

How Much Training Data Is Enough? Fine-Tuning LLMs for Short-Answer Scoring

Joshua A. McGrane

University of Melbourne
joshua.mcgrane@unimelb.edu.au

David Torres Irribarra

Pontificia Universidad Católica de Chile
davidtorres@uc.cl

Abstract

Fine-tuned models approached state-of-the-art agreement using small labelled sets. Open-weight models met operational criteria on seven of ten items, with a median of 50 responses per score point among passing items. A single marking exercise may supply enough data for automated short-answer scoring.

1 Introduction

Automated short-answer grading (ASAG) has a long history, but instruction-tuned large language models (LLMs) have reset the practical questions that assessment bodies face. Recent work shows that LLMs can approximate human scores on constrained short-answer items, whether prompted with a rubric and a small number of exemplars (Jiang and Bosch, 2024; Lee et al., 2024; Huang and Wilson, 2025) or fine-tuned on human-scored responses (Latif and Zhai, 2024; Ormerod and Kwako, 2024; Ormerod and Kehat, 2025). What it does not establish is whether such systems meet the agreement criteria that high-stakes assessment already requires of automated scoring, or how many human-scored responses are needed to meet those criteria.

Two reporting practices limit the available evidence. The first is that agreement is often reported primarily as quadratic weighted kappa (QWK). The evaluation frameworks developed within testing organisations, notably those of Williamson et al. (2012) and Loukina et al. (2020), require more than chance-corrected agreement with a single rater; machine and human mean scores must be sufficiently close, and machine-human agreement must not fall too far below that of two trained raters on the same item. QWK alone does not establish either requirement. QWK-only reporting can therefore overstate operational readiness, as the results below illustrate. The second is that fine-tuning studies train on whatever labelled data the benchmark

supplies, typically a thousand or more responses per item, which tells a practitioner what a large annotation effort achieves but not how agreement grows with the size of the labelled set, nor whether the choice of model matters once fine-tuning begins. Ormerod and Kwako (2024) illustrate both, as they fine-tuned four open-weight models on fixed ASAP-SAS training sets and reported QWK alone.

This paper treats the number of labelled responses as an explicit experimental factor and evaluates fine-tuning against the operational criteria of educational measurement. We fine-tuned GPT-4o-mini, as a closed-source comparator, and gpt-oss-20b and gpt-oss-120b, as small and large open-weight models, on the ten ASAP-SAS items at six training set sizes from 10 to 100 labelled responses per score point.¹ Four questions organise the analysis. First, how many labelled responses per score point are needed before a fine-tuned model satisfies a defensible set of operational criteria? Second, once fine-tuned, do closed-source, small open-weight and large open-weight models still differ? Third, can an item's operational readiness, and which agreement criterion it fails, be anticipated from its rubric and score distribution before collecting additional labelled responses? Fourth, how does agreement reached with labelled sets of this size compare with that of the published systems for the same items?

2 Data

2.1 The ASAP-SAS benchmark

The Automated Student Assessment Prize Short Answer Scoring (ASAP-SAS) dataset (Shermis, 2015) comprises ten short-answer items from standardised assessments in United States schools: three science items (1, 2 and 10), two biology items (5 and 6) and five English Language Arts (ELA)

¹The companion prompting experiment on the same items and models is reported elsewhere.

items (3, 4, 7, 8 and 9). Items 1, 2, 5 and 6 are scored 0 to 3 and the remainder 0 to 2 (Table 1).² The two biology items are extremely skewed, with 78% and 85% of responses at the lowest score point and fewer than 3% at the highest; the Science and ELA items are comparatively balanced.

2.2 Evaluation set and training pool

For each item, 600 responses were randomly selected from the public training release, stratified by score to reproduce its score-point distribution, and held out as the evaluation set for every model at every training set size. The evaluation set was drawn from the training release rather than the public test split because the training release is the only part of the dataset in which every response carries two independent human scores. Two of the criteria of Section 3.4 cannot be computed without the second rating: the degradation criterion requires human-human QWK on the same responses as machine-human QWK, and PRMSE requires an estimate of the rater error variance. The remaining responses (658 to 1,193 per item) form the pool from which fine-tuning examples were drawn, less those used as in-context exemplars in a companion experiment. There is no validation set because no hyperparameter was tuned, i.e., the procedure of Section 3.2 was fixed before the first run.

Agreement statistics were computed against the first rater. Loukina et al. (2020) recommend at least 1,000 double-scored responses for PRMSE, but suggest that 500 may suffice when the interrater correlation exceeds 0.65; every item meets these latter conditions.

3 Method

3.1 Models

Three models were fine-tuned. GPT-4o-mini (gpt-4o-mini-2024-07-18) is closed-source and was fine-tuned through the OpenAI fine-tuning API.³ gpt-oss-20b and gpt-oss-120b are open-weight mixture-of-experts models released by OpenAI in 2025, with approximately 21 and 117 billion parameters. They were selected to compare closed-source and

open-weight models from one developer, including two open-weight models of different sizes.

3.2 Fine-tuning procedure

Each model was fine-tuned at six training set sizes, $s \in \{10, 20, 30, 50, 70, 100\}$, where s is the target number of labelled training responses per score point on each item. At each s a single adapter (for GPT-4o-mini, a single fine-tuned model) was trained on the pooled training sets of all ten items, each response carrying its own item’s system message (Section 3.3), and that model was evaluated on every item. We fixed s per score point rather than in total because it is the sparsely populated upper score points, not the modal one, that limit agreement on skewed items. Training responses were drawn by score-stratified random sampling with a fixed seed, with nested draws so that each training set contained the next smaller set.

The open-weight models were fine-tuned with rank-stabilised LoRA adapters (Hu et al., 2022; Kalajdziewski, 2023) on the attention projections (rank 8, alpha 16), for three epochs at a learning rate of 2×10^{-4} and an effective batch size of 16, on a single 80 GB A100. gpt-oss-20b was trained with its base weights in bfloat16; gpt-oss-120b does not fit at that precision, so its base weights were quantised to 4 bits and frozen in the manner of QLoRA (Detrmers et al., 2023), with the adapter in full precision. Loss was computed on the completion tokens only, matching the default of the OpenAI API, through which GPT-4o-mini was fine-tuned for two epochs with otherwise default hyperparameters. The base model at $s = 0$ was the zero-shot reference.

3.3 Prompt construction and training examples

Each training example, and each evaluation prompt, consisted of a system message and a user message. The system message was written for each item as a translation of the published ASAP-SAS rubric into explicit scoring instructions and was held fixed across models and training set sizes. It comprised a role statement (an automated scorer with ‘knowledge’ of the subject matter), a task statement giving the item stem, any source materials and the list of correct response elements, with explicit exclusions and a note that students may express an element in different words, and a scoring statement mapping the number of correct elements to score points and requiring a single integer in reply. The stems, ma-

²The released rubrics for the items marked ‘key elements’ enumerate a set of required elements and map the number present to a score point; we retain the dataset’s own term. Holistic rubrics call for a single overall judgement of the response.

³GPT-4o itself was not fine-tuned because API pricing made a sweep over training set sizes prohibitive.

Item	Domain	Scale	Rubric style	Human-human QWK [95% CI]	Training pool
1	Science	0–3	key elements	0.949 [0.939, 0.957]	1072
2	Science	0–3	key elements	0.915 [0.901, 0.930]	658
3	ELA	0–2	holistic	0.761 [0.735, 0.784]	1193
4	ELA	0–2	holistic	0.749 [0.719, 0.777]	1042
5	Biology	0–3	key elements	0.955 [0.941, 0.965]	1175
6	Biology	0–3	key elements	0.962 [0.944, 0.975]	1177
7	ELA	0–2	key elements	0.971 [0.964, 0.978]	1184
8	ELA	0–2	holistic	0.862 [0.842, 0.879]	1184
9	ELA	0–2	holistic	0.833 [0.815, 0.851]	1183
10	Science	0–2	key elements	0.884 [0.867, 0.900]	1025

Table 1: ASAP-SAS items, rubric style, human-human agreement and training pool size. Human-human QWK is computed from the two independent raters on the full public training release, with 2,000-sample bootstrap intervals. Training pool is the number of responses remaining after the 600 evaluation responses are removed.

terials and rubrics are those released in 2012; the system messages that paraphrase them are ours and available on request. The user message contained the student response. In the fine-tuning files the assistant turn was the human score; at evaluation it was generated at temperature 0.

No scored exemplar appeared in any prompt at any training set size; the labelled responses entered the model only through gradient updates, and the evaluation prompt was identical across all s , including $s = 0$. The zero-shot condition was therefore a rubric-only baseline in the strict sense, and the fine-tuned conditions differed from it only in the trained weights. For the gpt-oss models the ‘reasoning’ channel was suppressed at training and inference so that the completion was the bare score.

3.4 Agreement criteria

Every (model, item, s) condition was evaluated on the 600 evaluation responses against the criteria of [Williamson et al. \(2012\)](#), which must be met simultaneously: QWK against the first human rater of at least 0.70; degradation of QWK relative to human-human QWK on the same item of no more than 0.10; a standardised mean difference (SMD) between machine and human scores, using the pooled standard deviation, of magnitude no greater than 0.15.

Alongside these we report the proportional reduction in mean squared error (PRMSE) of [Loukina et al. \(2020\)](#). We use 0.70 as a minimum reference level, consistent with [McCaffrey et al. \(2022\)](#) and [Casabianca et al. \(2023\)](#). Higher thresholds may be needed depending on assessment stakes and the contribution of automated scores to reported results; McCaffrey et al. suggest 0.95 when automated scores must be interchangeable with human ratings. Let h_1 and h_2 be the two human scores, m

the machine score, $\bar{h} = (h_1 + h_2)/2$ and $n = 600$. Then

$$\sigma_e^2 = \sum (h_1 - h_2)^2 / 2n,$$

$$\sigma_\tau^2 = \text{Var}(\bar{h}) - \sigma_e^2 / 2,$$

$$\text{MSE} = \text{mean}((m - \bar{h})^2) - \sigma_e^2 / 2,$$

$$\text{PRMSE} = 1 - \text{MSE} / \sigma_\tau^2.$$

PRMSE measures the reduction in prediction error relative to predicting the mean true score, after adjusting for rater error, and is sensitive to differences in the location and scale of machine scores. It is a separate criterion, not a fourth component of the Williamson et al. set.

3.5 Statistical analysis

To model agreement as a function of training set size we fitted a linear mixed model to the Fisher- z transform of QWK in each condition, $z = \text{atanh}(\text{QWK})$, with each condition weighted by the inverse of a bootstrap variance estimate on the z scale from 1,000 resamples of the evaluation responses. Because a single log-linear curve through the zero-shot and fine-tuned conditions would misspecify the zero-shot point as the limit of a curve whose slope is identified only on the fine-tuned side, we fitted a piecewise specification with a separate term for the zero-shot condition,

$$z \sim \text{zs} + \text{model} + \text{zs:model} \\ + \text{model:log}(1 + s_{\text{ft}}) \\ + (1 + \text{log}(1 + s_{\text{ft}}) \mid \text{item}),$$

where zs indicates the zero-shot condition and s_{ft} equals s for fine-tuned conditions and zero otherwise. The piecewise specification was preferred to a single log-linear curve by a wide margin ($\Delta\text{AIC} = -40.7$, $\Delta\text{BIC} = -24.0$). The

model was fitted by REML in lme4 (Bates et al., 2015), with Satterthwaite degrees of freedom from lmerTest (Kuznetsova et al., 2017) and Tukey-adjusted pairwise contrasts on the fine-tuning slopes. The same fixed and random effects were refitted to PRMSE, degradation and $|\text{SMD}|$ on their original scales without weights. The QWK and PRMSE slope contrasts and the QWK zero-shot model contrasts were also evaluated by an item-level cluster bootstrap ($B = 500$). With only ten items, we deferred to the bootstrap when it disagreed with model-based inference. The degradation and $|\text{SMD}|$ fits describe population curves, with model-based confidence intervals. Item-level results are reported as the smallest observed s at which a (model, item) condition satisfies all three Williamson et al. criteria, and as a model-based estimate from a log-linear fit to each (model, item) series.

4 Results

4.1 Agreement as a function of training set size

Figure 1 plots the marginal curves for the four criteria against s for the three models, as fixed-effect estimates from the piecewise model; Table 2 gives fitted QWK across training set sizes and all four outcomes at $s = 100$. Four features warrant comment. First, agreement rose steeply with few labelled responses. From a zero-shot baseline near 0.60, ten responses per score point raised marginal QWK to approximately 0.68, twenty carried it above the 0.70 floor, and one hundred brought it to approximately 0.79, against a mean human-human QWK of 0.888. The gain was steepest over the first ten to fifty responses.

Second, the zero-shot condition was not the origin of the fine-tuning curve. The zero-shot main effect was significant ($F(1, 184.5) = 16.9$, $p < 0.001$), and each pretrained baseline sat above the value that back-extrapolating the fine-tuning curve to $s = 0$ would predict, by 0.18, 0.13 and 0.07 on the z scale for GPT-4o-mini, gpt-oss-20b and gpt-oss-120b ($p < 0.001$, 0.016 and 0.18). By $s = 10$ every model already exceeded its zero-shot value, so whether agreement dips between $s = 0$ and $s = 10$ or the curve simply flattens below the smallest training set size cannot be told from these data. The displacement did not differ by model ($p = 0.35$ for the zero-shot by model interaction).

Third, marginal PRMSE tracked QWK but sat

s	GPT-4o-mini	gpt-oss-20b	gpt-oss-120b
QWK by training set size			
0 (zero-shot)	0.620	0.599	0.567
10	0.676	0.681	0.687
20	0.714	0.717	0.724
30	0.735	0.738	0.744
50	0.760	0.762	0.768
70	0.776	0.777	0.783
100	0.791	0.792	0.798
Other outcomes at $s = 100$			
PRMSE	0.635	0.641	0.655
Degradation	-0.092	-0.092	-0.087
$ \text{SMD} $	0.096	0.079	0.077

Table 2: Marginal agreement from the fixed effects of the four piecewise models. QWK is back-transformed from the Fisher- z scale; the other outcomes are fitted without weights on their original scales. The zero-shot estimate is separate from the fine-tuning curve. Figure 1 shows confidence intervals.

lower: at $s = 100$ the refit gave 0.635, 0.641 and 0.655, with upper bootstrap limits of 0.72 to 0.74. The average item did not reach the 0.70 threshold for any training set size (Section 5.3).

Fourth, fitted degradation first met the -0.10 floor at $s = 100$ among the observed training sizes for all three models, making it the binding population criterion. For $|\text{SMD}|$, GPT-4o-mini first met the 0.15 ceiling at $s = 50$, whereas both open-weight models met it at every observed fine-tuned size. These comparisons concern fitted means; the confidence intervals in Figure 1 show the uncertainty around threshold crossings.

4.2 Comparison of fine-tuning learning rates

The estimated learning rates were similar across models, with no statistically detectable pairwise differences. The fine-tuning log-slopes on the z scale were 0.114 (SE 0.016) for GPT-4o-mini, 0.110 (0.016) for gpt-oss-20b and 0.112 (0.016) for gpt-oss-120b. No pairwise difference exceeded 0.005, all three Tukey-adjusted contrasts returned $p > 0.97$, and the cluster bootstrap was consistent (all pairwise $p_{\text{boot}} > 0.81$). The PRMSE refit showed the same pattern, with slopes of 0.162, 0.152 and 0.151 and all pairwise $p_{\text{boot}} > 0.67$. The zero-shot baselines differed in point estimate, GPT-4o-mini sitting 0.08 above gpt-oss-120b on the z scale (Tukey $p = 0.015$), but the contrast was not significant using cluster-robust inference ($p_{\text{boot}} = 0.29$).

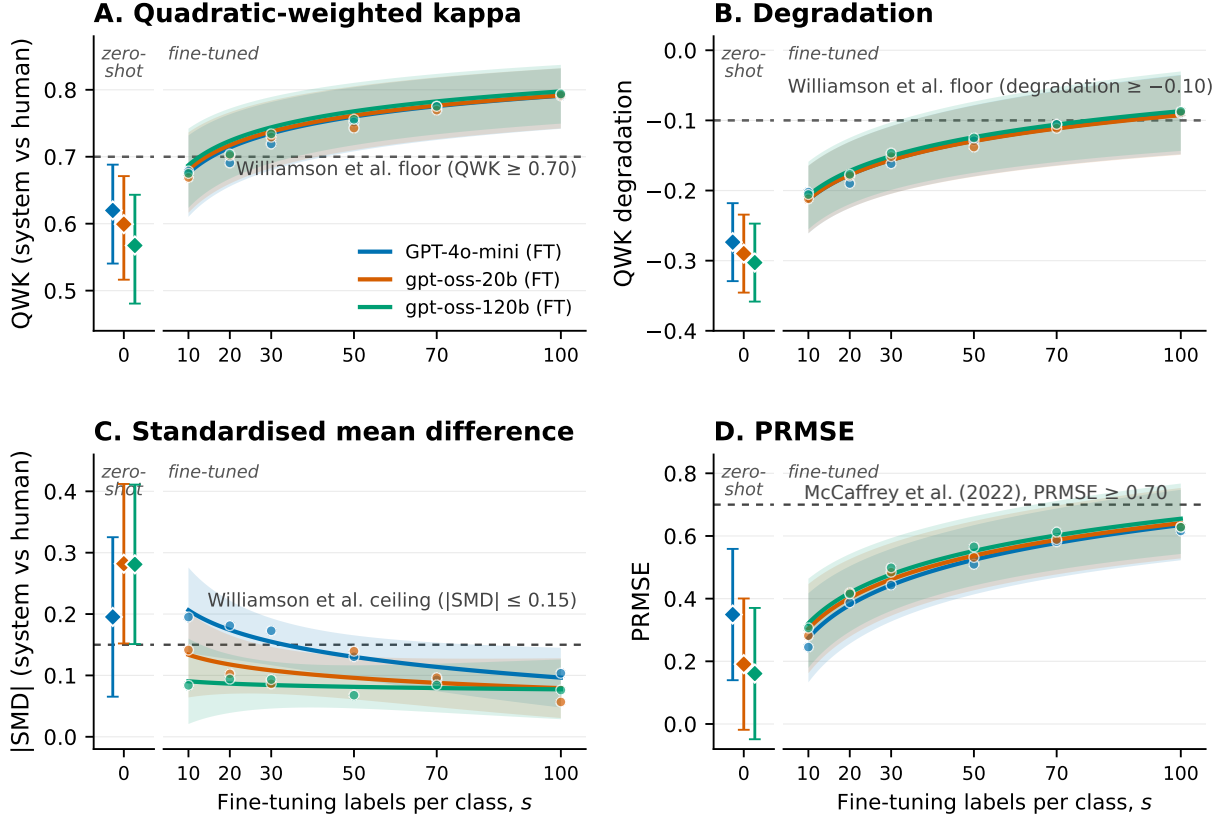


Figure 1: Marginal curves for the four criteria against labelled responses per score point: QWK (A), degradation of QWK below the human-human QWK (B), absolute standardised mean difference (C) and PRMSE (D), from the piecewise mixed model fitted to each outcome (QWK on the Fisher- z scale and back-transformed), with 95% confidence ribbons from the fixed effects. Zero-shot estimates are plotted on a separate axis at $s = 0$, disconnected from the fine-tuning curves, which are drawn over the observed training set sizes only. Points are the means of the ten observed item conditions at each training set size for each model. Dashed rules mark the Williamson et al. thresholds (QWK ≥ 0.70 , degradation ≥ -0.10 , $|SMD| \leq 0.15$) and the minimum PRMSE reference level of 0.70 (McCaffrey et al., 2022; Casabianca et al., 2023); its suitability depends on the intended use (Section 3.4).

4.3 Item-level results against the operational criteria

Figure 2 shows, for each of the 30 (model, item) conditions, the smallest s at which PRMSE reaches 0.70, with a marker where the three Williamson et al. criteria are satisfied; Table 3 gives the smallest observed s at which those three criteria are met. Within $s \leq 100$, gpt-oss-20b and gpt-oss-120b each satisfied all three on seven of ten items (1, 2, 3, 4, 6, 9 and 10), and GPT-4o-mini on five (1, 2, 6, 9 and 10). Across the passing items the median observed s was 50 for all three models; the model-based estimate for gpt-oss-20b was 42.

GPT-4o-mini’s two additional failures, on items 3 and 4, were failures of standardised mean difference: its QWK crossed the 0.70 floor, but the fine-tuned model shifted the score distribution relative to the human raters in a way that the open-weight models, at comparable QWK, did not.

4.4 Items that fail the criteria

Three features help explain whether items met the criteria: agreement generally rose with the number of labelled responses; high human-human agreement imposed a demanding degradation threshold; and extreme score skew limited the available examples at upper score points. These constraints can overlap, as item 6 illustrates. Items 3, 4, 7 and 8 generally had lower QWK across the published systems in Table 4, suggesting recurring item-level difficulties across scorers.

Items 5, 7 and 8 failed for every model at every training set size, for different reasons. Item 5 has an enumerated rubric and an evaluation-set human-human QWK of 0.951, so the degradation criterion requires machine-human QWK of at least 0.851. Fine-tuned agreement exceeded the absolute 0.70 floor but failed this stricter threshold. Its two upper score points also have the fewest available

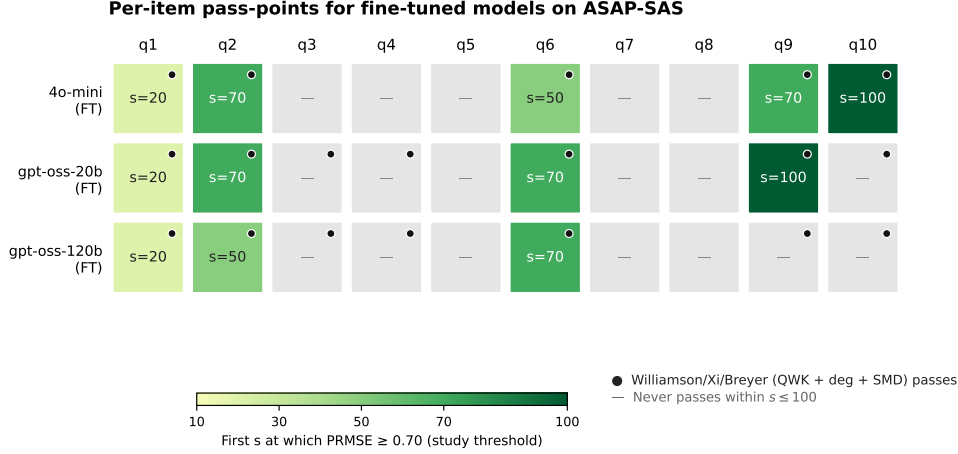


Figure 2: Smallest training set size at which PRMSE reaches 0.70 for each model and item (cell shading), with a marker where the three Williamson, Xi and Breyer criteria are satisfied. Grey cells do not reach the criterion within $s \leq 100$.

Item	4o-mini	oss-20b	oss-120b	Binding criterion
1	70	50	30	degradation
2	50	50	50	degradation
3	—	70	50	absolute 0.70
4	—	50	70	absolute 0.70
5	—	—	—	degradation
6	50	100	100	degradation
7	—	—	—	degradation
8	—	—	—	degradation
9	20	10	30	degradation
10	30	70	100	degradation

Table 3: Smallest observed training set size (responses per score point) at which all three Williamson et al. criteria are satisfied, by item and model, and the criterion last satisfied (binding) on each item for the models that pass; for GPT-4o-mini on items 3 and 4 the binding criterion is the SMD. A dash indicates that the criteria are not met within $s \leq 100$.

responses (25 and 19), and the confusion matrices showed almost exclusively under-prediction of these rare high scores. Item 6 shares the skew, has more upper-score responses (44 and 30), and passed at $s = 50$ for GPT-4o-mini and $s = 100$ for the open-weight models. Item 7 also has high evaluation-set human-human QWK (0.964), and Jiang and Bosch (2024) report over-scoring under GPT-4 prompting; the criterion is demanding and the item appears sensitive to how the rubric is specified. Item 8 has a lower machine-human QWK target (0.771) but a holistic rubric on which Jiang and Bosch document machine scores ‘distributed across the entire range’. The fine-tuned models showed similar but less pronounced score dispersion. These results warrant examining the rubric

and retaining human scoring while agreement remains inadequate; they do not establish that further labelling would be ineffective.

4.5 Rubric structure as an item-level moderator

Rubric style accounted for 47% of the estimated item random-intercept variance. Adding the binary distinction in Table 1 (key elements versus holistic) as a main effect produced this reduction (likelihood-ratio $\chi^2(1) = 7.25$, $p = 0.007$). Allowing rubric style to moderate the slope did not improve fit ($\chi^2(1) = 0.28$, $p = 0.60$). Items with enumerated rubrics had higher estimated QWK throughout, with similar estimated gains from additional labelled responses (slopes of 0.106 to 0.111 against 0.116 to 0.120 for holistic items). Rubric style was associated with agreement levels, but did not by itself determine whether items met all three criteria.

4.6 Comparison with published systems

Fine-tuning reached near state-of-the-art agreement with substantially fewer labelled responses per item (Table 4). At $s = 100$, mean observed QWK was 0.79, within 0.01 of the 0.796 ensemble benchmark identified as state of the art by Ormerod and Kwako (2024). That benchmark comes from the pretrained-model ensemble of Ormerod (2022). Our models used at most 300 or 400 labelled responses per item, depending on the score scale, compared with 1,022 to 1,446 training responses plus development data in the ensemble study. At $s = 50$, with at most 150 or 200 responses per item, mean QWK

was 0.75, matching BERT-base and the average of the four fine-tuned open-weight LLMs reported by Ormerod and Kwako (2024). Their generative models used 984 to 1,278 training responses per item. The $s = 100$ mean also exceeded the competition average of 0.73 calculated from Shermis (2015).

A published small-sample comparison used an n -gram support vector machine trained on 270 responses per item, which reached 0.64, 0.56 and 0.64 on items 1, 2 and 10 (Horbach et al., 2018); at $s = 20$, with at most 80 labelled responses per item, our fine-tuned models reached 0.84, 0.76 and 0.75 on the same items.

These counts are per-item labelling requirements for our model trained jointly on all ten items. Different evaluation samples and reference scores make the comparisons indicative, rather than evidence of superiority or equivalence under a common test. Our results are against the first human rater; the public test set used by BERT-base and the ensemble supplies one reference score.

5 Discussion

5.1 The size of the labelled set

On this benchmark, passing items required a median of fifty labelled responses per score point to satisfy all three Williamson et al. criteria, and none required more than one hundred. The comparison studies used roughly 1,000 to 1,400 training responses per item, and a reader of that literature could reasonably conclude that adequate agreement requires annotation on that scale. The present results show that smaller per-item labelled sets can suffice. For an item scored 0 to 3, fifty responses per score point corresponds to 200 labelled responses. A single marking exercise may supply this quantity, bringing fine-tuned scoring within reach of assessment units and research groups with limited annotation resources, subject to item-level evaluation.

5.2 Model selection

The models had similar estimated QWK and PRMSE learning rates; deployment requirements and item-level criterion failures therefore inform model choice. gpt-oss-20b is a practical default: it met the Williamson et al. criteria on the same seven items as gpt-oss-120b, avoided the mean-score bias that caused GPT-4o-mini to fail two ELA items, and can be hosted locally so that stu-

dent responses remain within the institution. It is the smaller open-weight model and requires fewer local resources than gpt-oss-120b. The similar learning rates are consistent with Ormerod and Kwako (2024), whose four fine-tuned models had mean QWK values spanning 0.749 to 0.759, and with Ormerod and Kehat (2025), whose 130M-parameter Mamba and 8B-parameter Llama models reached QWK of 0.797 and 0.792 on ASAP 2.0 essay scoring. These comparisons support choosing a smaller model when it meets the required item-level criteria; they do not establish equivalence across all scoring outcomes.

5.3 The limitations of QWK-only reporting

Two results illustrate the limits of QWK-only reporting. First, acceptable QWK can coexist with excessive mean-score bias. On items 3 and 4, fine-tuned GPT-4o-mini cleared the 0.70 floor and would meet that QWK-only threshold, but its scores were shifted relative to the human raters by more than the SMD criterion allows. The second is item-level heterogeneity, which a single benchmark mean conceals. At $s = 100$, the per-item QWK in Table 4 ranged from 0.71 to 0.90 around a mean of 0.79, and three items failed the operational criteria at every training set size; PRMSE showed the converse, with the curve for the average item never reaching 0.70 while seven of the ten items reached it individually. A benchmark mean therefore obscures variation in whether individual items meet the operational criteria.

5.4 Benchmark contamination

ASAP-SAS has been public since 2012, so its items, rubrics and released responses may have been encountered during pretraining. Three observations qualify the interpretation of the results. First, the zero-shot baselines were poor, failing both the Williamson et al. criteria and the PRMSE threshold on every item. Second, our learning-curve slopes were estimated from fine-tuned conditions that shared the same pretrained model within each model family. This holds pretraining exposure constant within each curve, but does not establish that contamination leaves slopes or estimated training requirements unaffected. Third, the gpt-oss models had the weakest zero-shot performance despite their later release. This provides no clear evidence of effective recall of the released scores, but cannot rule out contamination. Replication on an unreleased item bank is therefore needed to

System	1	2	3	4	5	6	7	8	9	10	Mean
Human raters, this study	0.95	0.91	0.79	0.72	0.95	0.97	0.96	0.87	0.82	0.87	0.88
2012 competition, competitor mean	0.82	0.76	0.71	0.64	0.78	0.81	0.65	0.60	0.78	0.73	0.73
BERT-base, fine-tuned	0.85	0.77	0.69	0.72	0.85	0.84	0.68	0.60	0.83	0.72	0.75
Ensemble, pretrained language models	0.88	0.89	0.72	0.75	0.81	0.82	0.73	0.70	0.87	0.78	0.80
GPT-4, one exemplar per score point	0.72	0.72	0.63	0.52	0.77	0.80	0.50	0.55	0.70	0.87	0.68
Open-weight LLMs, fine-tuned, average of four models	0.87	0.81	0.68	0.64	0.80	0.83	0.73	0.69	0.77	0.71	0.75
This study, zero-shot, average of three models	0.73	0.66	0.54	0.29	0.62	0.77	0.53	0.57	0.51	0.69	0.59
This study, $s = 50$, average of three models	0.85	0.83	0.69	0.70	0.75	0.81	0.68	0.66	0.76	0.76	0.75
This study, $s = 100$, average of three models	0.90	0.85	0.72	0.73	0.81	0.88	0.74	0.71	0.77	0.81	0.79

Table 4: Per-item QWK on the ten ASAP-SAS items. Sources: competition, [Shermis \(2015\)](#); BERT-base and ensemble, [Ormerod \(2022\)](#); open-weight LLMs, [Ormerod and Kwako \(2024\)](#); GPT-4, [Jiang and Bosch \(2024\)](#). Means are arithmetic averages across items; values are rounded to two decimals. Evaluation sets differ: competition results use the private leaderboard; BERT-base and the ensemble use the public test set; the four generative LLMs use the study’s train/development/test partition; GPT-4 uses samples of 100 to 356 responses per item; this study uses 600 responses held out from the training release (Section 2.2).

assess whether the reported training requirements generalise.

6 Limitations

Two limitations bear on the item-level estimates. First, the largest training condition targeted 100 labelled responses per score point, but four items had insufficient examples: mildly on items 2 and 4, where one score point is capped at 85 and 77 responses, and severely on items 5 and 6, where increases in training set size mainly added responses at the modal score point, leaving the rare upper score points underrepresented. Second, the public release contains no respondent background information, so the SMD criterion cannot be supplemented by subgroup analyses of the kind [Huang and Wilson \(2025\)](#) report for English language learners, and no fairness claim across such groups is made. Results are from a single English-language benchmark of ten items, and the training set sizes should not be generalised to item banks with different score distributions or rubric styles.

7 Conclusion

The four questions posed in the introduction can be addressed on the ASAP-SAS items. First, each open-weight model satisfied all three Williamson et al. criteria on seven of ten items, and GPT-4o-mini on five. Among passing items, the median was fifty labelled responses per score point, and none required more than one hundred. Second, the models showed similar gains from additional labelled responses, with no detectable differences in learning-curve slopes. Deployment requirements and item-level performance favour gpt-oss-20b among the

models examined: it can run locally and met the criteria on the same items as the larger open-weight model. Third, rubric features, score distributions and human-human agreement help explain item-level difficulties, but do not provide a sufficient rule for predicting which items will pass. Fourth, joint training on small labelled sets reached near state-of-the-art agreement: mean QWK of 0.79, within 0.01 of the ensemble benchmark highlighted by Ormerod and Kwako (2024), with fewer labelled responses per item. A set that a single marking exercise may supply can therefore be sufficient for items that meet the operational criteria. Evaluation against the full set of criteria is needed to identify those items and determine where human scoring remains necessary.

References

- Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. [Fitting linear mixed-effects models using lme4](#). *Journal of Statistical Software*, 67(1):1–48.
- Jodi M. Casabianca, Daniel F. McCaffrey, Matthew S. Johnson, Kathryn L. Ricker-Pedley, Ourania Rotou, and Joseph Martineau. 2023. [Exploration of the proportional reduction in mean-squared error for evaluating automated scores](#). Research Memorandum RM-23-01, Educational Testing Service.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. [QLoRA: Efficient finetuning of quantized LLMs](#). In *Advances in Neural Information Processing Systems*, volume 36.
- Andrea Horbach, Sebastian Stennmanns, and Torsten Zesch. 2018. [Cross-lingual content scoring](#). In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*,

- pages 410–419. Association for Computational Linguistics.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Yue Huang and Joshua Wilson. 2025. [Evaluating LLM-based automated essay scoring: Accuracy, fairness, and validity](#). In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Works in Progress*, pages 71–83. National Council on Measurement in Education.
- Lan Jiang and Nigel Bosch. 2024. [Short answer scoring with GPT-4](#). In *Proceedings of the Eleventh ACM Conference on Learning @ Scale (L@S '24)*. ACM.
- Damjan Kalajdzievski. 2023. [A rank stabilization scaling factor for fine-tuning with LoRA](#). *Preprint*, arXiv:2312.03732.
- Alexandra Kuznetsova, Per B. Brockhoff, and Rune H. B. Christensen. 2017. [lmerTest package: Tests in linear mixed effects models](#). *Journal of Statistical Software*, 82(13):1–26.
- Ehsan Latif and Xiaoming Zhai. 2024. [Fine-tuning ChatGPT for automatic scoring](#). *Computers and Education: Artificial Intelligence*, 6:100210.
- Gyeong-Geon Lee, Ehsan Latif, Xuansheng Wu, Ninghao Liu, and Xiaoming Zhai. 2024. [Applying large language models and chain-of-thought for automatic scoring](#). *Computers and Education: Artificial Intelligence*, 6:100213.
- Anastassia Loukina, Nitin Madnani, Aoife Cahill, Lili Yao, Matthew S. Johnson, Brian Riordan, and Daniel F. McCaffrey. 2020. [Using PRMSE to evaluate automated scoring systems in the presence of label noise](#). In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 18–29. Association for Computational Linguistics.
- Daniel F. McCaffrey, Jodi M. Casabianca, Kathryn L. Ricker-Pedley, René R. Lawless, and Cathy Wendler. 2022. [Best practices for constructed-response scoring](#). *ETS Research Report Series*, 2022(1):1–58.
- Christopher Ormerod. 2022. [Short-answer scoring with ensembles of pretrained language models](#). *Preprint*, arXiv:2202.11558.
- Christopher Ormerod and Gitit Kehat. 2025. [Long context automated essay scoring with language models](#). In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Full Papers*, pages 35–42. National Council on Measurement in Education.
- Christopher Ormerod and Alexander Kwako. 2024. [Automated text scoring in the age of generative AI for the GPU-poor](#). *Chinese/English Journal of Educational Measurement and Evaluation*, 5(3):Article 5.
- Mark D. Shermis. 2015. [Contrasting state-of-the-art in the machine scoring of short-form constructed responses](#). *Educational Assessment*, 20(1):46–65.
- David M. Williamson, Xiaoming Xi, and F. Jay Breyer. 2012. [A framework for evaluation and use of automated scoring](#). *Educational Measurement: Issues and Practice*, 31(1):2–13.