Learning from Failure: Improving Meeting Summarization without Good Samples

Ke Wang*, Xiutian Zhao*, Wei Peng

Huawei IT Innovation and Research Center
{wangke215, zhaoxiutian, peng.wei1}@huawei.com

Abstract

Existing methods aligning language models with various human needs are reliant heavily on high-quality and task-specific data. However, industrial deployment of task-specific language models often encounter challenges in the availability of appropriate training samples. Taking meeting summarization for instance, public datasets are scarce, and private corpora are also hard to obtain due to privacy issues or resource-demanding annotation. To improve meeting summarization in the absence of positively-rated (i.e., "good") samples, we propose Score Tuning, a cold start tuning framework that leverages bad samples of distinguishable degrees to incrementally enhance the performance of summary generation without an initial presence of good samples. Our method utilizes *asynchronous* and *numerical* human feedback that measure the quality of generated summaries. Formulating data into triplets of (transcript, summary, score), our approach instructs a pre-trained model to learn the association between summary qualities and human-rated scores and hence to generate better summaries corresponding to higher scores. The experiment results show that our method is effective in improving meeting summarization on both English and Chinese corpora while requiring less annotated data and training resources compared to existing alignment methods. Additionally, we also preliminarily explore the transferability of our approach in machine translation tasks and demonstrate its potential for future development and usage in other domains.

Introduction

The widespread use of video-telephony applications and the surging trend of remote work have contributed to a boom in online chatting and virtual meetings (Rennard et al. 2023). As a result, automatic meeting summarization, which condenses long and unorganized meeting transcripts into concise and comprehensible summaries, is of increasingly potential social economic values and thus draws rising attention (Gillick et al. 2009; Shang et al. 2018). The task aims to liberate people from cumbersome transcript reading and generate easy-to-digest meeting minutes automatically, and recent progress of large language models (LLMs) has yielded promising results tackling this task. (Wu et al. 2023; Hu et al. 2023; OpenAI 2023).

To produce summaries that align with human preferences, LLMs almost certainly necessitate fine-tuning on large-scale human-annotated data (Brown et al. 2020; Ouyang et al. 2022; Touvron et al. 2023). Among various alignment techniques, supervised fine-tuning (SFT) and reinforcement learning from human feedback (RLHF) are two widely-adopted and well-established ones. SFT utilizes human annotated positively-rated data, while RLHF relies on training a reward model from human annotated pairwise positive ("chosen") and negative ("rejected") examples.

However, the specific data either SFT or RLHF requires could be unobtainable in real-world industrial environments. In our case of developing a meeting summarization system, we are constrained in both attainable data and annotation resources. Facing the absence of training data suitable for existing tuning methods, we are bound to utilize collectible yet atypical data. Experimentally, we find that it is acceptable to request simple rating feedback on one given summary each time from actual users. To explore the feasibility of harnessing such self-augmenting feedback data to continuously improve alignment, we hypothesize that: (1) besides pairs of "chosen" and "rejected" generations, human feedback in other forms could also provide informative representations. Taking summarization tasks for instance, it is reasonable to make *asynchronous* measurements (assigning a stand-alone rating given one case at a time) across generated summaries on different source transcripts; (2) having been exposed to quality-wise distinguishable "bad" samples with corresponding human-rated scores, a language model could learn the association between numerical scores and generation quality.

Following this intuition, we take an in-context learning approach akin to Chain of Hindsight (CoH) proposed by Liu, Sferrazza, and Abbeel (2023). Unlike CoH that take pairwise comparison feedback, our approach utilizes *asynchronous* and *numerical* human feedback data (hence named "Score Tuning"). Having formulated data into triplets of transcripts, summaries and corresponding human feedback scores, we tune the model by simultaneously minimizing the confidence-aware loss (given transcripts and summaries, comparing the differences among model-predicted and human-assigned scores) and hindsight-score loss (given transcripts and human-assigned scores, comparing the differences among newly and initially model-generated sum-

*These authors contributed equally.

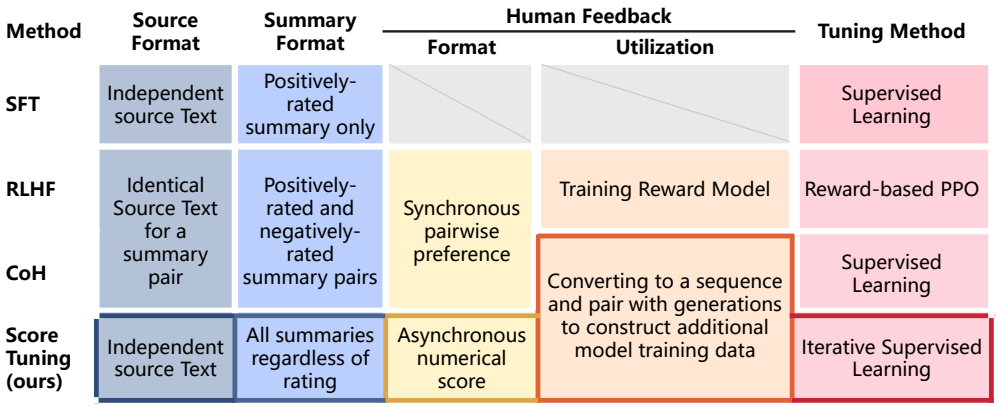

Figure 1: Comparison of tuning elements for summarization tasks by different approaches. RLHF and Chain of Hindsight (CoH) share the same form of input data: "synchronous" pairwise comparisons. While the utilization of human feedback is similar to CoH, Score Tuning takes all examples regardless of being specifically classified as "good" or "bad", as long as the qualities of examples are distinguished by human feedback.

maries). Presumably, given a transcript and a fixed high score, the score-tuned model could produce high quality summaries even without initially been exposed to "good" samples.

We conduct experiments to test the hypothesized effectiveness of our method in improving summarization on both an in-house dataset and two public corpora, and both automatic and human evaluation yield positive results. We also compare and discuss the trade-offs among different tuning strategies, and hope to draw more attention on utilization of atypical data in real-world practices.

In summary, the contributions of this study are three-fold:

- We propose Score Tuning, a cold start tuning framework that utilizes asynchronous and numerical human feedback to continuously improve meeting summarization.
- Having constructed a real-world meeting summarization dataset with human feedback, we benchmark our method against the in-house dataset as well as public corpora of English and Chinese. The results demonstrate that the method achieves better performance with reduced annotation demand than existing methods.
- Additionally, we migrate the Score Tuning to machine translation task; the experiments yield results that suggest a potential transferability of our method to other NLP tasks, laying the groundwork for future research.

Related Work

Abstractive Meeting Summarization. Automatic meeting summarization aims to condense information from a large piece of meeting transcript and produce a concise and comprehensible minutes or digest automatically (Gillick et al. 2009; Kumar and Kabiri 2022). While earlier works focus on *extractive* methods that create summaries by directly selecting and concatenating unedited sentences from source text, the development of neural networks has encouraged a growing trend in *abstractive* methods that implement encoder-decoder architectures on source text to generate summaries (Shang et al. 2018; Rennard et al. 2023). The lengthy, multi-speaker, spoken-language natures of meeting text pose many challenges for summarization, several strategies are proposed to address different aspects of those challenges (Li et al. 2019; Zhu et al. 2020; Koay et al. 2021; Zou et al. 2021).

Meeting corpora with human-written summaries are surprisingly scarce and largely in English (Hu et al. 2023). AMI (Janin et al. 2003) and ICSI (Carletta et al. 2005) are two early yet widely-used English meeting corpora. More recently, Nedoluzhko et al. (2022) release ELITR, a dataset of 120 English and 59 Czech meeting transcripts. QMSum (Zhong et al. 2021) contains 1,808 query-summary pairs on 232 meetings in English. Wu et al. (2023) introduce a meeting summarization dataset VCSum in Chinese consisting 1,359 meeting segment transcripts and human-written summaries. ORCHID (Zhao, Wang, and Peng 2023) is a Chinese debate summarization corpus covering 1,218 debate matches that resemble meeting-style text.

Elements for Different Tuning Techniques. As demonstrated in Figure 1, different tuning approaches require various elements. SFT requires solely positively-rated examples (i.e., human-written gold-standard summaries), yet this approach is not compatible with our case for (1) the access to internal meeting transcripts and minutes within our organization is limited, and (2) crowd-sourcing annotation on owned meeting transcripts is forbidden due to privacy restrictions.

On the other hand, RLHF-based frameworks ask for positively-rated and negatively-rated example pairs as inputs and train a reward model to guide the model weights update. However, this specific human feedback form, *synchronous pairwise comparison* (selecting a preference given two choices simultaneously), poses a great challenge in data collection. While such data are obtainable by employing annotators, it is neither intuitive nor user-friendly to display alternative summaries to actual users and ask them to

pick and choose. Without continuous user feedback, our system would be forced to rely on time-consuming and labor-intensive crowd-sourcing annotation to augment data and iterate model, which foreshadows an intolerable maintenance cost. Consequently, we decline those frameworks as well as Chain of Hindsight (CoH) (Liu, Sferrazza, and Abbeel 2023; Wei et al. 2023), since it asks for similar data input as RLHF-based approaches.

Learn Summarization from Human Feedback. Previous works have explored using human feedback to train summarization model with reinforcement learning (Böhm et al. 2019; Ziegler et al. 2020; Stiennon et al. 2020), and most of which learn a reward model based on the PPO algorithm (Schulman et al. 2017). While RLHF approach has achieved promising results in aligning with human preference, reward model training is nevertheless resource-demanding and tuning-challenging. More recently, a flourishing literature has suggested impressive in-context learning abilities of LLMs (Brown et al. 2020; Liu et al. 2021, 2022; Shinn et al. 2023; Wei et al. 2023; Gao et al. 2023; Roit et al. 2023). Among them, CoH Liu, Sferrazza, and Abbeel (2023) coverts human feedback into sequence and fine-tuning models by utilizing such feedback, which our work is most similar to. CoH takes natural language feedback, our method instead incorporates numerical human rating and implement a 2-stage loss-minimization workflow.

Score Tuning

Method Overview

Formally, let M denote a meeting transcript produced by an automatic speech recognition (ASR) system, consisting of a list of utterances. In a cold start scenario where no gold-standard meeting summary is available, we have several relatively low-quality generated summaries C. Human feedback scores S are integers between 1 and 100 assigned to those summaries based on one or more criteria (for simplicity, we assume that there is only one evaluation dimension, namely the overall quality). Therefore, each meeting transcript M has multiple corresponding low-quality summaries and score pairs $(C_1, S_1), ..., (C_n, S_n)$. The goal of Score Tuning is to generate a sample corresponding to a score of 100 (the upper bound of S_i), presumably the sample with best quality.

Inspired by the capabilities of evaluating the quality of generated results (Luo, Xie, and Ananiadou 2023) and reflecting on their own performance (Shinn et al. 2023) demonstrated by LLMs, we are motivated to instruct a summarization model to continuously improve summary generation by simultaneously measuring the quality of generated results and utilizing the differences among summaries of distinguishable qualities. Specifically, our method consists of three steps, as shown in Figure 2.

Step 1: Collect Human Feedback

We employ an ASR system to transcribe meeting recordings without further post-editing. We then instruct our preliminary stage model to sample and generate multiple summaries from obtained meeting transcripts. The annotators are asked to provide asynchronous and numerical feedback on each generated summary (see §Data Construction).

Step 2: Train the Model by Minimizing Confidence-Aware Loss and Hindsight-Score Loss

The iterative supervised training consists of two kinds of losses to be minimized: (1) given a meeting transcript and the corresponding generated summaries, the model is instructed to provide a rating on each summary. The *Confidence-aware Loss* is calculated by the differences between the model-predicted scores and human feedback scores; (2) given a meeting transcript and the human feedback scores corresponding to the generated summaries in Step 1, the model is instructed to produce a new summary for each score assigned. The *Hindsight-score Loss* is calculated by the differences between the newly produced summaries and the ones generated in Step 1.

Confidence-Aware Loss. Having a conditional generative language model, we reformulate the score prediction task based on the generation task as follows: the model input $M||C_i$ is

> **[Transcript]**: a transcript text. **[Summary]**: a corresponding output summary in step 1. **[Score]**:

where [Transcript], [Summary], and [Score] are special tokens in the tokenizer, and the model output is a value between 1 and 100. Let S_i be the human-rated score of the i-th generated summary, and $\hat{S}_i$ be the predicted score by the model. The *Confidence-aware Loss* is defined as:

$$L_{ca} = \sum_{i=1}^{N} w_i(-\sum_{t=1}^{T} \log p(S_t|S_{<t}, M||C_i)) \quad (1)$$

where T is the length of the output sequence, $M||C_i$ is the input sequence, and $p(S_t|S_{<t}, M||C_i)$ is the probability of predicting the correct token y_t at time step t, given the input sequence $M||C_i$ and the previous tokens $S_{<t}$. N is the number of summaries, and w_i is the confidence weight for the i-th summary. The confidence weight w_i is calculated as:

$$w_i = \frac{1}{\sigma_i^2 + \epsilon} \quad (2)$$

where σ_i^2 is the variance of the predicted scores for the i-th summary, and ϵ is a small constant to avoid division by zero. The variance σ_i^2 is estimated by applying Monte Carlo dropout during inference. The intuition behind this loss function is that if the model is confident about its prediction (i.e., low variance), then the loss will be large if the prediction is incorrect. On the other hand, if the model is not confident about its prediction (i.e., high variance), then the loss will be small even if the prediction is incorrect. This encourages the model to make confident predictions for summaries that it can accurately predict, while being cautious for samples that it cannot accurately predict.

Hindsight-Score Loss. The model takes a transcript text and a quality score as input:

> **[Transcript]**: a transcript text. **[Score]**: a human rated score. **[Summary]**:

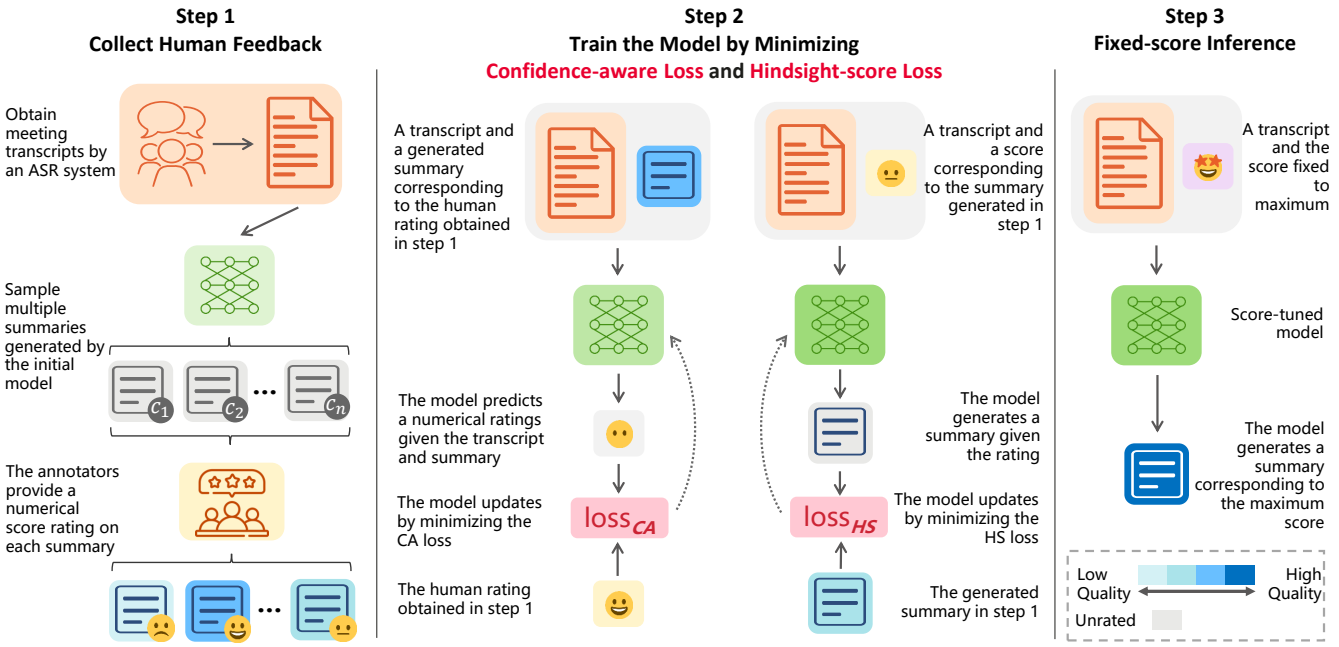

Figure 2: Illustration of the high level design of Score Tuning. Notably, human feedback in Step 1 is rating the qualities of generated summaries rather than providing human-written ones. Also, distinguished from RLHF-based methods and CoH, our method: (1) collects *asynchronous* and *numerical* human feedback on model-generated summaries rather than binary preference pairs, (2) minimizes the *confidence-aware loss* and *hindsight-score loss* to learn the association between summary qualities and scores from samples of various qualities even without an initial presence of positively-rated samples, and eventually (3) produces a summary corresponding to the fixed highest score, presumably of the best quality.

and outputs a summary corresponding to the score. Let $M||S_i$ denote the i-th input sample, consisting of a transcript text and a quality score, and let C_i denote the corresponding output summary. The *Hindsight-score Loss* for a given sample i can be defined as:

$$L_{hs} = -\frac{1}{N}\sum_{i=1}^{N}\log p(C_i|M_i||S_i) \tag{3}$$

where N is the number of samples and $p(C_i|M_i||S_i)$ represents the probability of generating the transcript text given the model input. Minimizing this loss enable the model to strengthen the association between scores and the corresponding summaries with different degrees of badness, leading to better summary generation patterns.

Step 3: Fixed-Score Inference

Finally, given a meeting transcript, the model is asked to generate a summary corresponding to the fixed highest score (100), ideally the best quality one.

Data Construction

To benchmark our method on meeting summarization, we build an in-house dataset appropriate to the task; additionally we augment VCSum dataset (Wu et al. 2023) as mentioned in §Related Work. It should be noted that there are no gold-standard summaries in the in-house dataset, while VCSum dataset has a test set that can be used for evaluation.

ASR Transcription. Firstly, we collect meeting recordings contributed by project sponsor groups within our organization. To accommodate with 4,096 input length limit of the pre-trained model, we slice recordings into segments with similar lengths, yielding a total of 18,963 segments across 2,000 meetings conducted mainly in Mandarin Chinese. Next, a commercially established ASR system (iFLYTEK) is employed to obtain machine transcripts. Notably, ASR transcription errors, such as ill recognized words or misplaced punctuation, are not manually post-edited; we deliberately preserve those error to mimic an industry deployment environment, in which manual corrections are not available. Training on relatively low quality source text helps us build a more robust summarization system.

Summary Generation. Having harvested transcripts, we employ a pre-trained language model to generate multiple summaries ($n = 3$ in this case) for each meeting segment. For VCSum, we use the ready segmentation summaries. Since the transcripts in the dataset are accompanied with human-written summaries, the gold summaries are joined with two model generated summaries to form a trio.

Human Feedback Collection. Our annotators are divided into two groups: Group A provides conventional synchronous pairwise preferences appropriate for RLHF and CoH, following prior studies (Ziegler et al. 2020; Stiennon et al. 2020; Liu, Sferrazza, and Abbeel 2023); Group B is instructed to give asynchronous numerical scores (1 to 100 rat-

	VCSum Dataset									TL;DR Dataset						
	Automatic					Human Evaluation				Automatic			Human Evaluation			
Method	**R-1**	**R-2**	**R-L**	**R-Avg.**	**Δ**	**Acc.**	**Cohr.**	**Cvg.**	**Oval.**	**R-1**	**R-2**	**R-L**	**Acc.**	**Cohr.**	**Cvg.**	**Oval.**
Pre-trained	59.16	24.72	28.45	37.44	-	3.40	3.62	3.17	3.35	27.11	11.23	22.54	3.12	3.17	3.08	3.11
SFT	60.31	26.20	29.49	38.67	+1.23	3.58	3.83	3.80	3.75	32.31	12.32	27.33	3.81	4.20	4.01	3.61
RLHF	62.11	27.36	31.45	40.31	+2.87	3.81	4.12	3.94	3.92	32.71	13.82	28.76	4.22	4.34	4.31	3.98
CoH	61.83	27.96	32.34	40.71	+3.27	3.89	4.21	3.64	4.02	33.41	15.82	28.79	4.30	4.44	4.21	4.02
Score Tuning	**64.67**	**29.34**	**34.21**	**42.74**	**+5.30**	**4.31**	**4.43**	**4.40**	**4.73**	**34.27**	**16.20**	**31.23**	**4.65**	**4.56**	**4.71**	**4.76**
w/o L_{CA}	63.80	28.18	33.56	41.85	+4.41	-	-	-	-	-	-	-	-	-	-	-
w/o L_{HS}	63.87	28.52	33.97	42.12	+4.68	-	-	-	-	-	-	-	-	-	-	-

Table 1: Abstractive summarization results of 5-run average ROUGE scores and human evaluation on VCSum dataset and TL;DR dataset. *Acc.*, *Cohr.*, *Cvg.* and *Oval.* denote accuracy, coherence, coverage and overall respectively.

ing scale). For Group B, we split summary pairs (which are presented to Group A) and make them all independent cases, so the annotators of Group B are shown to one transcript-summary case at a time in a randomized order. Both groups are asked to evaluate considering following aspects: (1) *accuracy* (to what degree the statements in the summary are part of the post), (2) *coherence* (how easy the summary is to read on its own), (3) *coverage* (how much important information from the original post is covered), and (4) *overall quality*. Concretely, a score S is a vector containing 4 dimensions $S = \{s_{accuracy}, s_{coherence}, s_{coverage}, s_{overall}\}$.

Quality Control. Several measurements are taken to ensure the quality of the datasets. Firstly, we filter out incomplete meeting recordings during collection and discard duplicate generated summaries. We also provide detailed instructions on human feedback criteria for annotators. Despite the measures taken, quantitative feedback nonetheless causes more challenges than categorical one in terms of reaching a good inter-rater agreement. Following Hallgren (2012); Koo and Li (2016), we calculate intraclass correlation coefficient (ICC) to assess the inter-rater agreement. Specially, we set: (1) two-way mixed-effects model, (2) average measures, and (3) absolute agreement, for we concern the extent to which scores assigned by different raters on the same case differ (McGraw and Wong 1996). Implementing R *irr* package (Gamer and Fellows 2019), we obtain resulting $ICC_{In-house} = 0.57$ and $ICC_{VCSum} = 0.61$. Both ICCs indicate a moderate (between 0.50 and 0.75) inter-rater agreement on the numerical feedback.

Experiment and Results

Experiment Setup

The main task of this study is *abstractive summarization*. Formally, let $M = \{U_1, U_2, \cdots, U_N\}$ be a meeting transcript of N utterances, and the goal is producing a corresponding summary C. We test the proposed method on the in-house dataset and VCSum dataset. To investigate the potential transferability of our method to other NLP tasks, we also test an variation of Score Tuning on CWMT 2018 dataset (Bojar et al. 2018) for *machine translation* task.

Metric. Following prior studies (Shang et al. 2018; Zhong et al. 2022; Liu, Sferrazza, and Abbeel 2023), we choose the well-established **ROUGE** (Lin 2004) scores as automatic evaluation metrics for *abstractive summarization* and report standard F_1 scores of ROUGE-1, ROUGE-2 and ROUGE-L. Regarding human evaluation for *abstractive summarization*, we sample 100 examples and employ two evaluators to rate the summary results of different methods. The evaluators were asked to rate the summaries on a scale of 1 to 5, with 5 being the best, giving consideration to the same four dimensions used in human feedback collection (see §Human Feedback Collection). For *machine translation* task, conventional **BLEU** (Papineni et al. 2002) is used as the metric.

Baselines. We choose the open-source ChatGLM-6B (Du et al. 2022) as our backbone pre-trained language model for its decent bilingual performance in Chinese and English.

- **Supervised Fine-Tuning (SFT)**: we retain only positively-rated ("accepted") summaries in the datasets for the supervised fine-tuning.
- **Reinforcement Learning from Human Feedback (RLHF)** (Stiennon et al. 2020): since our limited data size prevents us from training a sufficient reward model, we import `Ziya-LLaMA-7B-Reward` (IDEA-CCNL 2021), an open-source English-Chinese bilingual reward model.
- **Chain of Hindsight (CoH)** (Liu, Sferrazza, and Abbeel 2023): we fine-tune the model using data constructed from pairs of positively-rated ("chosen") and negatively-rated ("rejected") examples. Since there are no "good" samples in our in-house dataset initially, for comparison purpose, we consider samples with scores greater than 60 as positively-rated samples.

Chinese Meeting Summarization Results

Evaluation on VCSum Dataset. As shown in Table 1, our method achieves a significant improvement of +5.30 average ROUGE score in comparison with pre-trained baseline on VCSum dataset, outperforming all other methods in all four dimensions. Interestingly, our method showed the most significant improvement in overall quality. We speculate this result was contributed by training with an *overall quality* score. However, we also suspect that evaluators would assess more loosely on overall quality and judge more strictly on specialized metrics such as *accuracy* and *coverage* for they

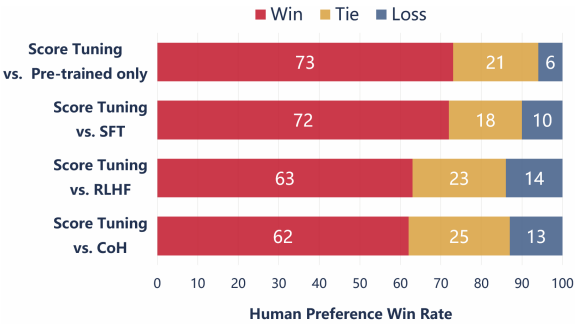

Figure 3: Abstractive summarization human evaluation results for score-tuned model compared to the models tuned by other methods on our in-house dataset.

Method	En-Zh	Zh-En	Avg.	Δ
Pre-trained	27.80	26.41	27.11	-
SFT	28.81	26.98	27.90	+0.79
RLHF	28.88	27.53	28.21	+1.09
CoH	29.06	27.55	28.31	+1.20
Score Tuning	**30.15**	**28.56**	**29.35**	**+2.24**
w/o L_{CA}	29.10	27.61	28.36	+1.25
w/o L_{HS}	30.11	28.13	29.12	+2.01

Table 2: Machine translation automatic evaluation results: 5-run average BLUE scores on CWMT 2018 dataset. *En* and *Zh* denote English and Chinese respectively.

are more objective and hence easier to measure. Nonetheless, the factors for this finding are inconclusive and worth further investigation.

Evaluation on the In-house Dataset. Since no gold-standard reference summaries are available in our in-house dataset, we rely on human evaluation to compare the performances. The evaluators are presented with pairs of outputs generated by different models and are asked to choose a preferred one or inconclusive (i.e., "tie"). The win rate is calculated by the percentage of times a model's outputs are chosen over its opponents'. As shown in Figure 3, our method achieves a significantly higher win rates than all other methods. Overall, these results demonstrate the effectiveness of our method in generating high-quality summaries, even when good samples are not available.

English Meeting Summarization Variation Results

To explore whether our strategy is language-independent, we follow the main experiment with an English meeting summarization task. Specifically, we use the TL;DR dataset (Völske et al. 2017), which contains about 3 million posts from `reddit.com` across a variety of topics, as well as summaries of the posts written by the original poster. Following the setup of Liu, Sferrazza, and Abbeel (2023), we use the filtered version containing 123,169 posts, provided by Stiennon et al. (2020). We variate our method by replacing the manual scoring with a reward model trained on the RLHF dataset constructed by Stiennon et al. (2020). For human evaluation, evaluators (proficient in English) rate summaries the same four dimensions (see §Data Construction). As shown in Table 1, Score Tuning outperforms the baseline methods on both automatics and human evaluation, which demonstrates the generalizability of the method across languages.

Machine Translation Variation Results

To test whether Score Tuning is effective in other text generation tasks, we devise another variation of our method by replacing human numerical feedback scores with BLEU scores on *machine translation* task. Notably, this variation trains a policy optimizing BLEU (Ranzato et al. 2016; Wu et al. 2016; Bahdanau et al. 2017) rather than a policy favors human preference. As shown in Table 2, compared to SFT, the improvements made by RLHF and CoH are not very high (increased by 0.3 and 0.41, respectively), which indicates that the quality of the translation measured by BLEU scores calculated with the reference sentence depends on the quality of the reference sentence. In other words, the increase in BLEU is not consistent with the improvement in translation quality. Nevertheless, our method remains competitive and outperforms other baseline methods.

Discussion and Limitations

Ablation Study. The ablation results in Table 1 and Table 2 show that removing either the L_{ca} or L_{hs} loss term result in a decrease in performance, indicating that both components are important for achieving better alignment. In particular, the L_{ca} term contributes greater improvement, suggesting a benefit of incorporating the ability to predict the scores of generated outputs.

Continuous Improvement through Iteration. To investigate whether the method can continuously improve the model by iterating Step 1 and 2, we take score-tuned models instead of the pre-trained one in Step 1 as the starting model and evaluate the iterated models. Specifically, we iterate the initial model for four rounds on use the in-house dataset. Starting from the second round, each round uses the summaries generated by sampling from the model trained after the previous round and re-annotate them.

As demonstrated in Figure 4, we observe that the initial improvement of our method is the largest, and the increase gradually decreases thereafter. This demonstrates the effectiveness of our method in cold start scenarios where data is scarce. Interestingly, the SFT and RLHF methods also show some improvement as the rounds progresses, yet it is worth noting that a drop occurs for RLHF. Overall, the results of each round of iterative training and corresponding evaluations suggest an effectiveness of our method in achieving continuous enhancement through repeated cycles of tuning and annotation. Our approach demonstrates the potential for continuous improvement, even in challenging scenarios where data is limited.

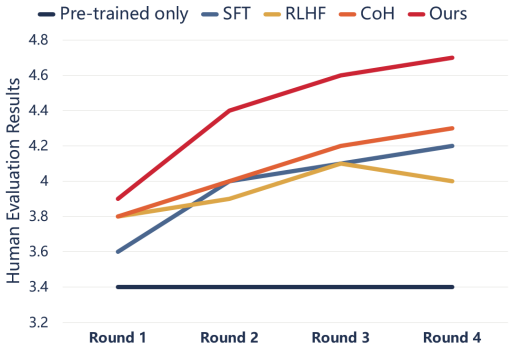

Figure 4: Human evaluation results by iteration

Asynchronous and Numerical Human Feedback. In asynchronous comparisons where the cases "to be compared with" are not concurrently presented, absolute numerical rating is a necessity. Then, the comparability and transitivity of numerical scores given by different raters is one of the major issues we must address. Thus, a reliable inter-rater agreement is considered essential for effectiveness of our approach. However, while we did not obtain great ICCs (see §Quality Control), our method demonstrates effectiveness.

Although utilizing such human feedback is consistency-wise challenging, it is not without its merits. Pairwise comparison forces raters to choose a preference, often not taking "neutral" or "equal" judgements into account. For instance, an annotator could be indecisive when exposed to a broken yet factually accurate summary and a fluent one with a lot hallucinations. Also, numerical rating empowers fine-tuning with more control. For example, by replacing the meaning of score (*overall quality*) with "summary lengths", a score-tuned model could conceivably output summaries with different lengths conditioned on corresponding scores.

Conclusion and Future Work

This study presents Score Tuning, a cold start tuning framework features utilizing asynchronous and numerical human feedback to improve alignment. Having constructed an in-house dataset, we benchmark our approach against in-house and public datasets, in comparison with typical existing tuning methods. The positive results demonstrate an effectiveness of our method in improving model performance on abstractive summarization in both English and Chinese settings. Furthermore, we test an variation of Score Tuning on machine translation, also yielding promising results. We hope that our real-world case calls more attention on leveraging rich human feedback beyond binary comparisons.

We suggest three directions for future studies of the most relevant: (1) to introduce controllable and multi-aspect scores in the framework; (2) to investigate broader transferability of the method on other tasks where numerical human feedback is well-grounded; and (3) to examine the interchangeability of absolute numerical and binary comparison human feedback. In other words, whether it is feasible to construct valid pairwise comparison data by joining cases with different scores into pairs and vice versa.

Acknowledgments

We appreciate the constructive discussions and informative feedback of all reviewers. Spacial thanks to Rui Zhang for his insightful comments.

Ethical Statement

Annotator and Evaluator Information. We recruit six annotators and evaluators who are proficient in language and reasoning skills, and the demographic information of whom is shown in Table 3. A smaller sample labelling task was given and examined before the actual annotation of this study to ensure quality. The annotators typically provide 12-15 cases of human feedback hourly.

Demographic Characteristics	Value
Total Participants	6
Age	[23, 29]
Gender (Female/ Male/ Prefer not to Answer)	3 / 3 / 0
Mandarin Chinese Proficiency	all native
English Proficiency	all proficient
Education	undergraduate[a]

Table 3: The demographic information of the annotators and evaluators. [a]All evaluators have received at least undergraduate level education.

Privacy and Licensing. Our employment of the ASR system is under a commercial-use license from iFLYTEK[1]. Names of meeting attendees and all personal information are anonymized during the annotation process, in accord with the General Data Protection Regulation (GDPR) guidelines. Our usage of the pre-trained model ChatGLM2 (Zeng et al. 2022) is under Apache-2.0 license and for research purpose only; we do not and will not implement the model for any commercial purpose.

References

Bahdanau, D.; Brakel, P.; Xu, K.; Goyal, A.; Lowe, R.; Pineau, J.; Courville, A.; and Bengio, Y. 2017. An Actor-Critic Algorithm for Sequence Prediction. arXiv:1607.07086.

Böhm, F.; Gao, Y.; Meyer, C. M.; Shapira, O.; Dagan, I.; and Gurevych, I. 2019. Better Rewards Yield Better Summaries: Learning to Summarise Without References. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3110–3120. Hong Kong, China: Association for Computational Linguistics.

Bojar, O.; Federmann, C.; Fishel, M.; Graham, Y.; Haddow, B.; Koehn, P.; and Monz, C. 2018. Findings of the 2018 Conference on Machine Translation (WMT18). In Bojar, O.; Chatterjee, R.; Federmann, C.; Fishel, M.; Graham, Y.; Haddow, B.; Huck, M.; Jimeno-Yepes, A.; Koehn, P.; Monz, C.; Negri, M.; Névéol, A.; Neves, M. L.; Post, M.; Specia, L.;

[1]https://global.xfyun.cn/products/real-time-asr

Turchi, M.; and Verspoor, K., eds., *Proceedings of the Third Conference on Machine Translation: Shared Task Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, 272–303. Association for Computational Linguistics.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165.

Carletta, J.; Ashby, S.; Bourban, S.; Flynn, M.; Guillemot, M.; Hain, T.; Kadlec, J.; Karaiskos, V.; Kraaij, W.; Kronenthal, M.; Lathoud, G.; Lincoln, M.; Lisowska, A.; McCowan, I.; Post, W.; Reidsma, D.; and Wellner, P. 2005. The AMI Meeting Corpus: A Pre-announcement. In Renals, S.; and Bengio, S., eds., *Machine Learning for Multimodal Interaction, Second International Workshop, MLMI 2005, Edinburgh, UK, July 11-13, 2005, Revised Selected Papers*, volume 3869 of *Lecture Notes in Computer Science*, 28–39. Springer.

Du, Z.; Qian, Y.; Liu, X.; Ding, M.; Qiu, J.; Yang, Z.; and Tang, J. 2022. GLM: General Language Model Pretraining with Autoregressive Blank Infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 320–335.

Gamer, M.; and Fellows, I. 2019. Various Coefficients of Interrater Reliability and Agreement.

Gao, T.; Yen, H.; Yu, J.; and Chen, D. 2023. Enabling Large Language Models to Generate Text with Citations. arXiv:2305.14627.

Gillick, D.; Riedhammer, K.; Favre, B.; and Hakkani-Tür, D. 2009. A global optimization framework for meeting summarization. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2009, 19-24 April 2009, Taipei, Taiwan*, 4769–4772. IEEE.

Hallgren, K. A. 2012. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology*, 8(1): 23.

Hu, Y.; Ganter, T.; Deilamsalehy, H.; Dernoncourt, F.; Foroosh, H.; and Liu, F. 2023. MeetingBank: A Benchmark Dataset for Meeting Summarization. In Rogers, A.; Boyd-Graber, J. L.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, 16409–16423. Association for Computational Linguistics.

IDEA-CCNL. 2021. Fengshenbang-LM. https://github.com/IDEA-CCNL/Fengshenbang-LM. Accessed: 2023-08-10.

Janin, A.; Baron, D.; Edwards, J.; Ellis, D.; Gelbart, D.; Morgan, N.; Peskin, B.; Pfau, T.; Shriberg, E.; Stolcke, A.; and Wooters, C. 2003. The ICSI Meeting Corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '03, Hong Kong, April 6-10, 2003*, 364–367. IEEE.

Koay, J. J.; Roustai, A.; Dai, X.; and Liu, F. 2021. A Sliding-Window Approach to Automatic Creation of Meeting Minutes. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, 68–75. Online: Association for Computational Linguistics.

Koo, T. K.; and Li, M. Y. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2): 155–163.

Kumar, L. P.; and Kabiri, A. 2022. Meeting Summarization: A Survey of the State of the Art. *CoRR*, abs/2212.08206.

Li, M.; Zhang, L.; Ji, H.; and Radke, R. J. 2019. Keep Meeting Summaries on Topic: Abstractive Multi-Modal Meeting Summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2190–2196. Florence, Italy: Association for Computational Linguistics.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Liu, H.; Sferrazza, C.; and Abbeel, P. 2023. Chain of Hindsight Aligns Language Models with Feedback. arXiv:2302.02676.

Liu, X.; Ji, K.; Fu, Y.; Tam, W.; Du, Z.; Yang, Z.; and Tang, J. 2022. P-Tuning: Prompt Tuning Can Be Comparable to Fine-tuning Across Scales and Tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 61–68. Dublin, Ireland: Association for Computational Linguistics.

Liu, X.; Zheng, Y.; Du, Z.; Ding, M.; Qian, Y.; Yang, Z.; and Tang, J. 2021. GPT Understands, Too. arXiv:2103.10385.

Luo, Z.; Xie, Q.; and Ananiadou, S. 2023. ChatGPT as a Factual Inconsistency Evaluator for Abstractive Text Summarization. *CoRR*, abs/2303.15621.

McGraw, K. O.; and Wong, S. P. 1996. Forming inferences about some intraclass correlation coefficients. *Psychological methods*, 1(1): 30.

Nedoluzhko, A.; Singh, M.; Hledíková, M.; Ghosal, T.; and Bojar, O. 2022. ELITR Minuting Corpus: A Novel Dataset for Automatic Minuting from Multi-Party Meetings in English and Czech. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 3174–3182. Marseille, France: European Language Resources Association.

OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, 311–318. USA: Association for Computational Linguistics.

Ranzato, M.; Chopra, S.; Auli, M.; and Zaremba, W. 2016. Sequence Level Training with Recurrent Neural Networks. arXiv:1511.06732.

Rennard, V.; Shang, G.; Hunter, J.; and Vazirgiannis, M. 2023. Abstractive Meeting Summarization: A Survey. arXiv:2208.04163.

Roit, P.; Ferret, J.; Shani, L.; Aharoni, R.; Cideron, G.; Dadashi, R.; Geist, M.; Girgin, S.; Hussenot, L.; Keller, O.; Momchev, N.; Ramos, S.; Stanczyk, P.; Vieillard, N.; Bachem, O.; Elidan, G.; Hassidim, A.; Pietquin, O.; and Szpektor, I. 2023. Factually Consistent Summarization via Reinforcement Learning with Textual Entailment Feedback. arXiv:2306.00186.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal Policy Optimization Algorithms. arXiv:1707.06347.

Shang, G.; Ding, W.; Zhang, Z.; Tixier, A. J.; Meladianos, P.; Vazirgiannis, M.; and Lorré, J. 2018. Unsupervised Abstractive Meeting Summarization with Multi-Sentence Compression and Budgeted Submodular Maximization. In Gurevych, I.; and Miyao, Y., eds., *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, 664–674. Association for Computational Linguistics.

Shinn, N.; Cassano, F.; Labash, B.; Gopinath, A.; Narasimhan, K.; and Yao, S. 2023. Reflexion: Language Agents with Verbal Reinforcement Learning. arXiv:2303.11366.

Stiennon, N.; Ouyang, L.; Wu, J.; Ziegler, D. M.; Lowe, R.; Voss, C.; Radford, A.; Amodei, D.; and Christiano, P. 2020. Learning to Summarize from Human Feedback. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713829546.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; Bikel, D.; Blecher, L.; Ferrer, C. C.; Chen, M.; Cucurull, G.; Esiobu, D.; Fernandes, J.; Fu, J.; Fu, W.; Fuller, B.; Gao, C.; Goswami, V.; Goyal, N.; Hartshorn, A.; Hosseini, S.; Hou, R.; Inan, H.; Kardas, M.; Kerkez, V.; Khabsa, M.; Kloumann, I.; Korenev, A.; Koura, P. S.; Lachaux, M.-A.; Lavril, T.; Lee, J.; Liskovich, D.; Lu, Y.; Mao, Y.; Martinet, X.; Mihaylov, T.; Mishra, P.; Molybog, I.; Nie, Y.; Poulton, A.; Reizenstein, J.; Rungta, R.; Saladi, K.; Schelten, A.; Silva, R.; Smith, E. M.; Subramanian, R.; Tan, X. E.; Tang, B.; Taylor, R.; Williams, A.; Kuan, J. X.; Xu, P.; Yan, Z.; Zarov, I.; Zhang, Y.; Fan, A.; Kambadur, M.; Narang, S.; Rodriguez, A.; Stojnic, R.; Edunov, S.; and Scialom, T. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288.

Völske, M.; Potthast, M.; Syed, S.; and Stein, B. 2017. TL;DR: Mining Reddit to Learn Automatic Summarization. In Wang, L.; Cheung, J. C. K.; Carenini, G.; and Liu, F., eds., *Proceedings of the Workshop on New Frontiers in Summarization, NFiS@EMNLP 2017, Copenhagen, Denmark, September 7, 2017*, 59–63. Association for Computational Linguistics.

Wei, J.; Hou, L.; Lampinen, A.; Chen, X.; Huang, D.; Tay, Y.; Chen, X.; Lu, Y.; Zhou, D.; Ma, T.; and Le, Q. V. 2023. Symbol tuning improves in-context learning in language models. arXiv:2305.08298.

Wu, H.; Zhan, M.; Tan, H.; Hou, Z.; Liang, D.; and Song, L. 2023. VCSUM: A Versatile Chinese Meeting Summarization Dataset. In Rogers, A.; Boyd-Graber, J. L.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, 6065–6079. Association for Computational Linguistics.

Wu, Y.; Schuster, M.; Chen, Z.; Le, Q. V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; Klingner, J.; Shah, A.; Johnson, M.; Liu, X.; Łukasz Kaiser; Gouws, S.; Kato, Y.; Kudo, T.; Kazawa, H.; Stevens, K.; Kurian, G.; Patil, N.; Wang, W.; Young, C.; Smith, J.; Riesa, J.; Rudnick, A.; Vinyals, O.; Corrado, G.; Hughes, M.; and Dean, J. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. arXiv:1609.08144.

Zeng, A.; Liu, X.; Du, Z.; Wang, Z.; Lai, H.; Ding, M.; Yang, Z.; Xu, Y.; Zheng, W.; Xia, X.; et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Zhao, X.; Wang, K.; and Peng, W. 2023. ORCHID: A Chinese Debate Corpus for Target-Independent Stance Detection and Argumentative Dialogue Summarization. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 9358–9375. Singapore: Association for Computational Linguistics.

Zhong, M.; Liu, Y.; Xu, Y.; Zhu, C.; and Zeng, M. 2022. DialogLM: Pre-trained Model for Long Dialogue Understanding and Summarization. arXiv:2109.02492.

Zhong, M.; Yin, D.; Yu, T.; Zaidi, A.; Mutuma, M.; Jha, R.; Awadallah, A. H.; Celikyilmaz, A.; Liu, Y.; Qiu, X.; and Radev, D. 2021. QMSum: A New Benchmark for Query-based Multi-domain Meeting Summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5905–5921. Online: Association for Computational Linguistics.

Zhu, C.; Xu, R.; Zeng, M.; and Huang, X. 2020. A Hierarchical Network for Abstractive Meeting Summarization with Cross-Domain Pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 194–203. Online: Association for Computational Linguistics.

Ziegler, D. M.; Stiennon, N.; Wu, J.; Brown, T. B.; Radford, A.; Amodei, D.; Christiano, P.; and Irving, G. 2020. Fine-Tuning Language Models from Human Preferences. arXiv:1909.08593.

Zou, Y.; Lin, J.; Zhao, L.; Kang, Y.; Jiang, Z.; Sun, C.; Zhang, Q.; Huang, X.; and Liu, X. 2021. Unsupervised Summarization for Chat Logs with Topic-Oriented Ranking and Context-Aware Auto-Encoders. arXiv:2012.07300.