A Transformations in LIT

In Table 5, we show the full list of transformation that LIT can generate. This is not the full capability of LIT, and more transformations are possible as long as the linguistic phenomena are allowed by the ERG grammar.

B Annotator Agreement

To confirm the quality of the generated sentences, we recruit experienced graduate students as our annotators. For each phenomenon, we randomly select 50 sentences and have three annotators to judge. Given a phenomenon, each annotator is asked to judge whether they deem the generated (and selected) sentence as grammatical. The gold labels (i.e., grammatical or not) are determined by majority vote. For the it-cleft phenomenon, LIT can generate sentences that emphasize the first theta argument (ARG1) or the second theta argument (ARG2) of the verb for the main clause. Annotation results are shown in the Table 6.

C Full experiments results

In this section, we show the detailed evaluation results from all models — `bert-base-uncased`, `bert-large-uncased`, `roberta-base`, and `roberta-large` — trained seperately on two scenarios — ORI and AUG.

Phenomenon	original sentence	generated sentence
Future	Two guards are standing at the exit.	Two guards will stand at the exit.
Future+It-cleft: AGR1	The boy is making snowballs.	It is the boy who will be making snowballs.
Future+It-cleft: AGR2	People don’t play sports.	It is not sports that will be played by people.
Future+Passive: AGR2	A woman drills rock.	Rock will be drilled by a woman.
It-cleft: ARG1	A boy is blowing bubbles	It is a boy who is blowing bubbles.
It-cleft: ARG1+Passive: ARG2	The man isn’t wearing a hat	It is not the man that a hat is being worn by.
It-cleft: ARG2	A woman is performing music.	It is music that is being performed by a woman.
Modality: may	A person is lounging in a pool	A person may be lounging in a pool.
Negation	Five people tend sheep.	Five people don’t tend sheep.
Negation+It cleft: ARG1	The woman is playing guitar.	It is not the woman who is playing guitar.
Negation+It cleft: ARG2	The man and woman are buying beer.	It is not beer that is being bought by the man and woman.
Negation+Passive: ARG2	A woman is riding a bike.	A bike is being ridden by no woman.
Passive: ARG2	Adults are playing soccer.	Soccer is being played by adults.
Past	There is two cats outside.	There were two cats outside.
Past+It cleft: ARG1	A woman is mopping.	It is a woman who was mopping.
Past+It cleft: ARG2	A boy is playing sports.	It is sports that was being played by a boy.
Past+Passive: ARG2	A man is reading a newspaper.	A newspaper was being read by a man.
Present	The large pothole in the road was due to bad winter weather.	The large pothole in the road is due to bad winter weather.
Present+It cleft: ARG1	The road developed a big hole.	It is the road that develops a big hole.
Present+It cleft: ARG2	A man ate a stick.	It is a stick which is eaten by a man.
Present+Passive: ARG2	Two girls pick flowers outside.	Flowers are picked by two girls outside.
Swap subj/obj	The people look at the mountain.	The mountain looks at the people.
Swap subj/obj+It cleft: ARG1	A woman is playing a board game.	It is a board game that is playing a woman.
Swap subj/obj+It cleft: ARG2	A girl in a pink top spins a ribbon.	It is a girl in a pink top that is spun by a ribbon.
Swap subj/obj+Passive: ARG2	A grown woman carries a scooter.	A grown woman is carried by a scooter.

Table 5: Examples for the full list of rules

Phenomenon	Major(%)	Una.(%)
Future + Passive: ARG2	98	74
It cleft: ARG1 + Passive: ARG2	90	76
Future + It cleft: ARG2	94	76
Past + Passive: ARG2	82	66
Future + It cleft: ARG1	98	94
Past + It cleft: ARG2	92	78
Past + It cleft: ARG1	100	94
Past	96	88
Present	100	100
Future	100	86
Modality: may	100	98

Table 6: The annotators' agreement table for phenomena used for training. We show the percentage of grammatical sentences deemed by majority of our annotators — Major(%), and the percentage of unanimous agreement — Una.(%)

		`f;p`	**`p;f`**	**`i;i`**	**`pa;pa`**	**`m;o`**
ORI	Acc@Ori	89.14/86.54	87.02/86.54	88.36/86.01	89.27/88.83	88.72/87.20
	Acc@Ctr	7.69/11.54	41.98/33.85	83.21/81.87	85.41/85.39	13.75/12.80
	Consistency	7.69/9.62	33.59/26.54	88.67/89.95	88.41/91.98	10.19/9.65
		`p;f +i`	**`p;f +pa`**	**`f;p +i`**	**`f;p +pa`**	
	Acc@Ori	86.81/86.47	88.52/85.37	91.89/87.41	91.53/87.50	
	Acc@Ctr	37.50/28.05	31.15/19.51	8.49/10.00	6.78/5.56	
	Consistency	34.72/24.42	22.95/19.51	7.34/10.74	5.08/9.72	
		`f;p`	**`p;f`**	**`i;i`**	**`pa;pa`**	**`m;o`**
AUG	Acc@Ori	90.05/86.54	86.26/86.92	88.05/87.19	89.70/89.11	88.43/87.66
	Acc@Ctr	99.55/97.12	98.47/98.85	87.74/84.83	85.84/85.67	99.21/97.31
	Consistency	89.59/83.65	84.73/85.77	93.72/93.30	87.55/93.70	87.64/85.34
		`p;f +i`	**`p;f +pa`**	**`f;p +i`**	**`f;p +pa`**	
	Acc@Ori	86.46/88.45	88.52/90.24	92.28/88.89	91.53/88.89	
	Acc@Ctr	76.39/65.35	100.00/95.12	46.72/38.15	89.83/97.22	
	Consistency	67.01/57.76	88.52/85.37	39.77/34.44	81.36/86.11	

Table 7: Consistency and accuracies of `bert-base-uncased` over different linguistic phenomena in MNLI. We first train two model separately on the original (ORI) training set and augmented (AUG) training set. Then, we evaluate the trained models on **dev-m.** and **dev-mm.** for each phenomena. In this table, we report accuracy on the original sentence pair (Acc@Ori), accuracy on the transformed sentence pair (Acc@Ctr), and the model's consistency. Each accuracy/consistency has the format (**dev-m./dev-mm.**)

		f;p	p;f	i;i	pa;pa	m;o
ORI	Acc@Ori	94.12/90.87	88.55/88.85	89.60/89.85	93.56/91.98	90.21/89.70
	Acc@Ctr	8.60/10.10	41.22/33.08	84.35/85.02	87.12/87.11	13.75/12.52
	Consistency	7.24/8.65	35.88/24.23	89.80/91.23	88.41/92.84	10.68/9.83
		p;f +i	p;f +pa	f;p +i	f;p +pa	
	Acc@Ori	88.89/87.46	93.44/90.24	93.05/92.96	91.53/95.83	
	Acc@Ctr	38.54/27.72	32.79/14.63	6.95/9.26	8.47/5.56	
	Consistency	32.99/23.10	29.51/12.20	5.41/8.15	3.39/6.94	
		f;p	p;f	i;i	pa;pa	m;o
AUG	Acc@Ori	90.50/92.31	89.31/88.46	88.36/89.26	91.85/90.54	88.63/88.87
	Acc@Ctr	100.00/99.04	98.85/99.23	87.95/88.87	88.84/85.96	99.31/98.70
	Consistency	90.50/91.35	88.17/87.69	95.06/93.50	93.56/90.83	87.93/87.76
		p;f +i	p;f +pa	f;p +i	f;p +pa	
	Acc@Ori	88.19/87.46	90.16/89.02	90.73/91.48	91.53/93.06	
	Acc@Ctr	79.51/70.63	100.00/96.34	59.46/59.26	96.61/93.06	
	Consistency	68.40/59.41	90.16/85.3791.48	52.51/56.67	88.14/88.89	

Table 8: Consistency and accuracies of bert-large-uncased over different linguistic phenomena in MNLI. We first train two model separately on the original (ORI) training set and augmented (AUG) training set. Then, we evaluate the trained models on **dev-m.** and **dev-mm.** for each phenomena. In this table, we report accuracy on the original sentence pair (Acc@Ori), accuracy on the transformed sentence pair (Acc@Ctr), and the model's consistency. Each accuracy/consistency has the format (**dev-m./dev-mm.**)

		f;p	p;f	i;i	pa;pa	m;o
ORI	Acc@Ori	91.40/91.83	90.08/90.38	90.83/91.23	94.42/91.98	91.30/91.93
	Acc@Ctr	8.14/9.13	35.50/23.46	85.58/83.84	88.84/87.11	13.06/11.69
	Consistency	10.41/7.69	30.15/18.46	90.42/89.26	91.85/89.40	8.11/9.18
		p;f +i	p;f +pa	f;p +i	f;p +pa	
	Acc@Ori	91.32/92.41	93.44/93.90	92.66/92.59	94.92/95.83	
	Acc@Ctr	29.51/20.79	27.87/13.41	7.34/7.04	8.47/4.17	
	Consistency	23.61/18.48	21.31/12.20	5.41/8.52	3.39/5.56	
		f;p	p;f	i;i	pa;pa	m;o
AUG	Acc@Ori	91.40/91.35	87.40/93.08	89.39/90.94	92.70/92.55	90.31/91.65
	Acc@Ctr	99.10/99.04	98.85/98.46	87.33/88.87	90.13/87.97	99.41/97.96
	Consistency	91.40/90.38	86.26/91.54	94.44/95.76	92.27/91.40	89.71/89.61
		p;f +i	p;f +pa	f;p +i	f;p +pa	
	Acc@Ori	89.58/92.74	91.80/97.56	91.89/91.85	93.22/93.06	
	Acc@Ctr	91.67/86.47	100.00/95.12	83.01/78.89	94.92/94.44	
	Consistency	82.64/81.19	91.80/92.68	75.68/72.22	88.14/87.50	

Table 9: Consistency and accuracies of roberta-base over different linguistic phenomena in MNLI. We first train two model separately on the original (ORI) training set and augmented (AUG) training set. Then, we evaluate the trained models on **dev-m.** and **dev-mm.** for each phenomena. In this table, we report accuracy on the original sentence pair (Acc@Ori), accuracy on the transformed sentence pair (Acc@Ctr), and the model's consistency. Each accuracy/consistency has the format (**dev-m./dev-mm.**)

		`f;p`	`p;f`	`i;i`	`pa;pa`	`m;o`
	Acc@Ori	93.21/93.75	91.60/92.31	91.86/92.12	95.28/94.84	90.90/93.41
ORI	Acc@Ctr	5.43/7.69	41.98/30.38	85.17/85.91	90.99/89.40	15.13/12.43
	Consistency	4.98/4.33	34.35/24.23	91.04/89.46	93.99/92.26	10.19/9.37
		`p;f +i`	`p;f +pa`	`f;p +i`	`f;p +pa`	
	Acc@Ori	90.97/92.74	93.44/93.90	94.21/95.93	94.92/98.61	
	Acc@Ctr	34.38/27.39	32.79/23.17	6.18/8.89	6.78/4.17	
	Consistency	28.82/23.43	29.51/21.95	5.79/9.26	5.08/5.56	
		`f;p`	`p;f`	`i;i`	`pa;pa`	`m;o`
	Acc@Ori	93.67/95.19	93.13/90.77	91.86/92.71	94.85/94.56	92.19/93.97
AUG	Acc@Ctr	99.10/98.08	99.62/98.85	89.80/90.54	91.85/91.12	99.11/98.52
	Consistency	92.76/93.27	92.75/89.62	95.06/94.68	94.42/92.55	91.49/92.49
		`p;f +i`	`p;f +pa`	`f;p +i`	`f;p +pa`	
	Acc@Ori	90.97/92.08	93.44/92.68	94.98/97.41	94.92/98.61	
	Acc@Ctr	87.50/82.51	98.36/97.56	76.06/71.11	94.92/94.44	
	Consistency	78.47/77.23	91.80/90.24	71.04/69.26	89.83/93.06	

Table 10: Consistency and accuracies of `roberta-large` over different linguistic phenomena in MNLI. We first train two model separately on the original (ORI) training set and augmented (AUG) training set. Then, we evaluate the trained models on **dev-m.** and **dev-mm.** for each phenomena. In this table, we report accuracy on the original sentence pair (Acc@Ori), accuracy on the transformed sentence pair (Acc@Ctr), and the model's consistency. Each accuracy/consistency has the format (**dev-m./dev-mm.**)

		`f;p`	`p;f`	`i;i`	`pa;pa`	`m;o`
	Acc@Ori	94.59	91.55	91.33	93.09	92.49
ORI	Acc@Ctr	5.41	32.39	90.94	89.46	8.62
	Consistency	5.41	25.35	95.70	93.96	4.42
		`p;f +i`	`p;f +pa`	`f;p +i`	`f;p +pa`	
	Acc@Ori	90.16	93.33	95.60	97.30	
	Acc@Ctr	48.36	20.00	4.40	2.70	
	Consistency	40.16	17.78	4.40	5.41	
		`f;p`	`p;f`	`i;i`	`pa;pa`	`m;o`
	Acc@Ori	92.79	90.85	91.56	92.57	93.26
AUG	Acc@Ctr	100.00	99.30	90.94	90.50	99.89
	Consistency	92.79	90.14	98.12	94.82	93.15
		`p;f +i`	`p;f +pa`	`f;p +i`	`f;p +pa`	
	Acc@Ori	90.98	91.11	94.51	91.89	
	Acc@Ctr	81.97	100.00	52.75	100.00	
	Consistency	74.59	91.11	47.25	91.89	

Table 11: Consistency and accuracies of `bert-base-uncased` over different linguistic phenomena in SNLI. We first train two model separately on the original (ORI) training set and augmented (AUG) training set. Then, we evaluate the trained models on SNLI development set for each phenomena. In this table, we report accuracy on the original sentence pair (Acc@Ori), accuracy on the transformed sentence pair (Acc@Ctr), and the model's consistency.

		`f;p`	`p;f`	`i;i`	`pa;pa`	`m;o`
	Acc@Ori	96.40	93.66	92.34	94.30	94.03
ORI	Acc@Ctr	7.21	47.89	91.25	89.98	6.96
	Consistency	5.41	42.96	96.41	91.54	3.65
		`p;f +i`	`p;f +pa`	`f;p +i`	`f;p +pa`	
	Acc@Ori	91.80	91.11	96.70	97.30	
	Acc@Ctr	55.74	26.67	7.69	2.70	
	Consistency	50.82	31.11	6.59	5.41	
		`f;p`	`p;f`	`i;i`	`pa;pa`	`m;o`
	Acc@Ori	94.59	92.25	92.58	93.78	93.48
AUG	Acc@Ctr	100.00	100.00	92.19	91.19	99.89
	Consistency	94.59	92.25	97.27	95.34	93.37
		`p;f +i`	`p;f +pa`	`f;p +i`	`f;p +pa`	
	Acc@Ori	92.62	91.11	95.60	97.30	
	Acc@Ctr	95.08	100.00	89.01	100.00	
	Consistency	87.70	91.11	84.62	97.30	

Table 12: Consistency and accuracies of `bert-large-uncased` over different linguistic phenomena in SNLI. We first train two model separately on the original (ORI) training set and augmented (AUG) training set. Then, we evaluate the trained models on SNLI development set for each phenomena. In this table, we report accuracy on the original sentence pair (Acc@Ori), accuracy on the transformed sentence pair (Acc@Ctr), and the model's consistency.

		`f;p`	`p;f`	`i;i`	`pa;pa`	`m;o`
	Acc@Ori	95.50	92.96	92.81	93.78	94.59
ORI	Acc@Ctr	4.50	46.48	92.42	90.67	5.97
	Consistency	3.60	39.44	97.42	93.44	2.32
		`p;f +i`	`p;f +pa`	`f;p +i`	`f;p +pa`	
	Acc@Ori	91.80	91.11	96.70	97.30	
	Acc@Ctr	54.92	22.22	4.40	2.70	
	Consistency	46.72	22.22	3.30	5.41	
		`f;p`	`p;f`	`i;i`	`pa;pa`	`m;o`
	Acc@Ori	96.40	93.66	92.97	93.61	94.92
AUG	Acc@Ctr	100.00	100.00	92.50	92.06	99.89
	Consistency	96.40	93.66	97.81	95.34	94.81
		`p;f +i`	`p;f +pa`	`f;p +i`	`f;p +pa`	
	Acc@Ori	92.62	91.11	96.70	97.30	
	Acc@Ctr	78.69	100.00	57.14	100.00	
	Consistency	71.31	91.11	56.04	97.30	

Table 13: Consistency and accuracies of `roberta-base` over different linguistic phenomena in SNLI. We first train two model separately on the original (ORI) training set and augmented (AUG) training set. Then, we evaluate the trained models on SNLI development set for each phenomena. In this table, we report accuracy on the original sentence pair (Acc@Ori), accuracy on the transformed sentence pair (Acc@Ctr), and the model's consistency.

		f;p	**p;f**	**i;i**	**pa;pa**	**m;o**
ORI	Acc@Ori	97.30	95.77	93.91	94.99	95.91
	Acc@Ctr	3.60	30.28	92.03	92.23	5.41
	Consistency	0.90	27.46	96.56	94.47	2.65
		p;f +i	**p;f +pa**	**f;p +i**	**f;p +pa**	
	Acc@Ori	95.08	95.56	96.70	94.59	
	Acc@Ctr	45.08	37.78	4.40	0.00	
	Consistency	41.80	37.78	1.10	5.41	
		f;p	**p;f**	**i;i**	**pa;pa**	**m;o**
AUG	Acc@Ori	95.50	94.37	92.89	93.96	94.48
	Acc@Ctr	100.00	100.00	92.58	93.09	99.89
	Consistency	95.50	94.37	98.12	97.75	94.36
		p;f +i	**p;f +pa**	**f;p +i**	**f;p +pa**	
	Acc@Ori	93.44	91.11	95.60	91.89	
	Acc@Ctr	95.08	100.00	68.13	100.00	
	Consistency	88.52	91.11	65.93	91.89	

Table 14: Consistency and accuracies of `roberta-large` over different linguistic phenomena in SNLI. We first train two model separately on the original (ORI) training set and augmented (AUG) training set. Then, we evaluate the trained models on SNLI development set for each phenomena. In this table, we report accuracy on the original sentence pair (Acc@Ori), accuracy on the transformed sentence pair (Acc@Ctr), and the model's consistency.