ZERO-SHOT AUTOMATIC ASSIGNMENT GRADING WITH LARGE LANGUAGE MODELS

Dmitry Pakutnik
Department of Intelligent Information Technologies,
Belarusian State University of Informatics and Radioelectronics
Minsk, Belarus
E-mail: d.pakutnik@bsuir.by

This work proposes a conceptual framework for zero-shot automatic assignment grading using large language models. The approach employs prompt-based reasoning to generate both grades and explanatory feedback, aiming to improve scalability, transparency, and pedagogical value in higher education assessment.

Introduction

Assessment plays a key role in higher education, helping to ensure that learning outcomes are measurable and consistent with academic expectations. However, grading continues to be time-consuming and inconsistent, despite decades of automation efforts. Most existing systems rely on predefined answers or large labeled datasets, limiting their flexibility across disciplines and languages.

Recent progress in large language models has introduced new possibilities for reasoning-based evaluation [1]. This paper proposes a conceptual framework for zero-shot automatic assignment grading, where language models assess student work through prompt-guided reasoning rather than training data. The approach aims to enhance scalability, transparency, and pedagogical feedback while reducing the burden on educators [2].

I. Background and Problem Definition

Before turning to newer language-based methods, it is helpful to briefly describe the main types of automated grading systems used today. Traditional automated grading systems are typically of two kinds:

1. **Rule-based evaluators**, which depend on rigid templates or pattern matching;
2. **Supervised learning models**, which require extensive annotated datasets to learn from human grading examples.

Both approaches are constrained by domain specificity and lack the flexibility to adapt to new courses or assignment types without substantial reconfiguration. They also tend to focus on correctness alone, neglecting the explanatory and formative aspects of feedback that support student learning.

Manual grading, on the other hand, remains the gold standard for interpretive accuracy but is slow, inconsistent, and subject to human bias. As class sizes increase and curricula diversify, maintaining both speed and fairness becomes unsustainable. Therefore, an adaptive, general-purpose evaluation system capable of reasoning about open-ended student responses is a critical need in higher education.

II. Conceptual Framework

The proposed **zero-shot LLM grading framework** is built upon the principle of linguistic reasoning rather than statistical training [3]. It positions the language model as an autonomous evaluator capable of interpreting assignment prompts, understanding the logic of student responses, and generating both evaluative judgments and explanatory feedback. The foundation of this approach lies in *prompt engineering*, where well-structured textual instructions define the assessment context, criteria, and expected response format.

Unlike traditional systems that depend on predefined answers or domain-specific models, the zero-shot framework allows for flexibility across disciplines and assignment types. The grading logic is expressed entirely in natural language, enabling the model to apply general reasoning skills to diverse academic inputs. This adaptability makes the framework suitable for subjects involving quantitative problem-solving, conceptual explanations, or written argumentation.

As illustrated in Fig. 1, the framework operates through four core stages:

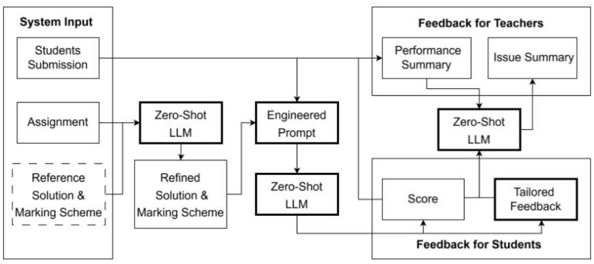

Figure 1 – Overview of the Zero-Shot LLM-Based Automated Assignment Grading System

1. **Input Assembly**
   The system consolidates essential components of the evaluation task – the assignment question, the student's response, and, when available, a grading rubric or reference criteria. These inputs provide the contextual basis for assessing both the content and quality of the submission.
2. **Prompt Construction**
   A structured prompt is formulated to guide the model's reasoning. The prompt defines the evaluator's role, outlines the assessment

dimensions–such as correctness, completeness, reasoning, and clarity–and specifies the required output: a numerical grade accompanied by a concise explanation. A conceptual example of an evaluation prompt is provided in Fig. 2.

Prompt for LLM

Overview: Your role is to assess and provide feedback on a student's response to a specific task in a <Course> assignment. Each task contains multiple components, and you are required to focus on evaluating the last question.
Background: <Question Background>
Question: <Question>
Student's Response: <Student Answer>
Marking Scheme: <Marking Scheme>

Instructions for the Model:

- **Focus:** Grade and provide comments exclusively on the last question.
- **Scoring:** Rate the student's answer from 1 to 10 according to the provided marking scheme.
- **Feedback:** Provide a brief and constructive critique. Highlighting strengths and areas for improvement according to the marking scheme. Feedback must be concise.

Figure 2 – Evaluation prompt for the Automated Assignment Grading system

3. **Evaluation and Feedback Generation**
   The language model processes the prompt and produces an integrated response consisting of a score and a written justification. The explanation highlights the logic behind the assessment, thereby functioning as immediate formative feedback for the learner.
4. **Consistency Moderation**
   To maintain reliability, multiple evaluations may be aggregated or cross-validated. Statistical normalization and human moderation can be applied to ensure alignment with institutional grading standards and to prevent systematic deviation.

Through these stages, the framework transforms the assessment process into an interaction driven by understanding and reasoning rather than data memorization. The use of natural-language prompts allows the evaluator's expectations to be communicated directly to the model, enabling adaptable and interpretable grading. The resulting system combines computational efficiency with pedagogical depth, ensuring that every assigned score is supported by a transparent and instructive rationale.

III. Advantages of the Zero-Shot Approach

The zero-shot approach introduces a new paradigm in automatic assignment grading by relying on linguistic reasoning rather than task-specific training or large labeled datasets. Its flexibility across domains allows the same framework to evaluate diverse subjects, from humanities essays to technical problem-solving, making it adaptable to a wide range of courses and assignment types.

Another key benefit is its reduced resource dependence. Traditional automated grading systems require extensive datasets and retraining, whereas the zero-shot framework operates using only task descriptions and rubrics embedded in prompts, making it more accessible for institutions with limited computational resources. It also ensures transparency and interpretability, as each evaluation includes a textual explanation outlining the reasoning behind the assigned grade, supporting both academic integrity and formative learning.

Additionally, the framework offers rapid adaptation and pedagogical support. By modifying prompts, the system can adjust to new assignments or updated rubrics without retraining, while the explanatory feedback guides students in identifying conceptual gaps and improving reasoning skills.

IV. Conclusion

The zero-shot automatic assignment grading framework represents a promising approach to modernizing assessment in higher education. By leveraging the reasoning and generative capabilities of large language models through structured prompts, it offers scalable, adaptable, and interpretable evaluation across diverse subjects and assignment types. The integration of explanatory feedback alongside numeric scores enhances the educational value of grading, supporting student learning and reflection while maintaining transparency and academic integrity.

This approach has the potential to transform grading from a time-consuming administrative task into a dynamic, formative process. With careful design, prompt optimization, and human oversight, zero-shot grading frameworks can improve consistency, reduce resource demands, and provide meaningful guidance to students, paving the way for more efficient and pedagogically effective assessment in universities.

1. Brown T., Mann B., Ryder N., Subbiah M., Kaplan J. D., Dhariwal P., Neelakantan A., Shyam P., Sastry G., Askell A., et al. Language models are few-shot learners // Advances in Neural Information Processing Systems. – 2020. – Vol. 33. – P. 1877–1901
2. Chiang C. H., Chen W. C., Kuan C. Y., Yang C., Lee H. Y. Large language model as an assignment evaluator: Insights, feedback, and challenges in a 1000+ student course [Electronic resource] // arXiv preprint arXiv: 2407.05216. – 2024. – Mode of access: `https://arxiv.org/abs/2407.05216`. – Date of access: 24.10.2025
3. Wei J., Bosma M., Zhao V. Y., Guu K., Yu A. W., Lester B., Du N., Dai A. M., Le Q. V. Finetuned language models are zero-shot learners [Electronic resource] // arXiv preprint arXiv:2109.01652. – 2021. – Mode of access: `https://arxiv.org/abs/2109.01652`. – Date of access: 24.10.2025