entailed trials are faster than on other trials $\chi^2(2) = 70.69, p < 0.001$). The rest of our analyses set aside the Trivially Entailed trials, comparing the non-trivially Entailed to the Not Entailed trials.

Our main question is whether participants are slower to process conclusions that are expected to be, but are not entailed by their premises compared to conclusions that are non-trivially entailed. Figure 10 shows these results. While there were 12 trial types, produced by combinations of 4 Quantifiers and 3 Containment relations, this question reduces the design into a much simpler 2 (Containment: Subset vs. Superset) X 2 (Quantifier: Upward- vs. Downward-Entailing). The Containment variable encodes whether the premise describes a subset, a superset, or the same set as the conclusion variable, but only the first two are relevant to comparing non-trivially entailed to unentailed conclusions. With respect to the quantifiers, *some* and *not all* are upward-entailing on their first argument, licensing inferences from a smaller subset to the conclusion, while *none* and *not all* are downward-entailing on that argument, licensing the reverse inference from a larger superset to the same conclusion.

To investigate whether participants took longer to read unentailed conclusions, we group the four quantifiers into a binary Entailment Direction variable (Up vs. Down) and look for its interaction with Containment (Subset vs. Superset). Following the random effect selection procedure above, our final model included only random intercepts by subject and item. We find that the interaction term significantly improves model fit ($\chi^2(2) = 10.99$, $p < 0.001$), with longer RTs on the Not Entailed trials. We also find a highly significant main effect of Containment ($\chi^2(2) = 24.92$, $p < 0.001$), with faster RTs when the preceding premise mentioned a subset of the noun phrase in the conclusion (*male spotted rats* $\prec$ *spotted rats*) compared to the cases where the premises mentioned a superset (*rats* $\prec$ *spotted rats*). This effect is consistent with lexical repetition priming (*76, 77*). When the premise had mentioned a subset, all of the lexical material in the conclusion is subsequently reduplicated, which seems to facilitate its access and integration into the sentence. In contrast, when the premise mentions a superset, the conclusion then contains a modifier that participants are seeing for the first time (e.g. *spotted*). Thus, the finding that conclusions following subset premises are read faster is both theoretically and statistically

independent of participants' processing of the logical relation between premise and conclusion. Finally, we also find a marginal main effect of Entailment Direction ($\chi^2(2) = 2.74$, $p = 0.098$), with the upward-entailing quantifiers (*some*, *not all*) taking marginally longer to read than the downward-entailing ones (*none*, *all*). Whether robust or not, this effect is not theoretically relevant for our purposes – it only reflects average differences in lexical access between these pairs of quantifiers, independent of any logical relation between the premises and conclusions.

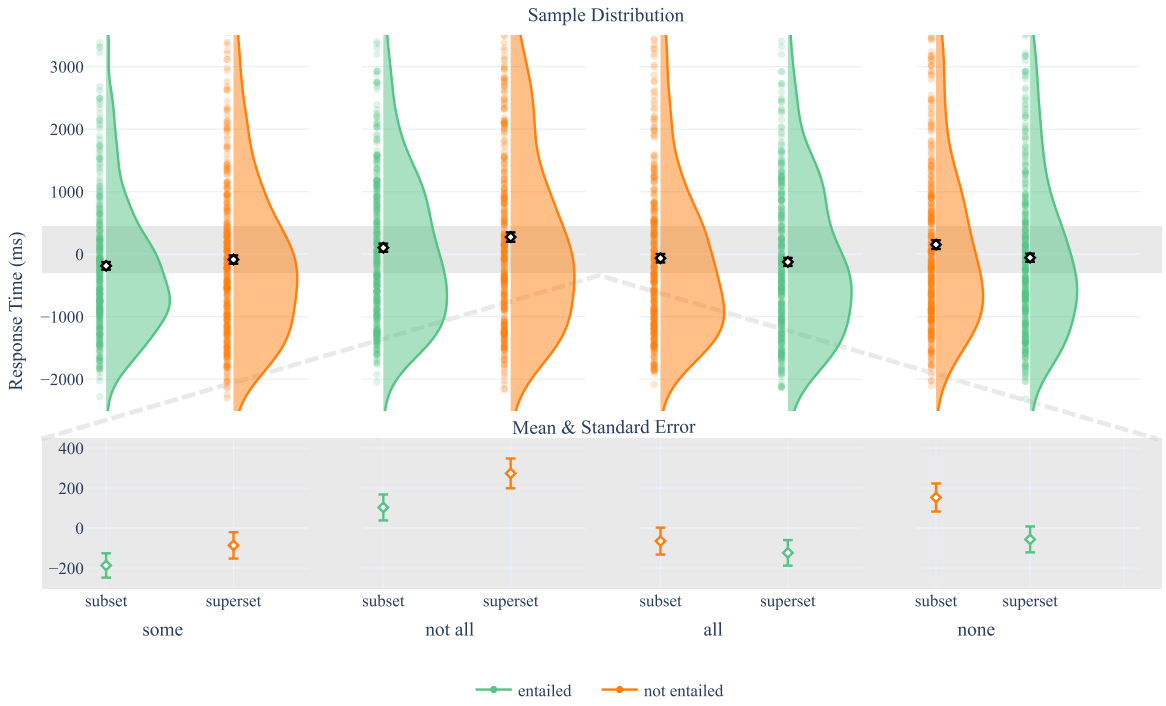


Figure 10: Experiment 2 results, partial residual graphs. [Subset: *male spotted rats* $\prec$ *spotted rats* , Superset: *rats* $\prec$ *spotted rats*]. The y-axis shows partial residual response times. These are computed by subtracting the response time within each combination of Quantifier and Containment from the mean of the corresponding level of the Containment variable. This visualizes how response times to each Quantifier differed above and beyond the main effect of Containment by showing cell means after subtracting out that main effect, illustrating the theoretically critical interaction that reflects whether each conclusion was entailed by the preceding premise or not.

Our preregistered plan was to examine the simple effects of the same Containment variable (Subset vs Superset) for each quantifier separately, coding each quantifier as a unique combi-

nation of Entailment Direction and a binary dummy variable of Negation (grouping Not All with None). The strongest possible result in support of our hypothesis would be a symmetrical crossover interaction, with simple effects of Containment going in one direction – with the conclusions in Subset vignettes read slower than in Superset vignettes for each of the upward-entailing quantifiers (None, All) – and vice versa for each of the upward-entailing (Not All, Some). However, we anticipated in our preregistration that this best-case scenario would not come to pass if we found – as indeed we did – a large main effect of Containment. In particular, the main effect of Containment, with Subset vignettes read faster than Superset vignettes, was larger than the interaction. This effectively exaggerates the effect in the upward-entailing quantifiers (*some*, *not all*) and masks it in the downward-entailing quantifiers (*none*, *all*). Given this main effect, the interpretation of the simple effects of Containment within each quantifier cannot be taken to reflect the logical inference directly, an important caveat to keep in mind when interpreting the simple effects we report below.

In analysing simple effects, we deviated from our preregistered plan, reasoning that it would be simpler to interpret a word-specific variable of Quantifier (Some, All, Not All, None) interacting with Containment than contrast coding these 4 quantifiers into two fully crossed variables. We therefore created 4 versions of the same model with the interaction of Containment $\times$ Quantifier, varying which quantifier is dummy-coded as the reference level. This approach allowed us to analyze simple effects, while including variance from the rest of the data in estimates of the random effects of subject and vignette. Unlike the main model above, we are not interested in the interaction term, but in whether the coefficient for the Containment variable is significant in the predicted direction in each of these analyses. With 4 comparable models including random intercepts for subjects and items, we find significant effects of Containment within each of the upward-entailing quantifiers (Some: $\beta = 0.11$, t = 4.17, $p < 0.001$; Not All ($\beta = 0.11$, t = 4.27, $p< 0.001$), a small but significant effect in the opposite direction than expected

in the downward-entailing All (β = 0.05, t = 2.01, p = 0.04; this is a case of the main effect of containment dominating the interaction), and no effect in the downward-entailing None (β = ≤ 0.01, t = -0.33, p = 0.74).

Experiment 3

The preregistration for Experiment 3 can be found on OSF (link anonymized for peer review). Except where otherwise noted, methods were identical to Experiment 2.

Methods

Participants

We tested another sample of 400 self-reported native English speakers, none of whom had participated in Experiments 1 or 2. Recruitment and compensation were identical. Seven participants were excluded for failing to answer the attention checks correctly and were not replaced.

Study design

Experiment 3 modifies the design of Experiment 2. Instead of manipulating the containment relation (*identical set*, *subset*, *superset*) between the subjects of the *premise* and of the *conclusion*, Experiment 3 manipulates the same relations between their predicates (Figure 11).

(1) *A group of scientists wanted to know what rats liked to eat.*
(2) *They gave rats a choice of different meats,*
(3) *as well as leafy and root vegetables, both fresh and frozen.*
(4) *They discovered that* QUANT *of the rats ate ((frozen) leafy) vegetables.*
(5) *Now that they knew that* QUANT *of the rats ate leafy vegetables,*
(6) *they decided to issue a recommendation based on their findings.*

Figure 11: An example Experiment 3 item.

This creates three trial types (Trivially Entailed, Entailed, Not Entailed), within each of the four quantifiers, just as in Experiment 2. However, this manipulation changes whether the premise entailed or failed to entail the conclusion in exactly half of the cases. Because *some* is upward-entailing with respect to both its subject and its predicate [↑, ↑], while *none* is downward-entailing in both [↓, ↓], the pattern of entailed conclusions in trials with *some* and *none* stays the same between Experiments 2 and 3. In contrast, the trials with *not all* and *all* flip which conclusion is entailed between experiments. While *all* licenses downward entailing inferences from set to subset in its subject (e.g. *All of the spotted rats liked* ... entails *All of the male spotted rats liked* ...), it licenses upward entailing inferences from set to superset in its predicate (e.g. ... *liked leafy vegetables* entails ... *liked vegetables*, but not ... *liked frozen leafy vegetables*; i.e. [↓, ↑]. Vice versa for *not all*, which is upward entailing on its subject and downward entailing on its predicate [↑, ↓] (Table 4).

	some	*not all*	*all*	*none*
SUBSET → to leafy veg. →	... **some** *of the rats ate* **frozen leafy vegetables**. *Now that they knew that* **some** *of the rats ate* **leafy vegetables** ...	... **not all** *of the rats ate* **frozen leafy vegetables**. *Now that they knew that* **not all** *of the rats ate* **leafy vegetables** ...	... **all** *of the rats ate* **frozen leafy vegetables**. *Now that they knew that* **all** *of the rats ate* **leafy vegetables** ...	... **none** *of the rats ate* **frozen leafy vegetables**. *Now that they knew that* **none** *of the rats ate* **leafy vegetables** ...
IDENTICAL → to leafy veg. →	... **some** *of the rats ate* **leafy vegetables**. *Now that they knew that* **some** *of the rats ate* **leafy vegetables** ...	... **not all** *of the rats ate* **leafy vegetables**. *Now that they knew that* **not all** *of the rats ate* **leafy vegetables** ...	... **all** *of the rats ate* **leafy vegetables**. *Now that they knew that* **all** *of the rats ate* **leafy vegetables** ...	... **none** *of the rats ate* **leafy vegetables**. *Now that they knew that* **none** *of the rats ate* **leafy vegetables** ...
SUPERSET → of leafy veg. →	... **some** *of the rats ate* **vegetables**. *Now that they knew that* **some** *of the rats ate* **leafy vegetables** ...	... **not all** *of the rats ate* **vegetables**. *Now that they knew that* **not all** *of the rats ate* **leafy vegetables** ...	... **all** *of the rats ate* **vegetables**. *Now that they knew that* **all** *of the rats ate* **leafy vegetables** ...	... **none** *of the rats ate* **vegetables**. *Now that they knew that* **none** *of the rats ate* **leafy vegetables** ...

- Trivially Entailed
- Entailed
- Not Entailed

Table 4: Trial types in Experiment 3, color coded by their entailment relation.

Taken together, Experiments 2 and 3 create a 2 × 2 factorial design. The crossed factors are whether the direction of entailment that a given quantifier licenses is up (from set to superset) or down (from set to subset), and whether this direction stays the same between experiments (*some*, *none*) or reverses across experiments (*all*, *not all*).

Results

Data Exclusion Following our preregistration, we excluded 32 individual trials with RTs on conclusions lines faster than 100 ms and 1236 trials with RTs slower than 10 seconds. We excluded an additional 1253 trials with RTs +/- 3 S.D. away from the mean for that line in that vignette. The remaining data included 27,304 trials on vignette lines from 393 participants.

Analyses Except where otherwise noted, our analysis strategy was identical to Experiment 2 and follows our preregistration.

Figure 12 shows the results of Experiment 3. As in Experiments 1 and 2, Trivially Entailed conclusions, in which the conclusion and the premise were identical, were read faster than all other trials, reflecting both the repetition of the lexical material between the two lines and the lack of need to make any further inference ($\chi^2(2) = 39.60$, $p < 0.001$). Next, as in Experiment 2, the rest of our analyses set aside the Trivially Entailed trials, comparing the non-trivially Entailed to the Not Entailed trials.

Once again, as in Experiment 2, our main question is whether participants are slower to process sentences where entailment relations have been violated than those where the premise non-trivially entails the conclusion. This question again reduces the 12 distinct combinations of Quantifier and Containment to a simpler 2 (Containment: Subset vs. Superset) X 2 (Entailment Direction: Upward- vs. Downward-Entailing) design. The main difference from Experiment 2 is that two out of the four quantifiers switch the entailment direction in which they license

inference, so that in Experiment 3 we group *all* with *some* as the Upward-Entailing quantifiers and *none* with *not all* as the Downward-Entailing. Otherwise, the model we fit to this data is identical to Experiment 2: including Containment, Entailment Direction, and their interaction. Following the random effect selection procedure above again led to our final model including only random intercepts by subject and by item.

As in Experiment 2, we again find a significant interaction between Containment and Entailment Direction ($\chi^2(1) = 7.18$, $p = 0.007$), with conclusions that are Not Entailed taking longer to read. We also find a significant main effect of Containment ($\chi^2(1) = 3.98$, $p = 0.05$), with faster reading times when the preceding premise mentioned a subset of the predicate noun phrase that subsequently appears in the conclusion (*frozen leafy vegetables* $\rightarrow$ *leafy vegetables*) compared to when the premise mentioned a superset (*vegetables* $\rightarrow$ *leafy vegetables*). As in Experiment 2, we interpret this effect as reflecting lexical repetition priming, with greater priming when more of the words in the conclusion have previously appeared in the premise, independent of the logical relation between the two lines. Finally, we also find a significant main effect of Entailment Direction ($\chi^2(1) = 23.07$, $p < 0.001$), with conclusions containing the downward-entailing quantifiers (*none*, *not all*) taking longer to read than the upward-entailing ones (*some*, *all*). This effect may reflect slower lexical access to some quantifiers than others or the generally slower processing of negative operators, but it too is independent of the varying logical relations between the different premises and conclusions.

To analyze the simple effect of Containment within each of the four quantifiers, we deviate from our preregistration in the same way and for the same reasons as for Experiment 2, above. We again created 4 versions of the same model with the interaction of Containment $\times$ Quantifier, varying which of the four quantifiers is dummy-coded as the reference level. With four comparable models including random intercepts for subjects and items, we find significant effects of Containment within one of the quantifiers that are upward-entailing on the predicate (All:

$\beta = 0.07, t = 3.16, p = 0.002$), but not the other (Some: $\beta = 0.035, t = 1.5, p = 0.126$). There was no significant effect of Containment within the downward-entailing quantifiers (None: $\beta \leq -0.0001, t = -0.02, p = 0.98$; Not All: $\beta = -0.02, t = -0.70, p = 0.49$). As was in Experiment 2, the simple effects here proved uninformative, illustrating only that the main effect of Containment within each quantifier drowns out the relevant interaction.

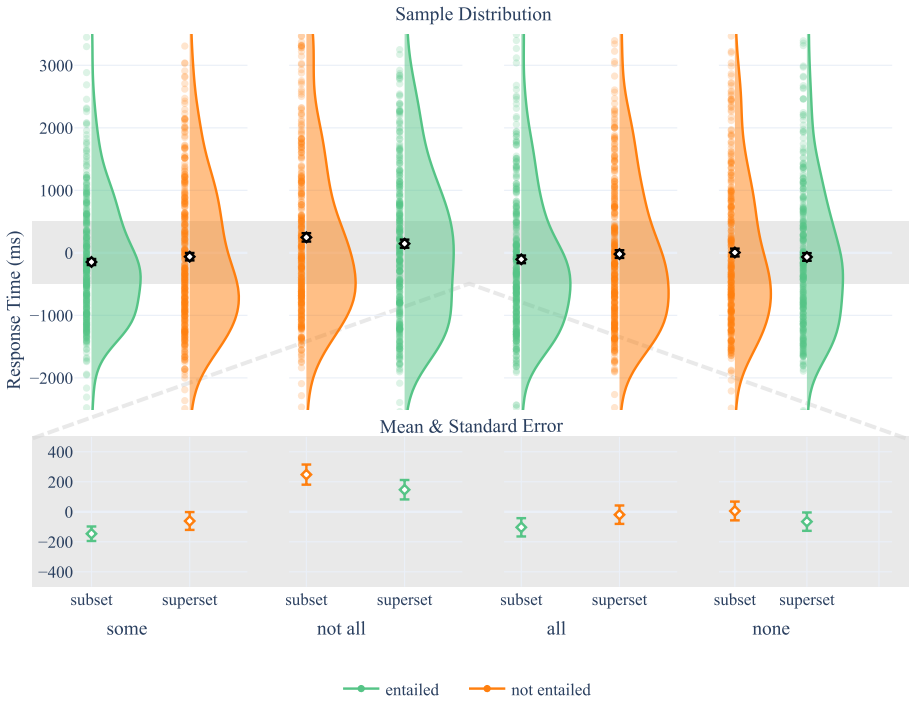


Figure 12: The results of Experiment 3. The y-axis shows partial residual reading times, derived by subtracting the reading time in each combination of Quantifier and Containment from the main effect of Containment. [subset: *male spotted rats* $\prec$ *spotted rats* , superset: *rats* $\prec$ *spotted rats*].

Comparison of Experiments 2 and 3 Recall that both *all* $[\downarrow, \uparrow]$ and *not all* $[\uparrow, \downarrow]$ license opposite inferences with respect to the subjects of sentences (e.g. *spotted rats...*) and the predicates (e.g. *...liked leafy vegetables*), while *some* $[\uparrow, \uparrow]$ and *none* $[\downarrow, \downarrow]$ license inferences in the same direction for subjects as for predicates (see Tables 3 and 4). Experiment 2 tested inferences over

relations between the subjects of sentences, and Experiment 3 tested inferences over relations between the predicates. A further test of whether people engage in automatic, correct reasoning is whether the interaction between the containment relations (Subset vs. Superset), and the quantifiers *all* and *not all* change between Experiments 2 and 3, while the interactions between Containment and *some* and *none* stay the same in both Experiments.

Our preregistered plan was to test for a three-way interaction between Containment, Experiment, and the Directionality of the inference licensed by the quantifier on the subject and the predicate (Same: *some* and *none* vs Opposite: *all* and *not all*). However, we realized after analyzing the data this way that this coding scheme is invalid. Because the inferences that are licensed change both *within* each level of the Directionality variable and *between* Experiments, the three-way interaction term we were interested in should actually cancel out in this coding scheme. That is, instead of predicting a significant three-way interaction, as we intended, normatively correct reasoning on each trial type in both experiments would actually predict a three-way interaction effect of exactly zero.

We therefore took a different approach to comparing the results of Experiments 2 and 3. First, we combined the data from Experiments 2 and 3 and added the variable of Experiment and all accompanying interactions to the model specifications we had used to analyze each of these experiments separately, with the same random intercepts for subject and item. The resulting fixed effects were Containment (Subset vs. Superset) $\times$ Quantifier (4 Quantifiers, treatment-coded with Some as the reference level) $\times$ Experiment (2 vs. 3). We then conducted two linear hypothesis tests on the coefficients of this model. One test groups together *all* and *not all*, the two quantifiers that license inferences in opposite directions across Experiments 2 and 3. It creates a restricted model in which the coefficient for the three-way interaction between Containment, Experiment, and Some-vs-All equals the coefficient for the three-way interaction between Containment, Experiment, and Some-vs-NotAll, and then tests this restricted model against the

full model fit to the data. A Chi-sq test comparing these two models revealed a significant difference ($\chi^2(1) = 9.17$, $p = 0.002$) between the restricted and full model, indicating that the effect of Containment did not vary between the two Experiments in the same way for *not all* as for *all*. The other linear hypothesis test grouped together the two quantifiers that have the same direction of entailment in both experiments, *some* and *none*, and asked the same question. This Chi-sq test found that the coefficient for the three-way interaction between Containment, Experiment, and None-vs-Some was not significantly different from zero ($\chi^2(1) = 2.53$, $p = 0.112$), indicating that the effect of Containment did not change between Experiments differently for *some* than for *none*.

These results suggest that participants reasoned exactly in accordance with the logical entailment profiles of all four quantifiers: for *some* and *none*, they slowed down upon reading the same logically invalid inferences about the subjects of the conclusion sentences as about their predicates. For *all* and *not all*, the direction of inference that is valid with respect to the subject is invalid with respect to the predicate, and vice versa. Participants correspondingly slowed down when reading just the invalid inferences in these cases, such that whether the delay in reading time was caused by inferences from subset to superset or from superset to subset varied according to which direction would violate the logically normative inference for which quantifier.

Experiment 4

Methods

Models We use two language models to measure the distributional probability estimates of Experiment 1-3's vignettes: (i) GPT-3 (`text-davinci-002`) (*66*) as one of the largest available language models tuned solely on a next-word prediction task, and (ii) the smallest GPT-2 (124M) (*66*) given recent work (*60*) that it aligns with human reading times better than GPT-3. Both these models allow the log-probability estimates of each token to be extracted directly.

We acknowledge that these models are no longer state-of-the-art on general tasks given the advent of models like Llama2 (*78*), Claude (*79*), Alpaca (*80*), and most notably, GPT-4 (*81*). However, these newer models are no longer models of textual distributional information available as they are further tuned on additional tasks that are not simply next-word prediction (e.g. instruction-tuning, cross-modal training, reinforcement learning from human feedback). As a result, this class of models are poorer predictors of human reading times than models trained only on next-word prediction (*67*). Furthermore, with GPT-4 researchers no longer have direct access to models' probability estimates of any given word, making it impossible to perform the same analysis.

Surprisal computation We input each full vignette from Experiments 1-3 into the two models, and extract the log-probabilities of tokens in the conclusion sentence. Each vignette is input independently, and models do not receive any updates after reading a vignette. For GPT-3, the extraction of log-probability values was done through the OpenAI Completions API with `echo = True, max_tokens = 0,` and `temperature = 1`.

We calculate the surprisal of the conclusion sentence by taking the negative sum of extracted log-probabilities of tokens in the conclusion sentence. Surprisal ($\mathbb{S}$) is the negative base-2 logarithm of probability and has been found to hold a linear relationship with human reading times (*60*). The procedure of computing the surprisal estimate for a sentence by summing its constituent tokens parallels the procedure of (*82*), where surprisal estimates for words are computed by summing the surprisal of constituent sub-word tokens. The procedure is described in Equation 1:

$$
\begin{aligned}
\mathbb{S}(S_n) &= -\log_2(\mathbb{P}(w_1) \cdot \mathbb{P}(w_2|S_1) \ldots \mathbb{P}(w_n|S_{n-1})) \\
&= \sum_{i=1}^{n} -\log_2 \mathbb{P}(w_i|S_{i-1}) \qquad (1) \\
&\text{where } S_n = \{w_1, ...w_n\}\text{, a sentence of } n \text{ words} \\
&\quad\ w_n \text{ is the } n^{th} \text{ word}
\end{aligned}
$$

Results

We parallel the analyses conducted for human reading times by applying the same effect structure to instead predict surprisal values on target items. Random effects in this case involve only the vignette.

Experiment 1 vignettes We again build a regression including an Identity variable to compare the surprisal values on Identity vignettes with the average of vignettes in the Entailment and Contradiction contradiction. Like human reading time results, we find significantly lower surprisal values assigned to Identity trials (Table 5).

	GPT-2		GPT-3	
Effect	χ^2	p	χ^2	p
Identity	84.183	< 0.001(***)	106.87	< 0.001(***)

Table 5: Significance of being in the Identity condition to surprisal values from Experiment 1 vignettes.

In modeling the binary effect of Trial Type (Contradiction vs. Entailment), Conclusion quantifier (Some vs. Not All) and their interaction (Table 6) however, while human reading times demonstrated a highly significant effect of Trial Type, surprisal values lack any significant effect from Trial Type. There is only a similarly significant effect of the Quantifier, observed in GPT-2's surprisal values. Parallel to how participants read conclusions with *not all* slower than

with *some* , GPT-2 on average assigns 0.823 bits more surprisal to conclusions with *not all* than those with *some*. This suggests that the reading time slowdown based on quantifiers can be partly accounted for by distributional information. However, within each quantifier separately, there is still no effect of trial type (Table 7).

	GPT-2		GPT-3	
Effect	χ^2	p	χ^2	p
Trial Type	0.5786	0.4469	1.3789	0.2403
Quantifier	84.7661	< 0.001(***)	0.2524	0.6154
Interaction	0.0389	0.8436	0.7809	0.3769

Table 6: Significance of each effect to modeling surprisal values from Experiment 1 vignettes.

	GPT-2		GPT-3	
Quantifier	t	p	t	p
some	0.494	0.6213	0.181	0.8567
not all	0.638	0.5236	1.399	0.1619

Table 7: Simple effect of Trial Type within each quantifier condition.

Experiment 2 and 3 vignettes As with human reading times, surprisal values are significantly lower on Trivially Entailed trials (Table 8). We set these trials aside to compare the non-trivially Entailed and Not Entailed trials for each experiment. We proceed to fit the same effect structure as was performed on human reading time data—using fixed effects of Entailment Direction (Up vs. Down), Containment (Subset vs. Superset) and their interaction. (Table 9). Parallel to human reading times, containment continues to have a highly significant main effect on surprisal values for both experiments. These suggest that lexical repetition effects are, perhaps unsurprisingly, predicted by distributional information.

For Experiment 2, Entailment Direction has marginal effects for both models—with a marginal main effect for GPT-2 and a slightly stronger main effect for GPT-3—but the direction of the effect is in opposite directions between the two language models. GPT-2 assigns more bits

	GPT-2		GPT-3	
Exp	χ^2	p	χ^2	p
Experiment 2	74.481	< 0.0001(***)	36.34	< 0.0001(***)
Experiment 3	82.658	< 0.0001(***)	59.109	< 0.0001(***)

Table 8: Significance of being in the Trivially Entailed condition, relative to either of the other conditions, predicting surprisal values on vignettes from Experiment 2 and 3.

	GPT-2		GPT-3	
Experiment 2				
Effect	χ^2	p	χ^2	p
Entailment Direction	4.7118	0.02996(*)	7.9944	0.0046 (**)
Containment	113.8404	< 0.0001(***)	22.9583	< 0.0001(***)
Interaction	0.0240	0.87689	0.0139	0.906296
Experiment 3				
Effect	χ^2	p	χ^2	p
Entailment Direction	10.3689	0.0012(**)	9.9132	< 0.0016(**)
Containment	151.7381	< 0.0001(***)	16.0478	< 0.001(***)
Interaction	0.5313	0.466045	0.8807	0.348004

Table 9: Significance of each fixed effect, predicting surprisal values on vignettes from Experiment 2 and 3.

of surprisal to conclusion lines with upward-entailing quantifiers (*some, none*) than downward-entailing quantifiers (*all, not all*), which appears to follow the pattern from human reading times. By deviation coding each quantifier, however, we observe that this is because GPT-2 assigns the least surprisal to *all* by such a large margin ($\beta = -0.92, p \leq 0.001$) that even when the greatest surprisal is assigned to *not all* ($\beta = 0.60, p = 0.006$), the average surprisal for downward-entailing quantifiers is less than that of upward-entailing quantifiers (none: $\beta = 0.33, p = 0.12$; some: $\beta = -0.02, p = 0.91$). GPT-3 shows the opposite pattern in assigning less bits of surprisal to upward-entailing quantifiers than downward-entailing. GPT-3's surprisal values diverge from the marginal effect of Entailment Direction observed for human reading times, where upward-entailing quantifiers took marginally longer to read.

The main effect of Entailment Direction for Experiment 3, however, is significant in a uniform direction in Experiment 3 for both models. Both GPT-2 and GPT-3 assign higher surprisal values to conclusion lines with downward-entailing quantifiers (*all, none*) than upward-entailing quantifiers (*some, not all*). This exactly matches the significant main effect of Entailment Direction observed for human reading times in Experiment 3, where downward-entailing quantifiers took longer to read. These results again support the interpretation that the main effects of Entailment Direction observed for human reading times in Experiment 2 and 3 are due to distributional information.

Importantly, the interaction terms for both experiments across both models were not significant. That is, whether the conclusions are Entailed or Not Entailed did not significantly affect the surprisal values either model assigned to them. This runs counter to the pattern of human reading time data, where the interaction term had a significant effect in both Experiment 2 and 3.

Overall, the model experiments support the conclusion that some components of human participants' reading times are indeed best explained by distributional information – both lexical repetition effects and specific quantifiers being read faster than others. In these cases we find that both model surprisal values and reading times share the same significant main effects. However, while whether a conclusion was entailed or not consistently had a strong effect on reading time, it did not have any significant effect on model surprisal values. This in turn supports the conclusion that reading time delays caused by the logical relations between premises and conclusions cannot be attributed to distributional information.

Appendix A: Target items

Experiment 1

Item 1

A group of scientists wanted to know whether spotted rats,
who are pickier eaters than other rats, liked a new kind of food.

They tested white, black, and spotted rats of both sexes.
The scientists discovered that some/all/none/not all of the rats loved the food.
Now that they knew that some/not all of the rats loved the food,
they decided to issue a recommendation based on their findings.

comprehension question: The researchers studied rodents.
correct answer: True

Item 2

A local furniture store sells both white and black metal chairs,
but most of their sales come from handcrafted wooden chairs and tables.
When the manager looked at the most recent sales report,
he saw that some/all/none/not all of the chairs were sold.
Now that he knew that some/not all of the chairs were sold,
he adjusted the store's order for next month.

comprehension question: The store won't ever get new furniture.
correct answer: False

Item 3

A watchmaker in Beijing fixes all kinds of watches, including antiques.
Rarely, customers bring European and American pocket watches,
but he mostly deals with wrist watches.
Recently, his apprentice told him that some/all/none/not all of the watches he got were antiques.
Now that the watchmaker knew that some/not all of the watches he got were antiques,
he realized it would take him a few hours to fix them.

comprehension question: The repair store is located in the US.
correct answer: False

Item 4

A college town bookstore sells some fiction,
including both American and foreign fiction,
but a lot of their business is selling textbooks.
This month, the accountant told the owner that some/all/none/not all of the books sold out.
Now that the owner knew that some/not all of the books sold out,
he changed how they'd advertise around campus next week.

comprehension question: There are students in the town where the bookstore is located.
correct answer: True

Item 5

In one national forest, there were a couple of rare pines:
a few in the north and a few more in the south.
Most of the other trees were oaks. After a recent fire,
a Forest Service worker heard on the radio that some/all/none/not all of the trees were burnt.

Now that he knew that some/not all of the trees were burnt,
he decided to inspect that part of the forest as soon as possible.

comprehension question: There were various tree species in the national forest.
correct answer: True

Item 6

Alice has an impressive collection of hats. She has a few formal hats,
some colorful and some black, but most of her hats are casual.
Alice's daughter often takes her hats and plays with them.
Yesterday, Alice's housekeeper told Alice that some/all/none/not all of the hats were gone.
Now that she knew that some/not all of the hats were gone,
She wondered what her daughter was doing at that time.

comprehension question: Alice is an only child.
correct answer: False

Item 7

Emery's Catering Service served food and drinks at a university event.
They served a few dry red wines and a few sweet red wines,
but they mostly served different varieties of white wine.
After the event, a bartender told management that some/all/none/not all of the wine had run out.
Now that they knew that some/not all of the wine had run out,
they would adjust what drinks they serve in the future.

comprehension question: Alcohol was served at the event.
correct answer: True

Item 8

Jack plays a lot of games. He has some board games, both classic and newer ones,
But mostly he has a lot of great collectible card games.
Jack's friends often borrow his games without permission.
Last night, Jack's roommate told him that some/all/none/not all of the games were gone.
Now that Jack knew that some/not all of the games were gone,
he could guess which of his friends had come by.

comprehension question: Jack lives alone.
correct answer: False

Item 9

An aquatic pet store, which sold mostly common freshwater fish,
recently also stocked Tropical and Northern saltwater fish.
Last night, the water filters went haywire. This morning,
the store's employees reported to the owner that some/all/none/not all of the fish survived.
Now that the owner knew that some/not all of the fish survived,
she reconsidered what fish to stock next time.

comprehension question: Equipment at the pet store never malfunctioned.
correct answer: False

Item 10

A big car company makes both gas-powered and electric cars.
They mostly make commuter cars, but also some luxury sports cars.
Recently, they had to test their cars for passanger safety.
The tests revealed that some/all/none/not all of the cars failed the safety standards.
Now that they knew that some/not all of the cars failed the safety standards,
they found themselves in a different position from their competitors.

comprehension question: The company tested their cars for safety.
correct answer: True

Item 11

A local food critic visited a new high-end restaurant,
which mostly offers meat dishes as well as
a few vegetarian dishes, with nut-free options for both.
In his review, he wrote that some/all/none/not all of the dishes were delicious.
Now that his readers knew that some/not all of the dishes were delicious,
They could revise their opinion of this new restaurant.

comprehension question: The restaurant sells high-end food.
correct answer: True

Item 12

Sue loves watching foreign movies.
She especially likes comedies and horror movies from Mexico and France,
She's not too picky about comedies, but more picky about horror movies.
The newest reviews of foreign movies claimed that some/all/none/not all of the movies are great.
Now that she knew that some/not all of the movies are great,
she radically revised her must-see list.

comprehension question: Sue's favorite movies are from Hollywood.
correct answer: False

Experiment 2

Item 1

A group of scientists wanted to know whether spotted rats,
who are pickier eaters than other rats, liked a new kind of food.
They tested white, black, and spotted rats of both sexes.
The scientists discovered that some/all/none/not all of the ((male) spotted) rats loved the food.
Now that they knew that some/all/none/not all of the spotted rats loved the food,
they decided to issue a recommendation based on their findings.

comprehension question: The researchers studied rodents.
correct answer: True

Item 2

A local furniture store sells both white and black metal chairs,
but most of their sales come from handcrafted wooden chairs and tables.
When the manager looked at the most recent sales report,
he saw that some/all/none/not all of the ((white) metal) chairs were sold.
Now that he knew that some/all/none/not all of the metal chairs were sold,
he adjusted the store's order for next month.

comprehension question: The store won't ever get new furniture.
correct answer: False

Item 3

A watchmaker in Beijing fixes all kinds of watches, including antiques.
Rarely, customers bring European and American pocket watches,
but he mostly deals with wrist watches.
Recently, his apprentice told him that some/all/none/not all of the ((European) pocket) watches he got were antiques.
Now that the watchmaker knew that some/all/none/not all of the pocket watches he got were antiques,
he realized it would take him a few hours to fix them.

comprehension question: The repair store is located in the US.
correct answer: False

Item 4

A college town bookstore sells some fiction,
including both American and foreign fiction,
but a lot of their business is selling textbooks.
This month, the accountant told the owner that some/all/none/not all of the ((foreign) fiction) books sold out.
Now that the owner knew that some/all/none/not all of the fiction books sold out,
he changed how they'd advertise around campus next week.

comprehension question: There are students in the town where the bookstore is located.
correct answer: True

Item 5

In one national forest, there were a couple of rare pines:
a few in the north and a few more in the south.
Most of the other trees were oaks. After a recent fire,
a Forest Service worker heard on the radio that some/all/none/not all of the ((northern) pine) trees were burnt.

Now that he knew that some/all/none/not all of the pine trees were burnt,
he decided to inspect that part of the forest as soon as possible.

comprehension question: There were various tree species in the national forest.
correct answer: True

Item 6

Alice has an impressive collection of hats. She has a few formal hats,
some colorful and some black, but most of her hats are casual.
Alice's daughter often takes her hats and plays with them.
Yesterday, Alice's housekeeper told Alice that some/all/none/not all of the ((black) formal) hats were gone.
Now that she knew that some/all/none/not all of the formal hats were gone,
She wondered what her daughter was doing at that time.

comprehension question: Alice is an only child.
correct answer: False

Item 7

Emery's Catering Service served food and drinks at a university event.
They served a few dry red wines and a few sweet red wines,
but they mostly served different varieties of white wine.
After the event, a bartender told management that some/all/none/not all of the ((dry) red) wine had run out.
Now that they knew that some/all/none/not all of the red wine had run out,
they would adjust what drinks they serve in the future.

comprehension question: Alcohol was served at the event.
correct answer: True

Item 8

Jack plays a lot of games. He has some board games, both classic and newer ones,
But mostly he has a lot of great collectible card games.
Jack's friends often borrow his games without permission.
Last night, Jack's roommate told him that some/all/none/not all of the ((classic) board) games were gone.
Now that Jack knew that some/all/none/not all of the board games were gone,
he could guess which of his friends had come by.

comprehension question: Jack lives alone.
correct answer: False

Item 9

An aquatic pet store, which sold mostly common freshwater fish,
recently also stocked Tropical and Northern saltwater fish.
Last night, the water filters went haywire. This morning,

the store's employees reported to the owner that some/all/none/not all of the ((Tropical) saltwater) fish survived.
Now that the owner knew that some/all/none/not all of the saltwater fish survived,
she reconsidered what fish to stock next time.

comprehension question: Equipment at the pet store never malfunctioned.
correct answer: False

Item 10

A big car company makes both gas-powered and electric cars.
They mostly make commuter cars, but also some luxury sports cars.
Recently, they had to test their cars for passanger safety.
The tests revealed that some/all/none/not all of the ((electric) sports) cars failed the safety standards.
Now that they knew that some/all/none/not all of the sports cars failed the safety standards,
they found themselves in a different position from their competitors.

comprehension question: The company tested their cars for safety.
correct answer: True

Item 11

A local food critic visited a new high-end restaurant,
which mostly offers meat dishes as well as
a few vegetarian dishes, with nut-free options for both.
In his review, he wrote that some/all/none/not all of the ((nut-free) vegetarian) dishes were delicious.
Now that his readers knew that some/all/none/not all of the vegetarian dishes were delicious,
They could revise their opinion of this new restaurant.

comprehension question: The restaurant sells high-end food.
correct answer: True

Item 12

Sue loves watching foreign movies.
She especially likes comedies and horror movies from Mexico and France,
She's not too picky about comedies, but more picky about horror movies.
The newest reviews of foreign movies claimed that some/all/none/not all of the ((French) horror) movies are great.
Now that she knew that some/all/none/not all of the horror movies are great,
she radically revised her must-see list.

comprehension question: Sue's favorite movies are from Hollywood.
correct answer: False

Experiment 3

Item 1

A group of scientists wanted to know what rats liked to eat.
They gave rats a choice of different meats,
as well as leafy and root vegetables, both fresh and frozen.
They discovered that some/all/none/not all of the rats ate ((frozen) leafy) vegetables.
Now that they knew that some/all/none/not all of the rats ate leafy vegetables,
they decided to issue a recommendation based on their findings.

comprehension question: The researchers studied rodents.
correct answer: True

Item 2

A furniture chain sells tables and wooden chairs,
But their best-selling products are white and black plastic chairs.
When a regional manager inquired about his stores' inventories,
he was told that some/all/none/not all of the stores had ((white) plastic) chairs left in stock.
Now that he knew that some/all/none/not all of the stores had plastic chairs left in stock,
he knew which chairs to order more of for next year.

comprehension question: The store won't ever get new furniture.
correct answer: False

Item 3

A college cafeteria served coffee with different kinds of dairy and non-dairy milk,
including sugar-free and sweetened almond milk.
The chef was curious what options were popular.
A staff member told him that some/all/none/not all of the students used ((sugar-free) almond) milk.
Now that he knew that some/all/none/not all of the students used almond milk,
it would make it easier to decide which other options to serve.

comprehension question: The college cafeteria only had vegan options.
correct answer: False

Item 4

A college town bookstore sells textbooks, American and Asian fiction,
as well as electronics and college-branded clothing.
This week, an accountant told the owner
that some/all/none/not all of the customers bought ((Asian) fiction) books.
Now that the owner knew that some/all/none/not all of the customers bought fiction books,
he changed how they'd advertise books around campus next week.

comprehension question: There are students in the town where the bookstore is located.
correct answer: True

Item 5

A local art store put in an order for new supplies.
They ordered canvasses, brushes, and different colors of oil and watercolor paints.
The owner asked the employee to check which paints were delivered.
The employee reported back that some/all/none/not all of the suppliers delivered ((black) oil) paint.
Now that the owner knew that some/all/none/not all of the suppliers delivered oil paint,
they will have some losses this month.

comprehension question: The art store ordered various painting supplies.
correct answer: True

Item 6

A new fashion designer in New York wanted to create accessories,
including bags, as well as tall and flat hats of different colors.
To learn about trends, the designer asked her friend about a recent fashion show.
Her friend told her that, in the show, some/all/none/not all of the models wore ((tall) brown) hats.
Now that she knew that some/all/none/not all of the models wore brown hats,
she started drafting the first designs for her new fall collection.

comprehension question: The fashion designer was male.
correct answer: False

Item 7

Emery's Catering Service served food and drinks at a university event.
They served wine and beer, including dark and light beer from both America and Europe.
Management asked a bartender to let them know what guests liked to drink.
After the event, the bartender told management that some/all/none/not all of the guests drank ((European) dark) beer.
Now that management knew that some/all/none/not all of the guests drank dark beer,
they would plan to adjust what drinks they serve next time.

comprehension question: Alcohol was served at the university event.
correct answer: True

Item 8

Jack opened a new bar with rare collectible card games,
cooperative board games, competitive board games, and karaoke.
After a month, a bartender told Jack
that when they came in, some/all/none/not all of the customers played ((cooperative) board) games.
Now that Jack knew that some/all/none/not all of the customers played board games,
he thought about how many more karaoke rooms they should set up.

comprehension question: Jack does not employ anyone.
correct answer: False

Item 9

An aquatic pet store, which sold mostly turtles and common freshwater fish,
recently also stocked more sensitive Tropical and Northern saltwater fish.
Last week, the store's employees tried new cleaning products and told the owner
that some/all/none/not all of the chemicals killed the ((Tropical) saltwater) fish.
Now that the owner knew that some/all/none/not all of the chemicals killed the saltwater fish,
she realized she needed to be even more careful in future.

comprehension question: The aquatic pet store only sold fish.
correct answer: False

Item 10

A local university started offering online and in-person physics classes.
They offered seminars at both introductory and advanced levels,
As well as some more hands-on laboratory courses.
The dean checked and saw that some/all/none/not all of the students enrolled in ((online) advanced) seminars.
Now that she knew that some/all/none/not all of the students enrolled in advanced seminars,
the department would have to reassign teaching assistants.

comprehension question: The university offered science classes.
correct answer: True

Item 11

A new music magazine covered rock and metal music. Recently, they started focusing
on progressive and industrial metal music from Norway, Sweden, and Finland.
To find out whether their readers liked this coverage, the editors ran a survey.
The survey showed that some/all/none/not all of the readers listened to ((Norwegian) progressive) metal.
Now that they knew that some/all/none/not all of the readers listened to progressive metal,
they planned to change how much coverage that genre would receive.

comprehension question: The magazine covered international music.
correct answer: True

Item 12

Traditional craftsmen, such as tailors and potters, were invited to present at a fair.
The fair showcased medieval and renaissance crafts from France and Germany.
A rich collector planned to go, and asked the organizers which crafts would be represented.
She was told that some/all/none/not all of the artisans made ((French) medieval) pottery.
Now that she knew that some/all/none/not all of the artisans made medieval pottery,
she asked whether the fair would be more diverse next year.

comprehension question: The fair centered around modern crafts.
correct answer: False

Appendix B: Filler items

Item 1

David noticed he had put on a lot of weight recently.
He examined his habits to try and figure out the reason.
He realized he'd been eating too much fast food lately.
He stopped going to burger places and started a vegetarian diet.
After a few weeks, he started to feel much better.

comprehension question: David eats a lot of bacon now.
correct answer: False

Item 2

Dan was learning how to carve a pumpkin.
He wanted to make a scary looking one.
Despite his best efforts, it always ended up looking silly.
After several tries, Dan gave up after he realized something.
Silly pumpkins were better because they wouldn't scare away the kids!

comprehension question: Dan really dislikes children.
correct answer: False

Item 3

Marcus needed clothing for a business casual event.
All of his clothes were either too formal or too casual.
He decided to buy a pair of khakis.
The pair he bought fit him perfectly.
Marcus was happy to have the right clothes for the event.

comprehension question: The pants Marcus bought were too loose.
correct answer: False

Item 4

John was a pastor with a very bad memory.
He tried to memorize his sermons many days in advance but to no avail.
He decided to learn to sing to overcome his handicap.
He then made all his sermons into music and sang them on Sundays.
His congregation was delighted and so was he.

comprehension question: John is a very traditional pastor.
correct answer: False

Item 5

Melody's parents surprised her with a trip to the big aquarium.
Melody took a nap during the two hour car ride to the aquarium.

When they arrived, Melody was energetic and excited.
At the aquarium Melody saw sharks, tropical fish and many others.
After five hours at the aquarium, Melody and her family drove home.

comprehension question: Melody slept at the aquarium.
correct answer: False

Item 6

The math teacher announced a pop quiz as class began.
While some students complained, he began passing out the quiz.
Max took out his pencil and began to work.
About 5 minutes later, he finished.
He stood up feeling confident and turned it in.

comprehension question: Max thought he did well on the quiz.
correct answer: True

Item 7

Janice was out exercising for her big soccer game.
She was doing some drills with her legs.
While working out and exercising, she slipped on the grass.
She fell down and used her wrist to break her fall.
She broke her wrist in the process and went to the hospital.

comprehension question: Janice is a sportswoman.
correct answer: True

Item 8

James got permission to use the office printer for personal business.
He printed up a stack of flyers for his band's gig on Saturday.
Sly, another of the workers, had once dreamed of being in a band.
Mad he never made it, he complained about James using the printer.
The boss told him to mind his own business and get back to work.

comprehension question: Sly never wanted to be a musician.
correct answer: False

Item 9

Andy was invited to a Halloween party.
Andy figured that for dramatic effect, he should color his hair.
Since Andy's costume was green, Andy decided on that color.
After the stylist finished the coloring, Andy regretted it.
Andy was disappointed with his new, bold, green hair color.

comprehension question: Andy wished he hadn't dyed his hair.
correct answer: True

Item 10

Sarah was on a vacation with her family enjoying the streets of Paris.
Except everywhere they went, it smelled like poop.
Finally, she jokingly asked her brother if he pooped his pants.
She was not sure what was worse, the fact that she asked or that he said yes.
He then proceeded not to do anything to fix it.

comprehension question: Sarah was in Paris.
correct answer: True

Item 11

Sally had a root canal this morning, as she had a damaged root.
After the procedure, the dentist wrote her a prescription.
She headed straight to the pharmacy to fill her medication.
She handed the prescription to the technician and waited patiently.
The technician called her name and she paid for the prescription.

comprehension question: Sally had enough money for her drugs.
correct answer: True

Item 12

Mark and Joe were brainstorming ideas for a children's show.
Mark suggested that a monster attacks the children.
Joe laughed because he thought Mark was joking.
Mark was confused because he thought it was a great idea.
Mark left in a huff for having his ideas mocked.

comprehension question: Mark and Joe were working on a show for adults.
correct answer: False

Item 13

Abby noticed toys all over her living room.
Abby immediately called her toddler to clean it up.
As she was cleaning up, they began singing the clean up song.
Abby decided to help her daughter clean up the toys.
Abby was proud that her daughter learned to clean up her toys.

comprehension question: Abby and her daughter cleaned up in the kitchen.
correct answer: False

Item 14

Soren ran through the airport, pulling his bags behind him.
The female voice above him announced final boarding for Soren's flight.
He yelled for them to wait as he neared his gate, waving his arms.
The attendant at the desk gave Soren a sad, sympathetic look.
Nearly out of breath, Soren presented his pass and boarded the plane.

comprehension question: Soren made it in time for the plane.
correct answer: True

Item 15

Walter was worried because his dog was showing a lot of aggression.
The dog recently began snapping at other dogs and even people.
Walter consulted with his veterinarian.
The vet discovered that the dog was suffering a flea infestation.
Some medication got rid of the fleas and the dog's bad mood, too.

comprehension question: Walter's dog was rabid.
correct answer: False

Item 16

Kayla parked her car in front of the convenience store.
She got out of the car, and began to walk towards the store.
Suddenly, she got hit in the head with an egg.
She looked up and saw teenagers on top of the store.
The teenagers ran away, and Kayla was left with an egg on her head.

comprehension question: Kayla wanted to do some shopping.
correct answer: True

Item 17

Bob stared in disbelief at the flooded basement.
All that could be seen were the stone arches above the doors.
The plumber told him it flooded because the sump pump was off.
Bob understood that was because the electricity had also been off.
He enabled the electricity and the pump drained the flooded basement.

comprehension question: Bob and the plumber fixed the problem.
correct answer: True

Item 18

Jane walked into the home improvement store, ready to complain.
Her new lawn mower would not crank, so she brought it back.
She talked to the clerk, who asked for her receipt.
She gave him the receipt, and she exchanged the mower for a new one.
Jane took the new mower home, and it cranked immediately.

comprehension question: Jane replaced her lawn mower.
correct answer: True