A Appendix

Robustness A realistic scenario would be creating a filtered set with some existing systems and using the created test set to evaluate new systems. To verify the effectiveness of the proposed method in this scenario, we calculate the variance over a random half of the submitted systems of WMT20 Chinese↔English (Zh↔En) task and evaluate the generated test set on the remaining half. For example, the WMT20 Zh→En task has 16 submitted results, we use a random half of the results (*dong-nmt.1207, Tencent_Translation.1249, WMT-BiomedBaseline.183, Online-G.1569, DeepMind.381, Huawei_TSC.889, Online-Z.1646, WeChat_AI.1525*) for calculating variance and generating a variance-aware test set, and then evaluate the remained results (*DiDiNLP.401, SJTU-NICT.320, Online-A.1585, zlabs-nlp.1176, Huoshan_Translate.919, OPPO.1422, Online-B.1605, THUNLP.1498*) over the original test set and variance-aware test set. As shown in Table 1, the results indicate that the proposed method is still effective in such a scenario. Since the aim of this paper is to provide new test sets for the community to conduct evaluation for new systems, we calculate the variance of all submitted results to make the calculated variance more reliable and the generated test sets more discriminative.

Table 1: Pearson correlations using original and variance-aware test sets on the competitive WMT20 Chinese ↔ English tasks. Using variance-aware test sets generated by a random half of systems (+*Half*) consistently improves the evaluation results on the remaining systems.

	Zh-En	En-Zh
BLEU	0.957	0.888
+*Half*	**0.961**	**0.893**
COMET	0.953	-0.840
+*Half*	**0.963**	**-0.821**
BLEURT	0.941	0.803
+*Half*	**0.955**	**0.811**
BERTS-P	0.938	0.902
+*Half*	**0.947**	**0.948**
BERTS-R	0.956	0.934
+*Half*	**0.965**	**0.947**
BERTS-F	0.949	0.924
+*Half*	**0.958**	**0.948**

Reproducibility Since the organizers of WMT has released the generation results of the participant MT systems, we directly use these results to conduct the experiments instead of seeking for the original MT models. The data for reproducing the results including original test sets and human ratings, can be easily accessed from WMT official website as follows:

- WMT16: `http://www.statmt.org/wmt16/results.html`.
- WMT17: `http://www.statmt.org/wmt17/results.html`.
- WMT18: `http://www.statmt.org/wmt18/results.html`.
- WMT19: `http://www.statmt.org/wmt19/results.html`.
- WMT20: `http://www.statmt.org/wmt20/results.html`.

The evaluation metrics in this paper adopted the publicly available implementation:

- BLEU: We used sacreBLEU implementation[1] and default hyperparameters. The evaluation signature is: `BLEU+case.mixed+lang+smooth.exp+tok.intl+version.1.4.14`.
- COMET: We used original implementation[2], `wmt-large-da-estimator-1719` evaluation model and default hyperparameters.
- BLEURT: We used original implementation[3], `BLEURT-base-128` checkpoint and default hyperparameters.
- BERTScore: used original implementation[4], default BERT model settings and hyperparameters.

Computational Resources All the GPU computation is done by a single NVIDIA GeForce 1080Ti GPU card with CUDA Toolkit 10.1, and Intel Xeon E5-4655 CPU for handling other computation.

[1] https://github.com/mjpost/sacrebleu

[2] https://github.com/Unbabel/COMET

[3] https://github.com/google-research/bleurt

[4] https://github.com/Tiiiger/bert_score

Data Licensing The *variance-aware test sets* were created based on the original WMT test set. Thus, we follow the original data licensing plan already stated by WMT organizers, which is that "The data released for the WMT news translation task can be freely used for research purposes, we just ask that you cite the WMT shared task overview paper, and respect any additional citation requirements on the individual data sets. For other uses of the data, you should consult with original owners of the data sets." (quoted from the "LICENSING OF DATA" part in the WMT official website[5]).

Accessibility and Maintenance We have released the test sets and the codes for creating the variance-aware test sets on GitHub. Besides, we will maintain the repository including fixing any potential issues and updating the test sets if the new WMT test sets are released. Anyone can also prepare a specific variance-aware test set based on their customized data since the codes will be open-sourced.

Limitations The major limitation of variance-aware filtering method is the evaluation quality of automatic metrics. If the metrics cannot give reasonable scores to the model's hypothesis, the calculated variance values could be meaningless. Although the BERTScore metric we used in this paper can evaluate the semantic-level overlap and is one of the state-of-the-art evaluation metrics, it is still inevitable to give inaccurate scores for the evaluation outside of general domain or document-level MT evaluation. In future works, it is reasonable to combine more kinds of evaluation scores to give a more accurate variance of test instances.

[5] http://www.statmt.org/wmt20/translation-task.html