

PEMUTA: Pedagogically-Enriched Multi-Granular Undergraduate Thesis Assessment

Jialu Zhang^{1*}, Qingyang Sun^{1*}, Qianyi Wang¹, Risa Higashita⁴, Hongwu Qin⁴, Xiaoqing Zhang^{1†}, Jianfeng Ren², and Jiang Liu^{1234†},

ABSTRACT

Undergraduate thesis (UGTE) serves as a critical indicator of a student's cumulative academic development throughout university education. Although large language models (LLMs) have advanced educational intelligence, existing approaches typically focus on holistic assessment with only a single evaluation score, overlooking nuanced variations across multifaceted criteria. This limits their ability to reflect structural criteria, pedagogical objectives, and multi-dimensional academic competencies. Meanwhile, pedagogical theories have long guided manual UGTE evaluation through multi-dimensional assessment of cognitive development, disciplinary thinking, and academic performance, yet such guidance remains underexplored in automated assessment. Motivated by the research gap, we pioneer the Pedagogically-Enriched Multi-granular Undergraduate Thesis Assessment (PEMUTA) framework, which effectively activates domain-specific knowledge in pretrained LLMs for UGTE assessment in engineering disciplines. Grounded in Vygotsky's theory and Bloom's taxonomy, PEMUTA evaluates UGTEs across six fine-grained dimensions: Structure, Logic, Originality, Writing, Proficiency, and Rigor (SLOWPR), alongside an overall holistic assessment. A hierarchical two-stage pipeline is designed to first conduct fine-grained evaluation, then synthesize a holistic judgment, enabling more accurate elicitation of domain-relevant assessment knowledge in pretrained LLMs. Few-shot and role-based prompting are further incorporated to improve alignment with expert evaluation practices without parameter tuning. A benchmark consisting of authentic UGTEs collected from multiple engineering disciplines with expert SLOWPR-aligned annotations is curated to support multi-granular UGTE evaluation. Extensive experiments demonstrate that PEMUTA achieves strong agreement with expert assessments under diverse engineering UGTE evaluation scenarios, showing its substantial potential for fine-grained, pedagogically-grounded UGTE assessment within engineering education.

KEYWORDS

Generative Artificial Intelligence, Education Intelligence, Automated Thesis Assessment Benchmark, Pedagogical Design, Large Language Model, Knowledge Elicitation

1 Introduction

Undergraduate thesis (UGTE) is a significant component of undergraduate education quality assessment, closely related to graduation eligibility and degree classification. It assesses students' ability to apply disciplinary knowledge and academic skills, such as analytical reasoning, research rigor, and originality, to real-world or theoretical problems^[1, 2]. UGTEs are typically long and highly structured, often exceeding 50 pages, and require evaluation across multifaceted criteria, including structural coherence, academic rigor, and methodological soundness. Manual assessment is therefore time-consuming and labor-intensive, making timely, consis-

tent, and personalized feedback difficult to provide^[3]. With increasing submission volumes and limited faculty resources, assisted UGTE assessment has received growing attention^[4, 5].

Early rule-based methods^[6, 7, 8] and recent deep neural networks^[9, 10] have promoted the development of automated essay scoring (AES), but they are ill-suited for UGTE assessment due to their limited capacity to capture high-level semantics and dependence on costly, large-scale annotated datasets. More recently, educational agents powered by large language models (LLMs) have advanced the development of education intelligence and gained increasing attention^[11, 12, 13, 14, 15]. LLMs are pre-trained on massive and diverse corpora; these models encode rich linguistic patterns and

¹ Research Institute of Trustworthy Autonomous Systems and Department of Computer Science and Engineering, Southern University of Science and Technology, Guangdong, 518055, China

² School of Computer Science, University of Nottingham Ningbo China, Zhejiang, 315100, China

³ School of Ophthalmology and Optometry, Wenzhou Medical University, Zhejiang 325035, China

⁴ Department of Electronic and Information Engineering, Changchun University, 130022, China

* These authors contributed equally to this work.

† Corresponding authors: Jiang Liu, email: liuj@sustech.edu.cn; Xiaoqing Zhang, email: xq.zhang2@siat.ac.cn.

domain knowledge, enabling strong contextual reasoning and long-form text comprehension. Crucially, they can be adapted to complex educational tasks via prompt-based instructions without requiring additional fine-tuning [4, 16].

As illustrated in Fig. 1, LLMs can be used as UGTE assessment agents that take assessment instructions and thesis texts as input and return a final score. Although this paradigm improves scalability and efficiency, three key limitations remain. 1) Most systems rely on holistic scoring and provide only one score for each submission, overlooking nuances among academic competencies such as analytical reasoning, research rigor, and intellectual originality. This limits result interpretability and criterion-referenced feedback [17, 18]. 2) Prompt-based methods often lack domain-specific and pedagogical grounding, providing limited guidance for LLMs to apply academic standards or align their judgments with educational frameworks such as Bloom's Taxonomy and Vygotsky's theory. As a result, they often miss key indicators in UGTE assessment, e.g., intellectual development, and yield feedback with limited diagnostic value [19, 20, 5]. 3) UGTE evaluation requires criterion-specific reasoning across multiple dimensions, which generic prompts often fail to support effectively [21, 22, 23].

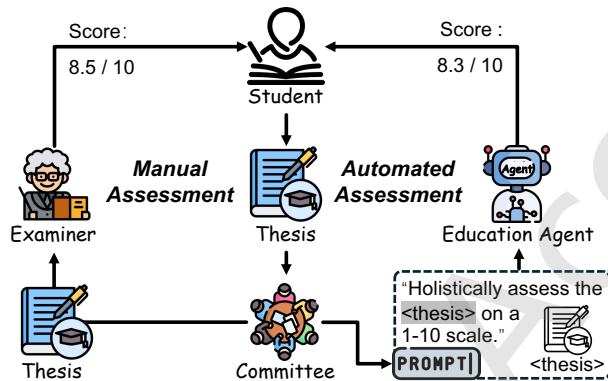


Fig. 1 Comparison of manual and automated methods for UGTE assessment. Manual assessment typically yields a holistic score. Current education agents improve scalability and consistency but still rely on generic prompts and lack criterion-specific feedback.

To tackle these challenges, we propose the Pedagogically-Enriched Multi-granular Undergraduate Thesis Assessment (PEMUTA), a novel framework that guides LLMs in mining domain-specific knowledge to deliver both criterion-level evaluations and synthesized holistic judgments. However, there remains a lack of standardized fine-grained evaluation dimensions tailored specifically for UGTE. To fill this gap, we draw inspiration from two foundational pedagogical theories widely adopted in manual UGTE assessment: Vygotsky's theory and Bloom's Taxonomy [20, 19]. Vygotsky's theory emphasizes the developmental and epistemological aspects of UGTE, and Bloom's Taxonomy offers a structured hierarchy of cognitive skills used in instructional design and assessment. Building on these insights, we define six pedagogically grounded dimensions: Structure, Logic, Originality, Writing, Proficiency, and Rigor, collectively abbreviated as SLOWPR. LLMs are prompted to evaluate UGTEs across multiple levels of granularity, leveraging their pretrained language and domain knowledge to produce criterion-aligned, multi-granular feedback. By integrating both cognitive and developmental perspectives, PEMUTA

improves alignment with expert evaluations while enhancing interpretability, pedagogical relevance, and diagnostic value.

Additionally, the multi-granular assessment typically involves processing rich, dimension-specific instructions, which poses challenges to the direct application of LLMs [24, 25, 26]. The extensive information embedded within detailed prompts may lead to attention distraction across subtasks, misalignment of criteria-specific knowledge across dimensions, and ultimately reduce overall coherence. To address this issue, we propose a hierarchical prompting strategy that guides the model through a structured two-stage evaluation process: first, performing dimension-level assessments, then synthesizing a holistic judgment. This design improves alignment with human evaluators while enhancing interpretability, consistency, and pedagogical relevance. To further improve alignment, we incorporate two in-context learning techniques: few-shot prompting, which supplies exemplars structured according to the proposed multi-granular framework [5, 27, 28], and role-play prompting, which conditions the model to act as an experienced academic evaluator [29, 30, 31]. These techniques together enhance assessment performance without additional fine-tuning. Extensive experiments demonstrate that PEMUTA achieves strong agreement with expert ratings across both fine-grained and holistic UGTE evaluation.

Our contributions can be summarized as follows. 1) We rethink automated UGTE assessment and pioneer PEMUTA, a pedagogically-enriched hierarchical prompting framework for structured evaluations across varying levels of granularity. 2) We integrate Vygotsky's theory and Bloom's Taxonomy into the assessment process, which guides the LLM to evaluate UGTEs across six critical SLOWPR dimensions, eliciting rich academic priors to produce criterion-based, pedagogically-grounded evaluations that reflect both academic performance and developmental potential. 3) We develop an In-context Enhanced Hierarchical Prompting scheme, integrating few-shot examples and role-playing to better align with expert judgment without fine-tuning. It decomposes UGTE assessment into dimension-specific evaluations before holistic synthesis, activating criteria-specific reasoning and enabling interpretable, multi-level assessment. 4) The Multi-granular Undergraduate Thesis Assessment (MUTA) dataset is curated to facilitate comprehensive UGTE evaluation in engineering disciplines. Experimental results indicate that PEMUTA achieves strong alignment with expert human assessment across both holistic and fine-grained dimensions, demonstrating consistent efficacy with various LLMs and engineering UGTE scenarios.

2 Related Work

2.1 Automated Essay Scoring

Automated essay scoring automatically evaluates and scores student essays in educational contexts, producing valid and efficient scores that are consistent with human ratings [32, 4, 5]. AES systems typically treat the task as a classification problem, wherein essays are mapped to predefined score levels [32, 18]. Early AES systems predominantly relied on heuristic or rule-based approaches and manually engineered features. These systems applied statistical learning algorithms such as linear regression, support vector machines (SVMs), decision trees, and ensemble models to estimate

overall scores through a weighted combination of sub-scores representing aspects such as grammar, vocabulary, organization, and coherence [32, 18]. E-rater, developed by the Educational Testing Service (ETS), is one of the most influential early systems. It incorporates approximately fifty predefined linguistic features derived from syntax, discourse structure, and lexical usage, combined with regression-based models for score prediction [6]. Another widely used system, IntelliMetric by Vantage Learning, utilizes over four hundred features spanning semantic, syntactic, and rhetorical levels. It processes input through multi-stage pipelines that include hierarchical linguistic parsing and statistical modeling [33]. These systems benefit from interpretability, since their predictions are based on explicit and well-defined features. However, they also face significant limitations. Hand-crafted features require extensive domain expertise and manual effort to design and adapt. Moreover, rule-based models often fail to capture deeper semantics, such as logical coherence, argument structure, and contextual relevance [10, 34]. These limitations have motivated a shift towards data-driven approaches that automatically learn high-level representations from large-scale corpora [35, 8, 36].

Deep learning has significantly advanced AES by learning scoring-related representations directly from text [37, 36, 34]. Early neural methods used recurrent neural networks and convolutional neural networks (CNNs) to capture semantic and syntactic patterns. Taghipour and Ng combined convolutional layers over word embeddings with a bi-directional long short-term memory (LSTM) to model local n-gram patterns and long-range dependencies [7]. Alikaniotis *et al.* introduced score-specific word embeddings and used a bi-directional LSTM to link contextual word usage with scoring outcomes [37]. Dong and Zhang proposed a hierarchical two-layer CNN to extract word- and sentence-level representations [8]. Recent transformer-based models and attention mechanisms further improve AES by modeling global context and long-distance dependencies [36, 34]. Wang *et al.* proposed a multi-scale BERT framework that jointly learns token-, segment-, and document-level representations for essay evaluation [10]. Faseeh *et al.* combined handcrafted linguistic features, contextualized RoBERTa embeddings, and an XGBoost regressor to enhance both predictive performance and interpretability [36]. Uto *et al.* further incorporated Item Response Theory and aggregated multiple AES models via the Generalized Many-Facet Rasch Model (GMFRM) [9]. However, these methods often require complex architectures, high computational costs, and large-scale annotated datasets, limiting their practical use and generalization in data-scarce educational settings [34].

Traditional AES methods have demonstrated notable progress, but they heavily rely on manual feature engineering or task-specific supervision, reducing their adaptability to complex undergraduate thesis assessment. Recent advances in large language models present a promising alternative by enabling flexible, prompt-based assessment paradigms.

2.2 LLMs in Automated Assessments

The emergence of LLMs has introduced a new paradigm in educational assessment [14, 11, 38]. Pre-trained on massive text corpora, these models encode rich linguistic knowledge and demonstrate strong capabilities in capturing deep semantics, discourse structure, and contextual coherence, all of which are

critical for effective essay evaluation [5]. Without requiring additional fine-tuning, LLMs can be adapted to essay scoring via prompt engineering. Typical prompting strategies include: 1) few-shot prompting, which provides task instructions along with a few labeled exemplars, and 2) zero-shot prompting, which uses only task instructions. These approaches significantly reduce the need for labeled datasets and hand-crafted features that traditional AES systems depend on [5, 11, 4, 39].

However, LLM-based solutions face several limitations. Most methods generate only a holistic score, which limits interpretability and provides weak support for pedagogical use [40, 41]. Yet submissions with identical holistic scores may differ substantially across evaluation dimensions [21, 22], indicating the need for interpretable, multi-granular UGTE assessment. Although LLMs can support rubric-aligned scoring, their effectiveness depends on well-structured prompts. Without domain-specific cues, they may rely on generic or irrelevant knowledge and produce misaligned assessments [14, 17]. This problem is more evident in thesis evaluation, where assessment requires discipline-specific knowledge, alignment with educational goals, and explainable reasoning [14, 39].

To address these challenges, recent studies focus on managing prompt structure [42, 43, 44]. Some researchers modularize detailed prompts into segments processed sequentially or conditionally, enabling the model to maintain focus on localized information before integrating global understanding [26, 23, 45]. Liu *et al.* introduced a hierarchical prompt learning strategy (HiPro), which quantifies inter-task affinity and constructs a task tree to capture task cross-granularity dependencies [45]. Cohn *et al.* further enhanced prompt design by integrating human-in-the-loop mechanisms. They combined few-shot and active learning techniques with chain-of-thought prompting to improve explanation quality in formative K–12 science assessments [23]. Together, these approaches highlight the crucial role of structured, context-aware prompt design in reducing distraction while enhancing the reliability, focus, and interpretability of LLM-based assessments.

3 Proposed Methodology

3.1 Overview of Proposed Method

This paper proposes PEMUTA, a pedagogically-enriched framework for multi-granular assessment of undergraduate theses. The system architecture is illustrated in Fig. 2. To support multi-granular evaluation, our proposed framework begins by designing a fine-grained evaluation scheme, SLOWPR, which establishes six assessment dimensions: Structure, Logic, Originality, Writing, Proficiency, and Rigor. These dimensions are grounded in the integration of Bloom's Taxonomy and Vygotsky's sociocultural theory, forming the basis for subsequent evaluation and data analysis, and ensuring both demonstrated academic performance and developmental potential are systematically captured. The conceptual rationale and detailed definitions of each SLOWPR dimension are presented in Section 3.2. In parallel, raw UGTEs in PDF format are converted into model-compatible plain text via a dedicated pipeline comprising two sequential stages: layout-aware content extraction and semantic document reconstruction. The output is a structured JSON representation that preserves both structural integrity and semantic coherence of the original

UGTE. Implementation details are described in Section 4.

Then, a hierarchical prompting scheme guides the LLM through a two-stage evaluation process. In the first stage, the model performs dimension-specific assessments, generating criterion-aligned scores and justifications for each SLOWPR aspect. To improve alignment with expert evaluations, two in-context learning techniques are incorporated: 1) *few-shot prompting*, supplying formatted UGTE exemplars matching the expected multi-granular output structure without accompanying fine-grained justifications or holistic feedback [5, 27, 28]; 2) *role-play prompting*, conditioning the model to assume the

role of a thesis committee member [29, 30, 31].

The final composite prompt consists of: (i) the preprocessed thesis, (ii) SLOWPR-grounded hierarchical instructions, (iii) few-shot formatted exemplars, and (iv) role-specific instructions. This prompt is subsequently submitted to a capable LLM for assessment. The resulting output is a standardized evaluation report comprising: (a) fine-grained scores with supporting justifications, and (b) a holistic score accompanied by revision suggestions. Overall, PEMUTA offers a scalable, interpretable, and pedagogically grounded solution for comprehensive UGTE assessment.

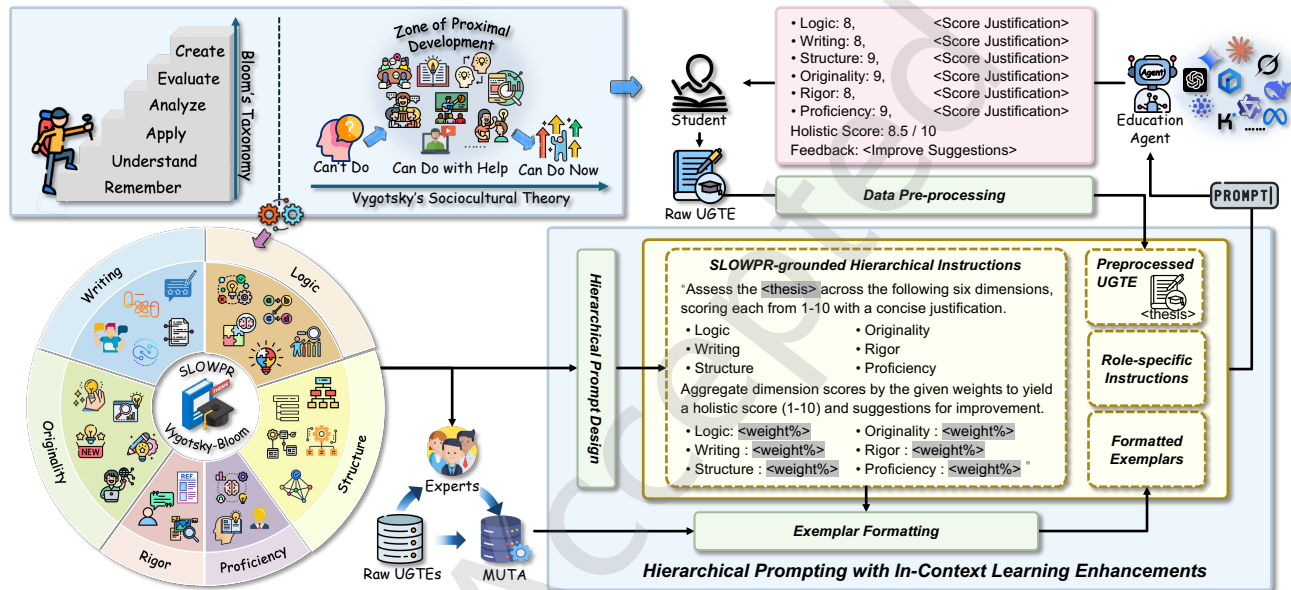


Fig. 2 Overview of the proposed PEMUTA framework for undergraduate thesis assessment across multiple levels of granularity.

3.2 Pedagogically-Grounded Fine-grained Dimensions

Pedagogical theories are often implicitly reflected in UGTE evaluations, guiding assessors in judging diverse fine-grained competencies such as reasoning quality, writing proficiency, and methodological rigor [20, 46, 47]. Existing LLM-based systems [5, 4] offer scalable and efficient solutions; however, they predominantly rely on simple standard prompts such as ‘Holistically assess the given thesis on a 1-10 scale.’ to generate a numerical score. Their responses often diverge from human judgments. We attribute this performance gap to the underutilization of domain-specific knowledge priors and the lack of structured reasoning in LLMs. Additionally, the holistic assessment paradigm often fails to provide actionable feedback that supports student learning and improvement. To address these limitations, we pioneer a multi-granular UGTE framework that evaluates UGTEs across both fine-grained dimensions and holistic outcomes. The definition of these fine-grained dimensions is informed by established pedagogical theories that reflect how students develop academic competence throughout their undergraduate education.

Vygotsky's sociocultural theory emphasizes the role of scaffolded learning in the transition from externally supported activities to independent academic competence [19]. Undergraduate students engage in various guided academic

practices, such as supervised research, collaborative problem-solving, and formal academic writing, that help internalize scholarly norms and skills. These experiences contribute not only to the acquisition of disciplinary knowledge but also to the broader educational goal of preparing graduates for professional practice and lifelong learning. As the capstone of undergraduate education, the UGTE should therefore be assessed not only in terms of its immediate academic quality but also in terms of how it reflects a student's developmental progression toward scholarly independence.

Table 1 Comparison of Bloom's Taxonomy and Vygotsky's Theory in the Context of UGTE Assessment

Theory	Strengths	Limitations
Bloom	Structured hierarchy of cognitive processes; Widely adopted in rubric design and educational assessment.	Presumes linear progression; Overlooks sociocultural or developmental variability.
Vygotsky	Highlights learning potential and scaffolded development; Captures dynamic progression over time.	Lacks formalized structures for assessment; Less directly applicable to rubric-based evaluation.

Complementing this developmental view, Bloom's Taxonomy offers a structured hierarchy of cognitive abilities: Remember, Understand, Apply, Analyze, Evaluate, and Create, which has been widely adopted in thesis assessment rubrics^[20]. However, its assumption of linear progression does not always align with real-world student performance, e.g., a student may exhibit sophisticated analytical reasoning while struggling with foundational application. Moreover, Bloom's framework pays little attention to sociocultural influences and individual learning trajectories. Consequently, UGTE evaluations based solely on Bloom's Taxonomy risk underrepresenting emerging capabilities and fail to recognize developmental efforts.

Table 1 summarizes the strengths and limitations of

Bloom's Taxonomy and Vygotsky's sociocultural theory in the context of UGTE assessment. These pedagogical theories provide complementary perspectives on cognitive performance and learning development that are highly relevant to UGTE evaluation. Bloom's Taxonomy offers a structured view of cognitive processes commonly reflected in rubric-based assessment, whereas Vygotsky's theory emphasizes scaffolded development and learning potential in supervised undergraduate research. Synthesizing insights from both perspectives, we develop a pedagogically informed fine-grained evaluation framework that reflects established UGTE assessment practices while remaining grounded in learning theory.

Table 2 Key Aspects Evaluated and Theoretical Justification of the Fine-grained Evaluation Dimensions

Dimension	Key Aspects Evaluated	Theoretical Justification (Motivation/Vygotsky/Bloom)
Structure	Organization of chapters; Sectional hierarchy; Structural cohesion	Motivation: A well-structured UGTE facilitates clear presentation of objectives, methodology, and outcomes, evidencing the student's capacity for systematic research planning and professional academic communication. Vygotsky: Students develop the structural competence for UGTE through scaffolded practice under supervision and the gradual internalization of organizational conventions. Bloom: Students <i>Understand</i> formal academic conventions and <i>Apply</i> them in structuring a coherent and systematically organized UGTE.
Logic	Consistency of reasoning; Alignment among research questions, methods, and conclusions	Motivation: Logical consistency ensures the alignment between research objectives, methodology, analysis, and argumentation. Students should demonstrate their capacity for coherent, evidence-based arguments. Vygotsky: Students develop logical reasoning through sustained academic interactions such as guided supervision and peer review, supporting the transition from assisted to independent argument construction in the UGTE. Bloom: Students <i>Analyze</i> complex research problems, methods, and data, <i>Evaluate</i> the validity of evidence and reasoning, and ultimately <i>Create</i> a UGTE with coherent reasoning chains.
Originality	Novel perspectives; Innovative theoretical or methodological contributions; Independent intellectual engagement	Motivation: As a culminating academic outcome, UGTE reflects not only mastery of knowledge but also students' intellectual independence and creative synthesis of disciplinary knowledge. Vygotsky: Through socially mediated learning such as supervisory interaction and academic dialogue, students internalize disciplinary reasoning and develop creative potential, enabling them to reformulate existing ideas into novel conceptual or methodological contributions. Bloom: Students <i>Analyze</i> established ideas, <i>Evaluate</i> existing alternatives, and <i>Create</i> new insights or solutions that demonstrate the creative competence expected in UGTE performance.
Writing	Clarity of expression; Grammatical precision; Adherence to academic style and disciplinary conventions	Motivation: The UGTE is a formal academic document whose quality depends not only on research content but also on clear, precise academic writing that meets disciplinary standards and avoids linguistic ambiguity or colloquialism. Vygotsky: Students' academic writing competence develops through iterative supervisory feedback and exposure to scholarly discourse, where external linguistic scaffolds are gradually internalized into autonomous scholarly expression. Bloom: Students <i>Remember</i> linguistic rules, <i>Understand</i> stylistic norms, and <i>Apply</i> them to construct a precise, discipline-appropriate UGTE.
Proficiency	Mastery of disciplinary knowledge; Use of technical terminology; Problem-solving and methodological competence	Motivation: UGTEs are designed to assess cumulative disciplinary proficiency, verifying whether students can independently apply accumulated theoretical and methodological knowledge to authentic academic or practical contexts. Vygotsky: Through scaffolded experiences such as supervision, academic discussion, guided use of disciplinary tools, and the transfer of knowledge across increasingly complex tasks, students internalize technical proficiency essential for UGTE completion. Bloom: Students <i>Understand</i> disciplinary concepts, <i>Apply</i> them in research or practical contexts, and <i>Analyze</i> domain-specific problems with methodological accuracy.
Rigor	Source reliability; Citation accuracy; Academic integrity and ethical compliance	Motivation: Academic rigor is a foundational criterion in UGTE evaluation, ensuring that work meets institutional and ethical research standards while maintaining scholarly credibility. Vygotsky: Through guided exposure to exemplary academic practices and institutional regulations, students internalize ethical and methodological rigor, leading to autonomous regulation of academic behavior throughout the UGTE process. Bloom: Students <i>Remember</i> ethical principles, <i>Understand</i> academic citation and integrity conventions, and <i>Apply</i> them consistently across all stages of UGTE research and writing.

The construction of SLOWPR follows a theory-informed design process. Commonly adopted UGTE evaluation criteria in institutional rubrics and supervisory practices are first identified, and then examined through the perspectives of cognitive development and scaffolded learning emphasized by Bloom and Vygotsky, respectively. Dimensions that consistently reflect both observable thesis quality and students' developmental academic competence are retained and operationalized through concrete indicators to support reproducibility and systematic evaluation. The resulting framework comprises six core evaluation dimensions, detailed as follows.

- **Structure:** Evaluates the overall organization and coherence of the UGTE, including the logical arrangement of chapters and the consistency of structural progression. This dimension reflects the student's ability to plan, manage, and present academic content systematically and coherently, which is commonly emphasized in institutional UGTE rubrics.
- **Logic:** Examines the internal consistency and reasoning flow across research objectives, methodology, evidence, and conclusions. This dimension focuses on the student's capacity to construct coherent arguments and maintain overall analytical rigor, reflecting higher-level cognitive processing expected in undergraduate research.
- **Originality:** Assesses the degree of intellectual independence and novelty demonstrated in the UGTE, such as innovative perspectives, problem framing, or methodological creativity. This dimension reflects the expectation that UGTEs demonstrate independent thinking beyond the reproduction of existing knowledge.
- **Writing:** Concerns the clarity, precision, and appropriateness of academic expression throughout the UGTE, including sentence formulation, terminology usage, and adherence to disciplinary writing conventions. This dimension focuses on the student's ability to communicate research ideas effectively and unambiguously in a formal academic context.
- **Proficiency:** Measures the student's command of disciplinary knowledge and methodological skills, including the use of technical concepts, analytical techniques, and domain-specific tools. This dimension assesses whether the student can apply relevant knowledge and methods to address the research problem in the UGTE.
- **Rigor:** Evaluates the extent to which the UGTE adheres to academic standards and scholarly norms, including the reliability of sources, correctness of citations, and consistency with academic integrity requirements. This dimension ensures that the thesis meets baseline expectations for ethical conduct and methodological soundness in undergraduate research.

These six dimensions constitute the fine-grained criteria for UGTE assessment, capturing both the cognitive depth reflected in UGTE content and development-related characteristics manifested through students' academic writing. Among them, Structure, Logic, Originality, and Writing serve as the core indicators of UGTE quality, whereas Rigor and Proficiency

function as supporting indicators that ensure disciplinary validity and academic compliance. Table 2 provides a compact summary of each dimension's evaluative focus and its theoretical rationale, thereby linking these operational assessment aspects to their underlying conceptual foundations.

3.3 Hierarchical Prompting with In-Context Learning Enhancements

Multi-granular assessment within a single input requires combining diverse criteria and task-specific instructions into an information-dense prompt. However, such prompts may distribute the model's attention across too many dimensions, weakening criterion-specific reasoning and reducing output coherence [24, 26, 25]. We therefore propose a hierarchical two-stage prompting mechanism that decomposes UGTE assessment into dimension-specific evaluation and holistic synthesis.

Firstly, the model conducts independent evaluations for each of the six SLOWPR criteria. For each dimension $d \in \{S, L, O, W, P, R\}$, the model generates a tuple (y_i^d, j_i^d) , where y_i^d denotes the numerical score and j_i^d the corresponding criterion-specific justification for the i -th sample. This decomposition reduces cross-dimensional interference and enables more focused activation of domain-relevant knowledge.

Secondly, the model synthesizes the six fine-grained assessments into a coherent, holistic evaluation. The overall score y_i^H is computed as a weighted sum:

$$y_i^H = \sum_{d \in \mathcal{D}} w^d \cdot y_i^d, \quad \mathcal{D} = \{S, L, O, W, P, R\}, \quad (1)$$

where $w^d \in [0, 1]$ represents the weight for each dimension, and $\sum_d w^d = 1$. These weights can be configured to reflect disciplinary emphasis or institutional rubrics, offering flexibility across different academic contexts. The model also produces formative feedback by integrating prior dimension-level evaluations with holistic insights, offering practical suggestions for student improvement.

To enhance alignment with expert judgment without fine-tuning, we augment SLOWPR-based hierarchical prompts with two in-context techniques: *few-shot prompting* and *role-play prompting*. Specifically, we select representative UGTE samples and format them to match the expected multi-granular output structure, including six SLOWPR dimension scores and a holistic rating. We exclude textual rationales, as the ground-truth annotations consist solely of numerical grades. Rather than relying on direct instructions, these few-shot exemplars provide pattern-based guidance, serving as an explicit chain-of-thought (CoT) trigger that helps the model internalize the structure and format of rubric-aligned evaluations. *Role-play prompting* functions as an implicit CoT trigger, shaping model behavior through persona conditioning. Prior studies show that assigning a role-specific persona helps activate domain-relevant knowledge and structured reasoning patterns [5, 29, 30, 31]. In addition, this role conditioning narrows the model's activation space, further reducing ambiguity and encouraging coherent, expert-like responses. Accordingly, we instruct the model to assume the role of an experienced UGTE committee member, prompting it to adopt a formal academic tone, discipline-appropriate vocabulary, and expert evaluative reasoning aligned with institutional expectations.

4 Construction of MUTA Dataset

To support the multi-granular UGTE assessment task, we constructed the Multi-granular Undergraduate Thesis Assessment (MUTA) dataset. This dataset comprises 210 authentic undergraduate theses: 105 collected from Computer Science undergraduates at Southern University of Science and Technology (SUSTech) during the 2022 and 2025 academic years, and 105 collected from Changchun University (CCU) in 2025. The CCU collection covers three disciplines, *i.e.*, Automation, Robotics, and Communication, with 35 UGTEs in each. The SUSTech subset is denoted as ‘MUTA-S’, while the three CCU subsets are denoted as ‘MUTA-C1’, ‘MUTA-C2’, and ‘MUTA-C3’, corresponding to Automation, Robotics,

and Communication, respectively. SUSTech and CCU represent research-intensive and application-oriented engineering education settings, respectively, allowing us to examine UGTE assessment across different institutional contexts and disciplines within engineering education. All research procedures were approved by the Ethics Committee of both universities and conducted in accordance with institutional guidelines and the Declaration of Helsinki guidelines. Written informed consent was obtained from all participants. All personally identifiable information, including student names, ID numbers, and supervisor details, was fully anonymized. The constructed MUTA dataset is available exclusively for non-commercial academic use.

Table 3 Summary of prior studies on thesis-related tasks. ‘UG’, ‘M’, and ‘P’ represent undergraduate, master’s, and PhD levels, respectively. ‘N/A’ denotes ‘not applicable’.

Paper	Year	Academic Level	Data Availability	Number of Theses	Number of Universities	Number of Disciplines	Holistic Assessment	Automated Fine-grained Assessment	Automated Full Thesis Evaluated	LLM-based Research	Training Free
Proposed	2025	UG	Restricted	210	2	4	Scoring	✓	✓	✓	✓
Yang <i>et al.</i> [48]	2025	M, P	Open-Source	2,000	–	–	Ranking	✗	✓	✗	✗
Ge <i>et al.</i> [49]	2024	M	Close-Source	275	–	–	–	✗	✗	✗	N/A
Lostanau <i>et al.</i> [50]	2022	UG	Close-Source	200	1	1	–	✗	✗	✗	✗
Haagsman <i>et al.</i> [21]	2021	UG	Close-Source	318	1	1	–	✗	✗	✗	N/A
Zahrotun <i>et al.</i> [51]	2018	UG	Close-Source	138	1	1	–	✗	✗	✗	✓
Gorrostieta <i>et al.</i> [52]	2017	UG,M,P	Open-Source	300	–	1	–	✗	✗	✗	✗
Gorrostieta <i>et al.</i> [53]	2016	UG,M,P	Open-Source	46	–	1	–	✗	✗	✗	✗
Xu <i>et al.</i> [54]	2016	M,P	Close-Source	2,007	–	–	Ranking	✗	✗	✗	✗
Li <i>et al.</i> [55]	2014	UG	–	10	–	–	Ranking	✗	✗	✗	✗
Mao <i>et al.</i> [56]	2014	UG	Close-Source	90	1	1	Ranking	✗	✗	✗	✗
Mosallam <i>et al.</i> [57]	2014	UG,M	Close-Source	361(UG) 202(M)	1	1	Scoring	✗	✗	✗	✗
Reynolds & Thompson [58]	2011	UG	Close-Source	190	1	1	–	✗	✗	✗	N/A

Table 3 compares MUTA with prior datasets and studies on thesis-related tasks. Existing UGTE studies primarily rely on high-level assessment records, such as rubric sub-scores, overall grades, or brief textual descriptors [49, 50, 21], and mainly analyze score-related factors rather than support automated full-text, multi-granular assessment. These approaches typically use shallow inputs like titles or coarse numerical ratings, without using full-text thesis content for detailed analysis [54, 51, 56]. For example, Xu *et al.* classified 2,007 samples into five quality levels using only score sets, omitting linguistic and structural information from full documents [54]. Yang *et al.* incorporated linguistic content with deep neural networks for master- and PhD-level thesis assessment, but their method produced only binary rankings and lacked support for multi-dimensional assessment or continuous score prediction [48]. In contrast, MUTA provides full-text UGTEs with expert annotations at both fine-grained and holistic levels, enabling comprehensive multi-granular evaluation.

In MUTA, each entry contains the original UGTE in PDF format together with evaluation ratings on a 0–10 scale at multiple levels of granularity. Each UGTE is independently evaluated by three of six experienced faculty members from the two contributing universities, all with at least three years of teaching or UGTE supervision experience. Following the SLOWPR framework, evaluators assign 0–10 ratings across

six fine-grained dimensions and one holistic score. The three scores for each fine-grained dimension are averaged to obtain the corresponding reference score, thereby incorporating assessments from multiple experts. Two holistic measures are subsequently derived: 1) the direct holistic score (DHS), calculated as the mean of the three expert-assigned holistic scores; and 2) the weighted holistic score (WHS), obtained by weighted aggregation of the averaged dimension-level ratings.

The statistical reliability of the SLOWPR framework is examined using one-way repeated-measures analysis of variance (ANOVA), the Kaiser–Meyer–Olkin (KMO) measure, Bartlett’s test of sphericity, and Cronbach’s α across all four subsets. Table 4 summarizes the results. Across all subsets, the ANOVA outcomes show statistically significant differences among the six SLOWPR dimensions. The largest F-statistic reaches 16.845 in MUTA-S, and the smallest value is 4.959 in MUTA-C2, with all p-values less than 0.001. These results confirm that the six proposed dimensions capture distinct scoring patterns in UGTE evaluation. The KMO values remain high across all subsets, with a maximum of 0.922 in MUTA-C2 and a minimum of 0.812 in MUTA-C3. Bartlett’s test is statistically significant in every group, with chi-square values ranging from 157.963 in MUTA-C3 to 621.949 in MUTA-S, and all p-values below 0.001. These findings demonstrate strong inter-dimensional correlation and support the dimensional

validity of the SLOWPR framework. Furthermore, Cronbach's α consistently remains above 0.9, reaching 0.969 in MUTA-C2 and not falling below 0.917 in MUTA-C3. This indicates stable internal consistency and shows that the six dimensions provide complementary and coherent aspects of UGTE quality.

Table 4 Statistical validation of the SLOWPR six-dimensional framework across four subsets. ANOVA examines mean-score differences among the six dimensions. KMO and Bartlett's tests assess suitability for factor analysis for construct validity. Cronbach's α measures internal consistency of the six dimensions.

Subset	ANOVA		KMO	Bartlett's		Cronbach's α
	F-statistic	p-value		χ^2	p-value	
MUTA-S	16.845	<0.001	0.909	621.949	<0.001	0.946
MUTA-C1	9.698	<0.001	0.819	194.566	<0.001	0.927
MUTA-C2	4.959	<0.001	0.922	268.686	<0.001	0.969
MUTA-C3	5.699	<0.001	0.812	157.963	<0.001	0.917

Table 5 presents the Pearson correlation coefficient (PCC), t-statistic, and p-value between DHS and WHS. WHS exhibits strong agreement with DHS, with PCC values above 0.89 across all subsets. The high t-statistics and extremely low p-values further confirm that the WHS reliably reflects human holistic judgments. Table 6 further compares both holistic

measures using the mean PCC between DHS/WHS and the six SLOWPR dimensions. Compared with DHS, WHS yields higher correlations with the fine-grained dimensions (Table 6) and smaller standard deviations across all subsets (Table 7), indicating a more stable and interpretable representation of overall performance. WHS is therefore adopted as the reference holistic measure for subsequent analyses.

Table 5 Statistical comparison between DHS and WHS across four MUTA subsets using PCC, t-statistic, and p-value.

Metric	MUTA-S	MUTA-C1	MUTA-C2	MUTA-C3
PCC	0.98	0.96	0.98	0.97
t-statistic	46.39	20.16	26.28	20.55
p-value	<0.001	<0.001	<0.001	<0.001

Table 6 Comparison of DHS and WHS on mean PCC with the SLOWPR fine-grained dimensions across four MUTA subsets.

Holistic Score Type	MUTA-S	MUTA-C1	MUTA-C2	MUTA-C3
Direct Holistic Score	0.87	0.80	0.89	0.82
Weighted Holistic Score	0.89	0.86	0.93	0.85

Table 7 Statistics of expert-assigned scores across fine-grained dimensions and overall holistic evaluation in the four subsets of the MUTA dataset.

Dimension	MUTA-S			MUTA-C1			MUTA-C2			MUTA-C3		
	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max
Structure	8.24±0.96	5.50	9.50	8.75±0.83	6.50	10.00	8.12±1.16	5.50	10.00	7.27±1.43	5.40	9.50
Logic	8.34±0.75	6.00	9.50	8.51±0.66	6.40	9.50	7.09±1.01	5.00	9.00	7.63±1.08	5.00	9.80
Originality	8.20±0.84	5.00	9.50	8.05±0.60	6.50	9.00	7.78±1.06	5.50	9.50	6.75±0.98	5.50	8.90
Writing	8.16±0.91	5.50	9.50	8.23±0.60	7.00	9.50	7.51±1.28	5.00	9.50	8.24±0.95	5.50	9.50
Proficiency	9.07±0.70	6.50	10.00	9.04±0.85	7.40	10.00	8.23±1.13	5.50	10.00	7.40±1.19	5.20	9.80
Rigor	8.43±0.92	6.00	10.00	8.66±0.62	7.50	10.00	7.65±1.17	5.80	10.00	7.59±1.24	5.00	9.00
Weighted Holistic	8.34±0.77	6.05	9.60	8.48±0.60	6.87	9.30	7.69±1.06	5.63	9.50	7.48±0.98	5.50	9.40
Direct Holistic	8.34±0.79	5.50	9.60	8.67±0.54	7.00	9.50	8.08±1.02	6.00	9.80	7.60±1.09	5.80	9.60

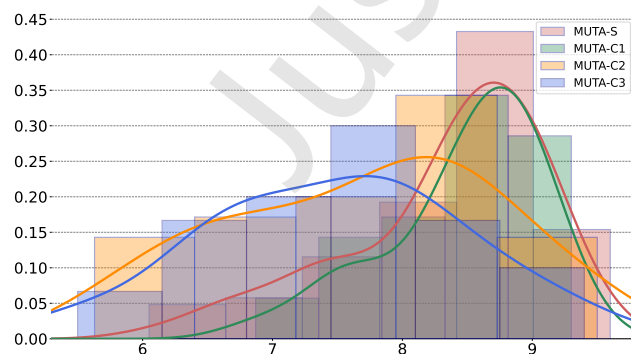


Fig. 3 Holistic score distributions of all four collected subsets.

Fig. 3 shows the distributions of the holistic scores across all four subsets. Each distribution is visualized with a histogram and its corresponding kernel density estimation (KDE) curve, both normalized by relative frequency to facilitate direct comparison across datasets. All subsets exhibit approximately Gaussian-like distributions, which supports the rationality of our SLOWPR framework and the WHS aggregation method.

Differences in peak locations and spread among the subsets also reflect variations in student performance across universities and disciplines. Table 7 summarizes the multi-granular scoring statistics. Using MUTA-S as an example, most holistic scores fall between 7.8 and 9.0, with the highest concentration in the 8.4–9.0 range. The mean WHS for MUTA-S is 8.34 with standard deviation 0.77, and values range from 6.05 to 9.60. The average scores of the six fine-grained dimensions are as follows: *Structure* – 8.24, *Logic* – 8.34, *Originality* – 8.20, *Writing* – 8.16, *Proficiency* – 9.07, and *Rigor* – 8.43. Among these, *Structure* exhibits the highest standard deviation ($\sigma = 0.96$), whereas *Proficiency* shows the lowest ($\sigma = 0.70$).

The collected theses are stored in PDF format, which is unsuitable for LLM processing. Most LLMs require UTF-8 plain text with a linear and logically ordered structure. In contrast, PDFs are encoded for visual presentation rather than linguistic parsing. Each page is rendered as a visual canvas with text blocks, figures, tables, and other elements positioned by absolute coordinates. Although visually organized for human readers, PDFs do not preserve the semantic and structural logic

required for reliable language model processing [59].

Common extraction tools such as pdfplumber often fail to recover this logic. In academic theses, repeated visual elements including headers, footers, and page numbers are frequently merged into the main text stream, interrupting sentence continuity and introducing noise. In addition, UGTEs often contain complex layouts such as interleaved text and figures, multi-level lists, and nested tables, which further degrade the coherence of extracted content. Consequently, the raw text obtained from PDFs is typically disordered and unsuitable for direct use in LLM-based evaluation. Each collected UGTE thus requires pre-processing to reconstruct a semantically coherent and structurally consistent plain-text representation.

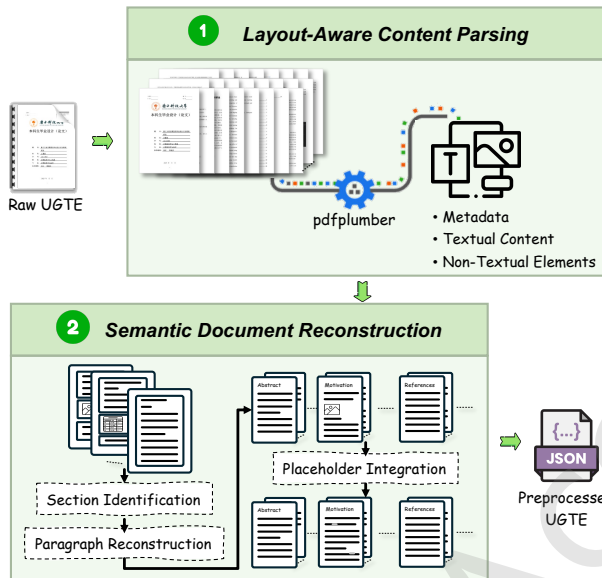


Fig. 4 Overview of the UGTE data pre-processing pipeline. Consisting of two sequential stages, i.e., layout-aware content parsing and semantic document reconstruction, it transforms PDF-formatted theses into structured, JSON-based plain-text representations suitable for LLM processing.

To address these issues, a text-centered pre-processing pipeline is developed, as illustrated in Fig. 4. Following common practice in existing automated UGTE and essay assessment studies [5], it converts PDF-formatted UGTEs into plain-text representations compatible with LLM interfaces. The overall process includes the following steps.

Layout-Aware Content Parsing: We use pdfplumber to extract content elements and layout metadata from PDF files. The extracted information includes text elements such as titles, paragraphs, and footnotes, non-text elements such as tables and figures, and metadata such as font attributes and absolute positions. These features help distinguish content types and structural roles. To ensure compatibility with LLMs that process text only, each non-text element is replaced with a semantic placeholder, together with its original caption and identifier. This helps preserve its semantic role and reference function without retaining raw visual content.

Semantic Document Reconstruction: Due to the absolute positioning mechanism used in PDFs, the output of layout-aware parsing is often fragmented or misordered. To restore logical coherence, section boundaries are identified using common academic headings such as ‘Abstract’ and ‘References’.

along with numeric section patterns. Fragmented text lines are then merged into complete paragraphs based on spacing, indentation, and punctuation cues. Semantic placeholders corresponding to figures and tables are inserted at appropriate textual locations according to their coordinates and surrounding references. This step preserves the document’s original argument flow and structure for subsequent processing.

Structured JSON Output: The final output consists of the reconstructed text and associated metadata, for instance, thesis title, section labels, and figure and table references. All information is serialized in a structured JSON format to support consistent downstream evaluation.

By adopting this text-centered pre-processing pipeline, each thesis is transformed into a semantically coherent and structurally faithful format suitable for LLM-based evaluation. While the internal graphical logic of diagrams such as flow charts is not directly evaluated at the visual level, their academic and logical roles are indirectly reflected through captions, in-text explanations, and methodological consistency.

5 Experimental Results

5.1 Experimental Settings

LLMs for Evaluation

To evaluate the effectiveness of PEMUTA, we test it on diverse state-of-the-art LLMs. Since multi-granular UGTE assessment requires document-level understanding, sufficient context length is essential. Therefore, models that cannot process full-length UGTEs, typically exceeding 8,000–10,000 tokens, are excluded.

Qwen-series: Developed by Alibaba, the Qwen series is a family of Transformer-based LLMs optimized designed for language understanding, generation, reasoning, and multilingual tasks. We select three representative models, i.e., *Qwen-Plus* [60], *Qwen2.5-14B-1M* [61], and *Qwen-Turbo-Latest* [61]. *Qwen-Plus*, with a maximum context window of 131K tokens, is officially recommended for its balanced trade-off among output quality, inference speed, and cost-efficiency. *Qwen2.5-14B-1M* and *Qwen-Turbo-Latest* extend the context window to 1 million tokens. Both *Qwen-Plus* and *Qwen-Turbo-Latest* adopt a Mixture-of-Experts (MoE) architecture to support faster inference and reduced computational cost. They exhibit strong capabilities in instruction following, commonsense reasoning, and multilingual understanding.

DeepSeek-series: Released by DeepSeek AI, this series features cost-efficient inference and strong reasoning performance. We employ two representative models, i.e., *DeepSeek-V3* [62] and *DeepSeek-R1* [63], both of which adopt MoE architectures and support 128K-token contexts for long UGTE inputs. *DeepSeek-V3* serves as a general-purpose model, excelling in writing, summarization, and multilingual translation, well-suited for holistic UGTE quality assessment. *DeepSeek-R1* is engineered for reasoning-intensive tasks and supports chain-of-thought (CoT) reasoning and self-verification, making it suitable for multi-granular UGTE assessment.

Evaluation Metrics

Three common evaluation metrics from automated essay scoring research [37, 5] are employed to assess the performance

of the proposed framework. Let N denote the number of evaluated UGTE samples.

Mean Absolute Error (MAE) computes the mean absolute difference between predicted scores $\hat{y}_i$ and ground-truth scores y_i , with $MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|$.

Mean Squared Error (MSE) penalizes larger errors more heavily and is defined as: $MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$.

Pearson Correlation Coefficient (PCC) measures the linear correlation between predicted and expert-assigned scores:

$PCC = \frac{\sum_{i=1}^N (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^N (y_i - \bar{y})^2} \cdot \sqrt{\sum_{i=1}^N (\hat{y}_i - \bar{\hat{y}})^2}}$, where $\bar{y}$ and $\bar{\hat{y}}$ are the means of the ground-truth and predicted scores, respectively.

MAE and MSE assess prediction accuracy, with lower values indicating better performance, while PCC reflects the alignment between model predictions in UGTE assessment and human expert evaluations, with higher values indicating stronger agreement. The standard deviations of sample-level absolute and squared errors are additionally reported to measure the dispersion of assessment errors across UGTE samples, providing insight into the scoring stability. Quadratic Weighted Kappa (QWK), common in traditional AES tasks, is excluded here because it is designed for discrete ordinal scales, while our scoring outputs are continuous-valued.

Implementation Details

All models are evaluated in a training-free setting without parameter tuning or model-specific adaptation, aligning with practical educational deployment. Inference uses both local and API-based settings. Qwen-2.5-14B-1M is deployed locally on four NVIDIA RTX 6000 Ada GPUs with tensor parallel size 4, maximum model length 262,144, token batch size 131,072, maximum sequence count 1, and FP8 quantization for memory-efficient stable inference. The other models are accessed via Alibaba Cloud Bailian APIs using official SDKs. All API-based models are directly invoked with the provider's default generation settings, without manually specifying or adjusting decoding parameters. API experiments are run on an Intel i9-10900K workstation with 64 GB RAM and Python 3.10. This setup maintains consistent runtime conditions across the API-based experiments. A fixed 30-second interval is applied between successive API requests to avoid rate limits.

For *few-shot prompting*, the number of exemplars is empirically fixed at two unless otherwise specified. Three representative UGTE samples from the MUTA dataset form the exemplar pool, from which two are randomly selected for each inference. To prevent data leakage, these samples are excluded from final metric computation. In role-play prompting, the instruction is: 'You are a university professor responsible for evaluating students' submitted undergraduate thesis.' For holistic scoring, the weights are adjustable across disciplines or assessment contexts. By default, *Structure*, *Logic*, *Originality*, and *Writing* are each assigned 0.2, whereas *Rigor* and *Proficiency* are each assigned 0.1.

5.2 Comparison Across State-of-the-art LLMs

To evaluate the effectiveness of the proposed PEMUTA framework, we conduct a comparison under a standard LLM-based assessment setting across multiple state-of-the-art LLMs on the MUTA dataset. Consistent with prior LLM-based

educational assessment studies [5], we adopt a minimal zero-shot prompting configuration as a reference setting. In this setting, the model is instructed to 'Holistically assess the given thesis on a 1–10 scale.' All comparisons are performed under identical LLM backends, with differences arising solely from the introduction of PEMUTA. This configuration represents a commonly used holistic evaluation condition. Table 8 presents holistic-level results: mean score, MAE, MSE, PCC, and mean scores across the six SLOWPR dimensions.

We have the following observations. First, across all evaluated models and four subsets, LLMs prompted with PEMUTA consistently produce improved holistic evaluation results compared to those generated under the standard prompt, with MAE improvements in the range 0.03 to 0.99, MSE reductions between 0.07 and 1.62, and PCC gains from values that are undefined due to zero variance in the reference scores to 0.88. These results indicate increased alignment with expert assessments across diverse UGTE subsets. Under the same LLM setting, outputs generated with PEMUTA also exhibit smaller standard deviations in holistic scores while maintaining these improvements, suggesting more consistent intra-model scoring behavior. The lower sample-level standard deviations of absolute and squared errors under PEMUTA further confirm that PEMUTA reduces the dispersion of individual prediction errors and enhances scoring stability. For example, on the MUTA-S subset using Qwen2.5-14B-1M, PEMUTA achieves an MAE reduction from 0.65 to 0.57, an MSE reduction from 0.72 to 0.56 while increasing the PCC from 0.19 to 0.31, together with the standard deviation decreases from 0.51 with the standard prompt to 0.40 with PEMUTA. Second, when instructed using PEMUTA, LLMs produce both holistic and dimension-specific evaluations, whereas those instructed by the standard prompt yield only overall ratings. Under PEMUTA prompting, the generated outputs facilitate multi-granular assessment, helping LLMs capture specific attributes of UGTE quality and produce more interpretable and diagnostic evaluations. Across subsets, the *Writing* dimension generally receives lower scores than *Proficiency*, suggesting that students typically demonstrate strong disciplinary knowledge but relatively weaker academic writing and presentation skills. Third, across UGTEs from multiple institutions and disciplines, scores vary across models and evaluation dimensions, reflecting model-specific tendencies in assessment. In general, Qwen2.5-14B-1M tends to assign higher scores on *Writing* and *Rigor* compared to other models, whereas DeepSeek-R1 often yields slightly lower scores in *Writing*. For instance, on the MUTA-S subset, Qwen2.5-14B-1M assigns mean scores of 8.07 for *Writing* and 8.97 for *Rigor*, while other models' scores are approximately 7.71–8.04 for *Writing* and 8.07–8.89 for *Rigor*. Fig. 5 provides a visual summary of mean scores at all granularity levels for subset MUTA-S and highlights these model-specific variations across dimensions.

Table 8 Comparison of the standard prompting strategy and the proposed PEMUTA framework across five state-of-the-art LLMs on four subsets of the MUTA dataset. Human expert scores are shown for reference.

Subset	Model	Prompt	Holistic				Structure	Logic	Originality	Writing	Proficiency	Rigor
			MEAN	MAE ↓	MSE ↓	PCC ↑	MEAN	MEAN	MEAN	MEAN	MEAN	MEAN
MUTA-S	Human Expert		8.34±0.77	–	–	–	8.24	8.34	8.20	8.16	9.07	8.43
	Qwen-Plus	Standard	8.51±0.70	0.85±0.56	1.04±1.42	0.02	–	–	–	–	–	–
		PEMUTA	8.24±0.42	0.61±0.49	0.61±0.85	0.23	8.74	8.17	8.33	7.75	8.39	8.07
	Qwen2.5-14B-1M	Standard	8.15±0.51	0.65±0.55	0.72±1.15	0.19	–	–	–	–	–	–
		PEMUTA	8.54±0.40	0.57±0.49	0.56±0.86	0.31	8.85	8.38	8.69	8.07	8.56	8.97
	Qwen-Turbo-Latest	Standard	8.49±0.76	0.85±0.58	1.06±1.38	0.07	–	–	–	–	–	–
		PEMUTA	8.40±0.36	0.60±0.44	0.55±0.79	0.25	8.04	8.00	8.05	7.71	8.52	8.41
	DeepSeek-V3	Standard	8.47±0.52	0.65±0.53	0.70±1.27	0.18	–	–	–	–	–	–
		PEMUTA	8.50±0.23	0.54±0.46	0.50±0.87	0.37	8.91	8.26	8.37	8.04	8.91	8.89
	DeepSeek-R1	Standard	8.66±0.52	0.69±0.57	0.80±1.34	0.15	–	–	–	–	–	–
PEMUTA		8.31±0.43	0.59±0.44	0.54±0.74	0.33	8.33	8.31	8.31	7.84	9.02	8.49	
MUTA-C1	Human Expert		8.48±0.60	–	–	–	8.75	8.51	8.05	8.23	9.04	8.66
	Qwen-Plus	Standard	7.97±0.33	0.82±0.36	0.80±0.62	−0.28	–	–	–	–	–	–
		PEMUTA	8.50±0.31	0.42±0.37	0.31±0.48	0.40	8.92	8.47	7.87	8.45	8.81	8.77
	Qwen2.5-14B-1M	Standard	7.88±0.60	0.92±0.63	1.22±1.40	−0.29	–	–	–	–	–	–
		PEMUTA	8.88±0.34	0.45±0.48	0.42±0.73	0.59	9.13	8.79	8.26	9.09	9.21	9.03
	Qwen-Turbo-Latest	Standard	8.58±0.22	0.52±0.49	0.49±0.81	−0.27	–	–	–	–	–	–
		PEMUTA	8.29±0.64	0.49±0.43	0.42±0.74	0.47	8.50	8.25	7.74	8.32	8.75	8.58
	DeepSeek-V3	Standard	8.00±0.00	0.67±0.32	0.55±0.45	–	–	–	–	–	–	–
		PEMUTA	8.41±0.42	0.49±0.37	0.37±0.43	0.32	8.77	8.28	8.00	8.33	8.67	8.69
	DeepSeek-R1	Standard	8.48±0.55	0.68±0.42	0.63±0.67	0.03	–	–	–	–	–	–
PEMUTA		8.23±0.55	0.50±0.38	0.38±0.53	0.48	8.69	8.09	7.72	7.95	8.94	8.44	
MUTA-C2	Human Expert		7.69±1.06	–	–	–	8.12	7.09	7.78	7.51	8.23	7.65
	Qwen-Plus	Standard	7.60±0.47	0.88±0.61	1.14±1.23	0.29	–	–	–	–	–	–
		PEMUTA	7.55±0.47	0.77±0.51	0.84±0.94	0.60	8.08	7.47	7.11	7.60	7.85	7.07
	Qwen2.5-14B-1M	Standard	7.92±0.55	0.86±0.65	1.15±1.50	0.32	–	–	–	–	–	–
		PEMUTA	8.09±0.38	0.78±0.62	0.98±0.23	0.59	8.43	7.89	7.94	8.05	8.37	7.92
	Qwen-Turbo-Latest	Standard	8.58±0.19	0.99±0.91	1.78±2.40	0.41	–	–	–	–	–	–
		PEMUTA	7.55±0.43	0.79±0.56	0.92±1.23	0.54	7.96	7.23	7.23	7.69	7.91	7.35
	DeepSeek-V3	Standard	8.40±0.30	0.98±0.86	1.66±2.15	0.11	–	–	–	–	–	–
		PEMUTA	8.15±0.23	0.84±0.70	1.18±1.52	0.50	8.47	8.05	8.37	7.87	8.41	7.57
	DeepSeek-R1	Standard	8.27±0.48	0.99±0.82	1.62±2.26	0.09	–	–	–	–	–	–
PEMUTA		7.50±0.30	0.88±0.59	1.12±1.28	0.36	7.92	7.52	6.93	7.38	7.98	7.53	
MUTA-C3	Human Expert		7.48±0.98	–	–	–	7.27	7.63	6.75	8.24	7.40	7.59
	Qwen-Plus	Standard	8.43±0.71	1.24±0.49	1.77±1.32	0.03	–	–	–	–	–	–
		PEMUTA	8.04±0.44	0.79±0.77	0.77±0.86	0.35	8.50	8.09	7.57	7.45	8.66	8.50
	Qwen2.5-14B-1M	Standard	8.04±0.60	0.91±0.49	1.06±0.96	0.09	–	–	–	–	–	–
		PEMUTA	8.06±0.40	0.70±0.50	0.73±0.91	0.38	8.29	8.13	7.35	8.48	8.00	8.13
	Qwen-Turbo-Latest	Standard	8.61±0.29	1.10±0.68	1.66±1.79	0.12	–	–	–	–	–	–
		PEMUTA	8.31±0.40	0.85±0.51	0.97±0.86	0.42	8.49	8.34	7.70	8.55	8.45	8.45
	DeepSeek-V3	Standard	8.45±0.22	0.99±0.59	1.32±1.27	0.02	–	–	–	–	–	–
		PEMUTA	7.98±0.36	0.64±0.43	0.59±0.63	0.50	8.15	8.11	7.62	8.28	7.88	7.59
	DeepSeek-R1	Standard	8.14±0.56	0.84±0.69	1.16±1.39	0.07	–	–	–	–	–	–
PEMUTA		7.93±0.42	0.68±0.51	0.71±0.87	0.31	8.23	7.97	7.73	7.65	8.10	8.10	

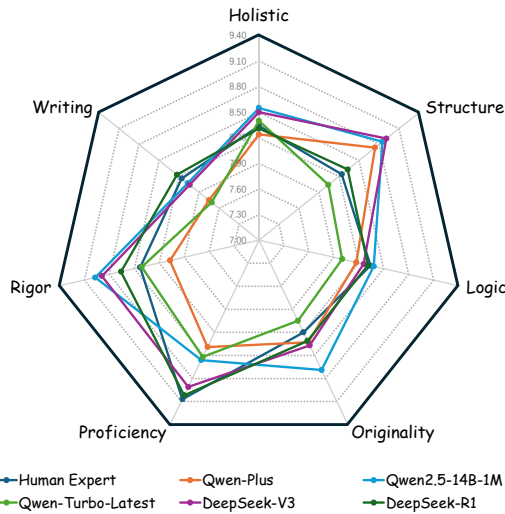


Fig. 5 Radar chart showing the holistic and dimension-specific mean scores for each model on the MUTA-S subset, with human expert annotations included for reference.

5.3 Ablation Studies

Ablation of Major Components

An ablation study is carried out on all four MUTA subsets to evaluate PEMUTA’s core components: Hierarchical prompting strategy (H-S), Role-Playing instruction (R-P), and Few-Shot exemplars (F-S). The comparison results are presented in Table 9. The holistic-level mean/std score, MAE, MSE, and PCC are employed for the performance comparison.

As shown in Table 9, the variant lacking all three components corresponds exactly to the standard prompt strategy and performs the worst across all models. In contrast, the full PEMUTA configuration, incorporating all three components, consistently outperforms the other variants, achieving the lowest error rates and highest correlation with human scores. Furthermore, removing either the Role-Playing instruction (R-P) or the Few-Shot exemplars (F-S) results in a measurable performance decline, demonstrating that both components contribute meaningfully and synergistically to aligning the assessment with human expert judgment.

Table 9 Ablation study of major components on the MUTA dataset across models. For brevity, ‘H-S’, ‘R-P’, and ‘F-S’ denote hierarchical prompting guidance, role-playing instruction, and few-shot exemplars. ‘Qwen2.5-14B’ and ‘Qwen-Turbo’ refer to ‘Qwen2.5-14B-1M’ and ‘Qwen-Turbo-Latest’, respectively.

	H-S R-P F-S				MUTA-S				MUTA-C1				MUTA-C2				MUTA-C3			
					MEAN	MAE↓	MSE↓	PCC↑	MEAN	MAE↓	MSE↓	PCC↑	MEAN	MAE↓	MSE↓	PCC↑	MEAN	MAE↓	MSE↓	PCC↑
Qwen-Plus	✗	✗	✗		8.51±0.70	0.85	1.04	0.02	7.97±0.33	0.82	0.80	-0.28	7.60±0.47	0.88	1.14	0.29	8.43±0.71	1.24	1.77	0.03
	✓	✓	✗		8.05±0.46	0.78	0.84	0.05	8.20±0.28	0.60	0.46	0.08	7.41±0.44	0.89	1.13	0.39	8.40±0.34	0.99	1.28	0.08
	✓	✗	✓		8.28±0.38	0.61	0.60	0.21	8.50±0.29	0.48	0.35	0.25	7.51±0.26	0.88	1.06	0.47	8.24±0.37	0.87	0.99	0.23
	✓	✓	✓		8.24±0.42	0.61	0.61	0.23	8.50±0.31	0.42	0.31	0.40	7.55±0.47	0.77	0.84	0.60	8.04±0.44	0.79	0.77	0.35
Qwen2.5-14B	✗	✗	✗		8.15±0.51	0.65	0.72	0.19	7.88±0.60	0.92	1.22	-0.29	7.92±0.55	0.86	1.15	0.32	8.04±0.60	0.91	1.06	0.09
	✓	✓	✗		8.57±0.39	0.60	0.63	0.21	8.78±0.28	0.54	0.50	0.13	8.01±0.41	0.80	1.07	0.42	8.15±0.46	0.86	0.95	0.21
	✓	✗	✓		8.55±0.43	0.59	0.59	0.29	8.92±0.43	0.55	0.50	0.53	8.11±0.35	0.78	1.04	0.56	8.19±0.40	0.74	0.86	0.35
	✓	✓	✓		8.54±0.40	0.57	0.56	0.31	8.88±0.34	0.45	0.42	0.59	8.09±0.38	0.78	0.98	0.59	8.06±0.40	0.70	0.73	0.38
Qwen-Turbo	✗	✗	✗		8.49±0.76	0.85	1.06	0.07	8.58±0.22	0.52	0.49	-0.27	8.58±0.19	0.99	1.78	0.41	8.61±0.29	1.10	1.66	0.12
	✓	✓	✗		7.99±0.32	0.76	0.78	0.08	8.74±0.24	0.50	0.48	0.09	8.45±0.38	0.95	1.52	0.41	8.58±0.40	1.02	1.54	0.21
	✓	✗	✓		8.13±0.56	0.68	0.73	0.24	8.16±0.46	0.51	0.43	0.39	7.52±0.38	0.80	0.96	0.53	8.34±0.16	0.93	1.09	0.41
	✓	✓	✓		8.40±0.36	0.60	0.55	0.25	8.29±0.64	0.49	0.42	0.47	7.55±0.43	0.79	0.92	0.54	8.31±0.40	0.85	0.97	0.42
DeepSeek-V3	✗	✗	✗		8.47±0.52	0.65	0.70	0.18	8.00±0.00	0.67	0.55	-	8.40±0.30	0.98	1.66	0.11	8.45±0.22	0.99	1.32	0.02
	✓	✓	✗		8.45±0.26	0.58	0.55	0.20	8.92±0.25	0.56	0.59	0.16	8.31±0.16	0.96	1.49	0.19	8.40±0.29	0.95	1.25	0.08
	✓	✗	✓		8.46±0.25	0.56	0.51	0.32	8.40±0.46	0.51	0.42	0.27	8.07±0.13	0.86	1.23	0.36	8.22±0.34	0.85	0.98	0.19
	✓	✓	✓		8.50±0.23	0.54	0.50	0.37	8.41±0.42	0.49	0.37	0.32	8.15±0.23	0.84	1.18	0.50	7.98±0.36	0.64	0.59	0.50
DeepSeek-R1	✗	✗	✗		8.66±0.52	0.69	0.80	0.15	8.48±0.55	0.68	0.63	0.03	8.27±0.48	0.99	1.62	0.09	8.14±0.56	0.84	1.16	0.07
	✓	✓	✗		8.90±0.36	0.70	0.88	0.16	8.73±0.27	0.51	0.49	0.06	8.07±0.47	0.97	1.37	0.15	8.23±0.41	0.94	1.08	0.10
	✓	✗	✓		8.23±0.42	0.62	0.56	0.31	8.37±0.73	0.55	0.55	0.38	7.61±0.46	0.89	1.14	0.29	8.00±0.55	0.80	0.98	0.12
	✓	✓	✓		8.31±0.43	0.59	0.54	0.33	8.23±0.55	0.50	0.38	0.48	7.50±0.30	0.88	1.12	0.36	7.93±0.42	0.68	0.71	0.31

Inference Latency Analysis

To examine practical evaluation efficiency, we analyze PEMUTA’s inference time under both local and API-based deployment settings. Across all configurations, PEMUTA evaluates each UGTE through a single model call, with its

structured reasoning completed within one prompt-response cycle. It does not require iterative model calls, external tool invocation, or multi-round interaction. Table 10 reports the approximate per-sample inference time under the full PEMUTA configuration across different UGTE subsets and LLM backends.

Table 10 Approximate inference time per sample (in seconds) under the full PEMUTA configuration.

Subset	Local Deployment	API-Based Deployment				
	Qwen2.5-14B	Qwen2.5-14B	Qwen-Plus	Qwen-Turbo	DeepSeek-V3	DeepSeek-R1
MUTA-S	7.37	38.98	24.94	12.30	18.04	60.96
MUTA-C1	9.03	39.29	16.44	15.11	19.65	65.36
MUTA-C2	8.40	37.66	20.81	19.50	16.99	65.68
MUTA-C3	8.87	42.20	25.78	13.81	17.11	53.58

Under local deployment, Qwen2.5-14B-1M achieves an average inference time ranging from 7.37 to 9.03 seconds per sample across all subsets. When using the same backbone model with identical prompts and inputs, API-based execution of Qwen2.5-14B-1M exhibits substantially higher latency. This increase can be attributed to network transmission overhead, cloud-side scheduling, and shared computational resources. As a result, for the same model architecture, local deployment demonstrates markedly higher inference efficiency, indicating its practical advantage for latency-sensitive evaluation scenarios. Within the API-based deployment setting, inference time varies considerably across models. Lightweight models such as Qwen-Turbo-Latest consistently deliver lower latency, whereas larger or reasoning-oriented models such as DeepSeek-R1 require significantly longer inference time.

Comparisons of Alternative Baseline Prompts

Table 11 Comparison of alternative baseline prompt designs for UGTE evaluation. Ratings assigned by human experts are provided for reference.

Type	Content	Mean	MAE ↓	MSE ↓	PCC ↑
Human Expert					
Standard Prompt	[Student Thesis] Holistically assess the given thesis on a 1–10 scale.	8.15±0.51	0.65±0.55	0.72±1.15	0.19
Chain-of-Thought Prompt	[Student Thesis] Holistically assess the given thesis on a 1–10 scale step-by-step.	8.09±0.58	0.68±0.57	0.80±1.20	0.21
Few-shot Prompt (2-shot)	[UGTE-Score Exemplars] [Student Thesis] Holistically assess the given thesis on a 1–10 scale.	8.17±0.36	0.68±0.47	0.67±0.86	0.10
Role-Playing Prompt	[Manual Scoring Criteria] [Student Thesis] You are a university professor responsible for holistically assessing the given thesis on a 1–10 scale according to the scoring criteria.	8.92±0.43	0.72±0.64	0.92±1.61	0.19
Role-Playing + Few-shot Prompt (2-shot)	[Manual Scoring Criteria] [UGTE-Score Thesis Exemplars] [Student Thesis] You are a university professor responsible for holistically assessing the given thesis on a 1–10 scale according to the scoring criteria.	8.24±0.36	0.64±0.46	0.63±0.87	0.15

An empirical comparison among five commonly used baseline prompt designs for LLM-based research in the education field is conducted. These include 1) a standard minimal prompt requesting a holistic score; 2) a chain-of-thought variant explicitly encouraging stepwise reasoning; 3) a few-shot prompt with UGTE-score pair examples; 4) a role-playing prompt with manual scoring criteria; and 5) a

combined role-playing plus few-shot prompt. All experiments are conducted with locally deployed Qwen2.5-14B-1M on the MUTA-S subset. As shown in Table 11, the standard prompt exhibits the most balanced performance across the evaluation metrics. It achieves an average score close to the human expert reference mean (8.15 vs. 8.34), with competitive MAE and MSE (0.65 and 0.72) and a PCC of 0.19. The *Chain-of-Thought Prompt* obtains a slightly higher PCC at 0.21, but it tends to underestimate scores relative to experts and shows higher per-sample deviations. The *Few-shot Prompt* has a slightly lower MSE, but its MAE is higher, and its PCC drops to 0.10. The *Role-playing Prompt* the same PCC as the standard prompt, but it shows an upward score shift and greater variability in errors. The combined *Role-playing + Few-shot Prompt* achieves the lowest MAE and MSE, yet its PCC decreases to 0.15, indicating weaker alignment with human assessment. Considering its accuracy, error stability, and correlation with human scores, the standard prompt is adopted as the baseline prompt for holistic UGTE assessment in this work.

Ablation of Few-shot Prompting

The PEMUTA framework incorporates few-shot prompting to provide exemplar-based demonstrations for UGTE assessment. These examples help guide the model toward human-like evaluative reasoning by presenting to domain-specific assessment patterns. To examine the effect of exemplar number, we vary the number of in-context exemplars from 0 to 3 and evaluate holistic-level performance using MAE, MSE, and PCC. Experimental results are summarized in Table 12.

As shown in Table 12, PEMUTA exhibits a clear performance improvement as the number of exemplars increases from 0 to 2 across all models. Compared with the 0-shot setting, the 2-shot configuration reduces MAE by 0.03–0.17, decreases MSE by 0.05–0.34, and improves PCC by 0.10–0.18. For instance, on Qwen-Turbo-Latest, MAE drops from 0.76 under 0-shot prompting to 0.60 under 2-shot prompting, while PCC increases from 0.08 to 0.25. These results indicate that exemplar-based prompting improves the alignment between model-generated scores and expert ratings. However, the improvement does not consistently continue beyond two exemplars. In several cases, such as Qwen-Plus, Qwen-Turbo-Latest, and DeepSeek-R1, increasing the number of exemplars to three leads to degraded performance. This decline may be attributed to two factors. First, adding more exemplars may dilute the attention allocated to the target UGTE, reducing the model's focus on diagnostic features and leading to less stable evaluation. This observation is consistent with prior reports on

long-context degradation [64]. Second, LLMs are sensitive to the selection and ordering of demonstrations, and may be affected by interference among exemplars [65, 66, 28]. As the number of exemplars increases, stylistic and structural differences among exemplars may introduce redundant or conflicting signals, which can weaken task guidance. These results highlight the role of demonstrations in providing contextual guidance,

validate the effectiveness of the few-shot design in PEMUTA for UGTE assessment, and underscore the importance of selecting an appropriate number of exemplars to balance contextual guidance and potential interference. Based on this analysis, the 2-shot setting is adopted as the default configuration for the remaining experiments unless otherwise specified.

Table 12 Impact of the number of few-shot exemplars on PEMUTA's holistic performance on the MUTA-S subset. As the number of exemplars increases, the effectiveness of the proposed method in leveraging exemplar-based inductive reasoning becomes more pronounced.

Model	0-shot			1-shot			2-shot			3-shot		
	MAE ↓	MSE ↓	PCC ↑	MAE ↓	MSE ↓	PCC ↑	MAE ↓	MSE ↓	PCC ↑	MAE ↓	MSE ↓	PCC ↑
Qwen-Plus	0.78	0.84	0.05	0.77	0.80	0.10	0.61	0.61	0.23	0.62	0.67	0.13
Qwen2.5-14B-1M	0.60	0.63	0.21	0.60	0.60	0.24	0.57	0.56	0.31	0.57	0.58	0.18
Qwen-Turbo-Latest	0.76	0.78	0.08	0.65	0.80	0.21	0.60	0.55	0.25	0.71	0.78	-0.04
DeepSeek-V3	0.58	0.55	0.20	0.54	0.57	0.24	0.54	0.50	0.37	0.52	0.52	0.37
DeepSeek-R1	0.70	0.88	0.16	0.64	0.76	0.25	0.59	0.54	0.33	0.70	0.93	0.22

5.4 Qualitative Analysis

To intuitively assess the effectiveness of the proposed PEMUTA in guiding LLMs toward pedagogically meaningful multi-granular UGTE assessment, a qualitative case study is presented in Fig. 6. The example is drawn from the MUTA dataset and evaluated using Qwen-Plus under five prompting

configurations: (1) Standard Prompt, (2) PEMUTA w/o H-S, which directly instructs the LLM to produce multi-granular assessment results without the proposed two-stage hierarchical prompting strategy, (3) PEMUTA w/o F-S, which removes few-shot prompting, (4) PEMUTA w/o R-P, which omits role-play prompting, and (5) Full PEMUTA.

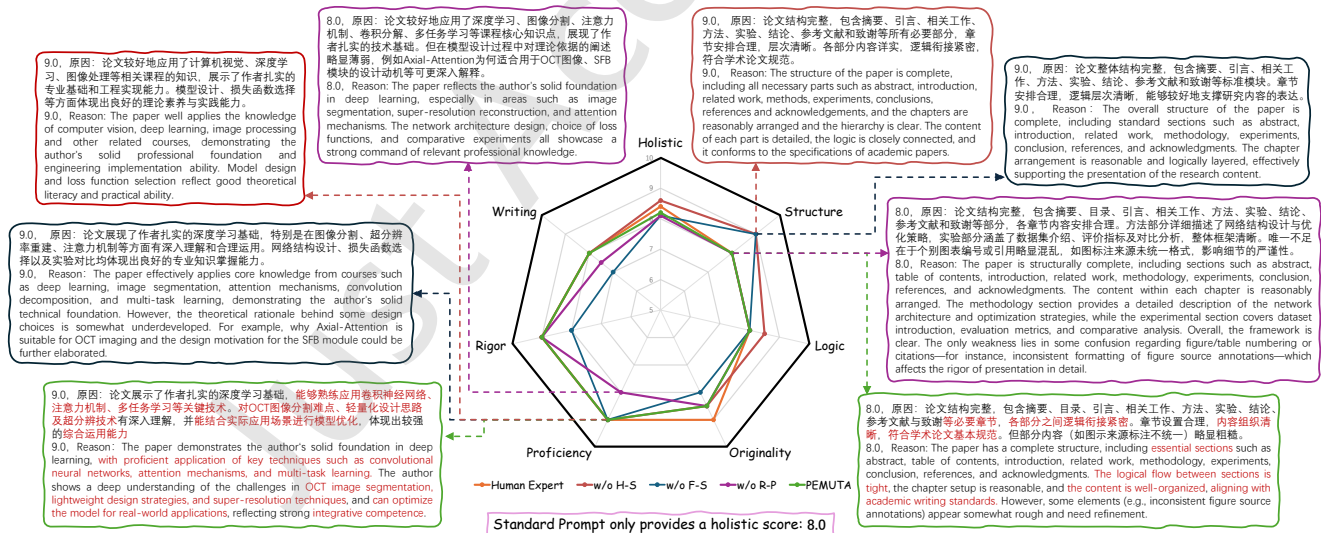


Fig. 6 Case study of multi-granular assessment for a sample from the MUTA dataset using Qwen-Plus under five prompting configurations: Standard Prompt, full PEMUTA, and three ablated PEMUTA variants. The radar plot shows model-assigned scores across six fine-grained dimensions and one holistic dimension for each multi-granular-capable configuration, alongside human expert ratings. The Standard Prompt, which produces only a holistic score, is presented separately. Representative score justifications are provided for the Structure and Proficiency dimensions. Textual segments uniquely generated by full PEMUTA, providing richer and more pedagogically aligned explanations, are highlighted in red.

As shown in Fig. 6, the standard prompt yields only a holistic score, whereas PEMUTA and its ablated variants are capable of multi-granular assessment. The radar chart visualizes the holistic rating together with model-predicted scores across the six SLOWPR dimensions. All four multi-granular-capable

setups produce dimension-relevant justifications, indicating that the PEMUTA prompting paradigm effectively activates latent domain knowledge within LLMs to support multi-granular evaluation. Among them, the full PEMUTA configuration achieves the highest alignment with expert judgments. A

closer inspection of the *Structure* and *Proficiency* dimensions further clarifies this advantage. In *Structure*, full PEMUTA delivers more specific and pedagogically grounded feedback, referencing segments such as ‘essential sections,’ ‘tight logical flow,’ and ‘organization consistent with academic writing standards.’ In *Proficiency*, full PEMUTA exhibits stronger diagnostic precision by identifying concrete techniques like ‘convolutional neural networks,’ ‘attention mechanisms,’ and ‘multi-task learning,’ and linking them to UGTE evaluative criteria such as ‘optimizing models for real-world applications’ and ‘reflecting strong integrative competence’.

In contrast, ablated variants tend to produce vague, superficial, or misplaced explanations. For instance, the ‘w/o H-S’ variant overestimates *Structure* while underestimating *Proficiency*, likely due to attention distraction from the *Logic* and *Writing* dimensions. This case study demonstrates that the two-stage hierarchical strategy, few-shot prompting, and role-play prompting synergistically enhance the interpretability and reliability of LLM-based UGTE assessment. The full PEMUTA configuration enables multi-granular evaluation that not only aligns closely with expert ratings but also provides dimension-specific, pedagogically informative feedback that reflects the essential criteria of UGTE assessment.

6 Conclusion and Future Work

We develop a pedagogically-enriched multi-granular UGTE assessment framework to effectively activate domain-specific knowledge in large language models for undergraduate thesis evaluation. PEMUTA adopts a hierarchical two-stage prompting strategy grounded in Vygotsky’s theory and Bloom’s Taxonomy, enabling structured assessment across six fine-grained dimensions: *Structure*, *Logic*, *Originality*, *Writing*, *Proficiency*, and *Rigor* (SLOWPR), followed by holistic synthesis. The hierarchical design guides the model to focus on one assessment dimension at a time, which helps reduce attention distraction and enables more precise use of embedded academic knowledge to generate evaluations that reflect both students’ academic performance and their developmental potential. Few-shot prompting and role-play prompting are further jointly incorporated to enhance pedagogical alignment and consistency with expert judgment, without model fine-tuning. Extensive experiments on engineering UGTEs demonstrate that PEMUTA consistently outperforms standard holistic prompting, achieving lower errors and higher PCC with expert ratings while delivering fine-grained, dimension-specific assessments that existing methods cannot provide.

PEMUTA provides advisory evaluation results that, while requiring expert validation for high-stakes decisions, offer students not only holistic scores but also dimension-level ratings, explanatory feedback, and revision suggestions to identify weaknesses in areas such as argument organization, methodological clarity, and writing quality. For educators, it reduces manual assessment workload and delivers consistent reference scores, while dimension-level feedback helps pinpoint common student deficiencies and guide supervision strategies, ultimately supporting both summative and formative assessment in engineering UGTE evaluation. While the present study focuses on engineering UGTEs, the configurable weighting of the SLOWPR dimensions allows disciplinary emphasis and

institutional assessment requirements to be reflected, providing a flexible basis for extension to broader disciplinary contexts.

Future work will extend PEMUTA to multimodal assessment by incorporating code artifacts and presentation recordings for more comprehensive evaluation of engineering competencies^[67]. We will also explore reliability enhancements through model editing, reinforcement learning, and post-hoc verification to reduce LLM hallucinations and improve assessment stability. Cross-institutional and cross-disciplinary collaboration will be pursued to further extend PEMUTA across varied linguistic and educational settings.

Acknowledgment

This work was supported in part by the Guangdong Provincial Teaching Quality and Teaching Reform Project under Grant SJZLGC202511; in part by the Medical and Health Science Program of Zhejiang Province under Grant WKJ-ZJ-26085; and in part by the Ningbo Municipal Bureau of Science and Technology under Grants 2024Z124 and 2025Z197.

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

Dates

Received: 31 May 2020; Accepted: 5 July 2020;
Published online: 7 September 2021

References

- [1] K.L. Chan. Statistical analysis of final year project marks in the computer engineering undergraduate program. *IEEE Trans. Educ.*, 44(3):258–261, 2001.
- [2] Xiaoqing Zhang, Jiaqi Wei, Shengjie Ye, Zunjie Xiao, Yan Hu, Jigen Tang, and Jiang Liu. Multimedia meets archaeology: A novel interdisciplinary teaching approach. In *Proc. Front. Educ. Conf. FIE*, pages 1–8, 2021.
- [3] Jialiang Lin, Jiaxin Song, Zhangping Zhou, Yidong Chen, and Xiaodong Shi. Automated scholarly paper review: Concepts, technologies, and challenges. *Inf. Fusion*, 98:101830, 2023.
- [4] 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. *Proc. Int. Learn. Analytics Knowl. Conf.*, page 293–305, 2025.
- [5] Yishen Song, Qianta Zhu, Huaibo Wang, and Qinhua Zheng. Automated essay scoring and revising based on open-source large language models. *IEEE Trans. Learn. Technol.*, 17:1880–1890, 2024.
- [6] Yigal Attali and Jill Burstein. Automated essay scoring with e-rater® v.2. *J. Technol. Learn. Assessment*, 4(3), Feb. 2006.
- [7] Kaveh Taghipour and Hwee Tou Ng. A neural approach to automated essay scoring. *Proc. Conf. Empirical Methods Natural Lang. Process.*, pages 1882–1891, 2016.
- [8] Fei Dong and Yue Zhang. Automatic features for essay scoring—an empirical study. *Proc. Conf. Empirical Methods Natural Lang. Process.*, pages 1072–1077, 2016.

- [9] Masaki Uto, Itsuki Aomi, Emiko Tsutsumi, and Maomi Ueno. Integration of prediction scores from various automated essay scoring models using item response theory. *IEEE Trans. Learn. Technol.*, 16(6):983–1000, 2023.
- [10] Yongjie Wang, Chuang Wang, Ruobing Li, and Hui Lin. On the use of bert for automated essay scoring: Joint learning of multi-scale essay representation. *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3416–3425, July 2022.
- [11] Milan Kostic, Hans Friedrich Witschel, Knut Hinkelmann, and Maja Spahic-Bogdanovic. LLMs in automated essay evaluation: A case study. *Proc. AAAI Symp. Ser.*, 3(1):143–147, May 2024.
- [12] Qi Lang, Minjuan Wang, Minghao Yin, Shuang Liang, and Wenzhuo Song. Transforming education with generative AI (GAI): Key insights and future prospects. *IEEE Trans. Learn. Technol.*, 18:230–242, 2025.
- [13] Jiang Liu, Qiushi Nie, Xingyue Wang, Zunjie Xiao, and Xiaoqing Zhang. Exploration of weakly supervised learning pedagogy under the background of generative artificial intelligence. *Computer Education*, (01):198–203, 209, 2025.
- [14] Qingyang Sun, Tianrui Wang, Peng Sheng, Qianyi Wang, Xiaoqing Zhang, and Jiang Liu. Exploration of Intelligent Evaluation Methods for Computer Course Project Reports Under the Education Intelligence Framework. *Computer Education*, (01):210–214, 2025.
- [15] Fuchun Sun, Runfa Chen, Tianying Ji, Yu Luo, Huaidong Zhou, and Huaping Liu. A comprehensive survey on embodied intelligence: Advancements, challenges, and future perspectives. *CAAI Artificial Intelligence Research*, 3:9150042, 2024.
- [16] Ying Wen, Ziyu Wan, Ming Zhou, Shufang Hou, Zhe Cao, Chenyang Le, Jingxiao Chen, Zheng Tian, Weinan Zhang, and Jun Wang. On realization of intelligent decision making in the real world: A foundation decision model perspective. *CAAI Artificial Intelligence Research*, 2:9150026, 2023.
- [17] Yaqiong He, Feng Jiang, Xiaomin Chu, and Peifeng Li. Automated Chinese essay scoring from multiple traits. *Proc. Int. Conf. Comput. Linguist.*, Mar. 2022.
- [18] Paraskevas Lagakis and Stavros Demetriadis. Automated essay scoring: A review of the field. *Proc. Int. Conf. Comput. Inf. Telecommun. Syst.*, pages 1–6, 2021.
- [19] Vania Dimitrova, Lydia Lau, and Rebecca O'Rourke. Semantic social scaffolding for capturing and sharing dissertation experience. *IEEE Trans. Learn. Technol.*, 4(1):74–87, 2011.
- [20] Steven Goh and Harry Ku. Work in progress - intervening to improve the dissertation of final year engineering research projects. *Proc. Front. Educ. Conf. FIE*, pages F3F–1–F3F–6, 2011.
- [21] Marjolein Haagsman, Basten Snoek, Anton Peeters, Karin Scager, Frans Prins, and Martijn van Zanten and. Examiners' use of rubric criteria for grading bachelor theses. *Assess. Eval. High. Educ.*, 46(8):1269–1284, 2021.
- [22] Ilona Rinne. Same grade for different reasons, different grades for the same reason? *Assessment & Evaluation in Higher Education*, 49(2):220–232, 2024.
- [23] Clayton Cohn, Nicole Hutchins, Tuan Le, and Gautam Biswas. A chain-of-thought prompting approach with LLMs for evaluating students' formative assessment responses in science. *Proc. AAAI Conf. Artif. Intell.*, 38(21):23182–23190, Mar. 2024.
- [24] Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. *Int. Conf. Mach. Learn.*, 202:31210–31227, 2023.
- [25] Bizhu Wu, Jinheng Xie, Keming Shen, Zhe Kong, Jianfeng Ren, Ruibin Bai, Rong Qu, and Linlin Shen. MG-MotionLLM: A unified framework for motion comprehension and generation across multiple granularities. *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 27849–27858, June 2025.
- [26] Chunhui Zhang, Yiren Jian, Zhongyu Ouyang, and Soroush Vosoughi. Working memory identifies reasoning limits in language models. *Proc. Conf. Empirical Methods Natural Lang. Process.*, pages 16896–16922, November 2024.
- [27] Bingsheng Yao, Guiming Chen, Ruishi Zou, Yuxuan Lu, Jiachen Li, Shao Zhang, Yisi Sang, Sijia Liu, James Hendler, and Dakuo Wang. More samples or more prompts? exploring effective few-shot in-context learning for LLMs with in-context sampling. *Findings Assoc. Comput. Linguist.: NAACL*, pages 1772–1790, June 2024.
- [28] Jiuhai Chen, Lichang Chen, Chen Zhu, and Tianyi Zhou. How many demonstrations do you need for in-context learning? *Proc. Conf. Empirical Methods Natural Lang. Process.*, pages 11149–11159, December 2023.
- [29] Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. Better zero-shot reasoning with role-play prompting. *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technologies*, pages 4099–4113, June 2024.
- [30] Xiaoxu Han, Wanqing Zhao, Ziyu Guan, and JinYE Peng. Act-llm: A whole-process chain for character-centric role-playing with large language models. *Expert Syst. Appl.*, 296:129024, 2026.
- [31] Jaewoo Ahn, Taehyun Lee, Junyoung Lim, Jin-Hwa Kim, Sangdoo Yun, Hwaran Lee, and Gunhee Kim. TimeChara: Evaluating point-in-time character hallucination of role-playing large language models. *Findings Assoc. Comput. Linguist.: ACL 2024*, pages 3291–3325, August 2024.
- [32] Zixuan Ke and Vincent Ng. Automated essay scoring: A survey of the state of the art. *IJCAI*, 19:6300–6308, 2019.
- [33] Lawrence M. Rudner, Veronica Garcia, and Catherine Welch. An evaluation of intellimetric™ essay scoring system. *J. Technol., Learn., Assessment*, 4(4), Mar. 2006.
- [34] Shengjie Li and Vincent Ng. Automated essay scoring: A reflection on the state of the art. *Proc. Conf. Empirical Methods Natural Lang. Process.*, pages 17876–17888, November 2024.
- [35] Yaqiong He, Feng Jiang, Xiaomin Chu, and Peifeng Li. Automated chinese essay scoring from multiple traits. *Proc. Int. Conf. Comput. Linguistics*, pages 3007–3016, 2022.
- [36] Muhammad Faseeh, Abdul Jaleel, Naeem Iqbal, Anwar Ghani, Akmalbek Abdusalomov, Asif Mehmood, and Young-Im Cho. Hybrid approach to automated essay scoring: Integrating deep learning embeddings with handcrafted linguistic features for improved accuracy. *Mathematics*, 12(21):3416, 2024.
- [37] Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. Automatic text scoring using neural networks. *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, pages 715–725, 2016.
- [38] Peiyi Zhang, Yazhou Zhang, Bo Wang, Lu Rong, Prayag Tiwari, and Jing Qin. Edu-Values: Towards evaluating the chinese education values of large language models. *Comp. Proc. ACM Web Conf.*, page 1519–1523, 2025.

- [39] Qingyang Sun, Jialu Zhang, Peng Sheng, Qianyi Wang, Tianrui Wang, Heng Li, Hanshu Zhan, Xiaoqing Zhang, and Jiang Liu. Evaluating the efficacy of a multifaceted prompt for use with LLMs to evaluate course project reports. *Frontiers of Digital Education*, 3(2):12, 2026.
- [40] Uwe Maier and Christian Klotz. Personalized feedback in digital learning environments: Classification framework and literature review. *Comput. Educ. Artif. Intell.*, 3:100080, 2022.
- [41] Wannapon Suraworachet, Qi Zhou, and Mutlu Cukurova. Impact of combining human and analytics feedback on students' engagement with, and performance in, reflective writing tasks. *International Journal of Educational Technology in Higher Education*, 20(1):1–24, 2023.
- [42] Jiwei Tang, Jin Xu, Tingwei Lu, Zhicheng Zhang, Yiming Zhao, LinHai LinHai, and Hai-Tao Zheng. Perception compressor: A training-free prompt compression framework in long context scenarios. *Findings Assoc. Comput. Linguist.: NAACL*, pages 4093–4108, April 2025.
- [43] Shuo Wang, Jing Li, Zibo Zhao, Dongze Lian, Binbin Huang, Xiaomei Wang, Zhengxin Li, and Shenghua Gao. TSP-Transformer: Task-specific prompts boosted transformer for holistic scene understanding. *Proc. IEEE Winter Conf. Appl. Comput. Vision*, pages 925–934, January 2024.
- [44] Jesse Mu, Xiang Li, and Noah Goodman. Learning to compress prompts with gist tokens. *Adv. Neural Inform. Process. Syst.*, 36:19327–19352, 2023.
- [45] Yajing Liu, Yuning Lu, Hao Liu, Yaozu An, Zhuoran Xu, Zhuokun Yao, Baofeng Zhang, Zhiwei Xiong, and Chenguang Gui. Hierarchical prompt learning for multi-task learning. *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10888–10898, June 2023.
- [46] Deborah Chetcuti, Joseph Cacciottolo, and Nicholas Vella. What do examiners look for in a phd thesis? explicit and implicit criteria used by examiners across disciplines. *Assessment & Evaluation in Higher Education*, 47(8):1358–1373, 2022.
- [47] Wee Chun Tan. Empowering examiners to develop doctoral assessment literacy: a situated learning perspective. *Frontiers in Education*, 9:1345661, 2024.
- [48] Xiaoli Yang, Xin Xie, and Rongyu Cui. Application research on quality evaluation of college graduation thesis based on artificial intelligence. *International Conference on Educational and Information Technology*, pages 329–332, 2025.
- [49] Zi-Gang Ge, Yuanyuan Zhao, and Yong Liu. Exploring quality issues in academic master theses of education majors and corresponding countermeasures. *Journal of Pedagogical Research*, 8(4):30–47, 2024.
- [50] Jean Patrick Lostaunau, Armando Soto, and Alfredo Barrientos. Thesis review and analysis automated system. *Conference of Open Innovations Association*, pages 143–151, 2022.
- [51] Lisna Zahrotun, Nila Hutami Putri, and Arfiani Nur Khusna. The implementation of K-Means clustering method in classifying undergraduate thesis titles. *International Conference on Telecommunication Systems, Services, and Applications*, pages 1–4, 2018.
- [52] Jesús Miguel García-Gorrostieta, Aurelio López-López, and Samuel González-López. Towards automatic assessment of argumentation in theses justifications. *Data Driven Approaches in Digital Education*, pages 54–66, 2017.
- [53] Jesús Miguel García Gorrostieta and Aurelio López-López. Argumentation identification for academic support in undergraduate writings. *Adaptive and Adaptable Learning*, pages 98–109, 2016.
- [54] Zhao Xu, Yong Jin, Shardrom Johnson, Weishan Gao, Hui Miao, and Qin Wei. Random inspection evaluation of shanghai graduate dissertation based on bayesian decision tree. *International Conference on Computer Science & Education*, pages 501–504, 2016.
- [55] Rong Li, Xiangjun Wang, and Dingyuan Li. Research on graduation thesis evaluation based on hopfield neural network. *International Symposium on Computational Intelligence and Design*, 2:18–21, 2014.
- [56] Hong ben Mao, Qi Jiang, Ai hua Zhou, and Shi fa Wang. Model and research on score evaluation model of graduation thesis. *International Conference on Computer Science & Education*, pages 735–738, 2014.
- [57] Yusra Mosallam, Lukas Toma, Mulu Weldegebreal Adhana, Costin-Gabriel Chiru, and Traian Rebedea. Unsupervised system for automatic grading of bachelor and master thesis. *Proceedings of the International Conference on eLearning and Software for Education*, pages 165–171, 2014.
- [58] Julie A. Reynolds and Robert J. Thompson. Want to improve undergraduate thesis writing? engage students and their faculty readers in scientific peer review. *CBE—Life Sciences Education*, 10(2):209–215, 2011.
- [59] John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S. Rosen, Gerbrand Ceder, Kristin A. Persson, and Anubhav Jain. Structured information extraction from scientific text with large language models. *Nat. Commun.*, 15(1):1418, 2024.
- [60] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.
- [61] An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, et al. Qwen2.5-1M Technical Report, 2025.
- [62] Aixian Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. DeepSeek-V3 Technical Report, 2024.
- [63] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, 2025.
- [64] Rishabh Agarwal, Avi Singh, Lei Zhang, Bernd Bohnet, Luis Rosias, Stephanie Chan, Biao Zhang, Ankesh Anand, Zaheer Abbas, Azade Nova, John D. Co-Reyes, Eric Chu, Feryal Behbahani, Aleksandra Faust, and Hugo Larochelle. Many-shot in-context learning. *Adv. Neural Inform. Process. Syst.*, 37:76930–76966, 2024.
- [65] Yiming Zhang, Shi Feng, and Chenhao Tan. Active example selection for in-context learning. *Proc. Conf. Empirical Methods Natural Lang. Process.*, pages 9134–9148, December 2022.
- [66] Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. Surface form competition: Why the highest probability answer isn't always right. *Proc. Conf. Empirical Methods Natural Lang. Process.*, pages 7038–7051, November 2021.
- [67] Lanxiao Wang, Wenzhe Hu, Heqian Qiu, Chao Shang, Taijin Zhao, Benliu Qiu, King Ngi Ngan, and Hongliang Li. A survey of vision and language related multi-modal task. *CAAI Artificial Intelligence Research*, 1(2):111–136, 2022.