A Hyperparameters

For inference of LLaMA-65B and LLaMA-30B to obtain the target precision curves, we use the deepspeed library (Rasley et al., 2020) with 4 A-100 GPUs. For training the fewshot recalibrator, we finetune LLaMA-7B using the AdamW optimizer and a cosine learning rate schedule. We use a warmup ratio of 0.03, learning rate of $2e-5$, and batch size of 16. We train for 4K steps for the MMLU experiments and 2K steps for the XNLI experiments. Our fine-tuning is conducted on 16 A100 GPUs of 40GB memory, and we use Deepspeed Stage 3 to ensure the 7B model fits on GPU. Our implementation of inference and finetuning are based on the Hugging Face library (Wolf et al., 2019).

B Additional Results (LLaMA-30B)

In addition to LLaMA-65B and PaLM2-Large, we also apply our fewshot recalibrator approach to LLaMA-30B to study the impact of model scales. See results in Table 6, Table 7, and Table 8. Compared to other base models (LLaMA-65B model and PaLM2-Large), we observe similar trends in the minimizing ECE and maximizing utility experiment: We find that our approach outperform all baselines in achieving the lowest calibration error with the highest win rate (Table 7). In addition, our approach outperform all baselines in selecting an abstention threshold that yields the highest utility score (Table 8). The only exception happens for the precision success rate experiment. Unlike the results of LLaMA-65B where our fewshot recalibrator outperform all the baselines including Domain Avg, for LLaMA-30B, Domain Avg achieves higher success rate than our fewshot recalibrator. The gap is particularly large for a target precision of 0.95. We hypothesis that this is because the LLaMA-30B suffers from lower accuracy compared to larger models. Thus, in the training data, the groundtruth precision curve of many custom distributions fail to hit the 95% precision level, leading to a sparsity of training data that hits the 95% precision level. As a result, when we try to infer about 95% precision level at inference time, the model predictions are more prone to error.

	Target Precision	**0.85**		**0.9**		**0.95**		
		Success	**Recall**	**Success**	**Recall**	**Success**	**Recall**	L_2
MMLU LLaMA-30B	**Sample Avg**	0.57	0.45	0.58	0.36	0.59	0.26	0.012
	Domain Avg	0.76	0.38	0.72	0.32	0.94	0.09	0.013
	Empirical	0.36	0.5	0.34	0.42	0.28	0.35	0.030
	FSC (ours)	0.75	0.35	0.68	0.26	0.52	0.16	0.007
	Oracle	1	0.46	1	0.38	1	0.28	0

Table 6: Precision Success Rate for LLaMA-30B on MMLU. Domain Avg achieves higher success rate than our fewshot recalibrator. The gap is particularly large for a target precision of 0.95. We hypothesizes that this is because the LLaMA-30B suffers from lower accuracy compared to larger models (LLaMA-65B). Thus, in the training data, the groundtruth precision curve of many custom distributions fail to hit the 95% precision level, leading to a sparsity of training data that hits the 95% precision level. As a result, when we try to infer about 95% precision level at inference time, the model predictions are more prone to error.

C Additional Results (Maximizing Utility)

Recall in §5.3, we report the utility score for 3 different settings (LLaMA-65B on MMLU, PaLM2-L on MMLU, and PaLM2-L on XNLI). Here, we provide additional pairwise comparison results that contains win/tie/lose rate of each baseline v.s. our approach in Table 9.

D Additional Results (Extrapolation)

Recall in §5.4, we show our fewshot recalibrator extrapolates well to unseen domains as demonstrated by the precision success rate experiments. Here, we provide more evidence, demonstrated

Method	ECE	win%	lose%
Base	0.093	0.2425	0.7575
Sample Avg	0.106	0.2325	0.7675
Domain Avg	0.109	0.192	0.808
Empirical	0.131	0.091	0.909
TS (Fewshot)	0.117	0.187	0.813
TS (all domains)	0.090	0.283	0.717
FSC(ours)	0.074	-	-
Oracle	0.016	0.9975	0.0025

Table 7: ECE for LLaMA-30B on MMLU. Our approach outperforms all the baselines in achieving the lowest calibration error with the highest win rate.

		$c = 0.4$				$c = 0.6$			
		Utility	**Win**	**Tie**	**Lose**	**Utility**	**Win**	**Tie**	**Lose**
XNLI PaLM2-L	**Abstain**	-0.352	0.3065	0.001	0.6925	-0.437	0.4595	0.002	0.5385
	Sample Avg	-0.326	0.231	0.212	0.557	-0.443	0.2445	0.1345	0.621
	Domain Avg	-0.329	0.185	0.145	0.67	-0.451	0.1985	0.0905	0.711
	Empirical	-0.329	0.279	0.0805	0.6405	-0.431	0.4105	0.1065	0.483
	FSC(ours)	-0.319	0	1	0	-0.428	0	1	0
	Oracle	-0.311	0.8125	0.13	0.0575	-0.416	0.8215	0.099	0.0795

Table 8: Utility Scores for LLaMA-30B on MMLU. Our approach outperforms all baselines in selecting abstention thresholds that yield the highest utility scores.

by the ECE results in Table 10. Same as the trend in the precision experiment, our approach outperforms all the baselines in achieving the lowest calibration error and more winning percentages in pairwise comparison.

		$c = 0.4$				$c = 0.6$			
		Utility	**Win**	**Tie**	**Lose**	**Utility**	**Win**	**Tie**	**Lose**
XNLI PaLM2-L	**Abstain**	-0.224	0.4	0.0005	0.5995	-0.24	0.398	0.0035	0.5985
	Curve agg	-0.206	0.183	0.3795	0.4375	-0.219	0.218	0.4975	0.2845
	Fewshot	-0.208	0.332	0.0775	0.5905	-0.225	0.299	0.246	0.455
	FSC(Ours)	-0.202	0	1	0	-0.218	0	1	0
	Oracle	-0.192	0.851	0.098	0.051	-0.213	0.709	0.22	0.071
MMLU PaLM2-L	**Abstain**	-0.162	0.484	0.0015	0.5145	-0.188	0.5085	0.0015	0.49
	Curve_agg	-0.171	0.188	0.2005	0.6115	-0.197	0.176	0.2355	0.5885
	Fewshot	-0.164	0.3095	0.0885	0.602	-0.19	0.4205	0.0885	0.491
	FSC(Ours)	-0.157	0	1	0	-0.189	0	1	0
	Oracle	-0.15	0.862	0.096	0.042	-0.18	0.823	0.124	0.053
MMLU LLaMA-65B	**Abstain**	-0.315	0.322	0.001	0.677	-0.39	0.401	0.002	0.597
	Curve_agg	-0.289	0.2715	0.2135	0.515	-0.388	0.225	0.1245	0.6505
	Fewshot	-0.293	0.3105	0.091	0.5985	-0.372	0.448	0.1305	0.4215
	FSC(Ours)	-0.284	0	1	0	-0.372	0	1	0
	Oracle	-0.277	0.787	0.139	0.074	-0.358	0.817	0.088	0.095

Table 9: Additional utility results, including the pairwise comparisons win/tie/lose rate compared to our approach. Overall, our fewshot recalibrator outperforms all baselines in achieving the highest utility scores, and more winning percentages.

Method	**ECE**	**Win**	**Lose**
Base	0.064	0.268	0.732
Sample Avg	0.052	0.4525	0.5475
Domain Avg	0.052	0.444	0.556
Empirical	0.093	0.115	0.885
TS (Fewshot)	0.095	0.1285	0.8715
TS (all domains)	0.061	0.3155	0.6845
FSC (ours)	0.049	-	-
Oracle	0.011	0.9965	0.0035

Table 10: Unseen ECE Evaluation. Our approach outperforms all the baselines in achieving the lowest calibration error and more winning percentages in pairwise comparison.