Supplementary Information for STL-LoRA: A Parameter-Efficient Single-Task Framework for Multi-Trait Arabic Automated Essay Scoring

Soad Abu Kwaik[1] and Hakan Koyuncu[2,*]

[1]Electrical and Computer Engineering Department, Altinbas University, Istanbul, Turkey
[2]Computer Engineering Department, Altinbas University, Istanbul, Turkey

[*]Corresponding author: hakan.koyuncu@altinbas.edu.tr

Contents

Supplementary Tables

Table S1: Comparative summary of representative AES literature showing the architectural evolution of the field. QWK scores are intentionally omitted because results span fundamentally different benchmarks (ASAP, ASAP++, AR-AES, QAES, TAQEEM, LAILA) that are not directly comparable. ⋆ = this work.

Study	Year	Lang.	Dataset	Scoring	Paradigm	Key Contribution
Foundational neural AES (holistic, prompt-specific)						
Taghipour & Ng	2016	En	ASAP	Holistic	STL	First LSTM-based end-to-end AES; established ASAP as the standard benchmark
Dong et al.	2017	En	ASAP	Holistic	STL	Hierarchical CNN capturing word- and sentence-level features
Yang et al.	2020	En	ASAP	Holistic	STL	Fine-tuned BERT for holistic AES; demonstrated transformer superiority over recurrent models
Multi-trait essay scoring (English, ASAP++)						
Mathias & Bhattacharyya	2020	En	ASAP++	Multi-trait	MTL	First neural MTL for multi-trait scoring; shared BiLSTM with trait-specific output heads
Ridley et al. (CTS)	2021	En	ASAP++	Multi-trait	MTL	Shared prompt-and-trait representations improve cross-prompt generalisation
Kumar et al.	2022	En	ASAP++	Multi-trait	MTL	Hierarchical MTL; holistic score as primary task with traits as auxiliary objectives
Do et al. (ProTACT)	2023	En	ASAP++	Multi-trait	MTL	Essay–prompt cross-attention and trait-similarity loss for prompt-aware trait scoring
Chen & Li (PMAES)	2023	En	ASAP++	Multi-trait	MTL	Prompt-mapping contrastive learning; inter-attribute correlation objective
Do et al. (ArTS)	2024	En	ASAP++	Multi-trait	Seq2Seq	Autoregressive trait score generation via T5; exploits inter-trait ordering dependencies
Xu et al.	2025	En	ASAP++	Multi-trait	MTL	Syntactic and POS-enriched prompt representations; prompt–essay relation features
Chen et al.	2025	En+Ar	ASAP++/LAILA	Multi-trait	MoE	Mixture-of-experts with specialised scoring, ranking, and prompt-adherence experts
Arabic AES						
Ghazawi & Simpson	2024	Ar	AR-AES	Holistic	STL	First AraBERT-based Arabic AES corpus and holistic scoring baseline
Bashendy et al.	2024	Ar	QAES	Multi-trait	STL	First publicly available Arabic multi-trait dataset (195 essays, 7 CAST traits)
Sayed et al.	2025	Ar	TAQEEM	Multi-trait	STL	TAQEEM shared task; demonstrates feature engineering competitive with neural fine-tuning
Bashendy et al.	2025	Ar	LAILA	Multi-trait	Multiple	Largest Arabic multi-trait corpus (7,859 essays); standardised LOPO evaluation protocol
LLM-based AES						
Ghazawi & Simpson	2025	Ar	AR-AES	Holistic	LLM	Systematic zero/few-shot LLM evaluation on Arabic AES; fine-tuned AraBERT still superior
This work						
STL-LoRA ⋆	2026	**Ar**	**LAILA**	**Multi-trait**	**STL**	**First PEFT (LoRA) approach for Arabic multi-trait AES; independent per-trait models achieve state-of-the-art on LAILA; first zero-shot frontier LLM baselines established**

Table S2: LAILA dataset statistics per prompt. Avg. len. is measured in words after preprocessing. Score columns show mean±SD.

Prompt	Essays	Avg. Len. (words)	Relevance (0–2)	Organisation (0–5)	Grammar (0–5)
P1	1,122	165	1.52±0.76	2.29±1.47	2.08±1.34
P2	1,168	178	1.80±0.55	2.59±1.17	2.46±1.15
P3	521	162	1.86±0.47	2.35±1.12	2.14±1.13
P4	500	155	1.72±0.64	2.13±1.16	1.94±1.12
P5	1,181	160	1.75±0.61	2.52±1.36	2.34±1.30
P6	1,162	163	1.84±0.51	2.64±1.33	2.32±1.21
P7	1,143	207	1.92±0.37	3.10±1.07	2.73±1.05
P8	1,062	190	1.83±0.53	2.81±1.20	2.53±1.17
All	7,859	175	1.78±0.58	2.60±1.29	2.36±1.22

Table S3: Four experimental systems in the 2×2 factorial design. ⋆ indicates the proposed system. Trainable parameters include encoder and classification head(s). STL trains 7 independent models per fold, requiring 7× the wall-clock time of MTL despite identical per-model cost under LoRA.

System	Paradigm	Strategy	Trainable	Runs/Fold
A1	STL	Full FT	135.2M × 7	7
A2 ⋆	STL	LoRA	0.89M × 7	7
B1	MTL	Full FT	135.2M	1
B2	MTL	LoRA	0.89M	1

Table S4: STL-LoRA: single-seed vs 5-seed ensemble. Average QWK across 8 LOPO folds. The ensemble significantly outperforms single-seed ($p = 0.012$, Wilcoxon).

Config.	Rel	Org	Voc	Sty	Dev	Mec	Grm	**Avg**
Single-seed	.498	.622	.646	.619	.592	.585	.627	.599
5-seed ens.	**.503**	**.656**	**.647**	**.669**	**.596**	**.614**	**.647**	**.619**
Δ	+.005	+.034	+.001	+.050	+.004	+.029	+.020	**+.020**

Table S5: Per-fold average QWK for STL-LoRA: single-seed vs 5-seed ensemble. $\Delta > 0$ favours ensemble.

Fold	Single-seed	5-seed ens.	Δ
P1	.541	**.580**	+.039
P2	.657	**.657**	+.000
P3	.665	**.692**	+.027
P4	.699	**.698**	−.001
P5	.461	**.467**	+.006
P6	.648	**.667**	+.019
P7	.563	**.565**	+.002
P8	.553	**.626**	+.073
Avg	.599	**.619**	**+.020**

Table S6: LoRA vs full fine-tuning within each paradigm. Average QWK across 8 LOPO folds (single-seed). LoRA outperforms full FT by +0.053 in STL ($p = 0.020$) and +0.052 in MTL ($p = 0.016$).

System	Rel	Org	Voc	Sty	Dev	Mec	Grm	**Avg**
Single-task learning (STL)								
A1: STL+Full	.357	.598	.597	.630	.497	.564	.581	.546
A2: STL-LoRA	**.498**	**.622**	**.646**	**.619**	**.592**	**.585**	**.627**	**.599**
Δ	+.141	+.024	+.049	−.011	+.095	+.021	+.046	**+.053**
Multi-task learning (MTL)								
B1: MTL+Full	.175	.606	.618	.664	.544	.614	.518	.534
B2: MTL+LoRA	**.427**	**.609**	**.616**	**.644**	**.587**	**.592**	**.630**	**.586**
Δ	+.252	+.003	−.002	−.020	+.043	−.022	+.112	**+.052**

Table S7: Zero-shot LLM baselines vs STL-LoRA (5-seed ensemble) per-trait. Average QWK across 8 LOPO folds. Best per trait in **bold**.

System	Rel	Org	Voc	Sty	Dev	Mec	Grm	**Avg**
GPT-5.2	.564	**.718**	.622	.613	.524	.354	.408	.529
Claude Sonnet 4-6	.571	.643	.592	.551	.494	.298	.338	.498
STL-LoRA	**.503**	.656	**.647**	**.669**	**.596**	**.614**	**.647**	**.619**

Table S8: Learned log-variance parameters $s_t = \log \sigma_t^2$ from MTL+LoRA, averaged over 8 folds. Near-uniform weights indicate no meaningful trait difficulty differences.

Param.	Rel	Org	Voc	Sty	Dev	Mec	Grm
s_t	−0.153	−0.152	−0.154	−0.152	−0.153	−0.154	−0.153
σ_t^2	0.858	0.859	0.857	0.859	0.858	0.857	0.858

Table S9: Per-trait QWK variability across five training seeds for STL-LoRA, averaged over 8 folds. The ensemble consistently outperforms both the per-seed mean and best single seed.

Metric	Rel	Org	Voc	Sty	Dev	Mec	Grm	**Avg**
Per-seed mean	.487	.621	.643	.618	.590	.603	.631	.599
Per-seed std	.145	.098	.054	.126	.081	.089	.193	.112
Best seed	.498	.622	.646	.619	.592	.585	.627	.599
5-seed ens.	**.503**	**.656**	**.647**	**.669**	**.596**	**.614**	**.647**	**.619**
Gain/mean	+.016	+.035	+.004	+.051	+.006	+.011	+.016	+.020
Gain/best	+.005	+.034	+.001	+.050	+.004	+.029	+.020	+.020

Table S10: 95% paired bootstrap CIs for pairwise QWK differences (10,000 resamples over 8 LOPO folds). CIs strictly above zero indicate significant improvements.

Comparison	Δ	95% CI	$P(\Delta > 0)$
LoRA advantage within paradigms			
STL-LoRA vs STL+Full	+.053	[+.016, +.094]	0.99
MTL+LoRA vs MTL+Full	+.052	[+.019, +.091]	0.98
STL vs MTL (LoRA only)			
STL-LoRA vs MTL+LoRA	+.013	[−.007, +.030]	0.90
Ensemble vs single-seed			
5-seed vs 1-seed	+.020	[+.006, +.039]	0.99
Relevance trait only			
STL-LoRA vs STL+Full	+.141	[+.028, +.271]	0.99
STL-LoRA vs MTL+Full	+.323	[+.162, +.474]	1.00
MTL+LoRA vs MTL+Full	+.252	[+.121, +.393]	1.00

Table S11: Nemenyi post-hoc pairwise comparisons following Friedman test ($\chi^2 = 11.25$, $p = 0.010$). CD = 1.297 at $\alpha = 0.05$.

Comparison	Rank diff.	CD	Sig.?
A2 vs B1	1.875	1.297	Yes*
B2 vs B1	1.500	1.297	Yes*
A1 vs B1	0.750	1.297	No
A2 vs B2	0.375	1.297	No
A2 vs A1	1.125	1.297	No
B2 vs A1	0.750	1.297	No

Table S12: Hyperparameter sweep for MTL+Full on two folds where Relevance catastrophe occurred (QWK = 0.000). Moderate regularisation rescues both folds; strong regularisation is fold-dependent.

Fold	Regularisation	WD	Dropout	Rel QWK
Fold P3				
P3	Baseline	0.01	0.1	0.000
P3	Moderate	0.10	0.1	0.351
P3	Strong	0.30	0.3	0.482
Fold P7				
P7	Baseline	0.01	0.1	0.000
P7	Moderate	0.10	0.1	0.453
P7	Strong	0.30	0.3	0.000
Reference (STL-LoRA, default settings)				
P3	LoRA default	0.01	0.1	0.487
P7	LoRA default	0.01	0.1	0.544

Supplementary Figures

Figure S1: Multi-task learning architecture. A shared AraBERT encoder feeds seven independent linear heads (768→1), one per CAST trait. Losses are combined via uncertainty weighting. Two variants are evaluated: B1 (MTL+Full FT) and B2 (MTL+LoRA).

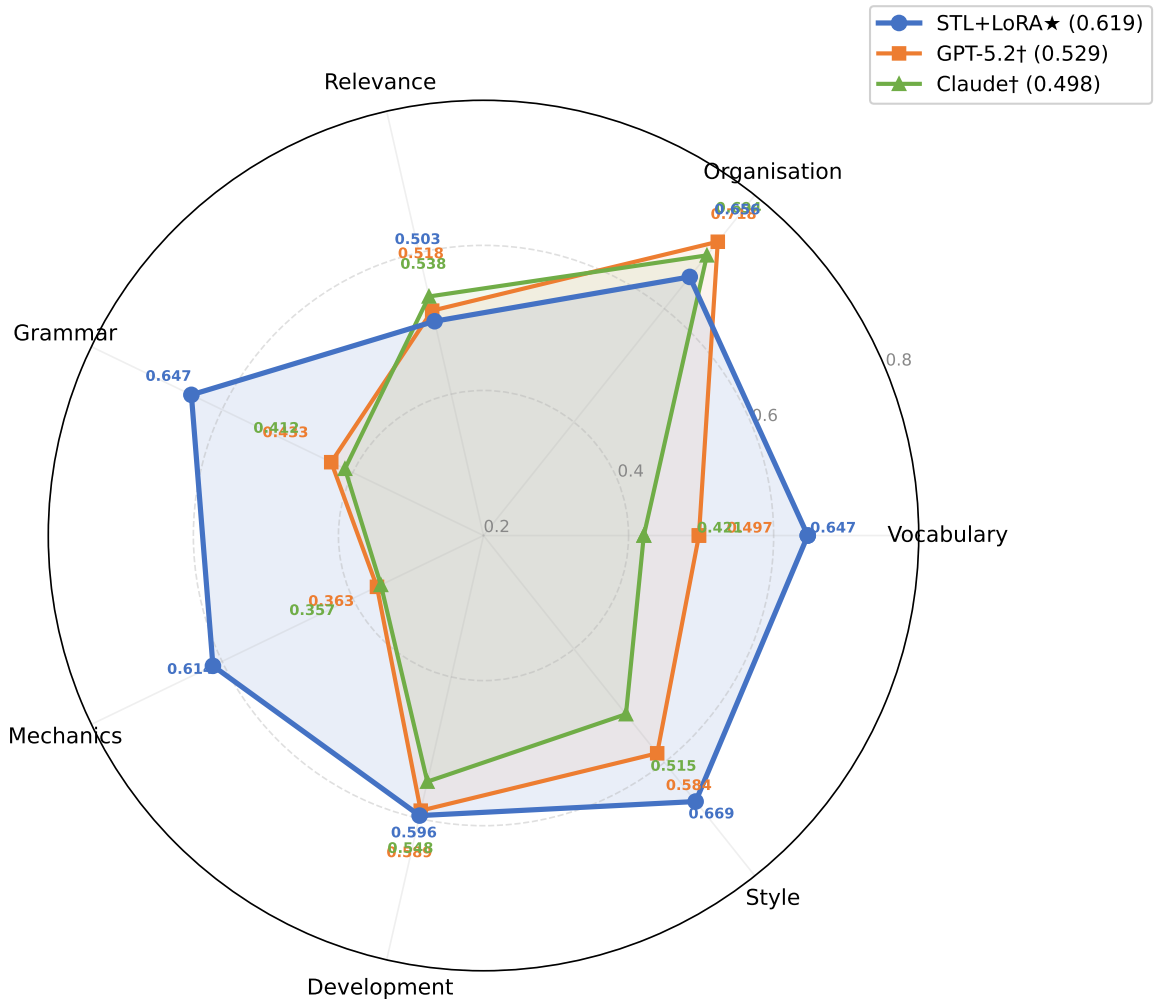

Figure S2: Radar chart comparing trait profiles of STL-LoRA, GPT-5.2, and Claude Sonnet 4-6. LLMs show complementary strengths on discourse-level traits (Organisation, Relevance) versus STL-LoRA's consistency on surface-level traits (Mechanics, Grammar).