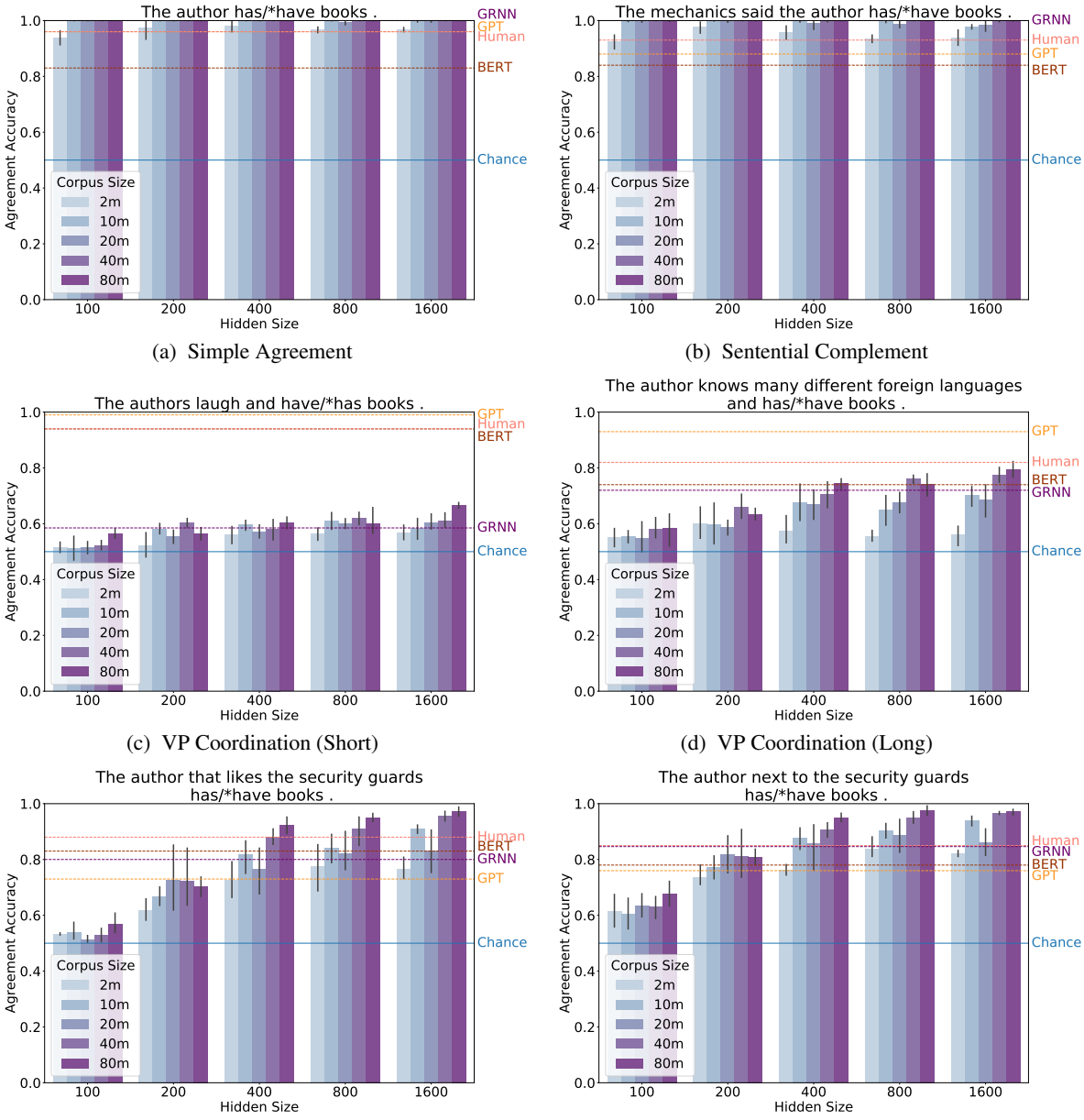


(a) Simple Agreement

(b) Sentential Complement

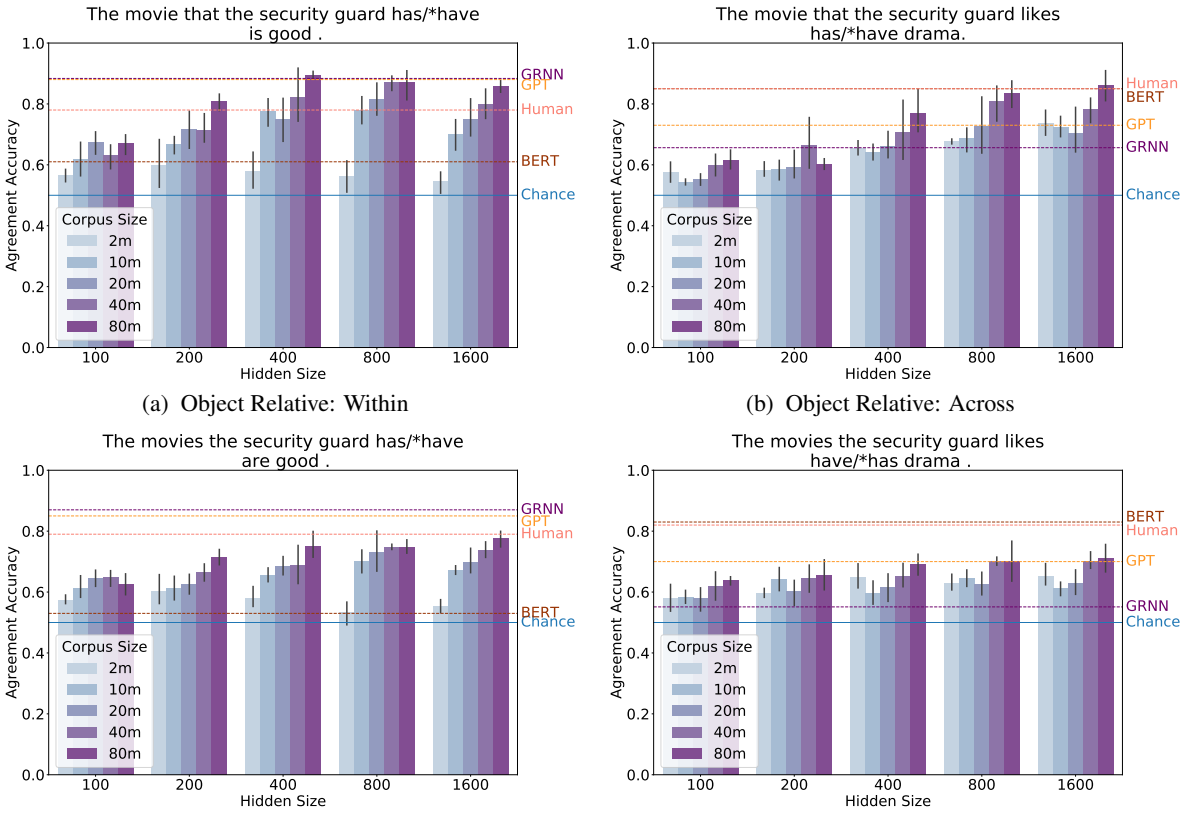
(c) VP Coordination (Short)

(d) VP Coordination (Long)

(e) Subject Relative

(f) Prepositional Phrase

Figure 3: Language model agreement performance when the verb is adjacent to its subject (3a,3b), when the verb is coordinated with another verb (3c,3d), when the verb and subject have an intervening relative clause (3e), and when the verb and subject have an intervening prepositional phrase (3f). The dashed horizontal lines show agreement performance of commonly-used large-scale models. Error bars reflect standard deviation across the five models in each category. GPT and BERT results are from those reported by Wolf (2019). Human results are those reported by Marvin and Linzen (2018).



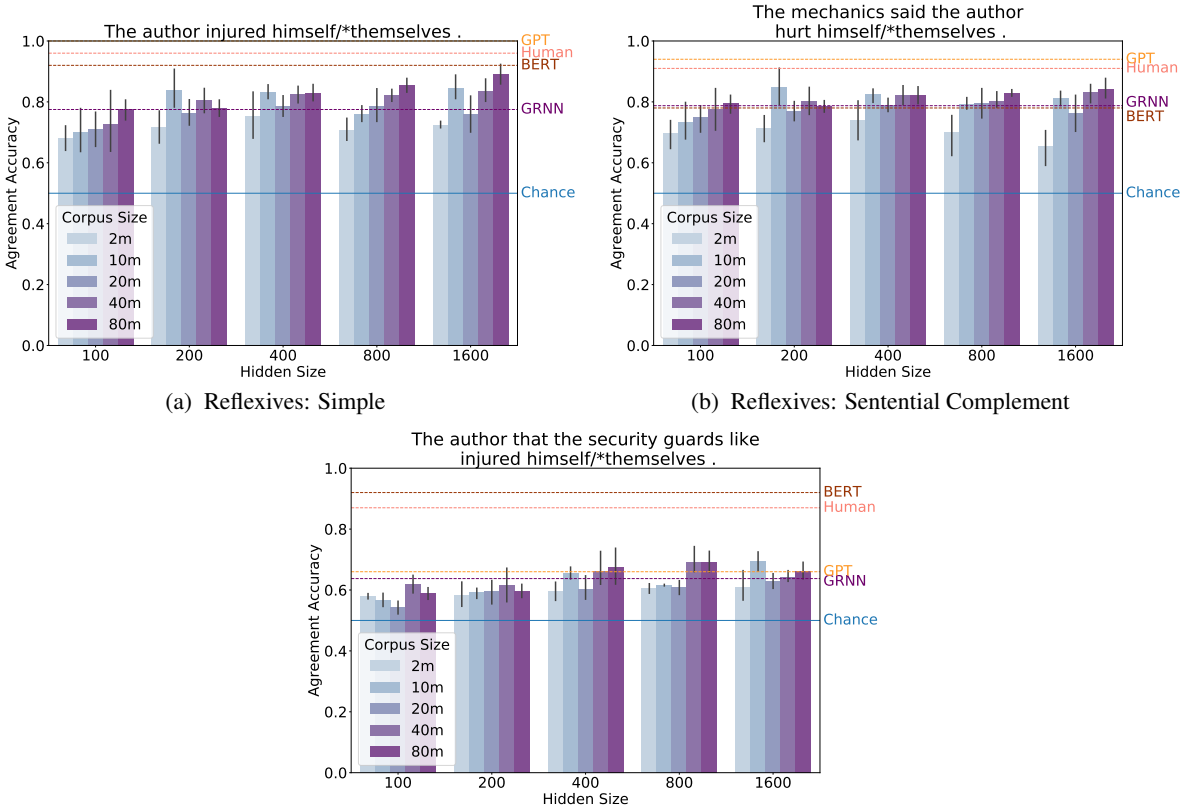
(a) Object Relative: Within

(b) Object Relative: Across

(c) Object Relative: Within (no that)

(d) Object Relative: Across (no that)

Figure 4: Language model agreement performance when the target verb is **within** an object-modifying relative clause (Left) and when an object-modifying relative clause intervenes between the target verb and its subject (Right). These results distinguish between when the relative clause has an overt relativizer (Top) and when it lacks an overt relativizer (Bottom). The dashed horizontal lines show agreement performance of commonly-used large-scale models. Error bars reflect standard deviation across the five models in each category. GPT and BERT results are from those reported by Wolf (2019). Human results are those reported by Marvin and Linzen (2018).



(a) Reflexives: Simple

(b) Reflexives: Sentential Complement

(c) Reflexives: Across

Figure 5: Language model agreement performance between reflexive pronouns and their antecedent in simple transitive sentences (5a), when agreement occurs within a sentential complement (5b), and when there is an intervening subject relative clause (5c). The dashed horizontal lines show agreement performance of commonly-used large-scale models. Error bars reflect standard deviation across the five models in each category. GPT and BERT results are from those reported by Wolf (2019). Human results are those reported by Marvin and Linzen (2018).

	2M→10M	10M→20M	20M→40M	40M→80M
VP Coordination (short)	4.0	0.4	0.5	0.6
VP Coordination (long)	$1.8e^3$	0.3	18.8	0.4
Subject Relative	26.2	0.9	45.1	1.6
Prepositional Phrase	$1.0e^5$	0.9	11.7	2.2
Object Relative: Within	$8.7e^5$	0.4	1.7	1.4
Object Relative: Across	0.4	0.4	1.5	1.1
Object Relative: Within (no that)	$1.4e^6$	0.8	0.5	1.4
Object Relative: Across (no that)	0.9	0.4	28.8	0.4
Reflexives: Simple	69.4	1.0	5.5	1.6
Reflexives: Sentential Complement	539	1.0	1.5	0.5
Reflexives: Across	18.3	6.4	9.7	0.4

Table 2: Strength of evidence of improvement in each construction produced by increasing the training data (averaged across model sizes). For this analysis, we excluded models with fewer than 400 hidden units (i.e. those with 100 or 200 hidden units) and which therefore might not make effective use of additional training data. Strength of evidence is quantified by Bayes factors. A Bayes factor $K < 1$ indicates that there is no difference between the two model groups, and $K > 10$ provides strong evidence that the model groups obtain different accuracies.

	10M→20M	20M→40M	40M→80M
VP Coordination (short)	-	$9e^{31}$	$1e^{22}$
VP Coordination (long)	$4e^{78}$	$1e^{10}$	$2e^{16}$
Object Relative: Across	$7e^{20}$	$2e^{10}$	$2e^{10}$
Object Relative: Within (no that)	$5e^{12}$	$8e^{13}$	$1e^{11}$
Object Relative: Across (no that)	$1e^{38}$	$4e^{10}$	$2e^{16}$
Reflexives: Simple	-	$6e^{11}$	$2e^{13}$
Reflexives: Sentential Complement	-	$1e^{12}$	$1e^{18}$
Reflexives: Across	-	$6e^{11}$	$3e^{23}$

Table 3: Training tokens needed for LSTMs to achieve human-like performance in each condition that does not presently reach human-like performance. Projections were obtained by assuming that doubling the data produces a constant rate of error reduction. Each column of the table shows the results from assuming a different rate of error reduction, estimated from the error reductions we actually observed from each of our training data doublings. Increases in error (negative improvement) for a given doubling are denoted with '-' since those would never produce human-like performance. These results demonstrate that the human-like performance data requirements we report in our paper are actually fairly low compared to other improvement rates we observed.

	10M→20M	20M→40M	40M→80M
VP Coordination (short)	-	$3e^{96}$	$1e^{60}$
VP Coordination (long)	$> 1e^{100}$	$2e^{19}$	$1e^{58}$
Subject Relative	-	$7e^{11}$	$8e^{13}$
Prepositional Phrase	-	$7e^{11}$	$4e^{13}$
Object Relative: Within	$2e^{39}$	$3e^{17}$	3^{17}
Object Relative: Across	$7e^{73}$	$1e^{19}$	$1e^{18}$
Object Relative: Within (no that)	$2e^{38}$	$1e^{46}$	$9e^{28}$
Object Relative: Across (no that)	$> 1e^{100}$	$1e^{23}$	$2e^{59}$
Reflexives: Simple	-	$5e^{18}$	$9e^{22}$
Reflexives: Sentential Complement	-	$2e^{23}$	$3e^{48}$
Reflexives: Across	-	$4e^{26}$	$8e^{88}$

Table 4: Training tokens needed for LSTMs to achieve 99.99% accuracy in each condition that does not presently reach 99% accuracy. Projections were obtained by assuming that doubling the data produces a constant rate of error reduction. Each column of the table shows the results from assuming a different rate of error reduction, estimated from the error reductions we actually observed from each of our training data doublings. Increases in error (negative improvement) for a given doubling are denoted with '-' since those would never produce better performance.