# S-GRADES – Studying Generalization of Student Response Assessments in Diverse Evaluative Settings

**Tasfia Seuti, Sagnik Ray Choudhury**
University of North Texas, Denton, TX, USA
TasfiaSeuti@my.unt.edu, Sagnik.Raychoudhury@unt.edu

**Abstract**

Evaluating student responses, from long essays to short factual answers, is a key challenge in educational NLP. Automated Essay Scoring (AES) focuses on holistic writing qualities such as coherence and argumentation, while Automatic Short Answer Grading (ASAG) emphasizes factual correctness and conceptual understanding. Despite their shared goal, these paradigms have progressed in isolation with fragmented datasets, inconsistent metrics, and separate communities. We introduce S-GRADES (Studying Generalization of Student Response Assessments in Diverse Evaluative Settings), a web-based benchmark that consolidates 14 diverse grading datasets under a unified interface with standardized access and reproducible evaluation protocols. The benchmark is fully open-source and designed for extensibility, enabling continuous integration of new datasets and evaluation settings. To demonstrate the utility of S-GRADES, we evaluate three state-of-the-art large language models across the benchmark using multiple reasoning strategies in prompting. We further examine the effects of exemplar selection and cross-dataset exemplar transfer. Our analyses illustrate how benchmark-driven evaluation reveals reliability and generalization gaps across essay and short-answer grading tasks, highlighting the importance of standardized, cross-paradigm assessment.

**Keywords:** Automated Essay Scoring (AES); Automatic Short Answer Grading (ASAG); Summative Evaluation

## 1. Introduction

Summative evaluation of student responses encompasses a diverse spectrum of assessment types, from extended essays demonstrating argumentative and analytical skills to concise short-answer responses testing factual knowledge and conceptual understanding. This diversity poses a central challenge in educational Natural Language Processing (NLP), as different response types demand distinct evaluation criteria. Research in this area has developed along two complementary trajectories. **Automated Essay Scoring (AES)** focuses on evaluating extended writing for coherence, organization, and argument quality, while **Automatic Short Answer Grading (ASAG)** targets concise factual responses that require concept-level verification rather than stylistic assessment (Burrows et al., 2015). ASAG tasks also include domain-specific settings such as conceptual assessment of chemistry (Sonkar et al., 2024), physics reasoning (Kortemeyer, 2024), and conceptual understanding of computer science (Xie et al., 2024). Despite their shared goal of automating human evaluation for long-form (non-MCQ) responses, AES and ASAG have largely progressed in isolation, supported by separate datasets, metrics, and research communities (§2). No previous work has unified diverse assessment tasks under a single blind-test evaluation.

We address this limitation by introducing **S-GRADES** (Studying Generalization of Student Response Assessments in Diverse Evaluative Settings), a unified, open-source benchmark [1] providing standardized evaluation infrastructure through a streamlined workflow (Figure 1). Our contributions

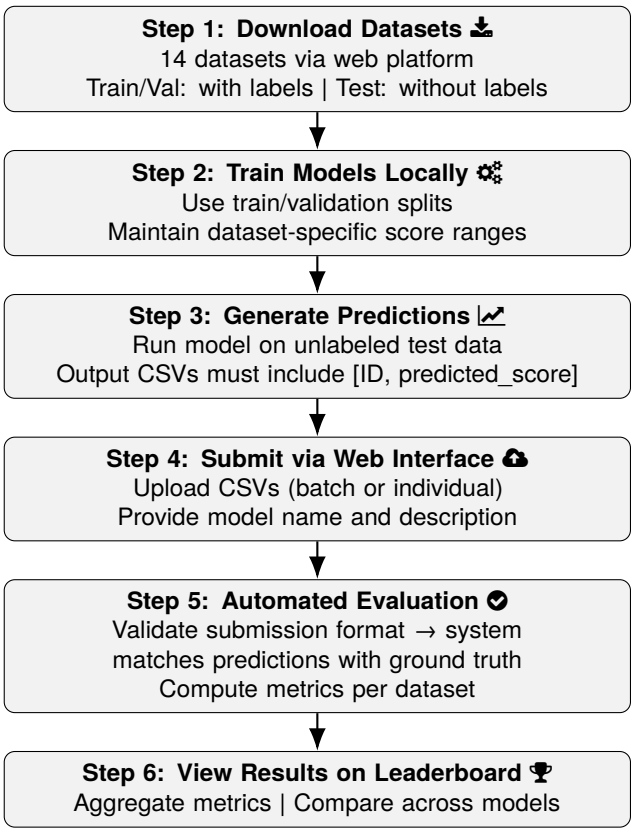


Figure 1: Overview of the **S-GRADES** benchmark user workflow.

[1] https://sgrades.eng.unt.edu



*Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 4684–4701
11–16 May 2026. 

are:

**Dataset Aggregation and Standardization.** The benchmark combines $14$ datasets in AES and ASAG tasks, covering general comprehension, domain-specific assessments, and language proficiency across subjects like English, chemistry, physics, life sciences, and computer science. It includes various response types (persuasive, narrative, expository, source-dependent) from public and proprietary sources. Each dataset undergoes consistent preprocessing and stratified data splits while maintaining original scoring scales to preserve task-specific characteristics (§3).

**Unified Web-Based Evaluation Infrastructure.** S-GRADES offers a web-based platform [2] for dataset access and performance benchmarking. Researchers can download datasets, train models locally, submit predictions, and receive evaluations. The system validates the submissions and calculates multiple metrics (§4).

**Systematic Evaluation of State-of-the-Art LLMs.** We evaluate [3] three frontier LLMs: GPT-4o mini, Gemini 2.5 Flash, and Llama 4 Scout, across six reasoning configurations (Liu et al., 2024) (inductive (few-shot) with random sampling, deductive (zero-shot), abductive (Chain-of-Thought), and a combination of them), answering the following questions (§5):

- **Model-Reasoning Alignment:** Which LLMs excel with which specific reasoning strategies and why certain approaches align better with their underlying architectures.
- **Cross-Dataset and Cross-Paradigm Generalization:** How do model performances change when few-shot exemplars are selected on one dataset and evaluated on another, both within the same paradigm (e.g., *ASAP2.0* (AES) $\rightarrow$ *ASAP-AES* (AES)) and across paradigms (e.g., *ASAP2.0* (AES) $\rightarrow$ *ASAP-SAS* (ASAG)). This tests whether reasoning strategies generalize across fundamentally different response types.
- **Exemplar Selection Stability:** How stable are the model predictions across multiple random few-shot sample selections in inductive reasoning setups.

Together, these contributions establish a foundation for standardized, reproducible research in LLM-based automated scoring and open new directions for reasoning-aware evaluation.

## 2. Related Work

**AES.** Previous work in AES has developed under three paradigms: feature-based statistical models, deep learning, and large language models (LLMs). Early systems, eg., PEG (Page, 1966) and e-rater (Attali and Burstein, 2006) relied on engineered features capturing surface traits (e.g., word frequency, syntax, sentence length), while IEA introduced latent semantic analysis for semantic similarity (Landauer et al., 1998; Foltz et al., 1999). Early DNN-based approaches, such as Taghipour and Ng (2016), showed that recurrent networks captured writing patterns, and BERT-based models improved contextual understanding (Ludwig et al., 2021). More recently, Mizumoto and Eguchi (2023) and Xiao et al. (2024) found GPT-4 and Claude closely align with human raters, though sensitive to rubric phrasing. Seßler et al. (2025) reported strong correlations ($r = 0.87$) for GPT-4, Gemini, and Claude but noted systematic overprediction for mid-range essays. However, a longitudinal evaluation by Pack et al. (2024) showed that GPT-4's reliability dropped from $0.843$ to $0.779$ over $90$ days, revealing instability and bias. These findings highlight persistent challenges in interpretability, reproducibility, and generalizability of LLM-based AES systems.

**ASAG.** Burrows et al. (2015) identified five methodological eras in ASAG: concept mapping, information extraction, corpus-based similarity, machine learning, and the current evaluation era focused on benchmarks. Early systems like C-rater (Burrows et al., 2015) relied on handcrafted or lexical features. Haller et al. (2022) traced the shift from word2vec and GloVe to transformers, noting improved semantic understanding but persistent dataset fragmentation. Domain-specific pretraining of BERT improved grading accuracy by $\approx 10\%$ on benchmarks such as ASAP-SAS (Sung et al., 2019), while in-context fine-tuning enhanced alignment with human raters (Fernandez et al., 2022). Despite advances in LLMs, systematic evaluations reveal reliability issues. Henkel et al. (2023) found that rubric-based prompting improved GPT-4 agreement, yet Chang and Ginter (2024) and Grévisse (2024) reported inconsistent score inflation or deflation across languages and domains. Schneider et al. (2023) showed ChatGPT's scores were highly prompt-sensitive, and Gao et al. (2023) found strong binary accuracy but weak partial-credit handling. Studies such as Jiang and Bosch (2024) (GPT-4), and Xie et al. (2024) (rubric–review pipeline) show that LLMs can achieve near-human agreement when properly prompted but remain sensitive to rubric design and task length.

Despite recent progress, AES and ASAG still suffer from poor reproducibility and generalization. AES is prone to prompt-specific overfitting and low interpretability, while ASAG is sensitive to domain shifts and rubric changes. Inconsistent dataset sizes and rubrics hinder standardized evaluation,

[2] https://github.com/nlpatunt/sgrades
[3] https://github.com/nlpatunt/sgrades-experiments

which S-GRADES helps address.

## 3. S-GRADES Benchmark Datasets

We summarize prior work on the fourteen datasets (Tab. 1) integrated into the platform.

**Essay Scoring Studies.** The **ASAP-AES** dataset from the 2012 Kaggle competition is a key benchmark for automated essay scoring, foundational for transformer-based AES. Its lack of public test labels limits reproducibility and comparison, as studies rely on cross-validation. Rodriguez et al. (2019) and Ormerod et al. (2021) set strong transformer baselines, showing both large models (BERT, XLNet) and efficient variants (ALBERT, MobileBERT, Electra) perform reliably with fewer parameters, despite cross-validation's limitations. Mathias and Bhattacharyya (2018) expanded ASAP-AES with trait-level annotations to form **ASAP++**, showing multi-task BiLSTM models enhance holistic score predictions.

Crossley et al. (2024) introduced **PERSUADE 2.0** with discourse and argumentation annotations on essays, emphasizing rhetorical assessment. Later, Crossley et al. (2025) introduced **ASAP 2.0**, a competition dataset with publicly available test labels. Top LLM systems (GPT-4, Claude) achieved 84% accuracy, though varying metrics (accuracy vs. QWK) complicated comparisons with previous AES studies.

**Short-Answer Grading Studies.** Jiang and Bosch (2024) evaluated prompting strategies on **ASAP-SAS**, a science short-answer dataset graded by detailed rubrics. They found that exemplar prompts consistently outperformed intermediate reasoning steps ($\Delta$QWK = +0.08), though performance varied substantially across question types (QWK range: 0.45–0.82), highlighting task-dependent prompting effectiveness.

Ivanova and Handschuh (2024) tested ChatGPT-3.5 on the **Mohlar** dataset, which contains introductory computer science answers (e.g., pointers, data structures), graded on a 5-point scale. LLM-human agreement lagged behind human-human levels, especially for technical items, revealing inconsistency in grading CS concepts.

Kortemeyer (2024) benchmarked GPT-4 on **SciEntSBank** (multi-domain science questions) and **BEEtlE** (introductory electricity and circuits), each with 2-way and 3-way classifications. GPT-4 performed strongly on SciEntSBank but struggled with BEEtlE's contradictory cases. Interestingly, omitting reference answers slightly improved accuracy, suggesting reliance on pre-trained knowledge over rubric cues.

**Domain-Specific Assessment Studies.**

Sonkar et al. (2024) introduced the **Rice_Chem** dataset for chemistry problem-solving, demonstrating that rubric-aware models with explicit scoring criteria surpass general BERT baselines, though long-form explanations still yield lower human–AI agreement. Xie et al. (2024) presented the **OS_Dataset**, a computer science dataset where students explain OS behavior using code tracing and system reasoning (e.g., registers, threads, memory). This dataset employs a rubric→grade→review pipeline that improved grading agreement (87% vs. 76%). Xiao et al. (2025) studied human–AI scoring with the **CSEE** dataset, which contains English writing tasks from Chinese high school students, including formal letters, short essays, and argumentative compositions (score range 0–16). This study found that LLM support reduced grading time by 38% while maintaining reliability, though institutional restrictions limit reproducibility. Gao et al. (2024) proposed the **ReGrading dataset (2JC)** for engineering short answers, combining binary and multi-score tasks, showing rubric-based models were accurate but overly lenient.

**Language Proficiency Assessments.** Qiu et al. (2024) studied IELTS writing assessment, showing fine-tuned transformers perform well on argumentative essays, with varying results by prompt type. Comparing GPT-4o and Llama-3, they found that GPT-4o remained stable regardless of exemplars, whereas Llama-3 improved without them, highlighting model-specific optimization needs.

## 4. Benchmark Development

### 4.1. Preprocessing & Standardization

We processed datasets into a standard tabular format to ensure unified evaluation while preserving scoring semantics.

**Unified Schema.** Each dataset includes three **core** columns:

1. **Unique Identifier:** A dataset-specific column such as `essay_id`, `ID`, `sis_id`, or `index` that uniquely identifies each student response.
2. **Response Text:** The column containing the student's written response, labeled as `essay`, `student_answer`, or `response`, depending on the dataset.
3. **Ground Truth Score:** The target score column with dataset-specific naming conventions such as `domain1_score`, `label`, `Score`, or `band_score`.

We kept and expanded each dataset with all available context from sources. For datasets with prompt texts, rubric descriptions, essay_set identifiers, reference materials, or extra metadata, we retained these as supplementary columns for complete context.

| Task Type | Dataset | Domain(s) | Score Range(s) | Public | Test Split Size |
|---|---|---|---|---|---|
| **AES** | ASAP-AES | Persuasive, Narrative, and Source-based Essays | Set 1: 2–12; Set 2: 1–6; Sets 3–4: 0–3; Sets 5–6: 0–4; Set 7: 0–30; Set 8: 0–60 | Yes | 1,298 |
| | ASAP++ | Argumentative and Expository Essays | 0–6 | Yes | 1,069 |
| | ASAP2.0 | Extended Essay Scoring (Multi-domain) | 0–6 | Yes | 4,946 |
| | Persuade_2 | Persuasive Writing (Argumentative) | 1–6 | Yes | 2,600 |
| | IELTS Writing (General) | General English Writing Tasks | 1–9 | Yes | 144 |
| | IELTS Writing (Task 2) | Academic English Writing Tasks | 1–9 | Yes | 491 |
| **ASAG** | ASAP-SAS | Science Short Answers | 0–3 | Yes | 3,409 |
| | ReGrading Dataset (2JC) | Conceptual Short-Answer (Engineering, Rubric-Based) | 0–8 | No | 198 |
| | CSEE | Short English Writing Tasks (extended answers, letters) | 0–16 | Yes | 2,654 |
| | Mohlar | Short Answer Questions (Computer Science concepts) | 0–5 | No | 455 |
| | BEEtlE | Physics and Electronics | 2-way: 0–1 (Correct/Incorrect); 3-way: 0–2 (Contradictory/ Correct/Incorrect) | Yes | 1,258 |
| | SciEntSBank | Life Science Conceptual Questions | 2-way: 0–1 (Correct/Incorrect); 3-way: 0–2 (Contradictory/ Correct/Incorrect) | Yes | 5,835 |
| | Rice_Chem | Chemistry Problem-solving | Q1: 0–8; Q2: 0–8; Q3: 0–9; Q4: 0–8 | Yes | Q1: 66; Q2: 66; Q3: 62; Q4: 60 |
| | OS_Dataset | Computer Science Conceptual Responses | Q1: 0–19; Q2: 0–16; Q3: 0–15; Q4: 0–16; Q5: 0–27 | No | Q1: 3; Q2: 8; Q3: 8; Q4: 8; Q5: 8 |

Table 1: Summary of datasets included in the S-GRADES benchmark. **For datasets that are not currently publicly hosted online, we obtained creator permission for research use and sharing.**

**Handling Large External Resources.** Some datasets included auxiliary materials such as source essays and images essential for evaluation. These were stored in a separate repository and linked via stable URLs or identifiers, ensuring dataset completeness and efficient distribution.

**Preserving Original Score Scales.** We retained each dataset's original scoring scales to respect contextual granularity. Ranges include binary labels (BEEtIE_2way, SciEntSBank_2way), extended numeric scales (ASAP-AES Set 8: 0–60, OS Q5: 0–27), and categorical, integer (0–3, 0–8), or band-based (IELTS 1–9) formats. All score definitions are documented.

**Data Splits.** We provide standardized data partitions (train/validation/test) for each dataset. Training and validation splits have complete ground-truth labels for development and tuning. Test data is available in two forms: a *public version* (with responses, IDs, and metadata) and a *private version* (with securely stored ground truth labels for evaluation).

## 4.2. Platform Architecture

S-GRADES provides a web-based evaluation platform built with `FastAPI`. It includes four components: dataset distribution, submission validation, evaluation engine, and leaderboard system.

**Dataset Distribution.** All curated datasets are available via a unified download interface, packaging all fourteen datasets with their train/validation/test splits into one archive. Test files keep metadata and identifiers but have ground-truth scores removed.

**Submission Validation.** The system includes instructions for uploading individual (per-dataset) and batch (all-dataset) results (Figure 2). Dataset-specific validator classes conduct multi-stage validation: 1) CSV format verification and security checks to prevent malicious uploads, 2) structural checks for correct column names, 3) ID uniqueness to avoid duplicates, and 4) score format validation confirming predictions match expected types (numeric vs. categorical) and ranges. Errors provide detailed messages on issues and expected formats.

**Evaluation Engine.** Predictions are automatically matched to ground truth by ID, with eight metrics computed for each pair. All results are logged with submission metadata (timestamp, IP, file hash) to ensure reproducibility and integrity.

**Leaderboard System.** Results are aggregated and shown on a public leaderboard (Figure 3), supporting filtering by metrics like QWK, Pearson correlation, and F1. It displays overall performance across datasets, allowing researchers to compare rankings.

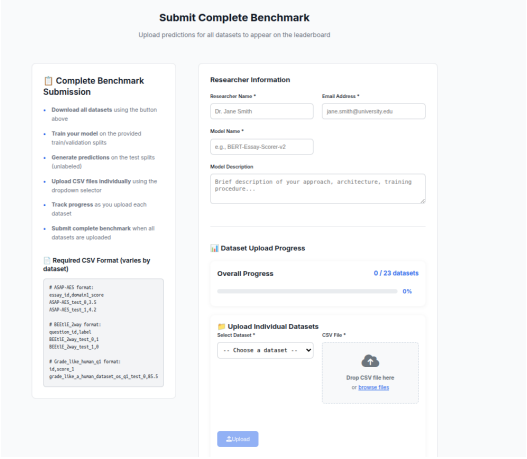


Figure 2: Complete benchmark submission interface.

**Security Measures.** Platform security is implemented with rate limits on submissions (5 per minute for single uploads and 2 for batch), and SQL injection attacks are prevented using SQLAlchemy ORM's parameterized queries.

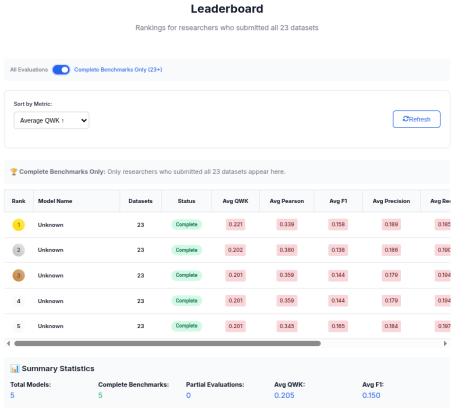


Figure 3: Public leaderboard displaying aggregated results across all datasets and evaluation metrics.

## 4.3. Evaluation Metrics

S-GRADES supports the computation of the following metrics:

**Agreement Metrics.** Quadratic Weighted Kappa (QWK) (McHugh, 2012) serves as the primary agreement measure, capturing inter-rater reliability between model predictions and human scores while accounting for the ordinal nature of grading scales. It penalizes larger discrepancies more heavily, rewarding models that maintain near-miss consistency.

$$\text{QWK} = 1 - \frac{\sum_{i,j} w_{ij} O_{ij}}{\sum_{i,j} w_{ij} E_{ij}},$$

where $O_{ij}$ and $E_{ij}$ are the observed and expected rating matrices, and $w_{ij} = \frac{(i-j)^2}{(k-1)^2}$ penalizes larger score differences across $k$ possible score levels. In contrast, the Pearson correlation coefficient ($r$) measures the strength of the linear relationship between predicted ($\hat{y}_i$) and true ($y_i$) scores.

$$r = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})^2}},$$

**Error Metrics.** MAE measures average prediction error in the original scoring scale, while RMSE highlights significant errors, identifying models with consistent over/underestimation.

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \quad \text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

**Classification-Oriented Metrics.** For datasets modeled as classification tasks (**BEEtlE**, **SciEntSBank**), we report accuracy, F1, precision, and recall. For regression datasets, predictions and true scores are rounded to the nearest integer class before computation.

These metrics provide a balanced evaluation of automated scoring performance. QWK captures ordinal agreement, Pearson's $r$ reflects rank consistency, and MAE/RMSE quantify absolute and squared errors on original scales. F1, precision, recall, and accuracy assess categorical correctness and, together, reveal both ranking reliability and calibration quality.

## 5. Experimental Setup

**Models.** We conduct a comprehensive evaluation across three state-of-the-art large language models: Llama-4-Scout (405B, open weight), Gemini 2.5 Flash (proprietary), and GPT-4o-mini (est. $<$30B, proprietary), representing different architectures and scale. All experiments were run through OpenRouter [4] with the decoding temperature set to $0.1$ for the main evaluation runs. The experiments yielded $\approx \mathbf{37K}$ scored responses from 14 datasets, covering short factual and long-form argumentative tasks.

**Reasoning Configurations.** We evaluate six reasoning configurations (as categorized in Liu et al. (2024)) covering individual and hybrid approaches. The *Deductive Reasoning (principle-based application)* is a zero-shot setup in which the model is provided with general scientific or scoring principles. Prompts include a few examples illustrating how principles lead to specific outcomes (e.g., correct, partial, or incorrect), but no labeled examples are provided. The *Inductive Reasoning (5 exemplars)* is a few-shot setup where the model learns grading patterns from five labeled examples *sampled from the training split of the same dataset*. The *Abductive reasoning (inference to the best explanation)* is similar to a Chain-of-Thought prompting setup, in which the models are asked to generate hypotheses. LLMs suggest grading explanations (partial understanding, misconception, correct reasoning), identify the most plausible one, and determine the correct classification or score. *Hybrid* modes combine two reasoning configurations. For example, *Ind + Ded* $\implies$ the prompt includes both scoring principles and exemplars (see Appendix D for detailed descriptions and examples).

**Prompt template.** Each reasoning strategy follows a structured, multi-part template ensuring consistency across models and datasets. It begins with a **system role** defining the evaluator's perspective (e.g., "You are an expert evaluator using *inductive reasoning*"), followed by detailed **reasoning instructions** outlining the step-by-step process (such as learning from examples, applying principles, or inferring explanations). **Illustrative examples** are then provided: five exemplars for inductive prompts; or 2–3 reasoning demonstrations for deductive and abductive prompts based on scientific or scoring principles. The **evaluation task** presents the target question and student answer to be classified or scored, and an explicit **output constraint** directs the model to produce only the final label or numeric score without justification (see Appendix D for full prompt templates).

**Ablation Experiments.** Two sources of variability exist in this experimental setup: (1) *exemplar selection*, where inductive setups randomly sample training examples that directly influence the prompt, and (2) *decoding randomness*, introduced by non-zero temperature. To isolate each source, we design two separate stability experiments. In the **Prediction Stability** experiment, we query each model three times on a subset of the test data, using a fixed set of exemplars in the inductive setups and a fixed temperature of 0.1 for all reasoning setups. The **Exemplar Selection Stability** experiments assess the impact of random exemplar selection in inductive setups by keeping the temperature fixed at $0$, but selecting different examples for each run. Finally, we test generalization in **In/Cross-Paradigm Generalization** experiment where we select representative datasets with consistent performance from each paradigm (ASAP2.0 and ASAP-AES for AES, ASAP-SAS and CSEE for ASAG), and evaluated the effect of using exemplars from one dataset for another dataset (in the inductive setups).

## 6. Results and Discussion

ASAG tasks differ in their output formulation; we distinguish between **regression-based** and

[4] https://openrouter.ai/

**classification-based** setups. While we compute all appropriate evaluation metrics, we report QWK as the primary metric, as it is most widely adopted.

**AES Tasks.** Figure 4 summarizes the QWK performance. Scores range from 0.10 (*GPT-4o-mini*, *ASAP++*, *Ded+Abd*) to 0.96 (*GPT-4o-mini*, *ASAP-AES*), showing wide variation driven by dataset complexity, reasoning type, and model capability.


Figure 4: QWK scores for AES datasets across models and reasoning strategies.

Across models and datasets, clear performance patterns emerge. GPT-4o-mini shows the highest internal consistency, reaching near-perfect agreement on ASAP-AES but declining on ASAP++ and ASAP2.0, indicating sensitivity to dataset structure. Gemini-2.5-Flash performs most evenly, excelling on Persuade-2, while LLaMA-4-Scout exhibits selective strengths but struggles on most datasets except ASAP-AES.

At the dataset level, ASAP-AES is the most predictable, with all models achieving QWK above 0.75, while ASAP++ is the most difficult due to prompt and rubric diversity. ASAP2.0 shows moderate complexity, Persuade-2 displays large model-dependent variation, and the IELTS datasets show medium difficulty reflecting stricter evaluation criteria and nuanced linguistic expectations typical of standardized language proficiency tasks.

Reasoning strategy proves the most influential factor. Hybrid modes outperform single strategies: Ind+Ded consistently leads, Ind+Abd excels in persuasive writing tasks, and Ded+Abd performs the worst. Among single modes, inductive reasoning provides the most stable baseline. GPT-4o-mini is the most stable across reasoning strategies, Gemini-2.5-Flash maintains moderate but consistent transferability, and LLaMA-4-Scout shows high variance.

**ASAG Regression.** Figure 5 presents QWK results for six ASAG regression datasets. Compared to AES tasks, ASAG results show greater variability and lower absolute performance, reflecting the greater difficulty of short-answer grading.


Figure 5: QWK scores for ASAG regression datasets across models and reasoning strategies.

As in AES, GPT-4o-mini performs reliably on structured datasets such as ASAP-SAS and CSEE, indicating strong alignment with explicit scoring rubrics. Gemini-2.5-Flash demonstrates the most balanced performance, leading on Rice_Chem and maintaining competitive results across domains, suggesting robust generalization. In contrast, LLaMA-4-Scout shows moderate capability on structured tasks but weaker results on short-context datasets like Mohlar, reflecting sensitivity to input length and ambiguity.

At the dataset level, ASAP-SAS remains the most stable benchmark, while Mohlar is the most difficult due to short, under-specified responses. Rice_Chem achieves mid-to-high performance, consistent with its structured, scientific nature. In contrast, CSEE, although classified as an ASAG dataset, contains short-form writing tasks that resemble essay responses. This essay-like structure likely contributes to its relatively stronger performance compared to other ASAG datasets. ReGrading Dataset (2JC) shows the lowest cross-model stability, highlighting the complexity of revision-based scoring (see Appendix A.1 for detailed dataset characteristics and statistics).

Reasoning strategy remains the primary determinant of performance. The hybrid Ind+Ded approach performs best overall, while Inductive reasoning remains a strong baseline for factual and definition-based tasks. The Ded+Abd strategy consistently underperforms, confirming its unsuitability for short-answer grading. In terms of robustness, Gemini-2.5-Flash exhibits the lowest cross-strategy variance ($\sigma < 0.06$), GPT-4o-mini shows moderate fluctuations, and LLaMA-4-Scout displays the highest sensitivity.

**AES vs. ASAG Regression.** We averaged results across all AES and ASAG regression datasets to capture overall task difficulty. While AES focuses on writing quality, ASAG emphasizes factual accuracy and rubric alignment, thus being more sensitive to reasoning quality. AES tasks have higher agreement (mean QWK $\approx$ 0.42–0.43 across models) than ASAG ($\approx$ 0.34–0.43), illustrating the challenges of grading short answers with limited context. **Gemini-2.5-Flash** performs consistently across both domains (QWK = 0.433 each), **GPT-4o-mini** excels in essays, and **LLaMA-4-Scout** is less stable.

**ASAG Classification.** For two ASAG datasets, BEEtIE and SciEntSBank, answers are labeled as correct/incorrect (2-way) or correct/incorrect/contradictory (3-way). As shown in Figure 6, classification yields more stable metrics than regression, indicating models more reliably distinguish correct from incorrect responses. Across datasets, Gemini-2.5-Flash achieves the highest F1 (0.72), surpassing GPT-4o-mini and LLaMA-4-Scout. Hybrid reasoning strategies (*Ind+Ded*, *Ind+Abd*) provide small but consistent gains.

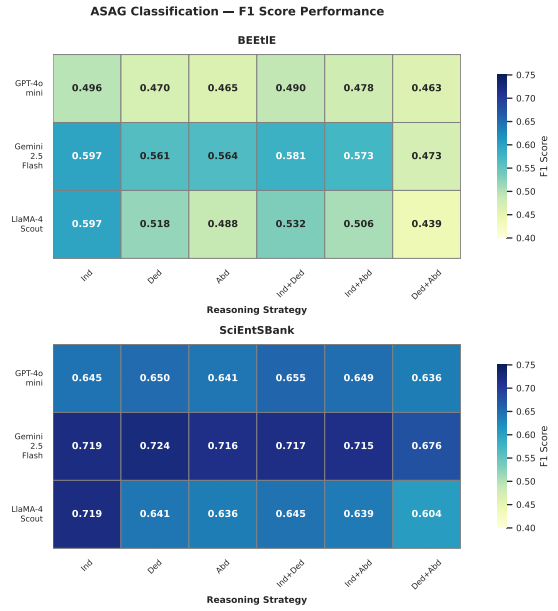


Figure 6: F1 scores across reasoning strategies for 2 classification-based ASAG datasets.

## 6.1. Prediction Stability

To evaluate whether model and strategy rankings are robust to repeated inference, we ran each model three times on a 21% subset of each dataset's test split. All experiments were conducted with the temperature set to 0.1, consistent with the original experimental setup.

For AES, the results are highly stable: model ranking and reasoning ranking change only slightly across dataset–strategy combinations. For ASAG regression, model rankings remain stable across datasets, with minor score fluctuations on *OS_Dataset* and *Regrading_Dataset_2JC*, where single-call estimates are inherently noisy due to small test sizes. For ASAG classification, model rankings remain largely consistent across both settings. Gemini-2.5-Flash leads across nearly all conditions, with the sole consistent exception being *Ded+Abd* on BEEtIE, where GPT-4o-mini marginally outperforms it. Overall, these observations indicate that the benchmark's main conclusions remain qualitatively consistent under repeated querying. Detailed per-call variance analysis is reported in Appendix E.

## 6.2. Exemplar Selection Stability

In inductive setups, exemplars are randomly selected. To test robustness to exemplar variability, we reran the evaluations on ASAP-AES with three random seeds (42, 123, 456) and a fixed temperature of 0, ensuring that all observed variance arises exclusively from exemplar selection rather than from decoding randomness. As shown in Table 2, Gemini-2.5-Flash achieves both the highest mean QWK (0.9520) and the greatest stability (QWK Std = 0.0047), demonstrating that its performance is largely invariant to which examples are shown. GPT-4o-mini (mean QWK = 0.9397, Std = 0.0099) and LLaMA-4-Scout (mean QWK = 0.9450, Std = 0.0111) exhibit slightly higher sensitivity to exemplar choice. Across all three seeds, QWK scores remain consistent within each model, suggesting that exemplar selection does not substantially alter evaluation outcomes under the inductive reasoning setup.

| Model | QWK Std |
|---|---|
| Gemini-2.5-Flash | 0.0047 |
| GPT-4o-mini | 0.0099 |
| LLaMA-4-Scout | 0.0111 |

Table 2: Exemplar sampling stability on ASAP-AES across three random seeds (42, 123, 456) at temperature = 0. Low standard deviations indicate that evaluation outcomes are largely invariant to the selection of examples.

## 6.3. Exemplar Generalization

We want to examine whether LLM reasoning abilities transfer in the inductive setups by selecting five samples from one dataset (base) and using them as exemplars in another dataset (target). Depend-

| Model | Reasoning | △QWK% |
|---|---|---|
| *In-paradigm generalizations* | | |
| **AES:** *ASAP2.0 → ASAP-AES* | | |
| GPT-4o-mini | Ind+Ded | -5.3 |
| LLaMA-4-Scout | Ind+Ded | -96.2 |
| Gemini-2.5-Flash | Ind+Ded | +9.4 |
| **ASAG**: *ASAP-SAS → CSEE* | | |
| GPT-4o-mini | Ind+Abd | -13.4 |
| LLaMA-4-Scout | Ind+Ded | -17.0 |
| Gemini-2.5-Flash | Ind | -9.0 |
| *Cross-paradigm generalizations* | | |
| **ASAG-AES:** *CSEE → ASAP-AES* | | |
| GPT-4o-mini | Ind+Ded | -2.8 |
| LLaMA-4-Scout | Ind+Ded | -95.5 |
| Gemini-2.5-Flash | Ind+Ded | +12.5 |
| **AES-ASAG:** *ASAP-AES → CSEE ASAP-SAS* | | |
| GPT-4o-mini | Ind+Abd | -13.2 |
| LLaMA-4-Scout | Ind | -47.2 |
| Gemini-2.5-Flash | Ind | -75.5 |

Table 3: Exemplar generalization results. For inductive (5-shot) reasoning setups, exemplars are selected from one dataset and used on another.

ing on whether the base and target datasets come from the same task family (ASAG/AES) or different ones, we refer to it as the "In" or "Cross" paradigm of generalization.

Table 3 presents the results (full results across all datasets in Appendix C). We use the best-performing dataset from prior evaluations as the base for most cross-generalization experiments and, for each transfer, apply the reasoning strategy that performs best on this base dataset. Within the AES in-paradigm setting, GPT-4o-mini exhibits a minor performance decline, Gemini-2.5-Flash shows an unexpected improvement, and LLaMA-4-Scout experiences a significant collapse, indicating instability in transferring scoring behavior across similar essay-scoring datasets. For the ASAG in-paradigm condition, all models exhibit moderate performance degradation under cross-dataset transfer, while retaining reasonable grading capability (QWK > 0.49). This suggests that, although models' performance is dataset-specific, their underlying reasoning partially generalizes within the short-answer domain. Interestingly, in the cross-paradigm transfer from ASAG to AES, exemplars drawn from the CSEE dataset yield measurable performance gains on ASAP-AES for the Gemini-2.5-Flash model, suggesting that certain structured, form-oriented writing exemplars, such as formal letters and compositions, may positively scaffold the evaluation of student responses for this model. In contrast, transfers from AES to ASAG consistently reduce performance across all models, aligning with expectations that rubric-driven essay exemplars offer limited utility for factual, short-answer evaluation tasks. Reasoning configuration affects performance, but is not the dominant factor.

## 7. Conclusion and Future Work

We introduce S-GRADES, a unified and extensible benchmark that consolidates 14 AES and ASAG datasets comprising over 37K graded student responses under a standardized evaluation framework. By enabling systematic and reproducible comparison across grading paradigms, S-GRADES addresses long-standing fragmentation in educational NLP and provides a foundation for studying cross-task generalization.

Using the benchmark as a testbed, we evaluate three state-of-the-art LLMs under multiple reasoning configurations and analyze the impact of exemplar selection and transfer. These experiments illustrate how standardized, cross-dataset evaluation surfaces meaningful differences in model robustness, reasoning sensitivity, and grading consistency. In particular, the results highlight persistent performance gaps between AES and ASAG, underscoring the need for a unified evaluation framework when assessing model generalization in student response grading.

Future work will extend S-GRADES to multilingual and multimodal settings, improve exemplar selection stability, and incorporate additional reasoning and rubric-grounded prompting strategies. We hope S-GRADES serves as a community resource for rigorous, transparent, and extensible evaluation of automated student assessment systems.

## 8. Limitations

While S-GRADES offers the most comprehensive reasoning-based evaluation to date, several limitations remain that motivate future directions.

- **Model Scope:** Our experiments included only three models: *GPT-4o-mini*, *Gemini-2.5-Flash*, and *LLaMA-4-Scout*. Expanding coverage to newer open-weight and proprietary models (e.g., Llama 3.1, Qwen 2.5, Mistral Large) will help validate architectural trends and reasoning scalability.
- **Exemplar Variance:** Performance varied substantially with random seed selection. Future studies should replicate across multiple random seeds and explore structured sampling to reduce noise and better isolate genuine reasoning gains.
- **Task Coverage:** Although S-GRADES consolidates 14 datasets, all are English and text-only.

Extending to multilingual, multimodal, and non-Western educational data is critical for assessing generalization beyond current benchmarks.

- **ASAG Domain Specificity.** Short-answer grading remains far harder than essay scoring (mean QWK: 0.34–0.41 vs. 0.42–0.43), driven by rubric and question variability. Even GPT-4o-mini's best cross-domain drop (-13.2%) exceeds AES losses (-2.8%). Improving ASAG transfer likely requires rubric-aware training, domain-specific pretraining, or hybrid human–AI grading with educator validation.
- **Methodological Extensions:** Future work should explore richer reasoning combinations (e.g., Inductive + Deductive + Abductive) to test whether multi-step reasoning can enhance generalization across datasets. Integrating these with fine-tuning or meta-learning may help mitigate ASAG transfer failures, especially through rubric-aware or normalized scoring methods.

Overall, these directions aim to advance reproducibility, reduce sampling artifacts, and enable more generalizable educational evaluation with reasoning-augmented language models.

## References


Y. Attali and J. Burstein. 2006. Automated essay scoring with e-rater v.2. *The Journal of Technology, Learning, and Assessment*, 4(3).

Steven Burrows, Iryna Gurevych, and Benno Stein. 2015. The eras and trends of automatic short answer grading. *International journal of artificial intelligence in education*, 25(1):60–117.

Li-Hsin Chang and Filip Ginter. 2024. Automatic short answer grading for finnish with chatgpt. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23173–23181.

Scott A Crossley, Perpetual Baffour, L Burleigh, and Jules King. 2025. A large-scale corpus for assessing source-based writing quality: Asap 2.0. *Assessing Writing*, 65:100954.

Scott A Crossley, Y Tian, P Baffour, Abigail Franklin, Margaret Benner, and Ulrich Boser. 2024. A large-scale corpus for assessing written argumentation: Persuade 2.0. *Assessing Writing*, 61:100865.

Nigel Fernandez, Aritra Ghosh, Naiming Liu, Zichao Wang, Benoît Choffin, Richard Baraniuk, and Andrew Lan. 2022. Automated scoring for reading comprehension via in-context bert tuning. In *International conference on artificial intelligence in education*, pages 691–697. Springer.

P. W. Foltz, D. Laham, and T. K. Landauer. 1999. The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, 1(2):939–944.

Rujun Gao, Xiaosu Guo, Xiaodi Li, Arun Balajiee Lekshmi Narayanan, Naveen Thomas, and Arun R Srinivasa. 2024. Towards scalable automated grading: Leveraging large language models for conceptual question evaluation in engineering. *arXiv preprint arXiv:2411.03659*.

Rujun Gao, Naveen Thomas, and Arun Srinivasa. 2023. Work in progress: Large language model based automatic grading study. In *2023 IEEE Frontiers in Education Conference (FIE)*, pages 1–4. IEEE.

Christian Grévisse. 2024. Llm-based automatic short answer grading in undergraduate medical education. *BMC Medical Education*, 24(1):1060.

Stefan Haller, Adina Aldea, Christin Seifert, and Nicola Strisciuglio. 2022. Survey on automated short answer grading with deep learning: from word embeddings to transformers. *arXiv preprint arXiv:2204.03503*.

Owen Henkel, Libby Hills, Bill Roberts, and Joshua McGrane. 2023. Can llms grade short-answer reading comprehension questions: An empirical study with a novel dataset. *arXiv preprint arXiv:2310.18373*.

Rositsa Ivanova and Siegfried Handschuh. 2024. Evaluating llms' performance at automatic short-answer grading.

Lan Jiang and Nigel Bosch. 2024. Short answer scoring with gpt-4. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*, pages 438–442.

Gerd Kortemeyer. 2024. Performance of the pre-trained large language model gpt-4 on automated short answer grading. *Discover Artificial Intelligence*, 4(1):47.

T. K. Landauer, P. W. Foltz, and D. Laham. 1998. An introduction to latent semantic analysis. *Discourse Processes*, 25(2–3):259–284.

Emmy Liu, Graham Neubig, and Jacob Andreas. 2024. An incomplete loop: Deductive, inductive, and abductive learning in large language models. In *Conference on Language Modeling (COLM)*, Philadelphia, USA.

S. Ludwig, C. Mayer, C. Hansen, K. Eilers, and S. Brandt. 2021. Automated essay scoring using transformer models. *Psych*, 3(4):897–915.

Sandeep Mathias and Pushpak Bhattacharyya. 2018. Asap++: Enriching the asap automated essay grading dataset with essay attribute scores. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Mary L. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3):276–282.

A. Mizumoto and M. Eguchi. 2023. Exploring the potential of using an ai language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2):100050.

Christopher M Ormerod, Akanksha Malhotra, and Amir Jafari. 2021. Automated essay scoring using efficient transformer-based language models. *arXiv preprint arXiv:2102.13136*.

A. Pack, Y. Li, L. Chen, M. Luo, Y. Xu, and Y. Liu. 2024. Large language models and automated essay scoring of english language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence*, 6:100234.

E. B. Page. 1966. The imminence of... grading essays by computer. *The Phi Delta Kappan*, 47(5):238–243.

Zhuang Qiu, Peizhi Yan, and Zhenguang Cai. 2024. Large language models for second language english writing assessments: An exploratory comparison. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 363–370.

Pedro Uria Rodriguez, Amir Jafari, and Christopher M Ormerod. 2019. Language models and automated essay scoring. *arXiv preprint arXiv:1909.09482*.

A. Schneider, M. Schuetz, T. Henneke, and J. Kuhn. 2023. Towards llm-based autograding for short textual answers. *arXiv preprint arXiv:2311.04621*.

Kathrin Seßler, Maurice Fürstenberg, Babette Bühler, and Enkelejda Kasneci. 2025. Can ai grade your essays? a comparative analysis of large language models and teacher ratings in multidimensional essay scoring. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, pages 462–472.

Shashank Sonkar, Kangqi Ni, Lesa Tran Lu, Kristi Kincaid, John S Hutchinson, and Richard G Baraniuk. 2024. Automated long answer grading with ricechem dataset. In *International Conference on Artificial Intelligence in Education*, pages 163–176. Springer.

Chul Sung, Tejas Dhamecha, Swarnadeep Saha, Tengfei Ma, Vinay Reddy, and Rishi Arora. 2019. Pre-training bert on domain resources for short answer grading. 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)*, pages 6071–6075.

K. Taghipour and H. T. Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1882–1891.

C. Xiao, J. Chen, S. Li, T. Wang, R. Zhao, Z. Tang, and Q. Liu. 2024. From automation to augmentation: Large language models elevating the essay scoring landscape. *arXiv preprint arXiv:2401.06431*.

Changrong Xiao, Wenxing Ma, Qingping Song, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Qi Fu. 2025. Human-ai collaborative essay scoring: A dual-process framework with llms. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, pages 293–305.

Wenjing Xie, Juxin Niu, Chun Jason Xue, and Nan Guan. 2024. Grade like a human: Rethinking automated assessment with large language models. *arXiv preprint arXiv:2405.19694*.

# A. Comprehensive Reasoning Strategy Analysis

This appendix provides detailed sensitivity analysis across all reasoning configurations, models, and datasets evaluated in S-GRADES. We quantify the extent to which performance varies with reasoning-strategy selection, complementing the primary results reported in Section 6.

## A.1. Strategy Sensitivity Metrics

To measure sensitivity to reasoning strategy selection, we compute the standard deviation ($\sigma$) of QWK scores across the six reasoning strategies (Ind, Ded, Abd, Ind+Abd, Ind+Ded, Ded+Abd) for each dataset-model pair. A higher $\sigma$ indicates that performance is more sensitive to strategy choice, while a lower $\sigma$ indicates consistent performance regardless of strategy.

Formally, for a model $m$ on dataset $d$, the per-dataset strategy variance is:

$$\sigma_{m,d} = \sqrt{\frac{1}{6}\sum_{s=1}^{6}(\text{QWK}_{m,d,s} - \overline{\text{QWK}}_{m,d})^2} \quad (1)$$

where $\overline{\text{QWK}}_{m,d}$ is the mean QWK across all six strategies for model $m$ on dataset $d$.

To summarize overall strategy sensitivity for a model across all datasets, we compute the mean strategy variance:

$$\sigma_{\text{strategy}} = \frac{1}{D}\sum_{d=1}^{D}\sigma_{m,d} \quad (2)$$

where $D$ is the number of datasets in the task type. Lower $\sigma_{\text{strategy}}$ indicates greater stability across reasoning strategies overall.

Figures 7 and 8 report per-dataset $\sigma_{m,d}$ values (Equation 1) for AES and ASAG regression tasks respectively.

To visually illustrate the per-dataset variance patterns, Figures 7 and 8 plot the per-dataset variance patterns as grouped bar charts. Each group represents one dataset, with bars corresponding to the three models. The figures were generated by computing $\sigma_{m,d}$ directly from the QWK scores in our experimental results using Equation 1, with each bar height representing how much a model's performance fluctuates across the six reasoning strategies on that dataset.

**Key Observations.** For AES tasks, GPT-4o-mini exhibits the lowest average strategy variance ($\sigma_{\text{strategy}} = 0.045$), indicating robust performance across reasoning configurations. This stability is most pronounced on ASAP-AES ($\sigma_{m,d} = 0.018$) and ASAP++ ($\sigma_{m,d} = 0.018$), where performance remains consistent across all six strategies. Gemini-2.5-Flash shows moderate sensitivity ($\sigma_{\text{strategy}} = 0.070$), while LLaMA-4-Scout exhibits the highest variance ($\sigma_{\text{strategy}} = 0.078$), particularly on IELTS Task 1 ($\sigma_{m,d} = 0.132$) where strategy selection critically impacts evaluation quality.

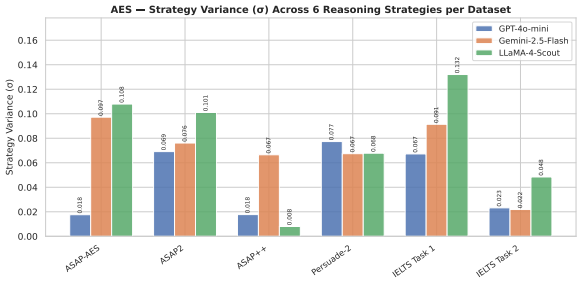


Figure 7: Strategy variance ($\sigma_{m,d}$) across 6 reasoning strategies per AES dataset. Higher values indicate greater sensitivity to reasoning strategy selection.


Figure 8: Strategy variance ($\sigma_{m,d}$) across 6 reasoning strategies per ASAG regression dataset. Higher values indicate greater sensitivity to the selection of reasoning strategies.

For ASAG regression, the pattern reverses: Gemini-2.5-Flash becomes the most strategy-sensitive model ($\sigma_{\text{strategy}} = 0.083$), with extreme variance on ReGrading Dataset (2JC) ($\sigma_{m,d} = 0.153$) driven by the catastrophic failure of pure inductive reasoning (QWK $= -0.20$). GPT-4o-mini maintains the lowest average variance ($\sigma_{\text{strategy}} = 0.060$) for structured datasets such as ASAP-SAS and CSEE, while LLaMA-4-Scout shows moderate sensitivity ($\sigma_{\text{strategy}} = 0.070$). These contrasting patterns suggest that strategy sensitivity is both model-dependent and task-dependent, with no single model consistently robust across all evaluation scenarios.

**Model Stability Rankings:** Across all tasks, GPT-4o-mini exhibits the smallest strategy variance (AES: $\sigma = 0.045$; ASAG Regression: $\sigma = 0.060$), indicating consistent performance regardless of reasoning approach. Gemini-2.5-Flash shows moderate variance (AES: $\sigma = 0.070$; ASAG Regression: $\sigma = 0.083$) with occasional catastrophic failures

(ReGrading Dataset (2JC) *Ind*: QWK $= -0.20$). LLaMA-4-Scout exhibits the highest variance in AES ($\sigma = 0.078$), particularly on IELTS Task 1 (range: 0.27–0.64).

**ASAG Classification Strategy Variance.** Figure 6 presents F1 scores for BEEtlE and SciEntSBank across all reasoning strategies. Unlike AES and ASAG regression tasks, classification datasets show narrower score ranges within each model (GPT: 0.27–0.44; Gemini: 0.21–0.52; LLaMA: 0.14–0.40), indicating moderate sensitivity to reasoning strategy selection. Notably, *Ded+Abd* consistently achieves the lowest F1 across both datasets and all models, while individual *Ind* and *Ded* strategies match or exceed hybrid approaches — a pattern distinct from AES tasks where hybrid strategies provide substantial gains.

## B. Dataset Specific Anomalies

The following cases represent extreme strategy sensitivity or unexpected performance patterns observed across datasets:

- **Gemini 2.5 Flash on ReGrading Dataset (2JC) ($\sigma_{m,d} = 0.153$):** Under purely inductive reasoning, the model shows a strongly negative correlation (QWK = -0.20), whereas deductive reasoning yields a positive correlation of 0.26, a performance upgrade of 0.46. This suggests that example-based reasoning fundamentally misaligns with this particular dataset's grading criteria without a deductive structure.
- **LLaMA-4-Scout on IELTS Task 1 ($\sigma_{m,d} = 0.132$):** Ind+Ded achieves QWK $= 0.64$ while Ind+Abd achieves only $0.27$, a 2.4$\times$ difference. This indicates extreme sensitivity to which secondary reasoning mode accompanies inductive reasoning on descriptive writing tasks.
- **Gemini-2.5-Flash on Rice_Chem** ($\sigma_{m,d} = 0.149$): driven by strong inductive performance (QWK = 0.70) but weak deductive performance (QWK = 0.51), suggesting chemistry content evaluation benefits strongly from example-based grounding for this model.
- **Gemini-2.5-Flash on CSEE ($\sigma_{m,d} = 0.109$):** Ded+Abd achieves only QWK $= 0.33$ while Ind+Abd and Ind+Ded both achieve $0.64$. It suggests that inductive examples are critical for evaluating structured short-form writing tasks such as formal letters and compositions.
- **LLaMA-4-Scout on ASAP++ ($\sigma_{m,d} = 0.008$):** This dataset exhibits the lowest variance across models, suggesting that reasoning strategy has minimal impact. The overall task difficulty appears to dominate, with all strategies performing similarly poorly.
- **GPT-4o-mini on ASAP-AES ($\sigma_{m,d} = 0.018$):** Remarkably stable across all strategies (range: 0.915–0.954), suggesting that GPT-4o-mini evaluates essays consistently regardless of the specific reasoning configuration used.

## C. Exemplar Generalization

Table 3 in Section 6 reports exemplar generalization results under inductive 5-shot prompting where training examples are drawn from a different dataset than the test set. Here we provide additional context on the experimental design and observed failure modes not discussed in the main paper.

Dataset pairs were selected based on task compatibility: ASAP2→ASAP-AES for within-domain AES transfer, and CSEE↔ASAP-SAS for cross-domain ASAG transfer. A single random seed (42) was used for exemplar selection, meaning results reflect one specific sampling, variability across seeds is characterized separately in Appendix **??**.

The most notable failure is LLaMA-4-Scout under cross-paradigm transfer (ASAP-AES→CSEE, $\Delta$QWK $= -95.5\%$; ASAP-SAS→CSEE, $\Delta$QWK $= -47.2\%$), which is disproportionately large compared to GPT-4o-mini and Gemini-2.5-Flash on the same transfers. This collapse suggests LLaMA-4-Scout is particularly sensitive to domain mismatch between exemplars and test content, likely because its inductive reasoning relies more heavily on surface-level patterns in the provided examples rather than abstract quality criteria. In contrast, the unexpected improvement of Gemini-2.5-Flash under CSEE→ASAP-AES transfer ($\Delta$QWK $= +12.5\%$) suggests that structured, form-oriented writing exemplars may positively scaffold student response evaluation for this model, an effect that warrants further investigation.

## D. Prompt Templates

Each reasoning strategy is implemented using a two-part prompt structure. The *system prompt* specifies how the model should reason about the task, while the *user prompt* provides the essay and clearly defines the expected output format. We describe the numeric essay-scoring variant here, as it represents the most general case. The classification variants follow the same structure, with the output restricted to one of three labels: `correct`, `incorrect`, or `contradictory`, depending on the task. For all strategies, we dynamically insert few-shot examples drawn from the training split of the corresponding dataset to guide the model's responses.

## D.1. Inductive Reasoning

Inductive reasoning learns scoring patterns directly from training examples, inferring criteria from observed score-response pairs.

**System Prompt: Inductive**

```
You are an expert essay scorer using INDUCTIVE REASONING.

INDUCTIVE PROCESS:
1. Learn scoring patterns from the examples below
2. Identify scoring criteria from the example patterns
3. Apply these learned patterns to score the new essay

SCORING EXAMPLES FROM TRAINING DATA:
[FEW-SHOT EXAMPLES INSERTED HERE]

From these examples, identify patterns in:
- What makes a high score vs low score
- How content quality affects scoring
- What level of development is expected

SCORING RANGE: [min, max]
TASK: Score the essay based on the patterns you learned.
```

**User Prompt: Inductive**

```
Based on the patterns you learned from the examples above, score this essay:

QUESTION/PROMPT:
[QUESTION]

ESSAY TO SCORE:
[STUDENT RESPONSE]

STOP. Do not write steps. Do not write explanations. Do not write reasoning.

Your ENTIRE response must be EXACTLY one number between [min] and [max].

Nothing else. Just the number.
```

## D.2. Deductive Reasoning

Deductive reasoning applies general scoring rules and principles to the specific essay, deriving a score through logical rule application.

**System Prompt — Deductive**

```
You are an expert scorer using DEDUCTIVE REASONING.

DEDUCTIVE PROCESS:
1. Start with GENERAL scoring rules/criteria
2. Apply rules to THIS SPECIFIC essay
3. Derive score logically from rule application

EXAMPLES OF DEDUCTIVE SCORING:
GENERAL RULE 1: Complete answers provide required details
APPLICATION: Student provides 3 pieces of information
DEDUCTION: Meets requirement → Base score awarded

GENERAL RULE 2: Answers must explain reasoning, not just state results
APPLICATION: Student gives correct number, no explanation
DEDUCTION: Partial credit only

GENERAL RULE 3: Scientific answers must correctly apply relevant laws
APPLICATION: Student mentions law but misapplies concept
DEDUCTION: Partial credit (understands principle, has errors)

SCORING RANGE: [min, max]
TASK: Essay scoring
```

**User Prompt: Deductive**

```
Apply deductive reasoning to score:

QUESTION/PROMPT:
[QUESTION]

ESSAY TO SCORE:
[STUDENT RESPONSE]

Apply general scoring rules to this specific essay and derive the score.

Provide ONLY a single number between [min] and [max] with no explanation. Just the number.
```

## D.3. Abductive Reasoning

Abductive reasoning treats the student response as an observation, generates possible explanations for the student's knowledge state, and infers the most plausible score from the best-fitting explanation. Although abductive reasoning is inherently inferential, the prompt requires the model to perform this reasoning internally while returning only the final numeric score, without exposing the reasoning traces.

**System Prompt: Abductive**

```
You are an expert scorer using ABDUCTIVE REASONING.

ABDUCTIVE PROCESS:
1. Observe what the student wrote
2. Generate explanations for their knowledge state
3. Identify which explanation best fits the evidence
4. Infer appropriate score from that explanation

EXAMPLES OF ABDUCTIVE SCORING:
OBSERVATION: Student provides amount, temperature, duration
POSSIBLE EXPLANATIONS:
- Student randomly listed items → irrelevant details expected
- Student identified genuinely missing info → important details expected
- Student did not read procedure → already-specified items expected
BEST EXPLANATION: Correctly identified critical missing details
SCORE: 2/3

OBSERVATION: Student wrote correct number, no explanation
POSSIBLE EXPLANATIONS:
- Student ran simulation but does not understand why
- Student guessed correctly
- Student understands deeply
BEST EXPLANATION: Obtained correct result, cannot explain process
SCORE: 8/15

SCORING RANGE: [min, max]
TASK: Essay scoring
```

**User Prompt: Abductive**

```
Use abductive reasoning to score:

QUESTION/PROMPT:
[QUESTION]

ESSAY (OBSERVATION):
[STUDENT RESPONSE]

STOP. Do not write steps. Do not write explanations. Do not write reasoning.

Your ENTIRE response must be EXACTLY one number between [min] and [max].

Nothing else. Just the number.
```

### D.4. Hybrid Strategies

The three hybrid strategies combine two reasoning phases sequentially within a single prompt. Each hybrid retains the full system prompt structure of both constituent strategies, with a combined instruction header directing the model to apply both reasoning processes before producing a score. Table 4 summarizes the process.

| Strategy | Phase 1 | Phase 2 |
|---|---|---|
| Ind+Ded | Learn patterns from examples (Inductive) | Apply general scoring rules (Deductive) |
| Ind+Abd | Learn patterns from examples (Inductive) | Infer best explanation (Abductive) |
| Ded+Abd | Apply general scoring rules (Deductive) | Infer best explanation (Abductive) |

Table 4: Hybrid strategy prompt composition. Each hybrid combines the system prompt blocks of two solo strategies sequentially.

**System Prompt Structure: All Hybrid Strategies**

```
You are an expert scorer using [STRATEGY A] then [STRATEGY B] REASONING.

PHASE 1 - [STRATEGY A] REASONING:
[Full Phase 1 block from Strategy A]

PHASE 2 - [STRATEGY B] REASONING:
[Full Phase 2 block from Strategy B]

COMBINED APPROACH:
1. Apply Strategy A reasoning
2. Apply Strategy B reasoning
3. Score based on both methods

SCORING RANGE: [min, max]
Use both reasoning approaches before scoring.
```

The user prompt for all hybrid strategies is identical to the solo strategies, constraining output to a single number within the valid scoring range. Full prompt code for all six strategies is available in our repository.[5]

## E. Prediction Stability Analysis

This appendix presents additional analysis of prediction stability across repeated model evaluations. The goal of this experiment is to measure how consistently LLMs produce the same predictions when evaluated multiple times on identical inputs. This analysis is distinct from the exemplar-sampling stability experiment reported in Section 6, which varies few-shot exemplar seeds; here, we measure run-to-run prediction variability under repeated querying of the same response.

[5] https://github.com/nlpatunt/sgrades-experiments

## E.1. Experimental Protocol

For each model–strategy combination, prediction stability was measured by generating **three separate predictions** for every student response.

To make repeated evaluation computationally feasible across models and reasoning strategies, stability analysis was conducted on a sampled subset comprising **21% of each dataset's test split**. The same sampled subset was reused across all models and strategies.

All runs used a **sampling temperature of 0.1**, consistent with the settings used in the primary experiments, which keeps model outputs nearly deterministic. As a result, any variation observed across repeated runs primarily reflects inconsistencies in the model's scoring behavior rather than sampling-induced randomness.

For strategies that include an inductive component (Ind, Ind–Ded, Ind–Abd), few-shot examples are drawn from the training split using **random seed 42**, ensuring prompt construction and example selection remain identical across repeated runs.

## E.2. Stability Metrics

We evaluate stability separately for numeric and categorical datasets.

**Numeric datasets.** For each essay $i$, let $\hat{y}_i^{(1)}, \hat{y}_i^{(2)}, \hat{y}_i^{(3)}$ be the scores produced by the three inference calls. We compute the sample standard deviation:

$$\sigma_i = \sqrt{\frac{1}{2}\sum_{k=1}^{3}\left(\hat{y}_i^{(k)} - \bar{y}_i\right)^2}, \qquad \bar{y}_i = \frac{1}{3}\sum_{k=1}^{3}\hat{y}_i^{(k)}. \tag{3}$$

We report the **pooled mean standard deviation** $\overline{\sigma}$ across all essays, weighted by essay count. Sub-question datasets (e.g., OS Dataset and Rice Chem) are first merged to avoid double-counting. **Lower values indicate more stable scoring.**

**Categorical datasets.** For categorical tasks, stability is measured using agreement across the three predictions. For essay $i$:

$$a_i = \frac{\left|\{k : \hat{y}_i^{(k)} = \mathrm{mode}(\hat{y}_i^{(1)}, \hat{y}_i^{(2)}, \hat{y}_i^{(3)})\}\right|}{3}. \tag{4}$$

The final predicted label is taken as the **mode** of the three predictions. For 2-way tasks, this corresponds to choosing between *correct* and *incorrect*; for 3-way tasks, it corresponds to choosing among *correct*, *incorrect*, and *contradictory*. If all three predictions differ, the response is **not discarded**; its agreement score is $1/3$, reflecting maximal disagreement across the three calls.

We report the **pooled mean agreement** across all categorical datasets. Higher values indicate more stable predictions.

## E.3. Results and Discussion

### E.3.1. Overall Stability Across Models and Strategies

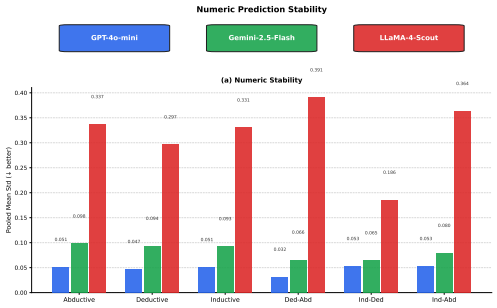


Figure 9: Pooled mean standard deviation ($\overline{\sigma}$) on numeric datasets across six reasoning strategies and three models. Lower is better.

Figure 9 summarizes stability for numeric scoring datasets using pooled mean standard deviation across three repeated calls. Clear differences emerge across the three models. GPT-4o-mini is the most stable overall, with pooled mean standard deviation values ranging from approximately 0.032 to 0.053 across reasoning strategies. Gemini-2.5-Flash shows slightly higher but still relatively low variability, with values ranging from about 0.065 to 0.098. In contrast, LLaMA-4-Scout exhibits substantially higher instability, with pooled mean standard deviation values between 0.186 and 0.391 across strategies.

Among the six strategies, *Ded+Abd* yields the lowest numeric instability for GPT-4o-mini, while *Ind+Ded* gives the lowest value for both Gemini-2.5-Flash and LLaMA-4-Scout. Overall, the gap between models is much larger than the variation introduced by strategy choice alone, indicating that model choice plays the dominant role in the stability of numeric predictions.

Figure 10 presents the corresponding results for categorical datasets using mean agreement across repeated calls. All three models are considerably more stable on categorical tasks than on numeric scoring tasks. **GPT-4o-mini and Gemini-2.5-Flash** achieve near-perfect agreement across all reasoning strategies, generally remaining between 98.6% and 99.3%. LLaMA-4-Scout is less stable than the other two models, but still maintains relatively high agreement, ranging from about 96.5% to 97.7%. These results suggest that categorical

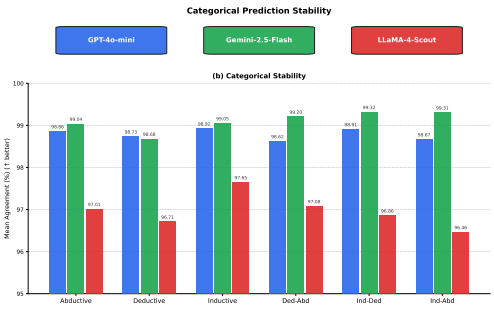


Figure 10: Pooled mean agreement (%) on categorical datasets across six reasoning strategies and three models. Higher is better.

grading is inherently less sensitive to variability in repeated inference than numeric scoring, likely because classification decisions involve a smaller output space and fewer fine-grained distinctions than essay score prediction.

Taken together, these findings show that prediction stability depends primarily on model choice and task type rather than reasoning strategy. Numeric scoring tasks are substantially more variable than categorical tasks across all models. GPT-4o-mini is the most stable on numeric tasks, Gemini-2.5-Flash leads on categorical tasks, and LLaMA-4-Scout is consistently the least stable under repeated evaluation.

### E.3.2. Dataset-Level Stability Patterns

Figures 11, 12, and 13 present per-dataset stability heatmaps for all three models. A clear pattern is that **ASAG predictions are consistently less stable than AES predictions** across all models.

Averaged across all datasets and strategies, ASAG mean standard deviation is 1.6× higher than AES for GPT-4o-mini (0.077 vs. 0.049) and Gemini-2.5-Flash (0.160 vs. 0.098), and 2.0× higher for LLaMA-4-Scout (0.599 vs. 0.305). This gap persists even after excluding OS_Dataset — which is an outlier due to its extremely small test split (3–8 essays per sub-question) — for GPT-4o-mini and LLaMA-4-Scout (1.3× in both cases), suggesting that the difference reflects broader task-level characteristics rather than only a single dataset artifact.

Within ASAG, **CSEE** and **Mohlar** are consistently the most unstable datasets after OS_Dataset, while **ASAP-SAS** and **Rice_Chem** remain stable across all models. Within AES, **ASAP-AES** shows the highest variability across models.

Overall, these heatmaps reinforce the broader conclusion that prediction stability is shaped by both model and task type. While the prompting strategy has some effect, the larger and more consistent difference appears between AES and ASAG, with short-answer grading showing substantially greater

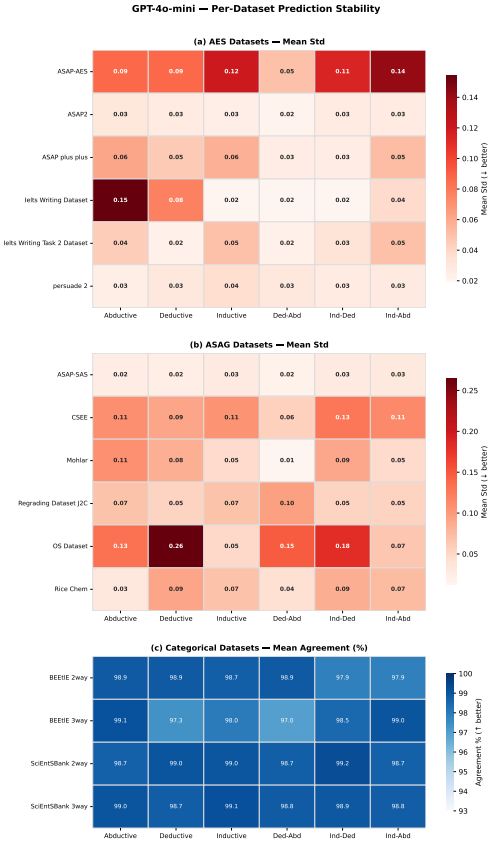


Figure 11: Per-dataset prediction stability for GPT-4o-mini across six reasoning strategies.

sensitivity to repeated evaluation. This supports the view that stability should be considered a task-dependent property of LLM-based scoring systems rather than a uniform model behavior.

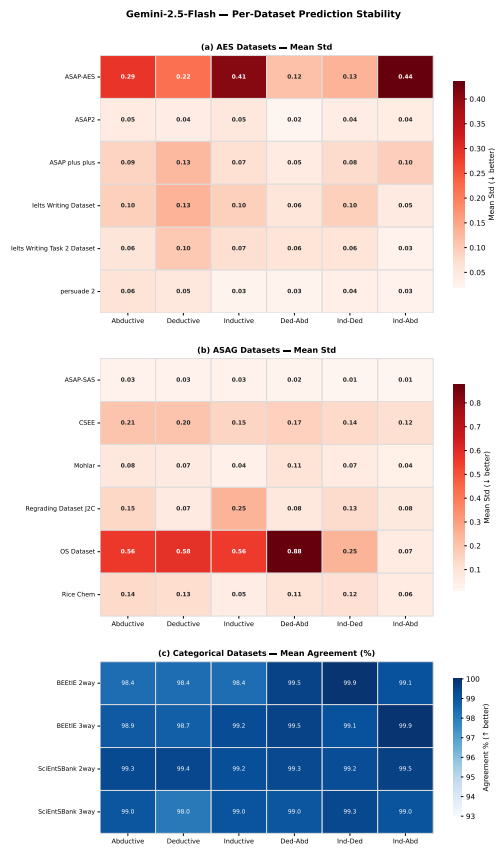


Figure 12: Per-dataset prediction stability for Gemini-2.5-Flash across six reasoning strategies.


Figure 13: Per-dataset prediction stability for LLaMA-4-Scout across six reasoning strategies.