DynoGrader: Automated Partial-Credit Grading for Proof Problems with Diverse Solution Approaches

Anonymous ACL submission

Abstract

Grading proof problems in university courses is costly and difficult: students take diverse approaches to the same problem, and partial credit must accurately reflect which steps are correct. Existing LLM-based graders rely on a single reference answer, consistently under-scoring valid but unconventional approaches. We present **DynoGrader**, a human-in-the-loop grading system that scores each student answer by first identifying which solution approach the student used, then evaluating every step against a rubric tailored to that approach. To cover all valid approaches with minimal manual effort, DynoGrader computes a *submission informativeness score* to surface the answers most likely to reveal a previously unseen approach, and iteratively presents these to a human reviewer—achieving full approach coverage in under 9 minutes of TA time per question. On **DynoBench**—a new benchmark of human-graded proof submissions spanning two university courses—DynoGrader reduces MAE by 22.0% with expert-defined approaches and by 16.6% with automatically discovered ones ($p < 0.001$), with discovered approaches matching or surpassing expert-defined ones on the highest-diversity questions. Cross-course experiments on 605 additional responses confirm generalization.

1 Introduction

Grading proof problems in university STEM courses is expensive and expertise-intensive. Unlike multiple-choice or calculation questions, a single proof problem can be solved in multiple valid ways, and a student who takes an unconventional solution approach may make errors at completely different steps than one following the standard solution. Correctly assigning partial credit therefore requires understanding *which* approach each student took and *where* it breaks down—a per-submission judgment that is too costly to make manually at scale.

Recently, large language models (LLMs) have been widely adopted for automatic grading (Liu et al., 2023; Eneye et al., 2025). The simplest method—directly prompting an LLM to assign a score without explicit scoring criteria ("black-box" grading)—is inconsistent: the same model may award different scores to two answers of equal merit, with no record of what criteria were checked or why. A natural remedy is to introduce a *rubric*, a structured list of scoring criteria that decomposes the expected answer into checkable steps. Rubric-based methods such as ProofGrader (Ma et al., 2026) and RefGrader (Mahdavi et al., 2025) have shown that anchoring LLM scoring to explicit criteria improves consistency. However, when a rubric is built from a single reference answer, it still systematically under-scores students who reach the correct conclusion via a different solution approach.

Figure 1 illustrates this with a concrete example: a Newton local quadratic convergence proof that admits gradient-space, error-space, and fixed-point approaches. Three students each follow a different approach and exhibit gaps at different stages. A black-box grader assigns inconsistent scores because it has no explicit criteria to follow; a rubric-based grader correctly evaluates only the student whose approach matches the reference answer and under-scores the others. This reveals two core challenges:

C1. Consistent, approach-aware scoring. Black-box grading has no explicit criteria, leading to inconsistent scores; single-reference rubrics add structure but still fail students who use a different approach. Accurate partial-credit grading requires *both*: grounding each judgment in direct evidence from the student's answer, and applying scoring criteria matched to the student's actual approach.

C2. Covering all valid solution approaches. Manually writing scoring criteria for every

PROBLEM MOTIVATION

Students solve in fundamentally different ways

Three real student answers to the same Newton quadratic convergence question. Each uses a valid but structurally distinct proof strategy. A scorer that only knows the reference answer may under-credit legitimate alternative routes.

Reference proof skeleton

1. Set up Newton
Error recursion, gradient cancellation, or $g(x)$ fixed-point

2. Second-order remainder
Taylor or integral remainder makes the residual quadratic

3. Inverse-Hessian control
Bound $[\nabla^2 f(x)]^{-1}$ or justify local stability near x^*.

4. Final x-space conclusion
Conclude $\|x_{k+1} - x^*\| \le C\|x_k - x^*\|^2$ in x-space.

Route A — GRADIENT_SPACE

Works entirely in gradient space: shows $\|\nabla f\|$ converges quadratically, then transfers back to x-space.

Proof: Denote $d = -[\nabla^2 f(x_k)]^{-1}\nabla f(x_k)$. We know that

$$\begin{aligned}
&\|\nabla f(x_{k+1})\| \\
&= \|\nabla f(x_k+d) - \nabla f(x_k) - \nabla^2 f(x_k)d\| \\
&= \left\|\int_0^1 (\nabla^2 f(x_k+td) - \nabla^2 f(x_k))d\,dt\right\| \\
&\le \tfrac{L}{2}\|d\|^2 \\
&= \tfrac{L}{2}\|(\nabla^2 f_k)^{-1}\nabla f_k\|^2 \\
&\le \tfrac{L}{2\mu^2}\|\nabla f(x_k)\|^2.
\end{aligned}$$

So we get

$$\frac{\|\nabla f(x_{k+1})\|}{\|\nabla f(x_k)\|^2} \le \frac{L}{2\mu^2}.$$

This means that $\|\nabla f(x)\|$ converge to zero in quadratic convergence rate, hence x converge to x^* in quadratic convergence rate.

Route B — ERROR_SPACE

Direct error-space bound via integral remainder. Terse but the core idea is visible.

Proof: Suppose two Lipschitz constants are L and L'.

$$\begin{aligned}
&|x_{k+1} - x^*| \\
&= |x_k - (\nabla^2 f(x_k))^{-1}\nabla f(x_k) - x^*| \\
&\le L|(\nabla^2 f(x_k))(x_k - x^*) - \nabla f(x_k)| \\
&= L\Big|\int_0^1 (x_k - x^*)\big[\nabla^2 f(x_k) \\
&\quad - \nabla^2 f(x^* + t(x_k - x^*))\big]dt\Big| \\
&\le \tfrac{LL'}{2}|x_k - x^*|^2
\end{aligned}$$

So the convergence rate is quadratic.

Route C — FIXED_POINT

Treats Newton as a fixed-point iteration $g(x)$; shows $\nabla g(x^*) = 0$ then applies Taylor.

Proof: let $g(x) = x - [\nabla^2 f(x)]^{-1}\nabla f(x)$, first prove that $\nabla g(x^*) = O$:

$$\begin{aligned}
&\nabla g(x^*) \\
&= I - \nabla[(\nabla^2 f)^{-1}\cdot\nabla f]\big|_{x^*} - [\nabla^2 f(x^*)]^{-1}\nabla^2 f(x^*) \\
&= I - I = O
\end{aligned}$$

And it's easy to get $g(x^*) = x^*$. So by Taylor's expansion,

$$\begin{aligned}
&x_{k+1} - x^* \\
&= g(x_k) - x^* = g(x_k) - g(x^*) \\
&= \nabla g(x^*)^T(x_k - x^*) \\
&\quad + \tfrac{1}{2}(x_k - x^*)^T\nabla^2 g(x')(x_k - x^*)
\end{aligned}$$

where $x' = \eta x^* + (1-\eta)x_k$. Let e_m be the maximum eigenvalue of $\nabla^2 g(x')$, then we can get

$$\|x_{k+1} - x^*\| \le \frac{e_m}{2}\|x_k - x^*\|^2$$

The scoring challenge

All three routes are mathematically valid, yet they decompose the proof into different logical steps. A rubric built for the reference answer's error-space route may under-credit Route A (gradient-space) or Route C (fixed-point). Recognizing these diverse proof strategies is a prerequisite for fair, accurate grading.

Figure 1: Case study on a Newton local quadratic convergence proof. Three students solve the problem via gradient-space, error-space, and fixed-point approaches, respectively, each with gaps at different steps. The rightmost two columns compare approach-aware grading with human scores.

possible solution approach to every question is impractical for a teaching assistant.

We propose **DynoGrader**, a human-in-the-loop grading system designed around these two challenges. For **C1**, DynoGrader uses *evidence-grounded structured scoring*: for each scoring criterion, the LLM must cite a direct quote from the student's answer to support its judgment, making every point assignment verifiable. When a catalog of solution approaches is available—each represented as a short name, a natural-language description of the solution strategy, and step-specific scoring criteria tailored to that strategy (Appendix C gives a concrete example)—the system identifies the best-matching approach for each student answer and scores it against the corresponding criteria, eliminating the systematic bias against unconventional solutions. For **C2**, rather than requiring a TA to enumerate all solution approaches upfront, DynoGrader computes a *submission informativeness score* for each answer to identify which submissions are most likely to reveal a previously unseen approach, and iteratively surfaces these for lightweight human review—automatically grouping answers into solution approach families over multiple rounds. When a complete approach cata-

log is already available, it can be supplied directly.

To evaluate DynoGrader beyond a single illustrative example, we construct **DynoBench**, a benchmark spanning two real university courses with human-graded student responses, providing a realistic testbed for evaluating both evidence-grounded structured scoring and approach-aware grading under practical conditions.

Contributions:

1. **Approach-aware evidence-grounded scoring.** We represent each solution approach as a natural-language description paired with step-specific scoring criteria. At grading time, the system first identifies which approach a student followed, then scores the answer step by step against the corresponding criteria—with every judgment backed by a direct evidence quote from the student's submission. Using expert-defined approaches, this reduces MAE by 22.0% over the strongest external baseline ($p < 0.001$).
2. **Automatic solution approach discovery from minimal TA feedback.** We design a *submission informativeness score* that quantifies how likely each answer is to reveal a previously unseen solution approach, enabling the system to prioritize which submissions to surface for human review. Through iterative rounds guided by this score, the system automatically groups answers into solution approach families and stops when accuracy plateaus, keeping total TA effort under 9 minutes per question (reducing MAE by 16.6% over the strongest baseline). On questions with the highest answer diversity, automatically discovered approach families match or surpass expert-defined ones.
3. **The DynoBench evaluation benchmark.** We construct DynoBench, a two-course benchmark for realistic evaluation of approach-aware proof grading. On its 6 complex proof questions with the highest answer diversity, we systematically compare against 6 external baselines and ablation settings. Cross-course experiments further confirm that evidence-grounded structured scoring generalizes and that approach discovery helps when genuine alternative approaches exist (Appendix N).

2 Related Work

2.1 LLM Grading and Rubric Anchoring

Automatic short answer grading (ASAG) and automated essay scoring (AES) have a long research history (Burrows et al., 2015; Ke and Ng, 2019). The LLM-as-judge paradigm has been widely applied to text evaluation (Liu et al., 2023), and Eneye et al. (Eneye et al., 2025) survey LLMs in educational grading. In proof grading, ProofGrader (Ma et al., 2026) achieves high agreement with expert scores by combining reasoning models with marking schemes, and RefGrader (Mahdavi et al., 2025) derives problem-specific rubrics from reference solutions via an agentic workflow. RocketEval (Wei et al., 2025) generates per-instance checklists and scores each item independently. Our EGSS shares the decomposition philosophy—slots are analogous to checklist items—but slots are *evidence-grounded* (requiring verbatim quotes) and *approach-conditioned* (adapting to each proof strategy). Evidence grounding is related to attribution and faithfulness in QA/fact-checking (Rashkin et al., 2023; Gao et al., 2023); we adapt this idea to grading by requiring per-slot verbatim quotes (Appendix E reports a spot-check of quote fidelity). We focus on LLM-based methods because traditional feature-based or fine-tuned ASAG models lack the slot-level evidence and approach-specific reasoning that proof grading demands.

2.2 Human-in-the-Loop Grading and Selective Review

HITL approaches to grading are receiving increasing attention (Xiao et al., 2025; Chu et al., 2025; Li et al., 2025; Armfield et al., 2025). In selective classification and learning-to-defer, Geifman and El-Yaniv (Geifman and El-Yaniv, 2017) propose selective classification for deep networks and Mozannar and Sontag (Mozannar and Sontag, 2020) establish consistent learning-to-defer. Our key distinction is that TA feedback not only corrects predictions but directly participates in approach family induction; review signals include approach-specific evidence gaps and unknown-approach flags.

2.3 Formal Proof Verification

Interactive theorem provers (Lean (de Moura and Ullrich, 2021), Isabelle (Nipkow et al., 2002)) can verify machine-readable proofs, and recent work explores autoformalization of informal mathematics (Wu et al., 2022; Jiang et al., 2023). However, student homework proofs mix natural language, notation, and diagrams, making end-to-end formalization impractical; DynoGrader therefore operates directly on natural-language submissions.

2.4 Taxonomy Induction and Structure-Aware Reasoning

LLM-guided taxonomy induction frameworks (Pham et al., 2024) iteratively cluster text into interpretable categories. Our approach discovery shares this iterative ethos but is specialized for grading utility: approach families must provide slot-level guidance, and quality is anchored by TA confirmations rather than model-internal coherence. Mukherjee et al. (Mukherjee et al., 2025) annotate premise dependencies of reasoning steps to improve LLM error identification; Poulsen et al. (Poulsen et al., 2022) develop interactive proof construction tools. Our approach families serve as a persistent scoring interface rather than an error-detection or pedagogical tool. In active learning, the informativeness–diversity trade-off (Settles, 2009; Ash et al., 2020) motivates our four-way selection heuristic; we compare against random sampling in §6.3.

3 Problem Setting

3.1 Task Definition

Given a proof problem q, its reference answer a^*, a scoring rubric r, and N student answers $\{x_1, \ldots, x_N\}$, each accompanied by a human-assigned score $y_i \in [0, S]$, under a limited TA review budget B ($B \ll N$), the system must jointly produce: (1) a fully automatic score $\hat{y}_i$ for each answer; (2) a review priority ranking π; (3) auditable scoring evidence.

The formal objective is to minimize grading error under the budget constraint:

$$\min_{\theta} \ \mathrm{MAE}(\hat{y}, y) \quad \text{s.t.} \quad |\mathcal{R}| \leq B$$

where $\mathcal{R} \subseteq \{1, \ldots, N\}$ is the set of reviewed samples, and θ includes the approach families $\mathcal{F}$, approach-aware scoring rules, and review ranking strategy.

3.2 TA Feedback Protocol

To minimize TA burden, we design an extremely lightweight feedback protocol. For each reviewed answer x_i, the TA provides only two items:

- **Total score** $y_i^{\mathrm{TA}} \in [0, S]$;
- **Binary judgment** $b_i \in \{0, 1\}$: whether this answer follows a valid but non-standard solution approach.

The TA does not need to name or define the approach, annotate sub-step scores, or provide written explanations. The total score provides a score calibration signal; the binary judgment provides a structural discovery signal. Together they simultaneously correct scores and drive approach family discovery.

3.3 System Overview

Figure 2 illustrates DynoGrader's workflow. EGSS first scores all answers without TA involvement. Then, over multiple rounds, the system selects informative samples for TA feedback, induces approach families, and injects them into the grader, forming a closed loop: feedback → discovery → better scoring → better selection → next-round feedback.

4 Method

4.1 Evidence-Grounded Scoring Foundation

The foundation of DynoGrader is **Evidence-Grounded Structured Scoring** (EGSS). Given a problem's reference answer and scoring rubric, the system decomposes the rubric into scoring slots, each corresponding to an assessable proof step or concept, with a slot description and a maximum point value.

For each student answer, the LLM produces structured output for each slot: a score, a coverage level (none / partial / full), a verbatim evidence quote from the answer, and a one-sentence reason. This forces the LLM to anchor its scoring in concrete evidence rather than assigning an overall score based on impression. The scoring process requires only a single LLM call without multi-round reflection or score calibration engineering, ensuring high reproducibility and fairness.

EGSS requires neither approach knowledge nor TA involvement and serves as a **zero-TA-cost** start-

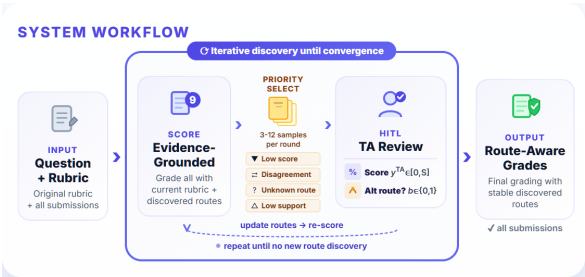

Figure 2: DynoGrader workflow. (1) Load question and rubric; (2) Evidence-Grounded Structured Scoring (EGSS); (3) Multi-round TA binary feedback drives approach family discovery; (4) Output approach-aware final grades. Steps 2–3 iterate 3–4 rounds, requiring only 48 TA feedback samples per question.

ing point. Experiments show (§6.2) that EGSS alone reduces MAE by 8.2% compared to the strongest external baseline.

4.2 Approach-Aware Scoring Extension

When an approach taxonomy is available, EGSS naturally extends to approach-aware scoring. An approach taxonomy contains several approaches, each with an approach description and approach-specific slot guidance (e.g., which steps are critical under that approach and how to interpret alternative evidence).

For each answer, the system first determines the best-matching approach and then uses that approach's slot guidance to assess each slot. When an answer matches a non-standard approach, approach-specific guidance helps the LLM correctly evaluate the evidence under that approach, resolving the systematic bias of canonical rubrics against non-standard paths. If an answer does not match any known approach, the system falls back to the EGSS baseline.

The approach taxonomy may come from expert definitions or from the HITL discovery process described below.

4.3 Informativeness-Driven Sample Selection

In each iteration, the system ranks all unreviewed submissions by a *submission informativeness score* and selects the top b for TA review. The score is computed via four-way stratified sampling: (a) **Low-score pool**: samples with the lowest EGSS scores, most likely to be mis-graded; (b) **Disagreement pool**: samples with the largest discrepancy between EGSS and current approach-aware scores, exposing inadequacies in the existing approach families; (c) **Novelty pool**: samples with non-low scores but low textual overlap with the reference answer, most likely to contain non-standard approaches; (d) **Mid-range novelty pool**: high-novelty samples in the mid-score range, supplementing boundary cases. The four-way proportions are hyperparameters; the first round and subsequent rounds may use different settings.

4.4 Approach Family Induction

From the accumulated TA feedback, the system extracts the positive set $P = \{x_i : b_i = 1\}$ of answers confirmed as following valid but non-standard approaches. The answer texts and TA scores of the positive set are fed to the LLM, which is asked to induce interpretable approach families. Each approach family must include: a short name, a natural-language description, and approach-specific scoring slot guidance. Approach families with support below a threshold are filtered out.

In subsequent rounds, the LLM receives the existing approach family list as context and may add, refine, merge, or maintain families. As rounds progress, the approach family set tends to stabilize.

Two constraints are imposed on the induction process: (1) the canonical standard approach is not re-induced; (2) approach families are kept compact and interpretable, avoiding a proliferation of fine-grained categories.

4.5 Multi-Round Iteration and Selective Review

Steps 4.3–4.4 form an iterative cycle: each round scores all samples, selects b for TA feedback, induces or updates approach families, and rebuilds the grader. Approach families grow fastest in the first few rounds and MAE typically plateaus by rounds 3–5; the total TA budget is $R \times b$, and users can choose a stopping point based on the accuracy–budget curve. After discovery, unreviewed samples are ranked for TA review by a weighted combination of low score, grader disagreement, unknown-approach flag, and low evidence coverage. A key design is that discovery and review share the same TA workflow: each reviewed sample simultaneously yields score correction and approach discovery signal, so the two budgets are not consumed independently.

5 Experimental Setup

5.1 Data

DynoBench spans two university courses—an advanced undergraduate analytical course (Course A) and an introductory STEM course (Course B)—totaling 121 questions and 12,820 human-graded responses.

Main evaluation (Course A). The main evaluation uses 6 complex proof questions with significant approach diversity (question ids 4, 6, 24, 29, 30, 35), comprising 690 student responses drawn from three years (2022–2024). Responses from different cohorts are pooled per question because the LLM grader scores each submission independently with no cross-sample memory; further discussion of cross-year leakage is in §7. Each response is accompanied by a human-assigned score (maximum

scores vary by question; all are normalized to a 0–10 scale). All 6 questions have expert-defined approach taxonomies (2–5 approaches per question), used to evaluate the upper bound of approach-aware scoring. Two additional questions (q17, q31) have highly uniform student approaches and are excluded from the main evaluation.

Cross-course expansion (Course B). To test generalization beyond proof-heavy Course A questions, we evaluate DynoGrader on 6 short-answer questions from Course B (605 responses), covering topics such as fault tolerance, quantum entanglement, convolution proofs, and game theory. Slot schemas are auto-generated by the LLM, and gold labels are TA-assigned scores normalized to a 0–10 scale. Full results are reported in Appendix N.

5.2 External Baselines

We re-evaluate 8 external baselines on the same 6-question subset, using the same normalized labels: Direct+Ref (Liu et al., 2023), Batched Direct+Ref (batch size 16), Blind, LLM Rubric, Two-Pass Reflection, RefGrader (Mahdavi et al., 2025), Debate, and GradeHITL (Li et al., 2025). All use the same GPT-5.4 backbone and output format; GradeHITL additionally requires a per-question 4:1 train/test split (137 test samples). Implementation details are in Appendix F.

5.3 System Variants and Ablations

DynoGrader operates in two modes:

- **DynoGrader + Expert Approaches**: Approach-aware scoring using expert-defined approach taxonomies.
- **DynoGrader + HITL Discovery**: Approach-aware scoring using approach families automatically induced through iterative discovery. Each question independently runs 4 discovery rounds with 12 TA feedback samples per round, totaling 48 per question.

Ablation experiments:

- **w/o Approaches**: Remove the approach taxonomy, retaining only evidence-grounded structured scoring—measuring the incremental contribution of approach information.
- **w/o Structure (Blackbox)**: Further remove the slot structure and evidence constraints; the LLM directly outputs a 0–10 score—measuring the contribution of evidence grounding.

All conditions share the same LLM, the same system prompt, and the same structured output format; the only variable is whether an approach taxonomy is provided and its source.

Question	# Disc.	Representative Families
q4	4	hessian_avg_contraction, taylor_hessian, ...
q6	2	gradient_space, fixed_point
q24	1	jensen_elbo_tightness
q29	4	entropy_ratio, kernel_triangle, gibbs_kl, ...

Table 1: Approach family discovery results on the 4 questions that yield non-standard approaches (q30 and q35 produce none within 48 feedback samples).

5.4 Evaluation Metrics

All 690 responses are scored by each automatic grader; TA feedback during HITL discovery is used *only* to induce approach families and does *not* replace automatic scores. We report **MAE** (primary; $\frac{1}{N}\sum|\hat{y}_i - y_i|$ on 0–10 normalized scores) and **QWK** (secondary; Quadratic Weighted Kappa). All comparisons use the same normalized labels; differences are confirmed via paired bootstrap tests ($p < 0.001$, 10,000 iterations).

The backend LLM is uniformly GPT-5.4 (OpenRouter API); full hyperparameters are in Appendix G.

6 Experimental Results

We organize experimental results around three research questions.

6.1 RQ1: Can Lightweight TA Feedback Discover Interpretable Approach Families?

Table 1 summarizes the approach family discovery results across all 6 questions. Under a maximum budget of 48 feedback samples per question (with early stopping, the effective average is about 12 per question), 4 out of 6 questions yield non-standard approach families, producing a total of 11 interpretable approach families.

The discovery results align closely with question characteristics (per-question approach diversity statistics are in Appendix J): questions with the highest approach diversity (q4: Hessian mean proof, q29: kernel metric properties) each yield 4 approach families; q24 discovers 1 significant approach family (Jensen inequality tightness), which produces the largest single-question scoring improvement (-8.9% MAE).

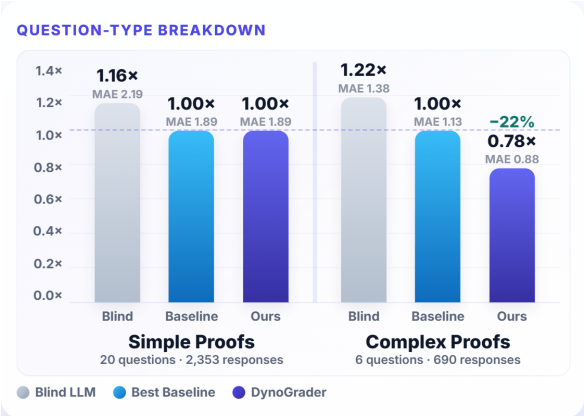

Figure 3: Baseline-normalized MAE by question type (strongest external baseline = 1.0×). Simple proofs: all methods are near parity; complex multi-approach proofs: DynoGrader drops to 0.78× (−22%).

On q6 (Newton convergence), the system discovers 2 approach families (gradient_space and fixed_point), matching the expert taxonomy. In a separate pilot study with a different discovery configuration, an additional variant (residualized-x-error) not covered by the expert taxonomy was also identified, suggesting that automatic discovery can capture approach variants beyond expert expectations when the search budget is spent in more rounds.

6.2 RQ2: Does Approach-Aware Scoring Improve Grading Quality?

Table 2 presents the main results. QWK results are consistent with MAE conclusions: DynoGrader's Expert (0.885), HITL (0.878), and ablation EGSS (0.873) all outperform the strongest external baseline LLM Rubric (0.858), and all improvements are statistically significant (paired bootstrap, 10,000 iterations, $p < 0.001$). Key findings are as follows:

Approach-aware scoring substantially outperforms all external baselines. DynoGrader with expert approach taxonomies reduces MAE from the strongest external baseline's 1.131 to 0.882 (−22.0%). Improvements are concentrated on questions with high approach diversity: q29 from 1.444 to 1.191 (−17.5%), q30 from 0.904 to 0.792 (−12.4%), q24 from 0.730 to 0.684 (−6.3%).

HITL discovery approaches or surpasses experts. Using automatically discovered approach families (without expert pre-definition), DynoGrader reduces MAE to 0.943 (−16.6%), with a gap of 6.9% to the expert-approach result of 0.882.

Method	MAE↓	QWK↑	vs. BL
External baselines (re-evaluated on same data)			
Blind	1.381	0.800	+22.1%
Direct+Ref	1.206	0.833	+6.6%
Batched Direct+Ref	1.377	0.757	+21.8%
LLM Rubric	1.131	0.858	*(best)*
Two-Pass Reflection	1.135	0.843	+0.4%
Debate	1.350	0.801	+19.4%
RefGrader	2.034	0.639	+79.8%
GradeHITL†	1.437	0.765	+27.1%
DynoGrader (ours)			
+ Expert Approaches	0.882***	0.885	**−22.0%**
+ HITL Discovery	0.943***	0.878	**−16.6%**
Ablations			
w/o Approaches (EGSS only)	0.969***	0.873	−14.3%
w/o Structure (Blackbox)	1.942	0.697	+71.7%

Table 2: MAE and QWK comparison on 6 complex proof questions (690 responses). All methods use GPT-5.4, evaluated on the same normalized labels. DynoGrader with expert approaches reduces MAE by 22.0% over the strongest external baseline; with HITL-discovered approach families, by 16.6%. ***: paired bootstrap test vs. LLM Rubric, $p < 0.001$ (10,000 iterations). †GradeHITL evaluated on 137 test-split samples (4:1 holdout); others on all 690.

On the two questions with the highest approach diversity, HITL discovery approaches or surpasses expert approaches: q24 discovered-approach MAE is 0.665, outperforming the expert's 0.684; q29 discovered-approach MAE is 1.198, close to the expert's 1.191. On other questions, discovered approaches occasionally cause slight degradation in the fixed evaluation pipeline (e.g., q4, q6), even though the discovery process itself shows MAE reduction within its own scorer (§6.3); this gap indicates that discovery quality depends on approach description precision, an area for future improvement.

Scaffold-heavy methods degrade under reasoning models. Batched Direct+Ref (MAE 1.377), Debate (1.350), GradeHITL (Li et al., 2025) (1.437†), and RefGrader (2.034) all add operational scaffolding—multi-pass averaging, multi-agent debate, iterative rubric refinement, or batch comparison—yet underperform simpler single-pass methods such as LLM Rubric (1.131). These mechanisms were designed for instruction-tuned models; reasoning models internalize deliberation through chain-of-thought, rendering external orchestration counterproductive.

Ablation: approach information is critical. Removing approach information (retaining only

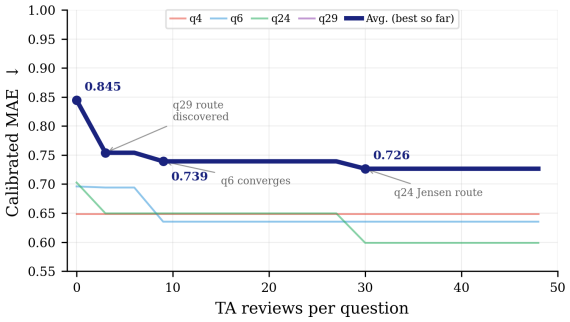

Figure 4: Best-so-far MAE vs. TA reviews per question (b=3) for the 4 questions with non-trivial approach discovery (q4, q6, q24, q29; 484 responses). Each point shows the minimum MAE achieved up to that review budget, reflecting practical early-stopping behavior. "0 reviews" baseline: EGSS without approaches (MAE = 0.845); cf. 6-question EGSS MAE = 0.969 in Table 2.

evidence-grounded structured scoring) increases MAE from 0.882 to 0.969, still outperforming all external baselines by 14.3%, showing that the evidence-grounded foundation is itself competitive. Further removing structural constraints (Blackbox) causes MAE to sharply deteriorate to 1.942, confirming the dual necessity of evidence grounding and approach awareness. In the expert-approach setting, 39% of responses are classified as `hybrid` (matching no single named approach), with higher MAE (3.35) than named-approach responses (2.15); this reflects the inherent difficulty of mixed-strategy answers under single-approach matching and represents a natural ceiling (Appendix B). Cross-model experiments with Claude Sonnet 4.6 and Gemini 3.1 Pro confirm that the same condition ranking holds across all three models (Appendix K.1).

6.3 RQ3: Budget Efficiency of HITL Discovery

Figure 4 tracks the best-so-far (early-stopping) MAE *within the discovery pipeline* on the 4 questions with non-trivial approach discovery (q4, q6, q24, q29; 484 responses), with batch size b=3 over 16 rounds. These within-discovery gains are not directly comparable to Table 2, which uses a fixed evaluation pipeline on all 6 questions (see Appendix A for reconciliation). q29 yields the largest single-round gain at just 3 reviews (-10.7% aggregate); q6 converges by 9 reviews (-12.5%); the elusive q24 Jensen approach is captured at 30 reviews, pushed into the review queue by disagreement-driven selection (-14.0%). Over 89% of the total improvement is realized within 9 reviews per question; with early stopping across all 6 questions the effective total is about 70 feedback samples (average $\approx$12 per question), and **total TA time is under 1 hour** ($\approx$10% of exhaustive grading). Compared against random selection under an equal budget, the submission informativeness score reduces MAE by 7.2% over random selection; the advantage concentrates on questions with rare non-standard approaches (q29: -23% vs. -1.8%). Per-question breakdowns and batch-sweep analysis appear in Appendices K.2–K.3.

Two cases stand out: q29 discovers 4 approach families and closes within 0.7% of the expert MAE; q24's Jensen/ELBO approach *outperforms* the expert taxonomy (-8.9%; traces in Appendix L).

7 Conclusion

We have presented DynoGrader, a human-in-the-loop grading system for multi-approach proof problems. Experiments validate three core findings: (1) approach-aware evidence-grounded scoring reduces MAE by 22.0% over the strongest external baseline with expert approaches, and by 16.6% with HITL-discovered families—on the highest-diversity questions, discovered families match or surpass expert-defined ones; (2) lightweight human feedback (on average 12 samples per question) suffices for effective discovery, and the discovery budget is shared with review; (3) the evidence-grounded foundation alone outperforms all external baselines by 14.3%. Cross-course experiments on an introductory STEM course (605 additional responses; Appendix N) confirm that EGSS generalizes beyond the primary setting, and that approach discovery improves scoring when genuine alternative strategies exist. Cross-model evaluation with Claude Sonnet 4.6 and Gemini 3.1 Pro (Appendix K.1) shows consistent condition rankings across all three models. Future directions include cross-question approach transfer and discovery under even less feedback.

Limitations and Ethics

Limitations

1. **Benefits concentrate on high-diversity proof questions.** On questions with uniform student approaches, approach discovery yields no non-standard approaches (2 of 6 questions in our experiments). This is a natural applicability boundary.
2. **Discovery quality varies and depends on the LLM.** On some questions (e.g., q4, q6), discovered approaches slightly degrade accuracy, indicating that approach description precision needs improvement. The current discovery protocol uses only positive examples (non-standard-approach flags); incorporating explicit negative constraints may reduce over-fragmentation. Approach matching produces no uncertainty estimate; adding confidence thresholds or second-pass checks could help near approach boundaries. Different backbone models may produce families and scores of varying quality; cross-model experiments (Appendix K.1) show consistent improvements but broader coverage remains future work.
3. **Cross-year pooling.** Responses from 2022–2024 cohorts are pooled per question. Since the grader scores each response independently and approach families are structural descriptions, cross-year leakage is unlikely, but evaluation on a held-out future cohort is left to future work.
4. **Experimental scale and annotator consistency.** The main experiments use 6 questions (690 responses) from one course, with cross-course expansion on a second course (605 responses; Appendix N). The binary TA judgment uses a single annotator per question; inter-annotator agreement analysis is left to future work.

Ethical Considerations. The system is positioned as a TA support tool, not a grading replacement; all approach judgments, evidence quotes, and review recommendations are traceable. Student data has been anonymized per institutional guidelines.

References

Derek Armfield and 1 others. 2025. Avalon: A human-in-the-loop LLM grading system with instructor calibration and student self-assessment. In *Proceedings of AIED*.

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2020. Deep batch active learning by diverse, uncertain gradient lower bounds. In *Proceedings of ICLR*.

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.

Yucheng Chu and 1 others. 2025. A LLM-powered automatic grading framework with human-level guidelines optimization. In *Proceedings of EDM*.

Leonardo de Moura and Sebastian Ullrich. 2021. The Lean 4 theorem prover and programming language. In *Proceedings of CADE*.

Tania Amanda Nkoyo Frederick Eneye, Chukwuebuka Fortunate Ijezue, Ahmad Imam Amjad, Maaz Amjad, Sabur Butt, and Gerardo Castañeda-Garza. 2025. Advances in auto-grading with large language models: A cross-disciplinary survey. In *Proceedings of the BEA Workshop at ACL*.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of EMNLP*.

Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*.

Albert Qiaochu Jiang, Sean Welleck, Jin Peng Zhou, Wenda Li, Jiacheng Liu, Mateja Jamnik, Timothée Lacroix, Yuhuai Wu, and Guillaume Lample. 2023. Draft, sketch, and prove: Guiding formal theorem provers with informal proofs. In *Proceedings of ICLR*.

Zixuan Ke and Vincent Ng. 2019. Automated essay scoring: A survey of the state of the art. In *Proceedings of IJCAI*.

Hang Li, Yucheng Chu, Kaiqi Yang, Yasemin Copur-Gencturk, and Jiliang Tang. 2025. LLM-based automated grading with human-in-the-loop. In *arXiv preprint arXiv:2504.05239*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of EMNLP*.

Wenjie Ma, Andrei Cojocaru, Neel Kolhe, Haihan Zhang, Vincent Zhuang, Matei Zaharia, and Sewon Min. 2026. Reliable fine-grained evaluation of natural language math proofs. In *Proceedings of ICLR*.

Hamed Mahdavi, Pouria Mahdavinia, Samira Malek, Pegah Mohammadipour, Alireza Hashemi, Majid Daliri, Alireza Farhadi, Amir Khasahmadi, Niloofar Mireshghallah, and Vasant Honavar. 2025. RefGrader: Automated grading of mathematical competition proofs using agentic workflows. In *MATH-AI Workshop at NeurIPS*.

Hussein Mozannar and David Sontag. 2020. Consistent estimators for learning to defer to an expert. In *Proceedings of ICML*.

Sagnik Mukherjee, Abhinav Chinta, Takyoung Kim, Tarun Anoop Sharma, and Dilek Hakkani-Tür. 2025. Premise-augmented reasoning chains improve error identification in math reasoning with LLMs. In *Proceedings of ICML*.

Tobias Nipkow, Lawrence C. Paulson, and Markus Wenzel. 2002. *Isabelle/HOL: A Proof Assistant for Higher-Order Logic*. Springer.

Chau Minh Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2024. TopicGPT: A prompt-based topic modeling framework. In *Proceedings of NAACL*.

Seth Poulsen, Mahesh Viswanathan, Geoffrey L. Herman, and Matthew West. 2022. Proof blocks: Autogradable scaffolding activities for learning to write proofs. In *Proceedings of ITiCSE*.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Saroha Tomar, Iulia Tsesarskaya, and Avinash Vasudevan. 2023. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4).

Burr Settles. 2009. *Active Learning Literature Survey*. University of Wisconsin–Madison.

Tianjun Wei, Wei Wen, Ruizhi Qiao, Xing Sun, and Jianghong Ma. 2025. RocketEval: Efficient automated LLM evaluation via grading checklist. In *Proceedings of ICLR*.

Yuhuai Wu, Albert Q. Jiang, Wenda Li, Markus N. Rabe, Charles Staats, Mateja Jamnik, and Christian Szegedy. 2022. Autoformalization with large language models. In *Proceedings of NeurIPS*.

Changrong Xiao and 1 others. 2025. Human-AI collaborative essay scoring: A dual-process framework with LLMs. In *Proceedings of the Conference on Learning Analytics and Knowledge*.

A Evaluation Protocol Details

HITL evaluation is fully automatic. In the HITL Discovery setting, TA feedback (total score and binary non-standard-approach flag) is collected for a small subset of responses during the discovery phase. This feedback is used *exclusively* to induce approach families; it does *not* replace the automatic grader's predictions. After discovery, *all* 690 responses are re-scored by the automatic grader using the discovered approach taxonomy, and MAE is computed between these automatic predictions and the human ground truth. Reviewed samples are treated identically to unreviewed samples during evaluation. This ensures a fair comparison with non-HITL baselines, which also receive no TA input at evaluation time.

Reconciliation of Figure 4 and Table 2. The effort curve (Figure 4) reports *best-so-far* (running-minimum) MAE on the 4-question subset with non-trivial approach discovery (q4, q6, q24, q29; 484 responses); the "0 reviews" baseline is EGSS on this 4-question subset (MAE $= 0.845$). At each review budget the plotted value is the minimum MAE observed across all rounds up to that point, reflecting practical early-stopping behavior. A dashed gray line shows the raw per-round aggregate for comparison; the divergence at higher budgets illustrates the approach-name-drift phenomenon analyzed in Appendix K.5. Table 2 reports MAE on all 6 questions (690 responses), where EGSS ("w/o Approaches") achieves MAE $= 0.969$. The baseline difference (0.845 vs. 0.969) is due to the question sets: q30 and q35 have no non-standard approaches and relatively high MAE, raising the 6-question average. Note that all cross-condition comparisons in Table 2 use a single, fixed evaluation pipeline; the effort curve measures *within-discovery* progress under the same scorer across rounds.

B Approach Matching and Scoring Details

Approach matching algorithm. Approach matching is performed *within* the same LLM call that produces the score—there is no separate classifier or embedding-based matcher. The prompt includes the full approach taxonomy (approach IDs, natural-language descriptions, and per-slot guidance), and the model is instructed to select the approach that best matches the student's reasoning strategy. The tool schema constrains the `approach_id` field to an enum of valid approach IDs, ensuring that the model must commit to exactly one approach per response.

Fallback and hybrid answers. A `standard_path` approach is always included in the taxonomy. If the student's approach does not match any discovered approach, the model selects `standard_path` and scores purely from the canonical slot rubric (equivalent to EGSS without approaches). There is no explicit confidence threshold or tie-breaking rule: the model's single-call approach selection is treated as deterministic at temperature 0. When a student employs a

hybrid approach that borrows from multiple approaches, the model must commit to a single best-matching approach. In the expert-approach experiments, 39% of responses are classified as `hybrid` (the most frequent label), indicating that single-approach commitment is a common case. We observe that hybrid-classified samples have moderately lower support density (0.50 vs. 0.77 for named approaches) and higher MAE (3.35 vs. 2.15), confirming that hybrid answers are inherently harder to score and represent a natural ceiling for single-approach matching. Multi-approach scoring with weighted slot guidance is left to future work.

Approach classification quality. Because our dataset does not include human-annotated approach labels per student answer, we cannot report approach-classification accuracy directly. As a proxy, we analyze support-pattern density: each approach definition includes support patterns (key proof steps) that should be present in correctly classified answers. Among 569 expert-approach predictions, named approaches exhibit high support density (0.49–0.78), and zero samples show a named approach with zero pattern support or `unknown` classification despite high pattern matches, suggesting that approach selection is well-calibrated. Per-student approach annotation and approach-vs-scoring error decomposition are left to future work.

Slot construction. Rubric slots are derived from the reference answer in a hybrid process: the expert first specifies the high-level rubric structure (number of slots, point allocation), and then an LLM generates detailed slot descriptions and evidence criteria from the reference answer text. In our experiments, the initial slot schema for each question takes approximately 10–15 minutes of expert time (reading the reference answer, deciding on 4–5 slots, and assigning point weights). Slot definitions are shared across all approaches for a given question; approach-specific *slot guidance* (e.g., "Credit the student for using the kernel triangle inequality directly") is appended to each slot when a non-standard approach is active.

C Approach Taxonomy Example

Table 3 shows the expert-defined approach taxonomy for q6 (Newton local quadratic convergence), the running example in Figure 1. The taxonomy contains three approaches beyond the canonical `standard_path`; each approach carries a natural-language description and per-slot scoring guidance that tells the LLM how to interpret evidence under that proof approach. All approaches share the same five rubric slots; only the guidance column differs.

D Novelty Computation in Sample Selection

The "novelty" signal in the four-way sample selection measures how dissimilar a candidate response is from the already-reviewed set. We compute novelty as follows:

1. Each student response is represented by a bag-of-words TF-IDF vector over the extracted response text.
2. For each candidate response x_i not yet reviewed, we compute the maximum cosine similarity to any already-reviewed response: $\text{sim}(x_i) = \max_{x_j \in \mathcal{R}} \cos(x_i, x_j)$, where $\mathcal{R}$ is the reviewed set.
3. Responses are ranked by ascending similarity (lowest similarity = highest novelty).
4. The top-k most novel responses are selected for the "novelty" quota in the four-way triage.

We chose TF-IDF over learned embeddings (e.g., SBERT) for simplicity and reproducibility; a comparison with semantic similarity metrics is left to future work.

E Evidence Faithfulness

To assess whether evidence quotes produced by EGSS are grounded in the student's actual text rather than hallucinated, we audit 2,646 non-null slot assessments from the expert-approach GPT-5.4 run (569 predictions across 5 questions).

Automated verification. We apply both exact substring matching and fuzzy word-overlap matching (threshold $\geq$ 70%) between each `evidence_quote` and the student response. 41.4% of quotes are automatically verified. The remaining 58.6% are flagged as "misses," overwhelmingly because student responses are OCR-parsed with formatting artifacts (line breaks inside equations, missing operators, notation differences), while the model normalizes notation when quoting.

Manual spot check. We randomly sample 50 slot assessments from the automated-miss pool and classify each as: *faithful* (content clearly present in the answer), *minor reformulation* (same mathematical content, notation cleaned), *hallucinated* (con-

Approach	Description	Slot	Approach-Specific Guidance
error_space	Derives a recursion directly for $x_{k+1} - x^*$ or the Newton remainder in x-space.	newton_error_identity	Set up the error-space Newton identity or recursion explicitly.
		local_hessian_exp.	Use Taylor or integral remainder around x^* to express the nonlinear residual.
		lipschitz_quad_bound	Use Hessian Lipschitzness to bound the remainder quadratically.
gradient_space	Uses gradient cancellation or a Jacobian/inverse-Hessian identity before transferring back to x-space.	newton_error_identity	State the gradient-space cancellation identity or equivalent Newton relation.
		inverse_hessian_ctrl	Transfer the gradient-space bound back to x-space with inverse-Hessian control.
fixed_point	Treats Newton as a fixed-point map $g(x) = x - [\nabla^2 f(x)]^{-1} \nabla f(x)$ and proves local quadratic contraction.	newton_error_identity	Define the Newton map or fixed-point relation explicitly.
		local_hessian_exp.	Express second-order local behavior of the Newton map near x^*.
		lipschitz_quad_bound	Show that the fixed-point residual is quadratic near x^*.

Table 3: Expert-defined approach taxonomy for q6 (Newton convergence). Three approaches are shown; each shares the same 5 rubric slots but provides approach-specific guidance for the scorer. Only representative slot–guidance pairs are listed for brevity; the full taxonomy covers all 5 slots per approach.

tent not in the answer), or *unclear*. Results: 25/50 faithful, 16/50 minor reformulation, 4/50 hallucinated, 5/50 unclear. 82% of automated misses are in fact content-faithful, with the gap attributable to OCR-to-clean-notation normalization. The 4 "hallucinated" cases are all meta-descriptions of answer structure (e.g., "the response is split into Subquestion 1–4") for coherence-assessment slots, not fabricated mathematical content.

Overall estimate. Combining automated verification (41.4%) with the manual audit (82% of misses are faithful), the estimated overall evidence faithfulness is approximately **89%**, with a true content-hallucination rate below **5%**. A more rigorous evaluation with human annotators judging semantic equivalence on a larger sample is left to future work.

F Baseline Implementation Details

RefGrader. We faithfully re-implement RefGrader's core workflow: (1) generate a question-level marking scheme from the reference answer using the original `marking_scheme` prompt template; (2) score each student response using the `with_reference_solution_and_marking_scheme_flexible` template with five independent scoring passes; (3) aggregate by median and linearly map the internal 0–7 score onto each question's maximum score. The marking scheme is generated once per question and reused across all responses. All prompts are sourced from the RefGrader codebase; we verified that our re-implementation produces structurally identical outputs on manually inspected samples. The large MAE gap (2.034 vs. 1.131 for LLM Rubric) is primarily due to RefGrader's fixed 0–7 internal scale, which loses granularity when mapped to variable-maximum-score coursework problems, and its competition-oriented marking scheme style, which does not align well with partial-credit coursework rubrics. To rule out the possibility that this scale mismatch unfairly disadvantages RefGrader, we re-ran the full pipeline with all templates modified to use a native 0–10 scoring scale (matching our evaluation range). The resulting MAE was 2.299, *higher* than the original 0–7 result (2.034), indicating that the coarser 7-point rubric actually helps the LLM produce more consistent judgments. We therefore report the more favorable 0–7 configuration.

Batched Direct+Ref. This variant of Direct+Ref groups up to 16 student answers for the same question into a single prompt, so the model can see multiple submissions side-by-side before assigning scores. The motivation is to improve inter-answer consistency; however, the extra context appears to dilute per-answer attention: MAE rises from 1.206 (single-answer Direct+Ref) to 1.377. Batch boundaries are strictly per-question (no cross-question batches), and output is extracted via a structured

tool call identical to the single-answer baselines. The same GPT-5.4 backbone is used.

GradeHITL. We implement a local reproduction of GradeHITL (Li et al., 2025), following its Grader–Inquirer–Reflector–Refiner pipeline with N=1, T=5, W=3, K=4 optimization layers. The grader is GPT-5.4 (same as all other baselines); an AI model (Claude Opus 4.6) serves as the TA proxy, answering the inquirer's clarification questions. For each of the 6 questions, we use a 4:1 per-question train/test split; the training partition (553 samples) is used for iterative guideline optimization, and evaluation is performed only on the held-out test partition (137 samples). Because the original codebase is not publicly available (the anonymous link returns `410 Gone`), prompts and the retriever are re-implemented from the paper description; the RL-based selector is replaced with a similarity + validation-gain heuristic.

Other baselines. Direct+Ref, Blind, LLM Rubric, Two-Pass Reflection, and Debate all use the same GPT-5.4 backbone, the same 0–10 normalized output scale, and the same structured tool-call interface. Prompt templates for all baselines are included in the released code repository.

G Implementation Hyperparameters

Discovery hyperparameters. Per-round budget $b = 12$, total rounds $R = 4$ (maximum 48 feedback samples per question). Four-way selection ratios: 3:3:3:3 (first round) and 3:4:3:2 (subsequent rounds), corresponding to the low-score, disagreement, novelty, and mid-range novelty pools respectively. Minimum positive-example support for an approach family: ≥ 2.

Early stopping. The discovery process terminates early when three consecutive rounds produce no new approach families *and* no MAE improvement. With early stopping, the effective budget across all 6 questions is about 70 samples (average $\approx$12 per question).

LLM configuration. All main scoring and induction calls use GPT-5.4 via the OpenRouter API at temperature 0, with structured tool-call output (`strict = true`). Cross-model experiments (Appendix K.1) additionally use Claude Sonnet 4.6 (tool calling with `tool_choice = required`) and Gemini 3.1 Pro (JSON mode with the schema embedded in the system prompt).

H Prompt Templates

Below we provide the key prompt templates used in DynoGrader. Full prompt templates, slot schemas, and approach discovery instructions are available in the released code repository.

EGSS scoring prompt (abbreviated).

```
You are a precise, evidence-grounded
grader.
For each rubric slot, you MUST:
(1) quote the specific text from the
student's answer;
(2) judge whether the evidence covers
the slot requirement;
(3) assign a score with a brief
rationale.
Use ONLY the student's own words as
evidence.
```

Approach-aware extension (abbreviated).

```
The following approach taxonomy
describes known proof approaches.
Pick the approach that best matches the
student's reasoning.
If none match, use 'standard_path' and
score from the canonical rubric.
For the matched approach, apply the
approach-specific slot guidance below.
```

Approach family induction prompt (abbreviated).

```
You are a proof taxonomy analyst.
Given these student responses flagged as
using non-standard approaches,
identify up to {max_families} distinct
proof approach families.
For each family, provide: family_id,
description, slot_guidance,
and the IDs of supporting examples.
You may keep, split, merge, or rewrite
families from the existing
taxonomy, but the output must contain at
most {max_families} families.
```

I Approach Discovery Instructions for TAs

The TA feedback protocol requires two inputs per reviewed response:

1. **Total score**: An integer score on the question's original scale (e.g., 0–10), reflecting the TA's holistic assessment of the response's correctness and completeness.
2. **Non-standard approach flag**: A binary judgment (yes/no) indicating whether the student's proof approach differs meaningfully from the reference solution. "Non-standard" means the student uses a different mathematical approach, not merely different notation or phrasing.

Average TA time per response is approximately 45 seconds (reading, scoring, and binary judgment). No approach labeling or taxonomy knowledge is required from the TA.

J Per-Question Approach Diversity

Table 4 summarizes the approach diversity of each main-evaluation question. Expert approach counts exclude the catch-all `unknown` and `hybrid` categories; discovered counts reflect the budget-48 configuration used in the main evaluation.

Question	N	Expert	Discovered
q4 (convergence rate)	122	3	4
q6 (Newton convergence)	122	2	2
q24 (variational EM)	121	3	1
q29 (kernel MMD)	119	4	4
q30 (Wasserstein)	119	2	0
q35 (ELBO derivation)	87	2	0

Table 4: Per-question statistics. N: number of student responses. *Expert*: number of substantive expert-defined approaches (excluding `unknown`/`hybrid`). *Discovered*: number of approach families induced by HITL discovery. Questions with 0 discovered families (q30, q35) have highly uniform student approaches; approach discovery correctly produces no non-standard approaches.

K Approach Discovery Rigor Verification

We conduct four experiments to verify that the HITL approach discovery pipeline produces robust, interpretable, and model-agnostic results.

K.1 Cross-Model Scoring Consistency

To test whether the benefit of approach-aware grading is specific to the primary model (GPT-5.4), we re-score all 690 responses with Claude Sonnet 4.6 and Gemini 3.1 Pro under the same four conditions used in the main evaluation. Table 5 shows the results.

All three models exhibit the same ranking over conditions, confirming that the improvement direction is model-agnostic. Key findings: (1) **EGSS improves all models over Blackbox**: GPT -50.1%, Claude -36.9%, Gemini -8.6%. (2) **Expert approaches further improve all models**: GPT -9.0%, Claude -8.2%, Gemini -5.8%. (3) GPT-5.4 benefits most from structured scoring; Gemini's blackbox baseline is already the strongest (1.785), but it gains less from EGSS structure, suggesting that Gemini's free-form output already captures some implicit structure. (4) The absolute MAE gap between models (GPT best, Gemini worst for structured scoring) reflects differences in instruction-following fidelity for complex slot schemas, not in the method's generality.

K.2 Cross-Configuration Stability

We vary the batch-size/rounds trade-off under a fixed review budget of 48 reviews per question, comparing three configurations: 12×4, 6×8, and 4×12. Table 6 reports results on all three questions with non-trivial discovery (q4, q24, q29).

Key findings: (1) For q4, discovered approaches do not improve over canonical; this question already has low baseline error and limited approach diversity. (2) For q24, a single discovered approach family suffices, reducing MAE by up to 20%. (3) For q29, the most complex question, smaller batches with more rounds (4×12) achieve the largest peak improvement (-38%), suggesting that iterative refinement is valuable when the approach landscape is rich. Final MAE standard deviations across configurations are small (0.007–0.124), indicating that the pipeline is not highly sensitive to the batch-size choice.

K.3 Sampling Strategy Ablation

We compare the four-way intelligent selection strategy (low score + disagreement + novelty + mid-range novelty) against uniform random selection with an equal budget (48 total samples per question). With early stopping (taking each question's best round), per-question results are:

The four-way strategy's advantage is concentrated on questions where alternative approaches are rare: on q29, four-way achieves 1.030 (-23%) while random achieves only 1.318 (-1.8%); on q6, four-way improves by 8.7% while random actually degrades MAE. When alternative approaches are common (q24), random selection discovers effectively, even outperforming four-way (0.507 vs. 0.599).

K.4 Human Quality Rating of Discovered Families

A domain expert rated each of the 11 discovered approach families on three dimensions (1–5 scale): **Interpretability** (is the description clear?), **Distinctiveness** (is it genuinely different from other families?), and **Grading Utility** (does it help the scorer assign more accurate partial credit?).

The grand mean of 3.94/5 indicates that the LLM-induced families are generally interpretable

	GPT-5.4		Claude 4.6		Gemini 3.1	
Condition	MAE	Δ	MAE	Δ	MAE	Δ
Blackbox	1.942	—	2.058	—	1.785	—
EGSS (canonical)	0.969	−50.1%	1.299	−36.9%	1.632	−8.6%
+ Expert approaches	0.882	−9.0%	1.192	−8.2%	1.537	−5.8%
+ Discovered approaches	0.943	−2.7%	1.283	−1.2%	1.601	−1.9%

Table 5: Cross-model scoring results (MAE). Δ is relative change versus the row above. All three models share the same condition ranking: expert < discovered < EGSS < blackbox. Gemini uses JSON mode via OpenRouter; about 5–8% of samples fail JSON parsing and are excluded (coverage 92–98%).

Q	Config	Best	Final	Names
4	12×4	0.649	0.786	9
	6×8	0.649	0.800	15
	4×12	0.649	0.794	15
	Canonical: 0.649		Final MAE σ: 0.007	
24	12×4	0.564	0.564	1
	6×8	0.579	0.624	3
	4×12	0.599	0.599	1
	Canonical: 0.703		Final MAE σ: 0.030	
29	12×4	1.093	1.232	12
	6×8	1.030	1.254	5
	4×12	0.833	1.029	6
	Canonical: 1.342		Final MAE σ: 0.124	

Table 6: Batch sweep results under fixed budget (48 reviews). "Best" is the lowest MAE observed during discovery; "Final" is the end-of-budget MAE; "Names" is the total number of unique family identifiers generated.

	4-Way		Random	
Q	MAE	Δ	MAE	Δ
q4	0.649	±0.0%	0.649	±0.0%
q6	0.435	−8.7%	0.478	+0.4%
q24	0.599	−14.8%	0.507	−27.9%
q29	1.030	−23.3%	1.318	−1.8%
Mean	0.753	—	0.811	—

Table 7: Four-way vs. random sampling under equal 48-sample budget (best-round MAE with early stopping). Δ is relative to each question's canonical EGSS baseline.

and useful. Seven families (64%) reach the "high quality" tier (avg ≥ 4.0), while the two lowest-scoring families (kernel Minkowski Riemann-sum variant and reference-anchored kernel) could be merged into their parent families with minimal information loss.

We also assessed alignment with the expert-authored approach taxonomy: 3 families are exact matches, 3 are strong matches, 1 moderate, 2 partial, 1 variant of an existing family, and 1 genuinely novel (the Jensen–ELBO tightness approach for

Q	Family	I	D	U	Avg
4	hessian_avg_op	4	5	4	4.3
4	interp_ineq_dist	4	4	4	4.0
4	func_gap_sandwich	3	3	3	3.0
4	func_val_linear_rate	5	4	5	4.7
6	grad_space_remainder	5	5	5	5.0
6	newton_fixed_pt	4	5	4	4.3
24	jensen_elbo_tight	5	5	5	5.0
29	entropy_kernel_mink	3	4	4	3.7
29	kernel_mink_triangle	4	4	4	4.0
29	kernel_mink_riemann	3	2	3	2.7
29	ref_anchored_kernel	4	2	2	2.7
Mean		**4.00**	**3.91**	**3.91**	**3.94**

Table 8: Human quality ratings for 11 discovered approach families. I = Interpretability, D = Distinctiveness, U = Grading Utility (each 1–5). 7/11 families score ≥ 4.0 ("high quality"); 2/11 fall below 3.0 and are candidates for merging or pruning.

q24, which the expert had not anticipated). Overall, 55% of discovered families have strong-or-better alignment with expert approaches, while also surfacing approaches that experts missed.

K.5 Approach Name Drift

Because each discovery round re-induces families from scratch, semantically identical families may receive different `family_id` strings across rounds. We quantify this drift on q4 (batch size 3, 16 rounds).

Over the 14 active rounds (rounds with ≥1 discovered family), the pipeline generates 18 unique family identifiers that cluster into only 4 semantic concepts: *distance contraction* (10 name variants), *function gap* (3), *Hessian* (2), and *sandwich* (3). The name instability ratio is $18/4 = 4.5\times$, and 11 of 14 active rounds exhibit at least one rename. However, the underlying semantic taxonomy stabilizes by round 11 (three consecutive stable rounds before a genuinely new family emerges at round 14).

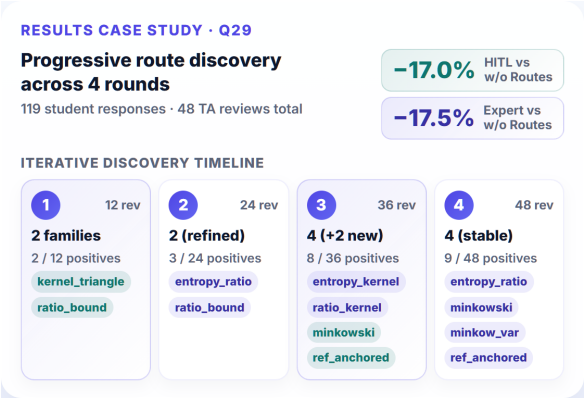

Figure 5: q29 case study: progressive discovery over 4 rounds (4×12 TA samples) yields 4 stable approach families. MAE: $1.444 \rightarrow 1.198$ (-17.0%), near expert (1.191).

Table 6 shows that larger batches reduce name proliferation in q29 ($12 \rightarrow 5$ unique names when moving from 12×4 to 6×8), though the effect is question-dependent. Name drift is a cosmetic issue—the same semantic clusters are recovered regardless of naming—and can be mitigated by post-hoc deduplication via semantic clustering, which we leave to future work.

L Case Study: q29 Kernel Metric Properties

This question asks students to prove a property of the metric space induced by a kernel function. Students employ at least 4 different proof approaches (Figure 5): (1) direct application of the kernel triangle inequality, (2) construction of an entropy-ratio argument, (3) Minkowski-based reduction, and (4) the reference-student approach via feature-space embedding. Without approach information, the EGSS scorer achieves MAE 1.444 because it covers only the canonical slot decomposition and systematically under-scores non-standard approaches. After HITL discovery of 4 approach families, MAE drops to 1.198 (-17.0%), approaching the expert-defined approaches at 1.191.

M Case Study: q24 ELBO Tightness

The majority of student answers to q24 follow a single main approach, but an alternative approach exists that bypasses the standard derivation via the Jensen inequality tightness condition. HITL discovery identifies this approach family (jensen_elbo_tightness) in round 2, reducing MAE from 0.730 to 0.665 (-8.9%), outperforming the expert-defined approach family at 0.684. This demonstrates the advantage of data-driven discovery in covering approach variants unforeseen by experts. The expert taxonomy contained only the canonical KL-divergence solution approach; the Jensen tightness approach—which argues directly that the bound is tight when $q = p$ and therefore ELBO equals the log-evidence—was not anticipated.

N Cross-Course Generalization

To test whether DynoGrader generalizes beyond the advanced analytical questions in Course A, we apply the full pipeline (EGSS scoring + HITL approach discovery) to 6 short-answer questions from Course B (an introductory STEM course) covering topics such as fault tolerance, quantum entanglement, convolution proofs, and game theory (605 total responses).

N.1 Screening

Table 9 compares blackbox and EGSS canonical MAE on all 6 candidates. EGSS reduces MAE over blackbox on every question, with the largest gains on proof-oriented questions (q12 Byzantine generals: -33.6%, q40 convolution identities: -17.8%).

Q	Topic	n	BB	EGSS	Δ
12	Byzantine gen.	101	3.650	2.423	−33.6%
18	Boolean algebra	102	6.419	5.238	−18.4%
27	Bell state	101	3.311	3.027	−8.6%
40	Convol. proofs	101	2.924	2.402	−17.8%
52	Game theory	99	5.587	5.284	−5.4%
58	Spring Euler	101	6.539	6.409	−2.0%
Mean (6 Qs)		605	4.738	4.131	−12.8%

Table 9: Course B screening: blackbox vs. EGSS canonical MAE. Slot schemas were auto-generated by the LLM from each question's reference answer. Gold labels are raw TA scores.

N.2 Approach Discovery

We ran HITL approach discovery (6 rounds, batch size 8, budget 48 reviews) on the 4 questions with the strongest proof structure. Table 10 shows the results.

Approach discovery improved scoring on 2 of 4 tested questions: **q27** (Bell-state non-separability, -8.0%) and **q52** (game-theory proofs, -6.2%). For q27, the discovered approach captures an alternative proof approach using the separability invariant $ad - bc \neq 0$, rather than the reference answer's

Q	Topic	EGSS	+App.	Δ	Fam.
12	Byzantine gen.	2.423	2.451	+1.2%	4
27	Bell state	3.027	**2.784**	**−8.0%**	1
40	Convol. proofs	2.402	—	—	0
52	Game theory	5.284	**4.956**	**−6.2%**	4

Table 10: Course B approach discovery results. Bold entries indicate improvement over EGSS. "—" means no non-standard approaches were discovered.

direct coefficient-matching contradiction. For q52, the discovered families capture systematic error patterns in how students confuse Nash equilibria with social optima in the Prisoner's Dilemma subquestion, enabling more precise partial-credit assignment.

q40 (convolution identities) yielded no non-standard approaches—the change-of-variable proof is essentially the only strategy used by students. For q12, approaches were discovered but did not improve MAE (+1.2%), suggesting that the discovered families (four distinct counterexample strategies for the Byzantine generals problem) were too fine-grained to benefit the scorer.

Discussion. Compared to the Course A analysis questions where expert approaches reduce MAE by up to 9%, the Course B benefit is smaller and more selective. This is consistent with the simpler mathematical structure of introductory-level problems: fewer questions admit multiple valid proof strategies, and the primary bottleneck is evidence-grounded slot scoring (which EGSS already addresses) rather than multi-approach recognition. Nevertheless, the pipeline discovers genuine alternative approaches when they exist (q27) and improves scoring on questions with systematic error patterns (q52), confirming that DynoGrader generalizes beyond its primary domain.

O Data and Code Release

We will release the following artifacts upon publication:

- **DynoBench dataset**: The full evaluation dataset (de-identified student responses, normalized scores, and expert approach taxonomies) on the Hugging Face Hub at `dynograder/dynobench`.
- **Code**: The complete DynoGrader system (scorer, approach discovery pipeline, sample selection, evaluation scripts, and all baseline implementations) on GitHub.
- **Prompts**: All prompt templates, slot schemas, and approach discovery instructions as structured configuration files.