A Appendix

A.1 Data Set

(Xie et al., 2017) The CLOTH dataset, comprising English cloze tests with sentences, missing words, answers, and distractors, serves as the benchmarking dataset in this study. To enhance model training efficiency, we refined the dataset by removing extraneous spaces, special characters, double quotes, and numbers. These modifications reduce potential parsing errors and improve data quality, as detailed in Table 1. This cleaned dataset is used for both training and evaluating our language model.

Dataset	**CLOTH**				**CLOTH-F (Filtered)**			
	Train	**Dev**	**Test**	**All**	**Train**	**Dev**	**Test**	**All**
# of Questions	76,850	11,067	11,516	99,433	69,939	9,696	10,422	90,057

Table 1: CLOTH Dataset

A.2 CEFR Table

Utilizing Cambridge English Language Assessment for Languages (CEFR), the proficiency level of these words is assessed. The modal level of these words is computed to determine the learners language proficiency, which is then translated into numerical values as detailed in Table 2. This process allows for a quantification of the learner's proficiency.

Level	A1	A2	B1	B2	C1	C2
Value	1	2	3	4	5	6

Table 2: CEFR Level and Value Conversion

A.3 DG Evaluation by GPT-4

We conducted a study to assess GPT-4's ability to evaluate distractor quality using 105 English cloze questions from Taiwan's General Scholastic Ability Test (GSAT), employed for university admissions. Using the CDGP system, distractors were generated for these questions. Subsequently, GPT-4 evaluated both the original and CDGP-generated distractors. The results, depicted in Figure 3, show GPT-4 favored GSAT distractors 92.4% of the time, demonstrating its effectiveness in quality assessment of multiple-choice questions and confirming its potential as a reliable tool for automated question assessment.

Method	**Percentage**
Ground truth	92.4%
CDGP Generated Results	7.6%

Table 3: Comparison of Ground Truth vs CDGP's Generated Results

A.4 Implementation Details

We employed the 7B version of the LLaMA2 model, optimized through the LoRA method for efficient usage under constrained computational resources. The model was fine-tuned using two NVIDIA GeForce RTX 3090 Ti GPUs. The initial QSG model training process involved training on 20,000 entries from the CLOTH dataset with the following parameters: a train batch size of 4 per device, gradient accumulation steps of 32, 3 training epochs, and an initial learning rate of 1e-4. The initial DG model training process involved training on 10,000 entries from the CLOTH dataset with the same parameters, except for an initial learning rate of 3e-4.

The personalized QSG model training process involved training on 4,000 entries from the CLOTH dataset, which included 2,000 entries of personalized level training data and 2,000 entries from the training data used for the initial QSG model. The personalized DG model training process involved training on 3,000 entries from the CLOTH dataset, which included 2,000 entries of personalized level training data and 1,000 entries from the training data used for the initial DG model. Both personalized models used the same parameter set: a train batch size of 4 per device, gradient accumulation steps of 32, 3 training epochs, and an initial learning rate of 3e-4.

Experiment	Model Configuration	Mean	Median	STD
A1 Sentence Difficulty	**Baseline Model**: 20000 training entries	1.88	1.67	0.886
	Enhanced A1 Model: 2000 new entries + 10% old entries	2.19	2.0	0.997
A2 Sentence Difficulty	**Baseline Model**: 20000 training entries	2.42	2.0	0.80
	Enhanced B2 Model: 2000 new entries + 10% old entries	3.0	3.0	0.893
B1 Sentence Difficulty	**Baseline Model**: 20000 training entries	2.82	2.50	0.770
	Enhanced B2 Model: 2000 new entries + 10% old entries	3.11	3.50	0.654
B2 Sentence Difficulty	**Baseline Model**: 20000 training entries	3.05	3.0	0.714
	Enhanced B2 Model: 2000 new entries + 10% old entries	3.71	4.0	0.593
A1 Sentence Difficulty	**Baseline Model**: 10000 training entries	2.10	2.0	0.933
	Enhanced A1 Model: 2000 new entries + 10% old entries	1.70	1.50	0.869
A2 Sentence Difficulty	**Baseline Model**: 10000 training entries	2.48	2.50	0.808
	Enhanced B2 Model: 2000 new entries + 10% old entries	2.54	2.5	0.852
B1 Sentence Difficulty	**Baseline Model**: 10000 training entries	2.9	3	0.667
	Enhanced B2 Model: 2000 new entries + 10% old entries	2.79	2.75	0.679
B2 Sentence Difficulty	**Baseline Model**: 10000 training entries	3.29	3.5	0.759
	Enhanced B2 Model: 2000 new entries + 10% old entries	3.67	4.0	0.505
A1 Distractor Difficulty	**Baseline Model**: 20000 training entries	1.56	1.50	0.775
	Enhanced A1 Model: 2000 new entries + 10% old entries	1.52	1.0	0.776
A2 Distractor Difficulty	**Baseline Model**: 20000 training entries	1.84	2.0	0.830
	Enhanced A1 Model: 2000 new entries + 10% old entries	1.85	2.0	0.730
B1 Distractor Difficulty	**Baseline Model**: 20000 training entries	2.02	2.0	0.807
	Enhanced A1 Model: 2000 new entries + 10% old entries	1.92	2.0	0.828
B2 Distractor Difficulty	**Baseline Model**: 20000 training entries	2.10	2.0	0.985
	Enhanced B2 Model: 2000 new entries + 10% old entries	2.15	2.0	1.130
A1 Distractor Difficulty	**Baseline Model**: 10000 training entries	1.70	1.67	0.848
	Enhanced A1 Model: 2000 new entries + 10% old entries	1.52	1.17	0.818
A2 Distractor Difficulty	**Baseline Model**: 10000 training entries	2.03	2.0	0.868
	Enhanced A1 Model: 2000 new entries + 10% old entries	1.86	2.0	0.790
B1 Distractor Difficulty	**Baseline Model**: 10000 training entries	2.17	2.0	0.881
	Enhanced A1 Model: 2000 new entries + 10% old entries	1.91	2.0	0.898
B2 Distractor Difficulty	**Baseline Model**: 10000 training entries	2.08	2.17	1.026
	Enhanced B2 Model: 2000 new entries + 10% old entries	2.40	2.5	1.189

Table 4: Experiment results comparing different training entries tuning for A1 and B2 difficulty levels across various experiments.

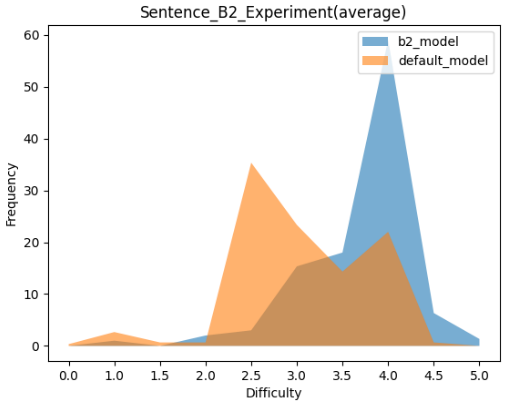

Figure 3: Distribution of Difficulty of Generated Question Stems

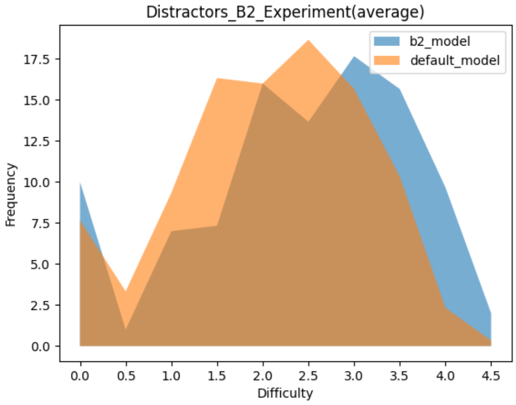

Figure 4: Distribution of Difficulty of Generated Distractors

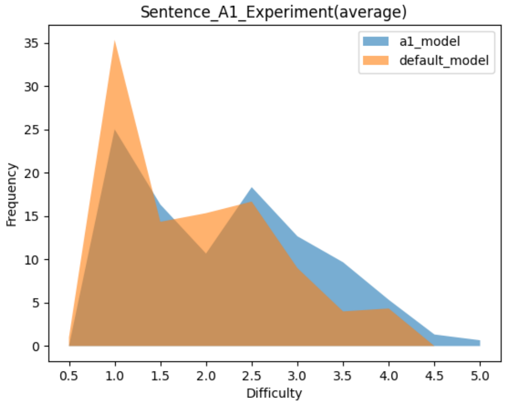

Figure 1: Distribution of Difficulty of Generated Question Stems

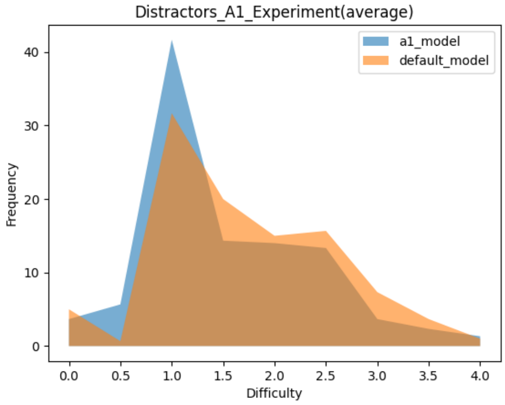

Figure 2: Distribution of Difficulty of Generated Distractors

References

Qizhe Xie, Guokun Lai, Zihang Dai, and Eduard Hovy. 2017. Large-scale cloze test dataset created by teachers. *arXiv preprint arXiv:1711.03225*.