From Prediction to Precision: Leveraging LLMs for Equitable and Data-Driven Writing Placement in Developmental Education

Miguel Da Corte
University of Algarve, Campus de Gambelas, Faro, Portugal
INESC-ID Lisboa, Portugal

Jorge Baptista
University of Algarve, Campus de Gambelas, Faro, Portugal
INESC-ID Lisboa, Portugal

Abstract

Accurate text classification and placement remain challenges in U.S. higher education, with traditional automated systems like ACCUPLACER functioning as "black-box" models with limited assessment transparency. This study evaluates Large Language Models (LLMs) as complementary placement tools by comparing their classification performance against a human-rated gold standard and ACCUPLACER. A 450-essay corpus was classified using Claude, Gemini, GPT-3.5-turbo, and GPT-4o across four prompting strategies: Zero-shot, Few-shot, Enhanced, and Enhanced+ (definitions with examples). Two classification approaches were tested: (i) a 1-step, 3 class classification task, distinguishing DevEd Level 1, DevEd Level 2, and College-level texts in one single run; and (ii) a 2-step classification task, first separating College vs. Non-College texts before further classifying Non-College texts into DevEd sublevels. The results show that structured prompt refinement improves the precision of LLMs' classification, with Claude Enhanced + achieving 62.22% precision (1 step) and Gemini Enhanced + reaching 69.33% (2 step), both surpassing ACCUPLACER (58.22%). Gemini and Claude also demonstrated strong correlation with human ratings, with Claude achieving the highest Pearson scores ($\rho = 0.75$; 1-step, $\rho = 0.73$; 2-step) vs. ACCUPLACER ($\rho = 0.67$). While LLMs show promise for DevEd placement, their precision remains a work in progress, highlighting the need for further refinement and safeguards to ensure ethical and equitable placement.

2012 ACM Subject Classification Social and professional topics → Adult education; Social and professional topics → Student assessment; Computing methodologies → Information extraction; Computing methodologies → Lexical semantics; Computing methodologies → Language resources; Computing methodologies → Natural language generation

Keywords and phrases Large Language Models (LLMs), Developmental Education (DevEd), writing assessment, text classification, English writing proficiency

Digital Object Identifier 10.4230/OASIcs.SLATE.2025.1

Supplementary Material *Software*: `https://gitlab.hlt.inesc-id.pt/u000803/deved`

Funding This work was supported by Portuguese national funds through FCT (Reference: UIDB/50021/2020, DOI: 10.54499/UIDB/50021/2020) and by the European Commission (Project: iRead4Skills, Grant number: 1010094837, Topic: HORIZON-CL2-2022-TRANSFORMATIONS-01-07, DOI: 10.3030/101094837).

1 Introduction and Objectives

Higher education institutions play a critical role in developing students' academic skills and ensuring equitable access to educational opportunities, particularly for those whose writing and reading abilities do not yet meet college-level standards. In the United States of America, this support is provided through a foundational literacy program known as Developmental

14th Symposium on Languages, Applications and Technologies (SLATE 2025).
Editors: Jorge Baptista and José Barateiro; Article No. 1; pp. 1:1–1:18
OpenAccess Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Education (DevEd). Placement in DevEd remains a persistent challenge, with traditional tools like ACCUPLACER[1] operating as "black box" automated systems. These assessments rely on machine-scored essays, emphasizing surface-level features (e.g., *word count*, *syntactic complexity*, *cohesion*) while offering limited insight into students' true writing proficiency. As a result, misplacement is common and could hinder student success [12] and progression in an academic program. In this paper, we focus on the writing skills of college-intending students at the onset of their academic career.

To address these limitations, institutions are exploring more transparent and precise assessment and course placement methods. LLMs have emerged as promising tools for placement and literacy instruction due to their writing assistance capabilities and potential for automated feedback [1, 33]. However, their reliability in assessing writing proficiency and alignment with institutional standards requires further exploration [3].

This study examines the role of LLMs in writing assessment, placement, and feature generalization within a DevEd setting. Building on previous work [10, 14], it expands an annotated corpus of 300 essays classified as Non-college level (DevEd Level 1 and DevEd Level 2) by both ACCUPLACER and human raters, adding 150 College-Level essays that were classified through the same procedures. The inclusion of College-Level texts enables a more comprehensive evaluation of LLM classification across these three proficiency levels. All writing samples were produced in English, which is the main language of instruction at the institution under study and the primary language assessed by ACCUPLACER.

In this study, we aim to:

(i) evaluate LLMs for DevEd placement by comparing their classifications to ACCUPLACER and a human-rated gold standard;
(ii) analyze the impact of a multi-step classification strategy on LLM precision across different prompting framework; and
(iii) determine whether prompt refinement mitigates classification bias, improving consistency and equity in placement outcomes.

To achieve these objectives, we propose the following research questions:

(i) How do LLM classifications compare to ACCUPLACER and human ratings in predicting placement outcomes?
(ii) What is the impact of a multi-step classification strategy on LLM precision across different prompting frameworks?
(iii) Can prompt refinement mitigate bias in LLM classification, enhancing placement accuracy and consistency across proficiency levels?

This research contributes to data-driven, equitable, and transparent DevEd placement by integrating LLMs with human-annotated corpora. By addressing systemic barriers in classification, placement, and pedagogy that hinder student success, this study explores alternative methods to refine entrance assessment tools and support linguistically underprepared students in DevEd and beyond [15]. The findings may inform more effective learning strategies, ultimately improving outcomes for students navigating writing-intensive coursework.

Our paper is structured as follows: Section 2 reviews LLMs in writing assessment and classification. Section 3 details the corpus, experimental setup, and LLMs used. Section 4 presents and discusses findings. Section 5 summarizes key contributions and future work.

[1] `https://www.accuplacer.org/` (Last accessed: 28th July 2025; all URLs in this paper were checked on this date.)

2 Related Work

Research on the use of LLMs in writing assessment, text classification, and automated feedback has gained increasing attention from higher education institutions and their policymakers [34, 41]. Several studies have evaluated LLMs, in particular to investigate their classification precision, interpretability, and feedback generation [18, 26, 42, 48] and have positioned them as potential alternatives to automated placement systems comparable to Accuplacer.

Despite this growing interest, we found that Accuplacer and similar standardized, automated systems remain widely used, yet they rely on "black-box" algorithms that prioritize surface-level linguistic features (e.g., *word count*, *syntactic complexity*) over deeper assessments of writing proficiency [16, 23] to better understand if students can communicate effectively. Some studies estimate that 30–50% of students are misplaced due to these limitations [30], raising concerns about the validity of such assessments [35]. Moreover, the proprietary nature of systems like Accuplacer restrict reproducibility, limit assessment transparency, and raise ethical research concerns [43].

While LLMs are also proprietary and subject to frequent updates, their integration in structured, prompt-based research provides a greater degree of process transparency and interpretability. Unlike Accuplacer, LLM-based classification allows researchers to design and control input conditions, evaluate model-generated justifications, and iteratively refine prompts to align with pedagogical goals. Nevertheless, we recognize that true reproducibility remains a challenge, particularly as models evolve over time. Although this caveat raises valid concerns, the promise of LLMs in advancing transparent placement practices remains a compelling one, continuing to gain traction in the field. Recent work highlights how carefully constructed LLM applications can improve placement accuracy while promoting fairness in high-stakes decisions [42, 47], which is one of the central aims of our ongoing study.

LLMs have shown promise in classification tasks across diverse text domains [42], demonstrating cost-effective assessment capabilities [38, 46]. For instance, GPT-4 and Gemini-Pro have excelled in sentiment classification for complex texts, surpassing traditional models with GPT-4 achieving the highest accuracy (0.76) [47]. GPT-4 has also outperformed Gemini in grammar correction, while Gemini excelled in sentence completion and logical reasoning [3]. In zero-shot text classification across 11 datasets, GPT-4 significantly outperformed GPT-3.5 [17]. Multi-prompting and iterative refinement strategies have been identified as effective in improving classification consistency by enabling LLMs to better internalize linguistic criteria [36], specifically, human-defined standards [29, 48].

Beyond classification, LLMs are being explored for Automated Essay Scoring (AES), where studies highlight their accuracy, consistency, and generalizability [28, 40]. When provided with clear rubrics and text examples, GPT-4 has demonstrated strong performance in AES tasks [48]. A study assessing 119 placement essays with four LLMs (Claude 2, GPT-3.5, GPT-4, and PaLM 2[2]) found that GPT-4 exhibited the highest intrarater reliability, though its performance fluctuated slightly over time [34]. Another study found GPT-4o to be the most consistent AES model (ICC above 0.7), while older GPT versions and Gemini 1.5 Flash displayed lower agreement (below 0.5) [41]. While LLMs perform well in rubric-based assessments, we noticed they struggle with discourse-level writing evaluation, particularly in areas requiring structural feedback, where human expertise is essential [34, 45].

Although the literature reveals promising results, we found that some challenges remain in writing classification due to LLMs prompt sensitivity and response variability. The literature posits that even slight variations in prompt phrasing can significantly impact classification

[2] `https://ai.google.dev/palm_docs/palm`

accuracy and consistency [6]. To address this, multi-step classification strategies have been proposed, where models first distinguish broad proficiency levels before refining subcategories [7, 22]. Bias in scoring also remains a concern. In an evaluation of 20 essays, for example, the GPT-4o1 model achieved the highest correlation with human ratings (Spearman's = .74) but, like GPT-4, exhibited a tendency to overrate texts [42].

LLM reliability varies with task complexity and domain. A study classifying 1,693 assessment questions based on Bloom's taxonomy [5] found traditional Machine Learning (ML) models outperforming Generative Artificial Intelligence (GenAI) models, with GPT-3.5-turbo achieving 0.62 precision versus 0.87 for traditional ML techniques, while Gemini Pro underperformed (0.43) [20]. Additionally, we learned that in [27] that Gemini Pro tends to struggle with fine-grained text comprehension tasks compared to GPT-4. Despite these limitations, LLMs have demonstrated potential in educational applications, particularly in tutoring. In an evaluation of 58,459 tutoring interactions, LearnLM[3] was preferred by expert raters over GPT-4o, Claude 3.5, and Gemini 1.5 Pro, highlighting its capacity for AI-driven learning support [45].

The studies we have here presented align with our objectives and research questions, emphasizing the complexities of assessing student writing consistently and accurately, and highlighting opportunities to improve traditional automated placement systems (e.g., Accuplacer) by leveraging LLMs and AI technologies. As these models evolve, improving classification precision, mitigating bias, and improving feedback quality will be paramount to ensure more reliable and equitable high-stakes assessments, particularly in DevEd [25, 28].

3 Methodology

To evaluate the precision of LLMs and their alignment with human raters for improving writing placement in DevEd, we first present a description of the corpus in Section 3.1, followed by a detailed explanation of the experimental setup in Section 3.2, along with an overview of the LLMs used and the classification experiments devised for this study.

3.1 Corpus

The corpus consists of essays written in English by college-intending students during the 2023–2024 academic year as part of Tulsa Community College's[4] standardized placement process. All essays were produced in a proctored environment without access to writing aids and were prompted by one of eleven possible reflective topics (e.g., *differences among people*, *unlimited change*) with minimal instructions.

A stratified random sampling method was employed to draw 450 essays from a larger pool of 1,000 placement essays. This sampling strategy minimized selection bias, complied with institutional data access protocols, and resulted in a corpus evenly distributed across three levels: DevEd Level 1, DevEd Level 2, and College-Level (150 texts each). The dataset combines an existing 300-essay developmental education corpus with 150 newly added College-Level texts, allowing for a more comprehensive evaluation of writing proficiency. Although demographic data was available for 67% of participants, it was excluded from the present analysis and will be examined in future research.

[3] `https://ai.google.dev/gemini-api/docs/learnlm`
[4] `https://www.tulsacc.edu`

For text-level classification, we reference an adapted version of the institution's official course descriptions:

- **DevEd Level 1**: Texts with frequent grammar, spelling, and punctuation issues, and lacking cohesion;
- **DevEd Level 2**: Texts with some structural improvements still needing targeted support; and
- **College Level**: Texts demonstrating academic-level writing with minimal errors.

All essays were also independently evaluated by two trained raters. Both raters are native English speakers who hold at least a bachelor's degree and have over five years of experience in higher education, particularly in DevEd curriculum and student writing assessment at U.S. community colleges. They were recruited through an open call for volunteers to participate in the writing assessment task. Prior to scoring, raters completed calibration sessions using classification guidelines developed by the authors of this paper [9] to ensure consistency in applying the three-level placement scale. Discrepancies were resolved through adjudication and score averaging, a widely accepted practice in classification tasks. Although only two raters were employed due to the intensive nature of the task and institutional resource constraints, their qualifications and training contributed to strong agreement. Interrater reliability [19] analysis yielded substantial agreement (K-alpha = 0.66), demonstrating a high level of consistency and reinforcing the dataset's reliability.

Table 1 presents key corpus statistics, including token distribution and classification assignments by Accuplacer and human raters.

Table 1 Corpus statistics and classification by level (by Accuplacer and human raters.

Levels	Tokens/text	Accuplacer	Human Raters
DevEd Level 1	≈ 190	150	118
DevEd Level 2	≈ 321	150	238
College Level	≈ 481	150	94
Total	**≈ 148,800**	**450**	**450**

It is pertinent to note that access to writing assessment corpora from standardized entrance exams is governed by strict protocols to protect at-risk, educationally disadvantaged student populations, ensure data privacy, and comply with institutional and ethical guidelines, which inherently limits their availability for research and broader analysis. This study adhered to these ethical considerations and followed the Institution's Review Board (IRB) protocols, under approval # 22-05.

3.2 Experimental Setup

We investigated the feasibility of using LLMs as complementary tools for placement decisions and writing instruction in DevEd. Specifically, we evaluated their classification performance against the gold standard corpus classification described in Section 3.1. Furthermore, the results of the LLM classification were then compared with the institution's existing placement system, Accuplacer, to assess its precision against alternative methods.

To evaluate classification performance, two classification experiments were envisaged:

- (i) a **1-step classification**, where students' texts were categorized into three classes: College-level, DevEd Level 1, and DevEd Level 2; and
- (ii) a **2-step classification**, where texts were first classified into College and DevEd categories and then further divided into DevEd Levels 1 and 2.

The transition from a **1-step** to a **2-step** classification focuses on potentially reducing initial complexity by first making a broad binary distinction (College vs. DevEd), ensuring that texts with clear college-level characteristics are correctly identified before refining classifications within the DevEd category. Since the differences between DevEd Levels 1 and 2 are often more subtle, this sequential approach allows for a more targeted classification between proficiency levels [18].

LLMs Selection

Four LLMs were tested and evaluated using their default hyperparameters. This study adopts the commonly accepted definition of large language models as pre-trained, transformer-based architectures [38]. Model selection was based on their wide availability via API access and their representation of varying tiers of generative performance, reasoning ability, and cost-efficiency – factors frequently highlighted in the literature as critical for educational applications and scalable implementation [18, 26, 47].

Model selection was also guided by practical considerations, including task complexity, processing speed, and data security – criteria commonly emphasized in applied LLM research and deployment frameworks. While the selected models align with current literature and offer broad applicability, this does not preclude the inclusion of additional or emerging models (e.g., DeepSeek) in future iterations of this research. The four LLMs included in this study are listed below in alphabetical order:

- **Claude**[5] [2], developed by Anthropic, is known to provide a more accessible API and structured responses, making it easier to evaluate for classification consistency [6]. It has also been reported to have better consistency in logical reasoning [4], which may enhance its performance in tasks requiring nuanced proficiency distinctions (e.g., DevEd Level 1 vs. Level 2). Claude 3.5 Sonnet (October 2024 version) was the model used.
- **Gemini**[6] [37], unlike the text-based LLMs GPT-3.5-turbo and GPT-4o, is designed for cross-modal reasoning and trained on diverse web sources, including educational content [44], which may enhance its ability to recognize academic proficiency markers and improve efficiency in handling texts with high linguistic variability [24, 27]. For this study, Gemini 2.0 Flash was used.
- **OpenAI's ChatGPT**[7]: (i) **GPT-3.5-turbo**, while computationally efficient [32], it is leveraged to assess any improvements in classification consistency with the more advanced, latest cost-efficient model [39], (ii) **GPT-4o**, in the task of DevEd placement context. The advanced model is explored for its strengths in text comprehension and structured response generation, which could not only help students identify writing weaknesses but also enhance the provision of targeted, context-aware feedback [27]. The experiments use the GPT models via the official OpenAI API.

LLMs Classification Experiments

To systematically examine how prompt structuring influences classification precision, we incorporated a four-stage prompting approach into the two experiments described in Section 3.2: (i) 1-step classification and (ii) 2-step classification.

[5] `https://docs.anthropic.com/en/home`
[6] `https://deepmind.google/technologies/gemini/`
[7] `https://chat.openai.com`

Prompt construction was guided by existing literature on LLM-based classification and rubric-based writing assessment [36, 48], ensuring that the design was not ad hoc. Prompts were iteratively refined to maximize linguistic clarity, align with defined classification objectives, and maintain compatibility across model architectures. These human-authored prompts were designed to capture patterns that distinguish proficient from non-proficient writing [31]. The complete series of prompts was archived for public access via GitHub[8] [11].

To evaluate whether increasing prompt specificity improved classification outcomes, we introduced a four-stage prompting sequence. Precision, the primary evaluation metric in this study, assessed the reliability of a model's positive predictions aligning with human-assigned classifications [21]. This focus is particularly relevant in DevEd contexts, where misclassification, whether through overplacement or underplacement, can significantly affect students' academic trajectories. The four prompting stages were as follows:

- (i) **Zero-shot**, in which the models classified texts following a prompt with a laconic approach. No sample texts were provided at this stage in the classification prompt. Due to the availability of comparable data, in terms of model precision, results from this approach establish the *baseline* for evaluation with the subsequent prompting strategies.
- (ii) **Few-shot**, in which the models received the same prompt but with basic definitions of each proficiency level, adapted from the Institution's official course descriptions and curriculum. No sample texts were provided at this stage either.
- (iii) **Enhanced**, in which the classification prompt included detailed descriptions of proficiency levels and key linguistic markers that encompassed grammatical accuracy, syntactic complexity, lexical richness, and discourse cohesion. No sample texts were provided at this stage.
- (iv) **Enhanced+**, this approach extends the **Enhanced** classification prompting by incorporating three manually assessed and classified sample texts (using Accuplacer's linguistic descriptors), each corresponding to the three proficiency levels already mentioned (DevEd Level 1, DevEd Level 2, and College-level). These samples were randomly selected from outside the corpus but drawn from the same database and time epoch.

The four-stage experimental framework aligned with the research questions in Section 1, allowing for a comparative analysis of how LLMs classify writing proficiency under varying levels of linguistic guidance. This approach evaluated classification performance with and without explicit linguistic criteria and concrete linguistic evidence (text samples). Across all prompting strategies, LLMs were also instructed to provide classification rationales, which, while more robust than Accuplacer, remain less comprehensive than human evaluations – an aspect to be explored in a subsequent study. A sample of this feedback is included in Appendix A.

4 Experimental Results

This section presents the results of the two classification experiments[9] described in Section 3.2: (i) 1-step classification and (ii) 2-step classification. In Sections 4.1 and 4.2, we analyze LLMs' performance across the four-stage prompting approaches, evaluating their effectiveness in distinguishing proficiency levels.

[8] https://gitlab.hlt.inesc-id.pt/u000803/deved

[9] All experiments were conducted over eight days, from February 18 to February 25, 2025.

4.1 1-step Classification: 3 classes

The 1-step classification experiment required LLMs to simultaneously differentiate among three proficiency levels – College-Level, DevEd Level 1, and DevEd Level 2 – without an intermediate binary decision. This method directly assessed the models' ability to distinguish college-ready writing from varying degrees of developmental writing in a single step.

Table 2 presents each LLMs' precision in classifying texts across these three levels under four prompting strategies. The results include precision figures for each level compared to the gold standard, as well as the overall model performance. A "+" next to the overall precision score indicates improvement relative to the Zero-shot approach (baseline), while a "−" denotes a decline. Results are analyzed in terms of overall precision first before comparing them to Accuplacer. Additionally, a confusion matrix of the best-performing model is included to provide a more detailed breakdown of classification outcomes relative to the gold standard.

Table 2 LLM and Accuplacer classification precision vs. gold standard, per level and overall.

LLM / Experiment	**Level 1**	**Level 2**	**College**	**Overall Precision**
Claude Zero-shot	44.44%	60.00%	72.09%	52.44%
Claude Few-shot	53.65%	61.73%	86.67%	59.1% (+6.66%)
Claude Enhanced	50.00%	66.83%	85.71%	60.89% (+8.45%)
Claude Enhanced+	50.89%	73.65%	73.08%	**62.22** % (+**9.78**%)
Gemini Zero-shot	41.16%	70.00%	64.38%	51.33%
Gemini Few-shot	42.86%	60.12%	81.08%	52.44% (+1.11%)
Gemini Enhanced	29.29%	64.04%	84.62%	42.89% (-8.44%)
Gemini Enhanced+	52.53%	63.76%	78.26%	59.56% (+8.23%)
GPT-3.5-turbo Zero-shot	28.78%	20.69%	81.82%	29.56%
GPT-3.5-turbo Few-shot	38.75%	65.04%	68.42%	48.44% (+18.88%)
GPT-3.5-turbo Enhanced	55.33%	59.26%	40.54%	53.33% (+23.77%)
GPT-3.5-turbo Enhanced+	43.85%	57.09%	43.75%	51.11% (+21.55%)
GPT-4o Zero-shot	36.59%	49.14%	88.24%	41.78%
GPT-4o Few-shot	43.95%	65.38%	78.26%	54.89% (+13.11%)
GPT-4o Enhanced	41.54%	69.53%	74.19%	54.00% (+12.22%)
GPT-4o Enhanced+	49.09%	70.76%	79.66%	61.33% (+19.55%)
Accuplacer	51.33%	68.67%	54.67%	58.22%

In the **Zero-shot** experiment, Claude achieved the highest classification precision (52.44%), followed closely by Gemini (51.33%), outperforming all other LLMs. The remaining models ranked in descending order of precision as GPT-4o > GPT-3.5-turbo.

In the **Few-shot** experiment, when given basic proficiency definitions, Claude remained the best-performing model (59.1%), showing noticeable gains (approximately +6.66%) from the Zero-shot run. The ranking of models in this experiment was: GPT-4o > Gemini > GPT-3.5-turbo.

In the **Enhanced** experiment, when refined linguistic descriptors were introduced, Claude maintained its leading position (60.89%). However, not much improvement (+1.79%) was observed from the previous setup (Claude Few-shot). The remaining models ranked as follows: GPT-4o > GPT-3.5-turbo > Gemini.

Lastly, in the **Enhanced+** experiment, Claude continued to perform best (62.22%), with a 10.22% improvement over its Zero-shot outcome, but only +1.33% improvement over the previous setup. The ranking, this time, was GPT-4o > Gemini > GPT-3.5-turbo. Notably, the highest gains in precision from the Zero-shot to the enhanced+ experiments were seen in GPT-3.5-turbo (21.55%) and GPT-4o (19.55%).

When compared to Accuplacer, three models, Claude (4%), GPT-4o (3.11%), and Gemini (1.34%), achieved slightly higher overall precision in the Enhanced+ experiments. Notably, Claude outperformed Accuplacer also across the Few-shot and Enhanced prompting strategies. While LLMs show potential for structured DevEd placement tasks compared to Accuplacer, they still present limitations, as evidenced by their low precision scores.

To complement this analysis, Figure 1 provides a visual representation of the classification precision scores distribution of LLMs and Accuplacer in the 1-step classification experiment.

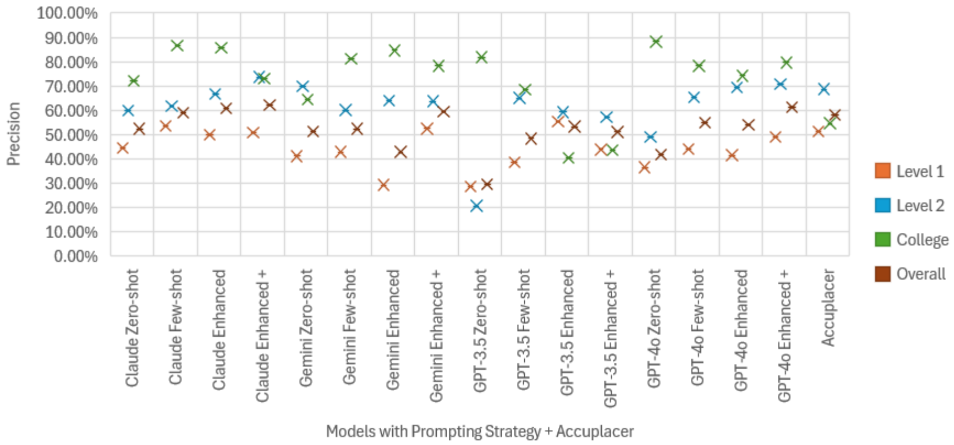


Figure 1 LLM and Accuplacer classification precision distribution per level and overall.

To better understand Claude's overall performance, given its highest classification precision in all prompting strategies, Table 3 displays the confusion matrix with the model's distribution of actual versus predicted classifications in its best performance, the enhanced+ experiment:

Table 3 Confusion matrix of predicted (LLM) vs. actual classifications (gold standard) of Claude in the Enhanced+ experiment.

↓ Actual / Predicted →	Level 1	Level 2	College	Total
Level 1	**114**	4	0	118
Level 2	108	**109**	21	238
College	2	35	**57**	94
Total	224	148	78	450

The confusion matrix analysis revealed that Claude excels in identifying Level 1 texts, correctly classifying nearly all (114/118) with only four misclassified as Level 2, indicating a strong grasp of beginning-level writing. However, its performance dropped significantly for Level 2, correctly identifying less than half of these texts. A strong tendency to underclassify was observed, as 108 Level 2 texts are misclassified as Level 1, and 21 as College-Level. Claude demonstrated good discrimination of advanced writing, correctly classifying 57 out of 94 College-Level texts, with most misclassifications occurring just one level below. Extreme errors were rare, with only two College-Level texts misclassified as Level 1. Overall, the model underclassified far more often than it overclassified (145 vs. 25 instances), suggesting it applied a stricter threshold for advancement compared to the human gold standard.

Based on the classification results obtained, further statistical validation is necessary. Pearson correlation coefficients (ρ) are computed next to assess the reliability and alignment of LLM-generated classifications with the gold standard.

Correlation of LLMs Classifications with Gold Standard

Pearson correlation coefficients (ρ) measured the degree of linear association between the classification levels assigned by LLMs and the gold standard in the 1-step experiment. A higher ρ value indicates a stronger correlation, meaning the LLMs' predictions align more closely with human-assigned levels.

Pearson correlation (ρ) scores were interpreted using SPSS guidelines[10], based on Cohen (1988) [8], which define correlation strength as follows:

$0.1 < |r| < 0.3$ indicates *small/weak* correlation (W);
$0.3 < |r| < 0.5$ corresponds to *medium/moderate* correlation (M); and
$0.5 < |r| \ldots$ denotes *large/strong* correlation (S).

The computed Pearson correlation coefficients (ρ) for each system are presented in Table 4.

Table 4 Pearson scores of LLMs and Accuplacer vs. gold standard. Correlation interpretation: substantial (S); moderate (M).

LLM/System	Zero-shot	Few-shot	Enhanced	Enhanced+
Claude	0.67 (S)	0.66 (S)	0.67 (S)	**0.75 (S)**
Gemini	0.60 (S)	0.62 (S)	0.62 (S)	0.63 (S)
GPT-3.5-turbo	0.40 (M)	0.54 (S)	0.51 (S)	0.39 (M)
GPT-4o	0.63 (S)	0.60 (S)	0.60 (S)	0.69 (S)
Accuplacer	**0.67 (S)**			

The results indicated that most LLMs exhibited a strong correlation with the gold standard, with Claude Enhanced+ achieving the highest alignment ($\rho = 0.75$), surpassing Accuplacer ($\rho = 0.67$). GPT-4o also outperformed Accuplacer but just by a smaller margin (0.02%) in the Enhanced+ run as well. In the Zero-shot and Enhanced prompting strategies, Claude matched Accuplacer's Pearson score, while the remaining models and configurations consistently underperformed.

Notably, GPT-3.5-turbo Zero-shot showed one of the weakest correlation scores ($\rho = 0.40$), but its performance improved significantly in the Few-shot ($\rho = 0.54$) and Enhanced ($\rho = 0.51$) settings before declining again in Enhanced+ ($\rho = 0.39$). This fluctuation suggests that models such as GPT-3.5-turbo may be more sensitive to prompt variations than others.

The strong correlation observed with Claude Enhanced+, particularly in such structured classification settings, highlights the potential of certain LLMs to approximate human-level proficiency in classification, though inconsistencies across models require further investigation.

4.2 2-step Classification: 3 classes

The 2-step classification experiment is followed in an attempt to address the risk of models disproportionately assigning texts to a particular category. Dividing a classification problem into two steps has shown to improve the overall precision of certain LLMs [29, 48]. To this extent, all 450 texts were initially classified as either College-Level or DevEd Level. Then, the DevEd-level texts, specifically, underwent a second classification, further distinguishing between DevEd Level 1 and DevEd Level 2. While the prompting methodology remained consistent (compared to the 1-step), level definitions and provided examples were slightly adjusted to align with the objectives of this classification experiment.

[10] https://libguides.library.kent.edu/SPSS/PearsonCorr

Table 5 presents the final precision scores[11] for each LLM (per level) across all four prompting strategies. A "+" next to the overall precision score indicates improvements relative to the Zero-shot approach (baseline), while a "−" represents the opposite. [13]

Table 5 LLM and ACCUPLACER classification precision vs. gold standard, per level and overall.

LLM / Experiment	Level 1	Level 2	College	Overall Precision
Claude Zero-shot	46.19%	74.07%	63.21%	56.89%
Claude Few-shot	51.72%	76.67%	57.48%	60.00% (+3.11%)
Claude Enhanced	52.88%	77.60%	58.97%	61.33% (+4.44%)
Claude Enhanced+	57.95%	77.36%	62.61%	66.00 % (+9.11%)
Gemini Zero-shot	48.83%	60.37%	82.61%	56.07%
Gemini Few-shot	57.79%	63.33%	76.92%	62.22% (+6.15%)
Gemini Enhanced	54.49%	61.92%	73.91%	59.78% (+3.71%)
Gemini Enhanced+	**66.67%**	**69.50%**	**73.85%**	**69.33% (+13.26%)**
GPT-3.5-turbo Zero-shot	45.29%	58.97%	43.62%	47.11%
GPT-3.5-turbo Few-shot	48.80%	64.52%	54.22%	50.89% (+3.78%)
GPT-3.5-turbo Enhanced	37.20%	65.00%	61.40%	46.44% (-0.67%)
GPT-3.5-turbo Enhanced+	46.24%	62.76%	47.06%	53.56% (+6.44%)
GPT-4o Zero-shot	40.81%	59.26%	72.09%	49.33%
GPT-4o Few-shot	34.24%	48.54%	82.35%	39.33% (-10.00%)
GPT-4o Enhanced	35.09%	51.38%	89.47%	41.33% (-8.00%)
GPT-4o Enhanced+	48.91%	73.83%	73.61%	61.11% (+11.78%)
Accuplacer	51.33%	68.67%	54.67%	58.22%

In the **Zero-shot** experiment, Claude demonstrated the highest classification precision (56.89%), closely followed by Gemini with only a 0.82% difference. The ranking of the remaining models in decreasing precision is GPT-4o > GPT-3.5-turbo.

In the **Few-shot** experiment, Gemini outperformed Claude by 2.22%. Both models showed modest improvements over their Zero-shot runs, with Claude achieving a 3.11% increase and Gemini, with almost twice the precision gain, achieving a 6.15%. The ranking of the remaining models was reversed from the Zero-shot experiment, as GPT-3.5-turbo outperformed GPT-4o. Notably, GPT-4o exhibited a significant 10% decrease in precision, indicating that added proficiency definitions did not necessarily enhance its performance.

In the **Enhanced** experiment, Claude achieved the highest precision score (61.33%), outperforming all other models. However, its improvement over the Few-shot setup was minimal (+1.33%), despite the inclusion of more precise linguistic descriptors and features in the prompt. Gemini ranked second, showing a 3.71% gain from its baseline but a 2.44% drop from the Few-shot setup, suggesting inconsistent benefits from increased prompt specificity. The remaining models ranked as follows: GPT-3.5-turbo > GPT-4o.

In the **Enhanced+** experiment, Gemini demonstrated the most significant gains, achieving an overall precision of 69.33%. Claude followed closely with a precision score of 66%, while GPT-4o ranked third with a precision score of 61.11%. GPT-3.5-turbo had the lowest precision in this setup, scoring 53.56%. All four models demonstrated performance gains from each of their baselines (Zero-shot) and the Enhanced run. However, Claude was the model to consistently improve its precision across all prompting strategies, suggesting that incorporating text samples alongside linguistic guidance enhances classification precision.

[11] This study is part of an ongoing research project. While the textual portion of the corpus cannot be released at this stage, a dataset containing the full classification results from the 2-step experiment – where overall precision scores exceeded those of the 1-step approach – has been made available for evaluation at `https://gitlab.hlt.inesc-id.pt/u000803/deved/` [13]. Full corpus access will be available upon completion of the study.

When compared to Accuplacer, Gemini and Claude demonstrated the highest overall precision in the Enhanced+ experiments, with notable gains of 11.11% and 7.78%, respectively. Both models also outperformed Accuplacer in the Few-shot and Enhanced strategies. GPT-4o showed a smaller improvement of 2.89% in the Enhanced+ strategy, while GPT-3.5-turbo consistently underperformed throughout. Despite their enhanced potential for structured DevEd placement, LLMs still exhibit some limitations, as reflected in their precision scores.

To complement this analysis, Figure 2 provides a visual representation of the classification precision scores distribution of LLMs and Accuplacer in the 2-step classification approach.

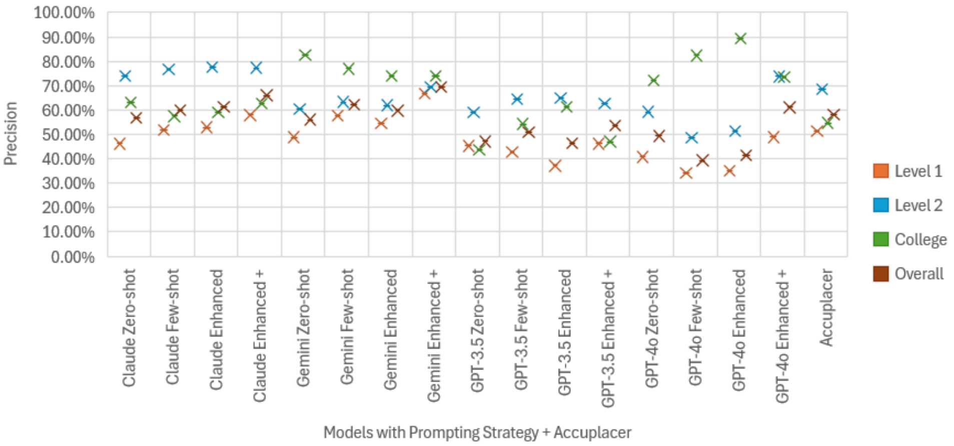


Figure 2 LLM and Accuplacer classification precision distribution, per level and overall.

Table 6 shows a confusion matrix detailing the model's distribution of actual versus predicted classifications in its best performance, the Enhaced+ experiment:

Table 6 Confusion matrix of predicted (LLM) vs. actual classifications (gold standard) in the Gemini Enhanced+ experiment.

↓ Actual / Predicted →	Level 1	Level 2	College	Total
Level 1	**84**	33	1*	118
Level 2	42	**180**	16	238
College	0	46	**48**	94
Total	126	259	65	450

Almost all errors (137 out of 138 misclassifications) occurred between adjacent levels, suggesting the model understands the ordinal relationship between levels. When Gemini made errors on Level 1 texts, it nearly always overclassified texts by just one level. The single outlier*, a text classified by Gemini as College-level but by human raters as DevEd Level 1, was closely examined. This text contained a single string of tokens, yet it was the shortest in the dataset.

Text 13: "Our ability to change ourselves is limitless."

Notably, Gemini distinguished it as a "statement" rather than a "text," which is the terminology used in all other justifications. The model's justification *verbatim* was:

> This is a concise, grammatically correct *statement* expressing a complex idea. It demonstrates strong control of language and does not contain any foundational errors. The sentence structure is simple but effective, conveying a clear and impactful message.

Gemini, additionally, showed some bias toward classifying texts as Level 2, which is also the level involved in most errors, both as the actual and predicted class. Overall, Gemini is more likely to underclassify than overclassify Level 2 texts, suggesting Level 2 is the most difficult to distinguish. On the contrary, the model never misclassified College texts as Level 1, showing excellent discrimination between these distant categories. As exhibited in Table 5, in the Enhanced+ run, Gemini maintains similar precision across all levels (≈67-74%), indicating balanced performance rather than excelling at one level at the expense of others.

Correlation of LLMs Classifications with Gold Standard

Similarly to the 1-step experiment, Pearson correlation coefficients (ρ) were also computed. Results are presented in Table 7.

Table 7 Pearson scores of LLMs and Accuplacer vs. gold standard. Correlation interpretation: substantial (S); moderate (M).

LLM/System	Zero-shot	Few-shot	Enhanced	Enhanced+
Claude	0.68 (S)	0.69 (S)	0.71 (S)	**0.73 (S)**
Gemini	0.64 (S)	0.62 (S)	0.58 (S)	0.69 (S)
GPT-3.5-turbo	0.59 (M)	0.54 (S)	0.49 (M)	0.45 (M)
GPT-4o	0.61 (S)	0.49 (M)	0.54 (S)	0.70 (S)
Accuplacer	**0.67 (S)**			

Compared to the 1-step classification, Pearson correlation calculations revealed that most LLMs exhibit a strong alignment with the gold standard. Claude Enhanced+ achieved the highest correlation ($\rho = 0.73$), followed closely by Claude Enhanced ($\rho = 0.71$) and GPT-4o ($\rho = 0.70$), outperforming Accuplacer ($\rho = 0.67$). There was a minimal difference (0.02% points) in Pearson scores observed between the 1-step and 2-step classifications for Claude Enhanced+, hinting at the model's stability in its classification performance.

The strong correlation observed in Claude and GPT-4o reinforces the idea that LLMs, especially in a 2-step classification, can closely align with human-assigned proficiency levels.

4.3 Precision Comparison: 1-step vs. 2-step classification experiments

Having conducted both 1-step and 2-step classification experiments using four prompting strategies, we highlight in Table 8 the top 3 major precision gains (green) and losses (red), per level and overall, to assess how LLMs respond to structured classification strategies. The figures represent the difference in precision between the 2-step and 1-step experiments, as overall precision scores were higher in the 2-step. This comparison directly supports the analysis of research questions 2 and 3 in Section 1, evaluating the impact of prompt refinement on classification precision and bias mitigation.

Most models improved in overall precision with the 2-step classification, particularly GPT-3.5-turbo Zero-shot (+17.55%), Gemini Enhanced (+16.89%), and Gemini Few-shot (+9.78%). Although not among the top three, Gemini Enhanced+ performed similarly to Gemini Few-shot, with only a 0.01% difference. Conversely, GPT-4o struggled to maintain positive precision gains across three out of the four prompting strategies, suggesting that a 2-step classification strategy may not be as effective for this particular model.

We observed significant gains in Level 1 in two of the four prompting strategies, particularly for Gemini (Few-shot: +14.93%, Enhanced: +25.20%) and GPT-3.5-turbo Zero-shot (+16.51%). Particularly for GPT-3.5-turbo Zero-shot, it also saw the sharpest trade-off,

Table 8 Precision gains and losses between 1 and 2-step experiments across prompting strategies.

LLM / Experiment	Level 1	Level 2	College	Overall
Claude Zero-shot	+1.75%	+14.07%	-8.88%	+4.45%
Claude Few-shot	-1.93%	+14.94%	-29.19%	+0.90%
Claude Enhanced	+2.88%	+10.77%	-26.74%	+0.44%
Claude Enhanced+	+7.06%	+3.71%	-10.47%	+3.78%
Gemini Zero-shot	+7.67%	-9.63%	+18.23%	+4.74%
Gemini Few-shot	+14.93%	+3.21%	-4.16%	+9.78%
Gemini Enhanced	+25.20%	-2.12%	-10.71%	+16.89%
Gemini Enhanced+	+14.14%	+5.74%	-4.41%	+9.77%
GPT-3.5-turbo Zero-shot	+16.51%	+38.28%	-38.20%	+17.55%
GPT-3.5-turbo Few-shot	+4.05%	-0.52%	-14.20%	+2.45%
GPT-3.5-turbo Enhanced	-18.13%	+5.74%	+20.86%	-6.89%
GPT-3.5-turbo Enhanced+	+2.39%	+5.67%	+3.31%	+2.45%
GPT-4o Zero-shot	+4.22%	+10.12%	-16.15%	+7.55%
GPT-4o Few-shot	-9.71%	-16.84%	+4.09%	-15.56%
GPT-4o Enhanced	-6.45%	-18.15%	+15.28%	-12.67%
GPT-4o Enhanced+	-0.18%	+3.07%	-6.05%	-0.22%

improving +38.28% in Level 2 but dropping nearly the same amount (-38.20%) in College-Level classification. Similarly, for Level 2, Claude Zero and Few-shot showed notable gains (+14.07% and +14.94%, respectively) but substantial losses at the College-Level (-29.19% and -26.74%) in the Few-shot and Enhanced runs.

Overall, most results suggest that breaking down classification into distinct steps improved precision. Among all models, Gemini demonstrated the most consistent performance across levels and prompting strategies. The trade-off between gains in lower proficiency levels and losses in College-Level precision suggests that some models, particularly Claude and GPT-3.5-turbo, may be more in-tuned with DevEd writing patterns, prioritizing features typical of lower-level texts while misclassifying more advanced writing as DevEd.

5 Conclusions and Future Work

We evaluated four LLMs as data-driven alternatives for equitable DevEd placement, offering (potential) competitive, cost-effective suggestions for higher education institutions traditionally reliant on "black-box" systems like Accuplacer. Our findings indicated that both classification structure and prompt refinement played a critical role in enhancing LLM precision when distinguishing between DevEd and College-level texts.

In terms of classification structure, across the 1 and 2-steps classification experiments, models exhibited higher precision in Level 2 and College-Level classifications but struggled with Level 1, likely due to the inconsistent structures and frequent grammatical errors in Level 1 texts. Compared to the 1-step experiment, the 2-step classification saw Gemini Enhanced+ achieved the highest precision (66.67%) in Level 1, reinforcing the notion that breaking classification into multiple steps enhances precision [29, 48], and the importance of incorporating clear linguistic criteria with sample texts in the classification prompts [36]. We noticed, too, that Claude and Gemini consistently underclassified Level 2 texts, suggesting they apply a stricter advancement threshold compared to human raters. This likely occurs because the models are more conservative in assigning texts to higher proficiency levels, meaning they are more prone to classifying Level 2 texts as Level 1 rather than correctly identifying them.

Claude and Gemini demonstrated a strong correlation with human ratings, with Claude achieving the highest Pearson scores ($\rho = 0.75$; 1-step, $\rho = 0.73$; 2-step), surpassing Accuplacer ($\rho = 0.67$). These findings confirm that LLMs have the potential approximate human-level proficiency classification, yet their precision remains a work in progress, particularly for high-stakes placement decisions. Given the social and economic impact of misplacement, safeguards are necessary to ensure an ethical and fair assessment of students' skills.

To refine the role of LLMs in placement decisions, our future research will: (i) explore the expansion of the dataset to increase generalizability; (ii) further refine prompting strategies to mitigate classification inconsistencies; (iii) test newer, more sophisticated LLMs for improved performance; (iv) investigate LLM-generated feedback to evaluate its effectiveness in justifying classification decisions and enhancing instructional usability in teaching practices. By addressing these areas, future studies will continue to contribute to more transparent, data-driven placement methodologies that balance automation with fairness and reliability.

References

1 Tufan Adiguzel, Mehmet Haldun Kaya, and Faith Kürşat Cansu. Revolutionizing education with AI: Exploring the transformative potential of ChatGPT. *Contemporary Educational Technology*, 15(3):ep429, 2023. doi:10.30935/cedtech/13152.

2 AI Anthropic. The Claude 3 model family: Opus, Sonnet, Haiku. *Claude-3 Model Card*, 1:1, 2024.

3 Nathan Atox and Mason Clark. Evaluating Large Language Models through the lens of linguistic proficiency and world knowledge: A comparative study. *Authorea Preprints*, 2024.

4 Jaehyeon Bae, Seoryeong Kwon, and Seunghwan Myeong. Enhancing software code vulnerability detection using gpt-4o and claude-3.5 sonnet: A study on prompt engineering techniques. *Electronics*, 13(13), 2024. doi:10.3390/electronics13132657.

5 Benjamin Samuel Bloom. *Taxonomy of educational objectives: Handbook II.* David McKay, 1956.

6 Loredana Caruccio, Stefano Cirillo, Giuseppe Polese, Giandomenico Solimando, Shanmugam Sundaramurthy, and Genoveffa Tortora. Claude 2.0 Large Language Model: Tackling a real-world classification problem with a new iterative prompt engineering approach. *Intelligent Systems with Applications*, 21:200336, 2024. doi:10.1016/J.ISWA.2024.200336.

7 Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. KnowPrompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. In *Proceedings of the ACM Web Conference 2022*, WWW '22, pages 2778–2788, New York, NY, USA, 2022. Association for Computing Machinery. doi:10.1145/3485447.3511998.

8 Jacob Cohen. *Statistical power analysis.* Hillsdale, NJ: Erlbaum, 1988.

9 Miguel Da Corte and Jorge Baptista. Enhancing writing proficiency classification in Developmental Education: the quest for accuracy. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6134–6143, 2024. URL: https://aclanthology.org/2024.lrec-main.542.

10 Miguel Da Corte and Jorge Baptista. Leveraging NLP and machine learning for English (l1) writing assessment in Developmental Education. In *Proceedings of the 16th International Conference on Computer Supported Education (CSEDU 2024), 2-4 May, 2024, Angers, France*, volume 2, pages 128–140, 2024. doi:10.5220/0012740500003693.

11 Miguel Da Corte and Jorge Baptista. LLM-based writing classification prompts for 1-step and 2-step classification experiments. Technical report, University of Algarve & INESC-ID Lisboa, 2025.

12 Miguel Da Corte and Jorge Baptista. Refining English writing proficiency assessment and placement in Developmental Education using NLP tools and Machine Learning. In *Proceedings of the 17th International Conference on Computer Supported Education - Volume 2: CSEDU*, pages 288–303. INSTICC, SciTePress, 2025. doi:10.5220/0013351500003932.

13 Miguel Da Corte and Jorge Baptista. Results for LLM-based 2-step classification. Technical report, University of Algarve & INESC-ID Lisboa, 2025.

14 Miguel Da Corte and Jorge Baptista. Toward consistency in writing proficiency assessment: Mitigating classification variability in Developmental Education. In *Proceedings of the 17th International Conference on Computer Supported Education - Volume 2: CSEDU*, pages 139–150. INSTICC, SciTePress, 2025. doi:10.5220/0013353900003932.

15 Jane Denison-Furness, Stacey Lee Donohue, Annemarie Hamlin, and Tony Russell. Welcome/Not Welcome: From Discouragement to Empowerment in the Writing Placement Process at Central Oregon Community College. In Jassica Nastal, Mya Poe, and Christie Toth, editors, *Writing Placement in Two-Year Colleges: The Pursuit of Equity in Postsecondary Education*, pages 107–127. The WAC Clearinghouse/University Press of Colorado, 2022. doi:10.37514/PRA-B.2022.1565.2.04.

16 Nikki Edgecombe and Michael Weiss. Promoting equity in Developmental Education reform: A conversation with Nikki Edgecombe and Michael Weiss. *Center for the Analysis of Postsecondary Readiness*, page 1, 2024.

17 Jessica López Espejel, El Hassane Ettifouri, Mahaman Sanoussi Yahaya Alassan, El Mehdi Chouham, and Walid Dahhane. GPT-3.5, GPT-4, or BARD? evaluating LLMs reasoning ability in zero-shot setting and performance boosting through prompts. *Natural Language Processing Journal*, 5:100032, 2023. doi:10.1016/J.NLP.2023.100032.

18 John Fields, Kevin Chovanec, and Praveen Madiraju. A survey of text classification with transformers: How wide? how large? how long? how accurate? how expensive? how safe? *IEEE Access*, 12:6518–6531, 2024. doi:10.1109/ACCESS.2024.3349952.

19 Deen Freelon. Recal OIR: ordinal, interval, and ratio intercoder reliability as a web service. *International Journal of Internet Science*, 8(1):10–16, 2013.

20 Mohammed Osman Gani, Ramesh Kumar Ayyasamy, Saadat M Alhashmi, Khondaker Sajid Alam, Anbuselvan Sangodiah, Khondaker Khaleduzzman, and Chinnasamy Ponnusamy. Towards enhanced assessment question classification: a study using Machine Learning, deep learning, and Generative AI. *Connection Science*, 37(1):2445249, 2025. doi:10.1080/09540091.2024.2445249.

21 Margherita Grandini, Enrico Bagli, and Giorgio Visani. Metrics for multi-class classification: an overview. *arXiv preprint arXiv:2008.05756*, 2020. arXiv:2008.05756.

22 Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. Tabllm: Few-shot classification of tabular data with Large Language Models. In *International Conference on Artificial Intelligence and Statistics*, pages 5549–5581. PMLR, 2023. URL: https://proceedings.mlr.press/v206/hegselmann23a.html.

23 Sarah Hughes and Ruth Li. Affordances and limitations of the ACCUPLACER automated writing placement tool. *Assessing Writing*, 41:72–75, 2019.

24 Muhammad Imran and Norah Almusharraf. Google Gemini as a next generation AI educational tool: a review of emerging educational technology. *Smart Learning Environments*, 11(1):22, 2024. doi:10.1186/S40561-024-00310-Z.

25 Elizabeth Kopko, Jessica Brathwaite, and Julia Raufman. The next phase of placement reform: Moving toward equity-centered practice. research brief. *Center for the Analysis of Postsecondary Readiness*, 2022. URL: https://files.eric.ed.gov/fulltext/ED626145.pdf.

26 Milan Kostic, Hans Friedrich Witschel, Knut Hinkelmann, and Maja Spahic-Bogdanovic. LLMs in automated essay evaluation: A case study. In *Proceedings of the AAAI Symposium Series*, volume 3, pages 143–147, 2024. doi:10.1609/AAAISS.V3I1.31193.

27 Gyeong-Geon Lee, Ehsan Latif, Lehong Shi, and Xiaoming Zhai. Gemini Pro defeated by GPT-4v: Evidence from education. *arXiv preprint arXiv:2401.08660*, 2023.

28 Stephanie Link and Svetlana Koltovskaia. *Automated Scoring of Writing*, pages 333–345. Springer International Publishing, Cham, 2023. doi:10.1007/978-3-031-36033-6_21.

29 Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in Natural Language Processing. *ACM Computing Surveys*, 55(9):1–35, 2023. doi:10.1145/3560815.

30 Ross Markle. Redesigning course placement in service of guided pathways. *Educational Considerations*, 50(2):8, 2025.

31 Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. Recent advances in Natural Language processing via Large pre-trained Language Models: A survey. *ACM Computing Surveys*, 56(2):1–40, 2023. doi:10.1145/3605943.

32 Yida Mu, Ben P. Wu, William Thorne, Ambrose Robinson, Nikolaos Aletras, Carolina Scarton, Kalina Bontcheva, and Xingyi Song. Navigating prompt complexity for zero-shot classification: A study of large language models in computational social science. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12074–12086, Torino, Italia, May 2024. ELRA and ICCL. URL: https://aclanthology.org/2024.lrec-main.1055/.

33 Sasha Nikolic, Carolyn Sandison, Rezwanul Haque, Scott Daniel, Sarah Grundy, Marina Belkina, Sarah Lyden, Ghulam M Hassan, and Peter Neal. ChatGPT, Copilot, Gemini, SciSpace and Wolfram versus higher education assessments: an updated multi-institutional study of the academic integrity impacts of Generative Artificial Intelligence (GenAI) on assessment, teaching and learning in engineering. *Australasian Journal of Engineering Education*, 29(2):126–153, 2024.

34 Austin Pack, Alex Barrett, and Juan Escalante. Large language models and automated essay scoring of english language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence*, 6:100234, 2024. doi:10.1016/J.CAEAI.2024.100234.

35 Dolores Perin, Julia Raufman, and Hoori Santikian Kalamkarian. Developmental reading and English assessment in a researcher-practitioner partnership. Technical report, CCRC, Teachers College, Columbia University, 2015.

36 Dylan Phelps, Thomas Pickard, Maggie Mi, Edward Gow-Smith, and Aline Villavicencio. Sign of the times: Evaluating the use of large language models for idiomaticity detection. In Archna Bhatia, Gosse Bouma, A. Seza Doğruöz, Kilian Evang, Marcos Garcia, Voula Giouli, Lifeng Han, Joakim Nivre, and Alexandre Rademaker, editors, *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 178–187, Torino, Italia, May 2024. ELRA and ICCL. URL: https://aclanthology.org/2024.mwe-1.22/.

37 Sundai Pichai and Demis Hassabis. Introducing Gemini: our largest and most capable AI model, 2023. Accessed: February 19, 2025. URL: https://blog.google/technology/ai/google-gemini-ai/#introducing-gemini.

38 Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

39 Areeg Fahad Rasheed, M Zarkoosh, Safa F Abbas, and Sana Sabah Al-Azzawi. TaskComplexity: A dataset for task complexity classification with in-context learning, FLAN-T5 and GPT-4o benchmarks. *arXiv preprint arXiv:2409.20189*, 2024. doi:10.48550/arXiv.2409.20189.

40 Eugénio Ribeiro, Nuno Mamede, and Jorge Baptista. Exploring the automated scoring of narrative essays in brazilian portuguese using transformer models. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese-Vol. 2*, pages 14–17, 2024. URL: https://aclanthology.org/2024.propor-2.4/.

41 Siti Bealinda Qinthara Rony, Tan Xin Fei, and Sasa Arsovski. Educational justice. reliability and consistency of Large Language Models for automated essay scoring and its implications. *Journal of Applied Learning and Teaching*, 8(1), 2025.

42 Kathrin Seßler, Maurice Fürstenberg, Babette Bühler, and Enkelejda Kasneci. 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, 2025. doi:10.1145/3706468.3706527.

43 Arthur Spirling. Why open-source Generative AI models are an ethical way forward for science. *Nature*, 616(7957):413–413, 2023.

44 Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

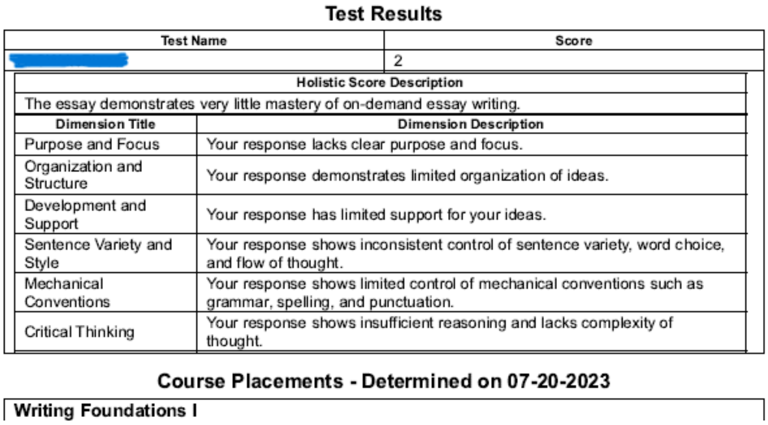
45 LearnLM Team, Abhinit Modi, Aditya Srikanth Veerubhotla, Aliya Rysbek, Andrea Huber, Brett Wiltshire, Brian Veprek, Daniel Gillick, Daniel Kasenberg, Derek Ahmed, et al. Learnlm: Improving Gemini for learning. *arXiv preprint arXiv:2412.16429*, 2024.

46 Yuqi Wang, Wei Wang, Qi Chen, Kaizhu Huang, Anh Nguyen, and Suparna De. Zero-shot text classification with knowledge resources under label-fully-unseen setting. *Neurocomputing*, 610:128580, 2024. doi:10.1016/j.neucom.2024.128580.

47 Zhiqiang Wang, Yiran Pang, Yanbin Lin, and Xingquan Zhu. Adaptable and reliable text classification using Large Language Models. *arXiv preprint arXiv:2405.10523*, 2024.

48 Changrong Xiao, Wenxing Ma, Qingping Song, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Qi Fu. 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, 2025. doi:10.1145/3706468.3706507.

A Accuplacer and LLMs' sample feedback



Test Results

Test Name	Score
[redacted]	2

Holistic Score Description
The essay demonstrates very little mastery of on-demand essay writing.

Dimension Title	Dimension Description
Purpose and Focus	Your response lacks clear purpose and focus.
Organization and Structure	Your response demonstrates limited organization of ideas.
Development and Support	Your response has limited support for your ideas.
Sentence Variety and Style	Your response shows inconsistent control of sentence variety, word choice, and flow of thought.
Mechanical Conventions	Your response shows limited control of mechanical conventions such as grammar, spelling, and punctuation.
Critical Thinking	Your response shows insufficient reasoning and lacks complexity of thought.

Course Placements - Determined on 07-20-2023

Writing Foundations I

Figure 3 Accuplacer's feedback for DevEd Level 1 = Writing Foundations I.

Gemini Enhanced+'s Feedback. Text: A 322; Classification: DevEd Level 1 Rationale:

> The text exhibits frequent errors in mechanical conventions, including misspellings ("themselfs," "simpley," "mojority," "to," "garentee," "sucurty," "revanew," "Anthoner," "belive," "unles"), punctuation (incorrect comma usage, missing apostrophes), and grammatical errors (subject-verb agreement, pronoun usage). Sentence variety and style are limited, with repetitive sentence structures. Development and support are present but lack coherence and depth. Organization and structure are attempted but weakened by abrupt transitions and lack of a clear thesis in the introduction. While the purpose and focus are identifiable, the arguments lack sufficient clarity. The text exhibits limited critical thinking, relying heavily on anecdotal evidence and generalizations.