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

Table 5: Examples for the full list of rules
### 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.(

<table>
<thead>
<tr>
<th>Phenomenon</th>
<th>Major(%)</th>
<th>Una.(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Future + Passive: ARG2</td>
<td>98</td>
<td>74</td>
</tr>
<tr>
<td>It cleft: ARG1 + Passive: ARG2</td>
<td>90</td>
<td>76</td>
</tr>
<tr>
<td>Future + It cleft: ARG2</td>
<td>94</td>
<td>76</td>
</tr>
<tr>
<td>Past + Passive: ARG2</td>
<td>82</td>
<td>66</td>
</tr>
<tr>
<td>Future + It cleft: ARG1</td>
<td>98</td>
<td>94</td>
</tr>
<tr>
<td>Past + It cleft: ARG2</td>
<td>92</td>
<td>78</td>
</tr>
<tr>
<td>Past + It cleft: ARG1</td>
<td>100</td>
<td>94</td>
</tr>
<tr>
<td>Past</td>
<td>96</td>
<td>88</td>
</tr>
<tr>
<td>Present</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Future</td>
<td>100</td>
<td>86</td>
</tr>
<tr>
<td>Modality: may</td>
<td>100</td>
<td>98</td>
</tr>
</tbody>
</table>

### 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.)

<table>
<thead>
<tr>
<th>Phenomenon</th>
<th>ORI</th>
<th>AUG</th>
</tr>
</thead>
<tbody>
<tr>
<td>f;p</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p:f</td>
<td></td>
<td></td>
</tr>
<tr>
<td>i;i</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pa;pa</td>
<td></td>
<td></td>
</tr>
<tr>
<td>m;o</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acc@Ori</td>
<td>89.14/86.54</td>
<td>90.05/86.54</td>
</tr>
<tr>
<td>Acc@Ctr</td>
<td>7.69/11.54</td>
<td>99.55/97.12</td>
</tr>
<tr>
<td>Consistency</td>
<td>7.69/9.62</td>
<td>34.72/24.42</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Phenomenon</th>
<th>ORI</th>
<th>AUG</th>
</tr>
</thead>
<tbody>
<tr>
<td>p;f +i</td>
<td></td>
<td></td>
</tr>
<tr>
<td>f;p +i</td>
<td></td>
<td></td>
</tr>
<tr>
<td>f;p +pa</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acc@Ori</td>
<td>86.21/86.47</td>
<td>86.46/88.45</td>
</tr>
<tr>
<td>Acc@Ctr</td>
<td>37.50/28.05</td>
<td>76.39/65.35</td>
</tr>
<tr>
<td>Consistency</td>
<td>34.72/24.42</td>
<td>67.01/57.76</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Phenomenon</th>
<th>ORI</th>
<th>AUG</th>
</tr>
</thead>
<tbody>
<tr>
<td>p;f +i</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p:f +i</td>
<td></td>
<td></td>
</tr>
<tr>
<td>f;p +pa</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acc@Ori</td>
<td>90.05/86.54</td>
<td>86.46/88.45</td>
</tr>
<tr>
<td>Acc@Ctr</td>
<td>99.55/97.12</td>
<td>76.39/65.35</td>
</tr>
<tr>
<td>Consistency</td>
<td>34.72/24.42</td>
<td>67.01/57.76</td>
</tr>
</tbody>
</table>
### 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.).

<table>
<thead>
<tr>
<th>Phenomenon</th>
<th>ORI</th>
<th>AUG</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>f;p</td>
<td>p;f</td>
</tr>
<tr>
<td>f;p +i</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acc@Ori</td>
<td>94.12/90.87</td>
<td>88.55/88.85</td>
</tr>
<tr>
<td>Acc@Ctr</td>
<td>8.60/10.10</td>
<td>41.22/33.08</td>
</tr>
<tr>
<td>Consistency</td>
<td>7.24/8.65</td>
<td>35.88/24.23</td>
</tr>
<tr>
<td>p;f +i</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acc@Ori</td>
<td>88.89/87.46</td>
<td>93.44/90.24</td>
</tr>
<tr>
<td>Acc@Ctr</td>
<td>38.54/27.72</td>
<td>32.79/14.63</td>
</tr>
<tr>
<td>Consistency</td>
<td>32.99/23.10</td>
<td>29.51/12.20</td>
</tr>
<tr>
<td>p;f +pa</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acc@Ori</td>
<td>93.40/91.48</td>
<td>93.23/90.38</td>
</tr>
<tr>
<td>Acc@Ctr</td>
<td>8.14/9.13</td>
<td>35.50/23.46</td>
</tr>
<tr>
<td>Consistency</td>
<td>10.41/7.69</td>
<td>30.15/18.46</td>
</tr>
<tr>
<td>f;p</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acc@Ori</td>
<td>90.50/92.31</td>
<td>89.31/88.46</td>
</tr>
<tr>
<td>Acc@Ctr</td>
<td>100.00/99.04</td>
<td>98.85/99.23</td>
</tr>
<tr>
<td>Consistency</td>
<td>10.50/91.35</td>
<td>88.17/87.69</td>
</tr>
<tr>
<td>f;p +i</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acc@Ori</td>
<td>88.19/87.46</td>
<td>90.16/89.02</td>
</tr>
<tr>
<td>Acc@Ctr</td>
<td>79.51/70.63</td>
<td>100.00/96.34</td>
</tr>
<tr>
<td>Consistency</td>
<td>68.40/59.41</td>
<td>90.16/85.01</td>
</tr>
</tbody>
</table>

### 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.).

<table>
<thead>
<tr>
<th>Phenomenon</th>
<th>ORI</th>
<th>AUG</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>f;p</td>
<td>p;f</td>
</tr>
<tr>
<td>f;p +i</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acc@Ori</td>
<td>91.40/91.83</td>
<td>90.08/90.38</td>
</tr>
<tr>
<td>Acc@Ctr</td>
<td>8.14/9.13</td>
<td>35.50/23.46</td>
</tr>
<tr>
<td>Consistency</td>
<td>10.41/7.69</td>
<td>30.15/18.46</td>
</tr>
<tr>
<td>p;f +i</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acc@Ori</td>
<td>91.32/92.41</td>
<td>93.44/93.90</td>
</tr>
<tr>
<td>Acc@Ctr</td>
<td>29.51/20.79</td>
<td>27.87/13.41</td>
</tr>
<tr>
<td>Consistency</td>
<td>23.61/18.48</td>
<td>21.31/12.20</td>
</tr>
<tr>
<td>p;f +pa</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acc@Ori</td>
<td>91.40/91.35</td>
<td>87.40/93.08</td>
</tr>
<tr>
<td>Acc@Ctr</td>
<td>99.10/99.04</td>
<td>98.85/98.46</td>
</tr>
<tr>
<td>Consistency</td>
<td>91.40/90.38</td>
<td>86.26/91.54</td>
</tr>
<tr>
<td>f;p</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acc@Ori</td>
<td>91.40/91.35</td>
<td>87.40/93.08</td>
</tr>
<tr>
<td>Acc@Ctr</td>
<td>99.10/99.04</td>
<td>98.85/98.46</td>
</tr>
<tr>
<td>Consistency</td>
<td>91.40/90.38</td>
<td>86.26/91.54</td>
</tr>
<tr>
<td>f;p +i</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acc@Ori</td>
<td>89.58/92.74</td>
<td>91.80/97.56</td>
</tr>
<tr>
<td>Acc@Ctr</td>
<td>91.67/86.47</td>
<td>100.00/95.12</td>
</tr>
<tr>
<td>Consistency</td>
<td>82.64/81.19</td>
<td>91.80/92.68</td>
</tr>
</tbody>
</table>
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.).

<table>
<thead>
<tr>
<th>Phenomena</th>
<th>ORI</th>
<th>AUG</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Acc@Ori</strong></td>
<td>93.21/93.75</td>
<td>93.67/95.19</td>
</tr>
<tr>
<td><strong>Acc@Ctr</strong></td>
<td>5.43/7.69</td>
<td>99.10/98.08</td>
</tr>
<tr>
<td><strong>Consistency</strong></td>
<td>4.98/4.33</td>
<td>92.76/93.27</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Phenomena</th>
<th>ORI</th>
<th>AUG</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Acc@Ori</strong></td>
<td>94.59</td>
<td>92.79</td>
</tr>
<tr>
<td><strong>Acc@Ctr</strong></td>
<td>5.41</td>
<td>100.00</td>
</tr>
<tr>
<td><strong>Consistency</strong></td>
<td>4.42</td>
<td>91.11</td>
</tr>
</tbody>
</table>

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.).

<table>
<thead>
<tr>
<th>Phenomena</th>
<th>ORI</th>
<th>AUG</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Acc@Ori</strong></td>
<td>94.59</td>
<td>92.79</td>
</tr>
<tr>
<td><strong>Acc@Ctr</strong></td>
<td>5.41</td>
<td>100.00</td>
</tr>
<tr>
<td><strong>Consistency</strong></td>
<td>4.42</td>
<td>91.11</td>
</tr>
</tbody>
</table>

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.
<table>
<thead>
<tr>
<th></th>
<th>f;p</th>
<th>p;f</th>
<th>i;i</th>
<th>pa;pa</th>
<th>m;o</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acc@Ori</td>
<td>96.40</td>
<td>93.66</td>
<td>92.34</td>
<td>94.30</td>
<td>94.03</td>
</tr>
<tr>
<td>Acc@Ctr</td>
<td>7.21</td>
<td>47.89</td>
<td>91.25</td>
<td>89.98</td>
<td>6.96</td>
</tr>
<tr>
<td>Consistency</td>
<td>5.41</td>
<td>42.96</td>
<td>96.41</td>
<td>91.54</td>
<td>3.65</td>
</tr>
<tr>
<td></td>
<td>p;f</td>
<td>+i</td>
<td>p;f</td>
<td>+pa</td>
<td>f;p</td>
</tr>
<tr>
<td>Acc@Ori</td>
<td>91.80</td>
<td>91.11</td>
<td>96.70</td>
<td>97.30</td>
<td></td>
</tr>
<tr>
<td>Acc@Ctr</td>
<td>55.74</td>
<td>26.67</td>
<td>7.69</td>
<td>2.70</td>
<td></td>
</tr>
<tr>
<td>Consistency</td>
<td>50.82</td>
<td>31.11</td>
<td>6.59</td>
<td>5.41</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acc@Ori</td>
<td>94.59</td>
<td>92.25</td>
<td>92.58</td>
<td>93.78</td>
<td>93.48</td>
</tr>
<tr>
<td>Acc@Ctr</td>
<td>100.00</td>
<td>100.00</td>
<td>92.19</td>
<td>91.19</td>
<td>99.89</td>
</tr>
<tr>
<td>Consistency</td>
<td>94.59</td>
<td>92.25</td>
<td>97.27</td>
<td>95.34</td>
<td>93.37</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>f;p</th>
<th>p;f</th>
<th>i;i</th>
<th>pa;pa</th>
<th>m;o</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acc@Ori</td>
<td>95.50</td>
<td>92.96</td>
<td>92.81</td>
<td>93.78</td>
<td>94.59</td>
</tr>
<tr>
<td>Acc@Ctr</td>
<td>4.50</td>
<td>46.48</td>
<td>92.42</td>
<td>90.67</td>
<td>5.97</td>
</tr>
<tr>
<td>Consistency</td>
<td>3.60</td>
<td>39.44</td>
<td>97.42</td>
<td>93.44</td>
<td>2.32</td>
</tr>
<tr>
<td></td>
<td>p;f</td>
<td>+i</td>
<td>p;f</td>
<td>+pa</td>
<td>f;p</td>
</tr>
<tr>
<td>Acc@Ori</td>
<td>91.80</td>
<td>91.11</td>
<td>96.70</td>
<td>97.30</td>
<td></td>
</tr>
<tr>
<td>Acc@Ctr</td>
<td>54.92</td>
<td>22.22</td>
<td>4.40</td>
<td>2.70</td>
<td></td>
</tr>
<tr>
<td>Consistency</td>
<td>46.72</td>
<td>22.22</td>
<td>3.30</td>
<td>5.41</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acc@Ori</td>
<td>96.40</td>
<td>93.66</td>
<td>92.97</td>
<td>93.61</td>
<td>94.92</td>
</tr>
<tr>
<td>Acc@Ctr</td>
<td>100.00</td>
<td>100.00</td>
<td>92.50</td>
<td>92.06</td>
<td>99.89</td>
</tr>
<tr>
<td>Consistency</td>
<td>96.40</td>
<td>93.66</td>
<td>97.81</td>
<td>95.34</td>
<td>94.81</td>
</tr>
</tbody>
</table>

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.
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.