IMPORTANT NOTE (DO NOT DELETE)
The rebuttal to the reviews **must** be restricted to:

- answering the specific "pressing" questions raised by the reviewers in slot 4 of the reviews; and
- pointing out factual errors in the reviews.

Furthermore, it **must be one page-only**.
The goal of the rebuttal is not to enter a dialogue with the reviewers. Therefore, it is **not allowed to include new results in the rebuttal, provide links to such results, or describe how you might rewrite your paper to address concerns of the reviewers**. It is the current version of your paper that is under review, not possible future versions.
The program committee will be instructed to ignore rebuttals violating these principles.

Rebuttal

We sincerely thank all reviewers for their valuable and constructive feedback. Below we answer the pressing questions in the order of their importance.

Reviewer 1, Question 1: Experiments on larger datasets like MNLI and QNLI.
Answer: Both MNLI, QNLI, and RTE (which is tested in the paper) are natural language inference tasks. Based on previously published results of BERT, the accuracy on RET is 72.5 while it is 83.5 for MNLI and 91.2 for QNLI. ***Though RTE is smaller in size, it is actually more difficult than MNLI/QNLI***. Therefore we have reasons to believe that the advantages of LCD still hold on these two tasks.

Reviewer 2, Question 1: The performance gap between ... give more analysis for this observation.
Answer: In general, the capacity of the student network increase as its number of layers increases, resulting in a smaller performance gap between all compression methods (e.g., as the strongest baseline, Bert-of-theseus$_4$ outperforms the second-best by 1.2 on average while it reduces to 0.8 for Bert-of-theseus$_6$). In a more technical view, Bert-of-theseus works by replacing certain teacher layers with a student layer and the replaced teacher layers are then not used. In contrast, LCD is designed with two branches so that teacher layers that were previously discarded by Bert-of-theseus are exploited. Naturally, LCD mitigates information losses for 4-layer student more effectively than 6-layer student.

Reviewer 2, Question 2: The knowledge learned ... affect the final performances on OOD dataset.
Answer: We primarily perform data augmentation via paraphrasing the original training set to double its size. We believe that not only the scale but also the linguistic diversity (which is encouraged in the paraphrasing process) of augmented data matters for good performance. We didn't conduct extensive ablation on how to augment data, but enlarging the size would generally improve performance.

Reviewer 3

Among all questions, ***Q1, Q4, Q5, Q6 have straight answers in the paper*** hence we only provide brief answers along with where they are described in the paper.

Reviewer 3, Question 2: This paper focuses on language model ...
Answer: Since our method has no constraint on the structure of network layer, it can be also applied to compressing other types of neural network, e.g., ResNet in the computer vision domain, LSTM in time series forecasting.

Reviewer 3, Question 3: The motivation of probability ...
Answer: One fundamental design in our method is the layer-wise interaction between student layers and teacher layers. A static interaction scheme can be one option, e.g., alternating student and teacher layers. However, the interaction in such a scheme is limited to certain participated layers and leads to inferior performance. Our probability scheduler is dynamic, ensuring rich and comprehensive knowledge transfer.

Reviewer 3, Question 7: What is the definition of ...
Answer: KD-free methods are those that do not include knowledge distillation objectives (e.g., KL divergence between logits) in addition to task-specific objectives. For LayerDrop, it contains layer-wise dropout during training, and compression is achieved by dropping layers on demand during inference. For Bert-of-theseus, it progressively replaces teacher layers with student layers during training.

Reviewer 3, Question 1: What's the metric presented in Table 1/2/3?
Answer: The metrics presented in Table1/2/3 for SST-2/MRPC/RTE/CoLA/STS-B are described in the first paragraph of Section 3.1.

Reviewer 3, Question 4: What is a module? ...
Answer: A module contains either one student layer or r (r is the compression ratio) teacher layers. To illustrate, Figure 1 presents an example where the teacher network has 8 layers and the student network has 4 layers. In this example, each module contains either 1 student layer or 2 teacher layers.

Reviewer 3, Question 5: From the section 2.1, is every layers ...
Answer: As stated in Section 2.1 and Figure 1, layers in different branches have no direct interaction with each other at the same optimization time step.

Reviewer 3, Question 6: Is all the student models ...
Answer: As stated in Section 3.3, all the student models are 4/6-layer Transformer.

Reviewer 4, Question 1: How does it compare to ... dropout methods?
Answer: As can be observed from the last ablation study in Secton 3.5, LCD performs better than individual distillation. LCD also outperforms structured dropout (i.e., LayerDrop and LC) by a larger margin.

Reviewer 4, Question 2: Is there ... alternatives?
Answer: The memory footprint for intermediate activations/gradients marginally increases (lower than 10%) while the memory of parameters/gradient/momentum in optimizer keeps the same as individual alternatives.