RESEARCH ARTICLE

Misclassification Analysis in Automated Bloom's Taxonomy Classifiers: A Data-Centric Perspective on Educational Software

Maleeha Mahboob[1,*]

[1] Virtual University of Pakistan, Islamabad, Pakistan

Abstract

Automated classification of assessment questions according to Bloom's taxonomy is increasingly used to support curriculum design and educational analytics. Many existing approaches rely heavily on instructional verbs as proxies for cognitive demand, despite longstanding concerns about their interpretive reliability. This paper adopts a data-centric perspective to examine why verb-centric Bloom-level classification remains fragile when applied to authentic multiple-choice question (MCQ) stems. The study is based on a custom, single-domain dataset of MCQ stems annotated according to the revised Bloom's taxonomy, with intentional class imbalance preserved to reflect realistic assessment practices. Through dataset diagnostics and systematic qualitative inspection, the analysis highlights plausible sources of misclassification risk arising from linguistic ambiguity, task–verb mismatch, conceptual proximity between adjacent cognitive levels, and domain-specific semantic effects. Rather than proposing new classification models, the paper introduces a formal error taxonomy and an analytical framework that explains the interactions producing structured, non-random misclassification patterns, without reporting performance scores or model comparisons. These findings expose recurring failure patterns in automated Bloom-level classification systems. From a software engineering perspective, the analysis informs system evaluation and reliability by clarifying where verb-centric classifiers are prone to systematic misinterpretation during assessment analytics and decision support. The main contributions are threefold: (1) a data-centric examination of Bloom-level misclassification grounded in authentic computer science MCQ stems; (2) a structured taxonomy of recurring error types linked to linguistic ambiguity, task interpretation, and dataset characteristics; and (3) clarification of the implications of these error patterns for system evaluation and reliability in educational software that relies on automated Bloom-level classification.

Keywords: bloom's taxonomy, automated assessment, cognitive level classification, educational data analysis, error analysis, dataset diagnostics, instructional verbs, question stems.

Submitted: 06 January 2026
Accepted: 27 March 2026
Published: 17 May 2026

Vol. 2, **No.** 2, 2026.
10.62762/JSE.2026.118512

***Corresponding author:**
✉ Maleeha Mahboob
maleeha.mahboob2@gmail.com

Citation
Mahboob, M. (2026). Misclassification Analysis in Automated Bloom's Taxonomy Classifiers: A Data-Centric Perspective on Educational Software. *ICCK Journal of Software Engineering*, 2(2), 138–155.

1 Introduction

Bloom's Taxonomy is a hierarchical framework that categorizes educational tasks according to the cognitive processes they require. In its revised form, it spans six levels ranging from basic recall and understanding to higher-order processes involving analysis, evaluation, and creation. By categorizing questions according to cognitive demand, it supports balanced curriculum design and helps instructors align learning objectives with evaluation strategies. For this reason, there has been growing interest in automating Bloom-level classification of assessment questions, particularly in large-scale and digital learning environments [1]. Many automated approaches rely, either explicitly or implicitly, on instructional action verbs as indicators of cognitive level. This reliance is understandable. Action verbs are visible, easy to extract, and widely associated with Bloom's taxonomy in pedagogical practice [2, 3]. As a result, both rule-based systems and machine learning models often treat verb usage as a primary signal for classification. However, the relationship between action verbs and cognitive demand is not always straightforward. In authentic assessment material, the same verb may be used to frame tasks that require different types of reasoning. Conversely, questions targeting similar cognitive processes may be expressed using different linguistic forms. These inconsistencies introduce ambiguity that is not easily resolved through surface-level analysis. Recent advances in machine learning have improved the performance of automated Bloom classifiers, yet performance gains alone do not necessarily indicate conceptual understanding of cognitive levels. In some cases, high accuracy may coexist with systematic confusion between adjacent Bloom categories. When such patterns are not examined, underlying weaknesses may remain hidden [4, 5].

These limitations motivate closer examination at the level of data and task formulation. Accordingly, this study adopts a data-centric perspective that emphasizes dataset characteristics, linguistic ambiguity, and contextual interpretation as primary sources of Bloom-level misclassification. In this context, a data-centric perspective focuses on examining the properties of the dataset itself, such as class distribution, wording of question stems, and sources of ambiguity, rather than emphasizing model design or predictive performance. Attention is given to class imbalance, conceptual proximity between cognitive levels, and the reuse of instructional verbs across tasks. The analysis is explanatory in nature. Rather than optimizing predictive models, it seeks to clarify why misclassification persists even when labeling decisions appear internally consistent. Beyond its educational role, Bloom-level classification is increasingly embedded within software systems that support assessment analytics, curriculum auditing, and decision support in learning platforms. In such systems, misclassification affects software-level outcomes. Errors propagate into evaluation reports, content balancing mechanisms, and automated feedback pipelines. From a software engineering perspective, Bloom-level misclassification therefore represents a form of semantic failure, where system outputs remain syntactically valid but conceptually misaligned with intended task complexity. Understanding these failure patterns is necessary for interpreting system reliability, assessing evaluation risk, and avoiding overconfidence in automated analytics.

These challenges arise from properties of the data itself. When linguistic ambiguity, class imbalance, and task under-specification are present, building more complex models may obscure rather than resolve misclassification behavior. In contrast, a data-centric analysis allows these structural issues to be examined directly. This approach supports clearer interpretation of failure patterns and provides insight that remains applicable across different modeling choices.

The primary objective of this study is to examine automated Bloom's taxonomy classification from a data-centric perspective, with particular emphasis on identifying and understanding recurring sources of misclassification. The work aims to analyze the dataset characteristics, including class distribution, linguistic overlap, and action-verb usage that influence classification behavior. Attention is also given to the ways in which task interpretation and contextual cues shape cognitive-level assignment, especially for adjacent Bloom categories. Through dataset diagnostics and structured error analysis, the study seeks to clarify why misclassification persists even when labeling decisions appear internally consistent [6, 7]. In doing so, the objective is not to optimize predictive models, but to provide interpretive insight that can support more cautious and informed use of automated Bloom-level classification in educational assessment settings. This paper addresses the unreliable inference of cognitive level when automated Bloom classifiers rely primarily on instructional verbs. This study is guided by a set of research questions

that focus on understanding, rather than optimizing, automated Bloom-level classification. The emphasis is placed on identifying structural and linguistic factors that contribute to misclassification when action verbs are treated as primary indicators of cognitive level.

RQ1: Why do action-verb-based approaches encounter difficulty when distinguishing between adjacent levels of Bloom's cognitive taxonomy in multiple-choice question stems?

RQ2: In what ways do dataset design characteristics, particularly domain specificity and intentional class imbalance, shape the patterns of Bloom-level misclassification?

RQ3: What recurring error types arise when cognitive intent is inferred mainly from surface-level linguistic cues instead of task structure and semantic context?

The conceptual overview of the research questions framing the analysis of limitations in verb-centric Bloom-level classification of MCQ stems is given in Figure 1.

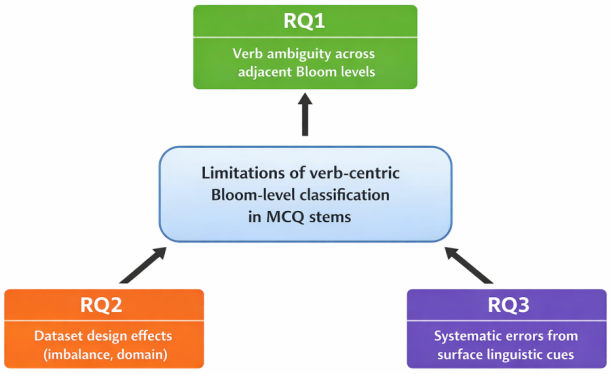

Figure 1. Conceptual overview of the research questions framing the analysis of limitations in verb-centric Bloom-level classification of MCQ stems.

The contributions of this paper are primarily analytical and data-centric in nature. They are intended to clarify limitations in current practice rather than to propose or validate a new classification system.

1. The study introduces an intentionally imbalanced, single-domain dataset of computer science MCQ stems annotated across all levels of Bloom's Taxonomy to support analysis of dataset-driven misclassification.
2. It presents a structured taxonomy of Bloom-level misclassification, identifying recurring error types associated with verb ambiguity, task–verb mismatch, and domain-specific semantics.
3. It provides a theory-driven explanation of adjacent-level confusion in verb-centric Bloom classification, clarifying limitations of inferring cognitive demand from surface-level linguistic cues.

The remainder of this paper is structured as follows. Section 2 reviews related work on automated Bloom's Taxonomy classification, with particular attention to misclassification behavior and prior data-centric and system-oriented analyses. Section 3 describes the dataset construction process. Section 4 presents a set of dataset-level diagnostics. Section 5 introduces formal error taxonomy for Bloom-level classification. Section 6 provides an analytical demonstration of Bloom-level misclassification, focusing on primary perspectives. Section 7 discusses the broader implications of the findings from a software engineering perspective, highlighting consequences for system design, evaluation, and reliability. Section 8 outlines the limitations of the study and identifies directions for future research. Finally, Section 9 concludes the paper by summarizing key insights.

2 Related Work

Bloom's taxonomy has long been used as a conceptual framework for organizing educational objectives and assessment practices. Since the revision of the original taxonomy, it has been widely applied to describe cognitive processes ranging from basic recall to higher-order reasoning. In educational research, Bloom's taxonomy is commonly employed to support curriculum alignment, assessment design, and learning outcome evaluation. Despite its pedagogical value, translating Bloom's cognitive categories into consistently identifiable linguistic patterns remains challenging in practical settings.

Early efforts to automate Bloom-level classification relied largely on rule-based and verb-driven approaches. These methods typically map instructional verbs to predefined cognitive levels using curated verb lists derived from pedagogical guidelines. While intuitive, such approaches have been shown to suffer from limited reliability. Several studies note that instructional verbs are reused across multiple Bloom levels and that verb choice alone does not always reflect the depth of reasoning required by a task. As a result, verb-based systems

Table 1. Summary of Prior Work on Bloom's Taxonomy Classification (2020–2025).

Reference	Year	Task focus	Methodology	Dataset characteristics	Key limitation	Gap addressed in this study
[10]	2020	Bloom-level classification of exam questions	TF–POS–IDF, word2vec with ML classifiers	Multi-domain exam questions (141 + 600 items)	Limited diagnostic error analysis	Does not explain why misclassification repeatedly occurs between adjacent Bloom levels despite acceptable overall accuracy.
[11]	2022	Automatic classification of learning objectives	BERT-based deep learning models	Large corpus of learning objectives (~21k)	Performance-focused evaluation	Does not analyze linguistic ambiguity or dataset properties that contribute to Bloom-level confusion.
[12]	2022	Mapping exam questions to Revised Bloom's taxonomy	NLP feature extraction and supervised classification	Exam question dataset	No structured error taxonomy	Lacks a formal categorization of recurring Bloom-level misclassification types and their underlying causes.
[13]	2023	Survey of Bloom taxonomy automation	Systematic literature review	Multiple datasets	Survey-only, no empirical diagnostics	Does not empirically relate dataset characteristics to observed misclassification behavior.
[14]	2025	Learning outcome classification using Bloom's taxonomy	GPT-4 prompting and BERT comparison	1000 annotated learning outcomes	Prompt sensitivity and disagreement regions	Does not explain the structural or data-level sources of persistent labeling disagreement.
[15]	2025	Bloom's taxonomy classification of assessment questions	Bayesian-optimized ensemble (TF–IDF + transformers)	600 assessment questions (balanced)	Predictive-performance oriented	Does not provide diagnostic interpretation of misclassification patterns beyond performance metrics.

often struggle when confronted with linguistically ambiguous or context-dependent questions. More recent work has explored machine learning and deep learning techniques for Bloom-level classification of learning objectives and assessment questions. These studies generally formulate the problem as a multi-class text classification task and report improvements over rule-based baselines. However, the majority of this literature emphasizes aggregate performance metrics, such as accuracy or F1-score, with limited attention given to the underlying causes of misclassification [8, 9]. In many cases, errors are reported but not examined in a systematic or diagnostic manner.

A smaller body of research has begun to recognize the influence of dataset characteristics and annotation practices on Bloom-level automation. Issues such as class imbalance, overlap between adjacent cognitive levels, and annotation subjectivity have been identified as factors that affect classification behavior. Domain specificity has also been shown to play a role, as models trained on general educational text may struggle when applied to specialized subject matter. Nevertheless, detailed analyses that explicitly link dataset properties to recurring error patterns remain relatively limited. In contrast to performance-oriented studies, the present work adopts a data-centric perspective. Rather than proposing a new classification model or optimizing predictive accuracy, it focuses on examining linguistic ambiguity, task interpretation, and dataset structure that contribute to Bloom-level misclassification. By situating this analysis within existing literature, the study aims to complement prior modeling efforts and to provide context for understanding the limitations observed in automated Bloom-level classification systems. Table 1 summarizes representative published studies (2020–2025) on automated Bloom's taxonomy classification and related educational text analysis, highlighting their task focus, methodologies, dataset characteristics, and commonly reported limitations.

The studies summarized in Table 1 reflect a strong emphasis on predictive modeling and performance-based evaluation. Earlier work primarily explored classical machine learning with handcrafted features, while more recent studies increasingly rely on deep learning and large language models. Despite these methodological differences, a shared pattern can be observed. Bloom-level classification is commonly treated as a conventional multi-class text classification task, and evaluation is largely reported through aggregate accuracy or F1-oriented measures. In many cases, misclassification behavior is reported only implicitly. Errors are rarely examined in relation to dataset composition, cognitive proximity between levels, or linguistic ambiguity within question stems. As a result, prior work offers limited insight into recurring failure patterns that persist across modeling choices, particularly in realistic assessment data characterized by class imbalance and cross-level verb reuse.

Recent studies published in 2025 increasingly explore Bloom-level classification using large language models

or optimized ensemble approaches, with emphasis placed on predictive accuracy, prompt sensitivity, or model comparison. While these works demonstrate improved classification capability, their primary focus remains on output performance rather than on systematic interpretation of misclassification behavior. In contrast, the present study does not introduce a new model or prompting strategy. Instead, it examines why misclassification persists across modeling choices by analyzing dataset structure, linguistic ambiguity, and task formulation. The contribution therefore lies in explanation rather than optimization, offering insights that remain applicable regardless of the specific classification technique employed.

The review of existing literature indicates several unresolved gaps in automated Bloom's taxonomy classification. First, much of the prior work emphasizes predictive performance while offering limited analysis of the underlying sources of misclassification, particularly those arising from dataset structure and linguistic ambiguity. Second, although the limitations of verb-based approaches are often acknowledged, they are rarely examined systematically through data-level diagnostics or structured error analysis. Third, the influence of class imbalance, conceptual proximity between adjacent Bloom levels, and domain-dependent semantics remains underexplored, despite their practical relevance in real assessment settings. This study addresses these gaps by adopting a data-centric perspective that shifts attention from model optimization to diagnostic analysis. Rather than proposing a new classification method, the work examines dataset characteristics, action-verb usage, and recurring error patterns through qualitative inspection. The study provides interpretive insight into why automated Bloom-level classification remains challenging and highlights the need for more cautious evaluation practices.

3 Dataset construction and annotation

3.1 Dataset scope and source

This study is based on a custom dataset consisting of multiple-choice question (MCQ) stems drawn from a single academic domain, specifically computer science. Only the question stems were retained for analysis. Answer options and distractors were excluded by design, as the intention was to focus on the instructional prompt itself. In most automated Bloom-level classification approaches, this prompt constitutes the primary source of linguistic and semantic cues. The dataset was assembled to reflect realistic educational assessment material rather than simplified or artificially generated text. Consequently, the question stems vary in length, phrasing style, and instructional formulation. Some prompts are concise, while others provide more contextual detail. Although the dataset is domain-specific, it spans a range of cognitive demands as defined by Bloom's revised taxonomy, allowing examination of surface-level cues corresponding to the underlying cognitive intent [16, 17]. Computer science MCQs were selected because they frequently rely on concise prompts that combine technical terminology with implicit task requirements. In such questions, action verbs often coexist with domain-specific semantics that influence cognitive interpretation. This makes computer science MCQs particularly suitable for examining verb ambiguity, task–verb mismatch, and adjacent-level confusion. The domain also reflects a common use case for automated assessment systems, where Bloom-level classification is applied to support curriculum analysis and software-driven evaluation.

It should be noted that the dataset was not curated with the aim of maximizing model performance. Instead, it was constructed to support diagnostic analysis of Bloom-level ambiguity and recurring misclassification behavior. This distinction is important for interpreting the analyses presented later in the paper. Across the dataset, MCQ stems are relatively concise. Most questions consist of a single sentence or a short prompt, with limited contextual elaboration. This characteristic mirrors real-world MCQ construction but also constrains the amount of linguistic and semantic information available for inferring cognitive intent. To illustrate analytical points, example question stems are referenced in paraphrased form only. Original items are not reproduced verbatim. The dataset itself is not publicly released due to source restrictions, but it may be made available for research purposes upon reasonable request.

The decision to construct a custom dataset was driven by analytical rather than scale-oriented considerations. Existing publicly available datasets often normalize class distributions, simplify question phrasing, or focus on learning objectives rather than authentic MCQ stems. Such properties reduce linguistic ambiguity and limit examination of misclassification behavior under realistic assessment conditions. In contrast, the present dataset preserves natural class imbalance, domain-specific phrasing, and borderline cases encountered in practice. These characteristics are essential for studying data-level sources of Bloom-level

misclassification, which constitute the primary focus of this work.

3.2 Annotation procedure and reliability

Each question stem was manually annotated according to Bloom's revised taxonomy, encompassing six cognitive process levels: Remember, Understand, Apply, Analyze, Evaluate, and Create. The annotation process emphasized the intended cognitive demand of each question rather than relying exclusively on the presence of specific action verbs. During annotation, attention was given to the task implied by the question stem as a whole. In several cases, questions containing identical or similar verbs were assigned different Bloom levels when the reasoning processes required for a correct response differed. This approach was taken to better approximate authentic assessment practice, where instructional verbs alone do not always determine cognitive complexity.

Bloom-level annotation is, by nature, subject to interpretation. To reduce inconsistency, standard definitions and commonly accepted examples associated with each cognitive level were followed. Importantly, ambiguous cases were not removed. These instances were retained deliberately, as they form a central component of the analysis and reflect challenges encountered in real-world assessment material. Ambiguous cases were handled conservatively. When a question stem exhibited characteristics associated with multiple cognitive levels, the label reflecting the dominant cognitive intent was assigned, and borderline cases were retained rather than excluded. This decision was made to preserve realistic ambiguity commonly encountered in assessment design, rather than artificially enforcing categorical separation.

The annotation process was primarily conducted by a single annotator with domain familiarity in computer science education, with consistency checks performed during dataset construction. Formal inter-annotator agreement measures were not computed, which is acknowledged as a limitation of the dataset. However, this choice aligns with the analytical focus of the study, where the presence of ambiguity itself constitutes an object of investigation rather than noise to be eliminated. The dataset construction and annotation workflow is given in Figure 2 which highlights design choices that preserve realistic ambiguity in MCQ stems.

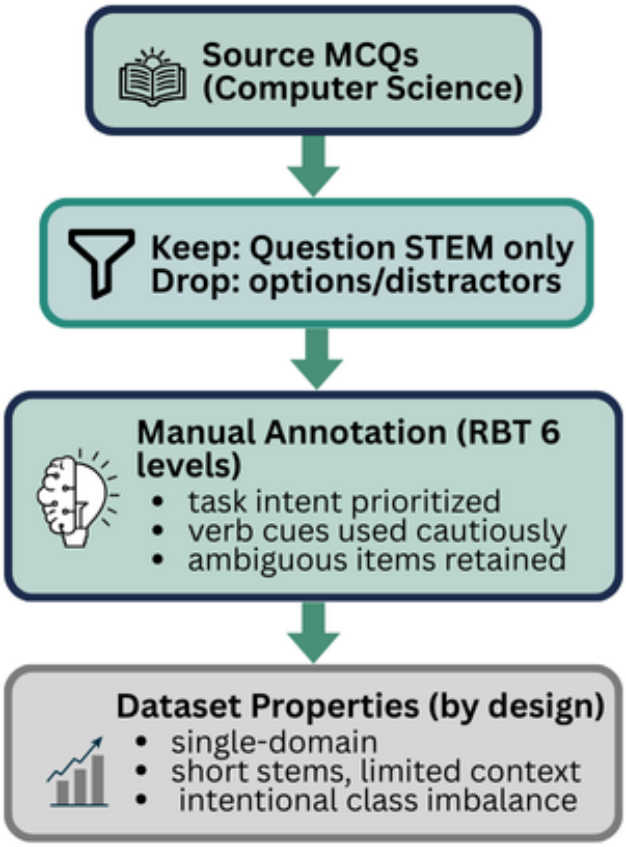

Figure 2. Dataset construction and annotation workflow, highlighting design choices that preserve realistic ambiguity in MCQ stems.

3.3 Class distribution and intentional imbalance

The resulting dataset exhibits a markedly imbalanced class distribution. Lower cognitive levels, particularly Remember and Understand, are represented by a substantially larger number of instances than higher-order levels such as Evaluate and Create. This imbalance was preserved intentionally. In practical educational settings, questions targeting higher-order cognitive processes tend to be less frequent. They often require greater effort to design and assess, and they are therefore used more selectively. Artificially balancing the dataset would risk obscuring these realities and might understate the challenges faced by automated Bloom-level classification systems in practice.

The imbalance is also relevant for subsequent error analysis. Minority classes frequently display linguistic overlap with adjacent cognitive levels, which increases the likelihood of systematic confusion. Rather than treating this as a limitation to be corrected, the dataset design treats imbalance as an analytical condition that helps reveal where and why misclassification occurs.

Annotation decisions were guided by the definitions of the Revised Bloom's Taxonomy rather than by action verbs alone. Each question stem was examined with emphasis on the cognitive operation required to determine the correct answer. Particular attention was given to distinctions between closely related levels such as Apply and Analyze, which frequently appear linguistically similar in short MCQ prompts. Questions were labeled as Apply when the solution required direct execution of a known rule, formula, or procedure in a familiar context. Typical examples included tracing a code fragment, computing a value using a known algorithm, or applying a defined concept to obtain a concrete result. In these cases, the reasoning process primarily involved procedural execution rather than structural interpretation.

In contrast, questions were labeled as Analyze when the task required identifying relationships among components, interpreting internal structure, or reasoning about interactions between multiple elements of a problem. Such cases typically involved decomposition of program behavior, comparison of alternative conditions, or reasoning about the effect of structural changes within a system. When both procedural execution and structural interpretation were present, the label was determined by the dominant cognitive process required for the correct answer. To provide a concise overview of the dataset composition, Table 2 summarizes the distribution of MCQ stems across the six levels of Revised Bloom's Taxonomy. It highlights the pronounced imbalance between lower-order and higher-order cognitive categories that was intentionally retained during dataset construction.

Table 2. Class distribution of the primary MCQ dataset.

Bloom Level	Number of MCQ Stems
Remember	253
Understand	744
Apply	115
Analyze	66
Evaluate	6
Create	5
Total	**1,189**

The dataset was constructed with a diagnostic objective, rather than a predictive one. Its role is to support qualitative and quantitative examination of Bloom-level ambiguity, action-verb overlap, and task–verb mismatches within realistic assessment data. For this reason, no filtering or simplification was applied to remove linguistically ambiguous or conceptually borderline questions. This design choice aligns with the central aim of the present paper, which is to examine the limitations of action-verb-driven Bloom classification from a data-centric perspective. Issues related to model optimization and cross-dataset generalization are therefore not addressed here and are instead deferred to a separate technical study.

4 Dataset diagnostics

4.1 Class distribution characteristics

An initial diagnostic examination of the dataset confirms a pronounced imbalance across Bloom's cognitive levels, which was an expected outcome of the dataset construction and annotation process [18]. Lower-order categories, particularly Remember and Understand, account for a substantial proportion of the available MCQ stems, whereas higher-order levels such as Evaluate and Create are represented by only a small number of instances. While this distribution reflects common assessment practices, its importance here lies in the diagnostic challenges it introduces rather than in its pedagogical justification. Figure 3 summarizes three dataset-level drivers of misclassification risk: class imbalance, adjacent-level overlap, and cross-level verb reuse. From a data-centric perspective, such imbalance influences cognitive boundaries expressed linguistically within the dataset. Minority classes are not only infrequent, but they also tend to share surface-level characteristics with adjacent cognitive levels. As a result, distinctions between neighboring categories become less clearly articulated in the text of the question stems. This condition increases the likelihood of systematic confusion during automated classification, particularly at higher levels of the taxonomy.

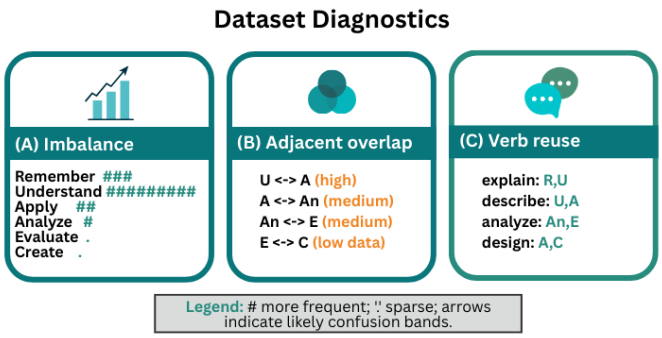

Figure 3. Dataset diagnostics highlighting three data-level drivers of misclassification risk: class imbalance, adjacent-level overlap, and cross-level verb reuse.

The retained class imbalance influences the observed

distribution of misclassification errors. Levels with higher representation naturally produce a greater number of absolute errors, which can create the impression that certain confusions occur more frequently. This effect was considered during analysis. Error patterns were interpreted in relation to relative class presence rather than raw counts alone. The imbalance therefore provides context rather than distortion, allowing misclassification behavior to be examined under conditions that reflect realistic assessment distributions. Importantly, class frequency alone does not fully explain these difficulties. Even when lower-order and higher-order categories are considered separately, overlaps in wording and instructional phrasing persist. The observed distribution therefore serves as a starting point for deeper diagnostic analysis, motivating closer examination of conceptual adjacency and linguistic cue reuse across Bloom levels in the subsequent subsections.

4.2 Overlap between adjacent cognitive levels

In addition to class imbalance, the dataset exhibits notable conceptual overlap between adjacent Bloom levels. This is particularly evident between Understand and Apply, as well as between Analyze and Evaluate. Question stems targeting these levels often differ only marginally in wording, despite requiring distinct forms of reasoning. For instance, questions intended to assess application may employ explanatory language typically associated with understanding. Similarly, evaluation-oriented questions may adopt analytical phrasing without explicitly indicating judgment or justification. Such similarities complicate the use of surface-level cues for reliable cognitive-level identification.

An overlap between Understand and Apply can be observed in questions that combine explanatory language with procedural reasoning. Consider a paraphrased example derived from the dataset. A question stem may state: Explain the output produced by the following code fragment when the variable x is equal to 5. The verb explain suggests an understanding oriented task. The student must actually execute the code step by step to determine the final value. The reasoning process therefore involves applying known execution rules rather than simply interpreting meaning. A related question in the dataset illustrates the neighboring Understand level. A stem may ask students to describe the purpose of a given code segment within a program. In this case the learner interprets the role of the code rather than executing it. Both questions share similar wording and instructional verbs. Their cognitive demands differ. One requires procedural execution. The other requires conceptual interpretation. This similarity in surface phrasing illustrates the conditions under which adjacent level confusion may arise during automated classification. This adjacency effect suggests that Bloom's taxonomy, while hierarchically structured in theory, does not always translate into clearly separable linguistic categories in practice. As a result, misclassification is more likely to occur between neighboring levels than across distant ones.

4.3 Action-verb distribution across classes

Action verbs are commonly treated as key indicators of Bloom levels in automated classification systems [19, 20]. However, analysis of verb usage within the dataset shows that many instructional verbs appear across multiple cognitive categories. Verbs such as explain, describe, analyze, and design are used in questions assigned to different Bloom levels, depending on context. In several instances, identical verbs frame tasks that differ substantially in reasoning depth. This pattern indicates that verb presence alone does not reliably signal cognitive demand. The reuse of instructional verbs across classes weakens the assumption of a direct mapping between verbs and Bloom levels. Instead, the dataset suggests that verbs serve as partial cues that require additional contextual interpretation to infer cognitive intent with greater confidence. To illustrate the reuse of instructional verbs across cognitive categories, Table 3 presents representative examples of action verbs that appear in multiple Bloom levels within the dataset.

Table 3. Illustrative overlap of action verbs across Bloom levels.

Action Verb	Bloom Levels Observed
explain	Remember, Understand
describe	Understand, Apply
analyze	Analyze, Evaluate
design	Apply, Create

4.4 Implications for automated Bloom-Level classification

Considered together, the observed class imbalance, conceptual proximity between cognitive levels, and cross-class verb reuse highlight several challenges for automated Bloom-level classification. Importantly, these challenges arise at the data-level, before any

Table 4. Taxonomy of Bloom-Level Misclassification Errors Observed in the Dataset.

Error Type	Description	Affected Bloom Levels	Representative Example from Dataset
Verb Ambiguity	The instructional verb signals a lower cognitive level than the task actually requires.	Understand ↔ Apply	"Explain the output of the following code segment." The task requires tracing execution, yet the verb suggests understanding rather than application.
Task–Verb Mismatch	The surface verb does not align with the underlying reasoning demand of the task.	Understand ↔ Apply	"Describe the steps performed when the loop executes for the given input." Correct response requires procedural execution rather than description alone.
Adjacent-Level Cognitive Overlap	The task lies near the boundary of neighboring Bloom levels, leading to labeling ambiguity.	Apply ↔ Analyze	"Determine the output produced when the function is called with x = 5." Conditional branching introduces analytical reasoning.
Domain-Specific Semantic Influence	Technical terminology implicitly raises cognitive demand beyond what the verb conveys.	Analyze ↔ Evaluate	"Evaluate the efficiency of the given sorting algorithm." Domain semantics elevate the reasoning requirement.
Surface-Level Cue Dominance	Abstract phrasing triggers higher-level interpretation despite limited task complexity.	Understand ↔ Analyze	"Identify the purpose of the following code fragment." Interpretation is required without structural decomposition.

modeling or algorithmic decisions are introduced. The dataset demonstrates that linguistic signals commonly relied upon by automated systems are inherently ambiguous. As a result, a system may appear acceptable under aggregate summaries while still producing consistent errors among conceptually adjacent classes. Such patterns may remain obscured when evaluation focuses solely on aggregate metrics [21]. These diagnostic observations motivate the error analysis presented in the following section, where misclassification patterns are examined in greater detail to identify recurring failure modes and their underlying sources.

5 Formal error taxonomy for Bloom-level classification

Bloom's taxonomy offers a well-established conceptual framework for categorizing cognitive processes. However, its application in automated classification systems remains challenging. Many existing approaches implicitly treat misclassifications as either random or primarily dependent on model choice. Closer inspection of the dataset suggests otherwise. Analysis indicates that errors tend to follow recurring and systematic patterns [22, 23]. The proposed taxonomy is grounded in dataset inspection and qualitative analysis, with emphasis placed on linguistic ambiguity, task interpretation, and contextual dependence rather than on algorithmic factors. Table 4 summarizes the proposed taxonomy of Bloom-level classification errors. The taxonomy was derived through systematic manual inspection of the annotated dataset rather than through analysis of the outputs of a trained classification model. Question stems were examined in relation to their assigned Bloom levels and compared across cases that exhibited similar linguistic patterns or task structures. During this inspection, recurring forms of ambiguity were noted, particularly when the instructional verb suggested a different cognitive level than the reasoning process required to answer the question.

These recurring patterns were grouped into broader categories representing common sources of Bloom-level confusion. The process therefore followed an inductive analytical approach in which error types were identified from repeated observations within the dataset. The resulting taxonomy reflects data-level characteristics such as verb reuse, task structure, and domain semantics rather than behavior of any specific classification algorithm. The representative examples shown in the table are paraphrased forms derived from actual question stems in the dataset. Original items are not reproduced verbatim due to dataset sharing restrictions. Each example therefore preserves the linguistic structure and task characteristics of the source question while avoiding direct reproduction

Figure 4. Analytical workflow used to derive the Bloom-level misclassification taxonomy through systematic inspection of annotated MCQ stems.

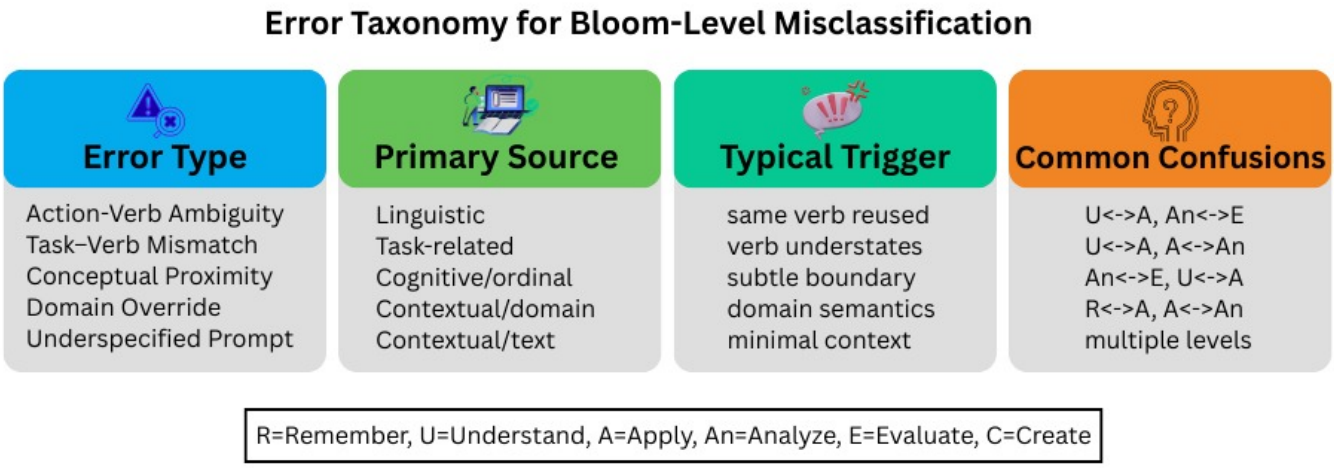

Figure 5. Visual summary of the proposed misclassification taxonomy, linking error types to their dominant source and typical adjacent-level confusions.

of assessment material. These paraphrased examples illustrate the types of task and verb interactions that produced recurring misclassification tendencies during dataset inspection. An analytical workflow used to derive the Bloom-level misclassification taxonomy through systematic inspection of annotated MCQ stems is presented in the Figure 4.

This process allowed recurring error types to be abstracted from data characteristics themselves, independent of model behavior. The taxonomy can be used as a diagnostic reference for examining Bloom-level misclassification beyond aggregate performance measures. For researchers, it provides a structured lens for analyzing error patterns across datasets and annotation schemes. For software designers, it offers guidance for identifying failure-prone conditions in automated assessment systems and for interpreting classification outputs with greater caution during system evaluation.

5.1 Linguistic and task-related sources of misclassification

A prominent source of misclassification arises from ambiguity in instructional verbs. Verbs commonly associated with Bloom's taxonomy do not correspond uniquely to individual cognitive levels. Terms such as explain, describe, and analyze are used across multiple levels, with their cognitive implication shaped largely by context [24]. Within the dataset, identical verbs appear in questions labeled as Understand, Apply, and Analyze. These questions often differ only in the depth of reasoning required. Such reuse weakens verb-based heuristics and contributes to confusion between neighboring categories when classifiers rely heavily on lexical cues. Another recurring error type occurs when the surface verb understates the actual cognitive demand of the task [25]. In these instances, the verb implies a lower-order process, while successful completion requires higher-level reasoning. For example, prompts framed using descriptive verbs may nonetheless require learners to apply concepts in novel situations or reason through constraints. Automated systems that prioritize verb cues over task structure are therefore susceptible to systematic underestimation of cognitive level. A visual summary of the proposed misclassification taxonomy, linking error types to their dominant source and typical adjacent-level confusions is provided in Figure 5.

5.2 Structural and contextual sources of misclassification

Misclassification is frequently observed between adjacent Bloom levels, reflecting their conceptual closeness. Distinctions between Analyze and Evaluate, or between Understand and Apply, are often subtle and highly dependent on context [26]. While such distinctions may be apparent to experienced educators, they are not always linguistically explicit. As a result, classification errors tend to cluster between neighboring levels rather than occurring randomly across the taxonomy. This pattern supports the view that Bloom's categories form a continuum of cognitive complexity, rather than discrete and easily separable classes [27]. A further pattern involves domain-semantic override, where subject-matter context shapes the cognitive interpretation of otherwise simple verbs. In specialized domains, including medicine and engineering, verbs such as identify or classify may require reasoning that extends beyond basic recall. In these cases, lexical cues alone are insufficient to capture the depth of cognitive processing involved. This limitation helps explain why classifiers trained on general-purpose datasets often encounter difficulty when applied to domain-specific content.

Some misclassifications stem from underspecified question stems. These prompts lack the contextual detail needed to clearly indicate the intended cognitive level, even during manual annotation [28]. Short or generic questions may reasonably align with multiple Bloom categories depending on interpretation. Rather than removing such instances, they were retained to reflect the nature of authentic assessment material. Their presence highlights the inherent ambiguity of Bloom-level classification and cautions against excessive confidence in automated labeling systems.

While the error taxonomy outlined above provides a structured way to categorize misclassification types in Bloom-level question analysis, it does not by itself explain these errors arising during automated classification. The taxonomy captures what goes wrong, but the underlying mechanisms remain implicit. To address this gap, the following section examines the error categories emerging in practice by analytically relating them to dataset composition, linguistic ambiguity, and the structural properties of MCQ stems. Rather than reporting experimental outcomes, the discussion focuses on explaining the conditions under which verb-centric Bloom classification is likely to fail, thereby grounding the proposed taxonomy in data design and cognitive considerations. The illustration of the three primary sources of misclassification identified in the proposed error taxonomy is given in Figure 6.

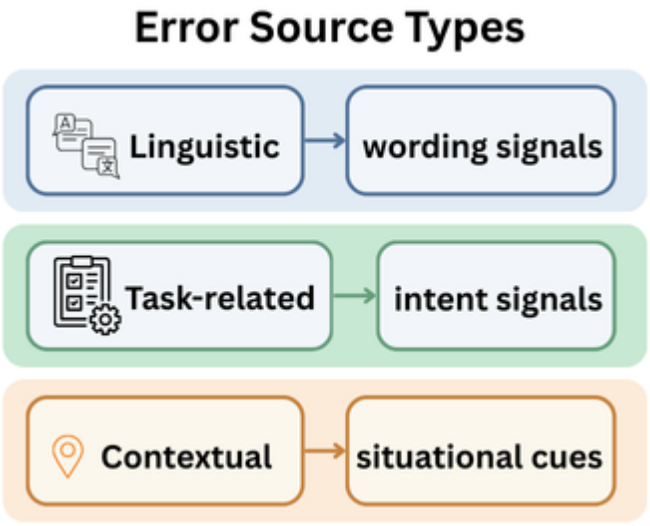

Figure 6. Primary sources of misclassification identified in the proposed error taxonomy.

6 Analytical demonstration of Bloom-level misclassification

This section provides a theory-driven examination of the error categories identified earlier. The analysis explains recurring misclassification patterns by relating them to dataset characteristics and linguistic ambiguity rather than to predictive model behavior. By grounding the analysis in data design and cognitive theory, this section clarifies the mechanisms that make verb-centric Bloom classification inherently fragile, particularly in the context of multiple-choice question (MCQ) stems.

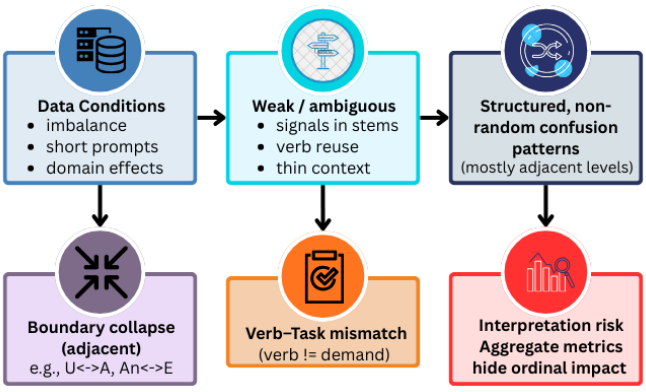

Figure 7. Conceptual framework linking dataset conditions to predictable adjacent-level confusion and interpretation risks.

Figure 7 illustrates the conceptual process through which dataset characteristics contribute to recurring Bloom-level misclassification. The framework begins with a set of data conditions present in the dataset. These include class imbalance across cognitive levels, short MCQ stems that contain limited contextual information, and the influence of domain-specific terminology. Each of these conditions constrains the linguistic signals available for identifying cognitive demand. Under these conditions, question stems often contain weak or ambiguous cues. Instructional verbs are frequently reused across multiple Bloom levels. Contextual detail may also be limited because MCQ prompts are intentionally concise. These linguistic characteristics reduce the clarity of cognitive boundaries expressed in the text of the question stems.

When automated classification relies primarily on such surface cues, confusion tends to emerge in a structured manner. Misclassification rarely occurs across distant cognitive categories. Instead, errors concentrate between neighboring levels where linguistic and conceptual overlap is strongest. Examples include confusion between Understand and Apply or between Analyze and Evaluate. The interaction of these factors produces a gradual weakening of the boundaries separating adjacent Bloom levels. As a result, distinctions that are conceptually meaningful in educational theory become difficult to detect through textual signals alone. The final component of the framework highlights the implications for system interpretation. When evaluation focuses mainly on aggregate accuracy, these structured errors may remain partially hidden even though they alter the intended cognitive distribution of an assessment.

6.1 Data and linguistic sources of Boundary confusion

As described earlier, the dataset exhibits a deliberate imbalance across Bloom's cognitive levels. Lower-order categories such as Understand and Remember are heavily represented, whereas higher-order levels such as Evaluate and Create occur only sparsely. This imbalance was not treated as a flaw to be corrected, but as a structural condition reflecting real assessment practices, where higher-order questions are relatively rare.

From an analytical standpoint, such imbalance has important implications for Bloom-level discrimination. When only a small number of instances exist for advanced cognitive categories, the linguistic signals associated with those categories become statistically weak. Verbs commonly associated with higher-order thinking appear infrequently and often overlap with verbs used at lower levels. As a result, the lexical boundary between adjacent levels becomes blurred. This effect is especially pronounced between Analyze and Evaluate, as well as between Apply and Create, where surface-level verb cues are insufficient to establish stable distinctions. It is important to note that this boundary collapse is not tied to any specific modeling choice. Rather, it emerges from the interaction between dataset composition and the hierarchical nature of Bloom's taxonomy itself. When higher-order levels are underrepresented, their semantic identity is diluted within the broader distribution of question stems. Any automated system that relies heavily on frequency-based cues, whether explicitly or implicitly, is therefore likely to struggle with these categories.

Verb ambiguity across Bloom levels remains a persistent source of misclassification because identical instructional verbs may correspond to different cognitive processes depending on task context. In isolation, these verbs do not uniquely encode the intended level of thinking. This ambiguity is particularly problematic for automated classification systems, as verb lists are often treated as primary indicators of cognitive level. When the same verb appears across Understand and Apply, or across Analyze and Evaluate, the classifier is forced to rely on surrounding lexical context to resolve the ambiguity. However, MCQ stems frequently provide limited contextual detail, especially when constrained by brevity or standardized formatting.

The consequence is a tendency toward adjacent-level confusion rather than random mislabeling. Questions are rarely shifted from Remember directly to Create. Instead, they are misassigned to neighboring levels that share overlapping linguistic markers. This pattern aligns with Bloom's hierarchical structure, where cognitive levels are conceptually related and often differ by degree rather than by kind. As a result, misclassification reflects structural ambiguity in the taxonomy rather than simple noise.

6.2 Task and contextual sources of systematic misclassification

A second major source of error lies in the mismatch between the action verb used in a question stem and the actual cognitive task required to answer it. In MCQ design, verbs are often selected to signal intent, but the constraints of the format limit the intent which can

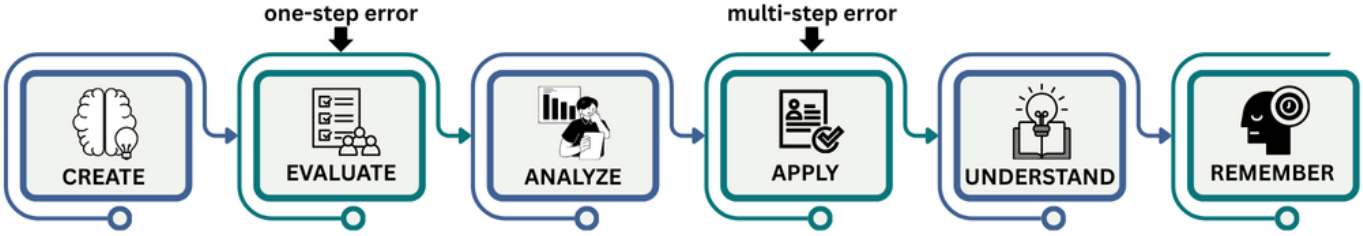

Figure 8. Ordinal structure of revised Bloom's taxonomy illustrating qualitative differences between adjacent (one-step) and non-adjacent (multi-step) misclassification errors.

be expressed. A question may use a verb associated with higher-order thinking while still requiring only surface-level reasoning to identify the correct option.

For instance, a stem that begins with analyze may simply ask the student to recognize a familiar pattern, rather than to decompose a problem or evaluate relationships among components. Similarly, verbs such as design or propose may be used in contexts where the answer options already contain the essential reasoning, reducing the task to selection rather than creation. In such cases, the verb suggests a higher cognitive level than the task actually demands. This task–verb mismatch is not accidental. It reflects common practices in assessment construction, where verbs are used aspirationally to elevate perceived difficulty. Automated systems that rely on verb cues are therefore exposed to a systematic distortion. The verb signals one cognitive level, while the task structure implies another. Without access to deeper semantic or procedural cues, the system has no reliable way to reconcile this discrepancy.

Domain specificity further complicates Bloom-level classification. Identical verbs can imply different cognitive operations depending on the subject matter in which they appear. For example, apply in a programming context may involve algorithmic reasoning, while apply in a theoretical context may require recall of a definition. Similarly, evaluate in a medical question may entail ethical judgment, whereas in a computer science question it may involve performance comparison. In the dataset examined here, domain-specific semantics often override generic verb meanings. The same verb may correspond to different levels of cognitive engagement depending on the underlying knowledge structure of the domain. This effect is especially visible when classification systems trained on single-domain data are confronted with questions drawn from another discipline.

From an analytical perspective, this highlights a fundamental limitation of verb-centric approaches. Verbs function as weak proxies for cognition when divorced from domain knowledge. Without explicit modeling of subject-specific semantics, automated systems are prone to systematic errors that cannot be resolved through lexical analysis alone. This issue persists regardless of dataset size or annotation quality, as it is rooted in the interaction between language and disciplinary practice.

6.3 Implications for interpreting classification outputs

A final concern relates to the way Bloom-level classification behavior is commonly interpreted. Even when aggregate summaries appear acceptable, this metric may conceal errors that are pedagogically significant. Misclassifying a Create-level question as Evaluate, or an Apply-level question as Understand, may have substantial implications for assessment balance, despite contributing minimally to overall error rates. Bloom's taxonomy is inherently ordinal. Adjacent-level confusion, while numerically minor, undermines the intended cognitive distribution of an assessment. When higher-order questions

Table 5. Error sources, and expected misclassification effects.

Analytical Factor	Error Source	Expected Misclassification Pattern
Class imbalance	Diluted linguistic signals	Boundary collapse between adjacent levels
Verb reuse across levels	Action-verb ambiguity	Confusion between neighboring cognitive categories
Verb–task inconsistency	Task–verb mismatch	Over- or under-estimation of cognitive level
Domain specificity	Domain-semantic override	Cross-domain misinterpretation of verbs
Ordinal evaluation focus	Aggregate accuracy reliance	Masking of pedagogically significant errors

are systematically shifted downward, assessments may appear balanced on paper while failing to evaluate advanced thinking in practice. The Ordinal structure of revised Bloom's taxonomy illustrating qualitative differences between adjacent (one-step) and non-adjacent (multi-step) misclassification errors is presented in Figure 8. Error pattern analysis is important because overall accuracy provides only a summary view of classifier behavior. In Bloom-level classification, many misclassifications occur between adjacent cognitive levels, where small shifts can alter the intended cognitive balance of an assessment. Such errors may have limited effect on aggregate accuracy. Their conceptual and system-level implications can still be significant. Examining error patterns allows systematic failure modes to be identified and interpreted in relation to dataset structure and task formulation, rather than being obscured by performance summaries.

This observation reinforces the central argument of this paper. Performance metrics alone are insufficient to judge the effectiveness of Bloom-level classification systems. Understanding how and why errors occur is equally important. Dataset design, verb distributions, and task structure must be considered alongside predictive outcomes if automated assessment tools are to be used responsibly. The sources of error and expected misclassification effects are illustrated in Table 5.

The analytical factors summarized in Table 5 correspond to the error categories introduced earlier in Table 4. Each factor represents a structural condition within the dataset that gives rise to one or more types of misclassification described in the taxonomy. The relationship between the two tables therefore reflects a progression from observed error patterns to their underlying sources within the data.

Factors associated with verb reuse and surface phrasing correspond to the error types labeled verb ambiguity and surface-level cue dominance. These errors occur when the instructional verb provides an incomplete signal of cognitive demand. Factors related to task structure correspond to the category described as task verb mismatch, in which the action verb does not accurately reflect the reasoning required to answer the question. Similarly, factors associated with conceptual proximity between cognitive levels correspond to the adjacent level cognitive overlap identified in the taxonomy. Domain-specific semantic effects map to the error type described as domain semantic influence, where disciplinary context alters interpretation of otherwise simple verbs. Through this mapping, Table 5 identifies the dataset level conditions that produce the error patterns summarized in Table 4. The taxonomy therefore describes the observable forms of misclassification, whereas the analytical factors explain the structural circumstances under which those patterns emerge.

7 Discussion

The findings of this study directly address the research questions introduced earlier in the paper. The analysis of adjacent-level confusion and verb-centric ambiguity responds to RQ1 by clarifying limitations in distinguishing closely related Bloom levels. The examination of dataset properties, including domain specificity and intentional imbalance, addresses RQ2 by showing their influence on observed misclassification patterns. The identified error taxonomy and recurring failure modes respond to RQ3 by characterizing systematic errors that arise when cognitive intent is inferred from surface-level linguistic cues.

The analysis presented in this study indicates that many of the challenges associated with automated Bloom-level classification arise at the level of data representation and task definition, rather than from shortcomings of particular modeling techniques. The dataset diagnostics and the error taxonomy suggest that Bloom's cognitive categories, although pedagogically meaningful, do not always correspond to clearly distinguishable linguistic patterns in authentic assessment material. A recurring observation concerns the reliance on instructional verbs as indicators of cognitive demand. Although such verbs provide useful cues, their repeated use across multiple Bloom levels limits their discriminative value. In practice, the same verb may frame tasks that differ considerably in reasoning depth, depending on context, constraints, and domain knowledge. As a result, verb-centric heuristics, whether rule-based or implicitly learned by classifiers, appear fragile under realistic conditions.

The findings also highlight the influence of conceptual proximity between adjacent Bloom levels. Misclassification tended to cluster along neighboring categories, reflecting the gradual and overlapping nature of cognitive processes. This behavior is consistent with educational theory, which recognizes that higher-order thinking often builds incrementally upon lower-order skills. Consequently,

while Bloom levels are conceptually distinct, their linguistic realization may be blurred. Domain semantics emerged as another important factor. Questions situated within specialized subject areas often require background knowledge that is not explicitly encoded in surface text. In such cases, cognitive demand cannot be inferred reliably from lexical cues alone. This has implications for deploying Bloom classification systems in domain-specific settings, where models trained on general datasets may exhibit systematic biases.

The absence of formal inter-annotator agreement represents an acknowledged limitation of the dataset and may influence the interpretation of individual Bloom level labels. Annotation of cognitive levels often involves judgment, particularly near the boundaries between adjacent categories such as Apply and Analyze. In such cases, different annotators may reasonably assign alternative labels when interpreting the reasoning demand implied by a question stem. This limitation may affect the stability of labels assigned to borderline questions. Some stems located near cognitive boundaries could plausibly shift to a neighboring category if reviewed by additional annotators. The effect would most likely appear in adjacent level distinctions rather than in distant category changes. For instance, a question labeled as Apply could potentially be interpreted as Analyze when emphasis is placed on structural reasoning rather than procedural execution. Despite this limitation, the analytical findings of the study remain relevant because the investigation focuses on recurring sources of ambiguity within the dataset itself. The analysis does not depend on precise numerical performance measures or model comparisons. Instead, it examines linguistic cues, task structure, and dataset composition that influence cognitive interpretation. From this perspective, the presence of some labeling uncertainty reflects the same interpretive challenges encountered by automated systems when Bloom levels are inferred from short question prompts.

The findings of this study suggest several practical steps for researchers and developers who design automated Bloom-level classifiers. First, dataset preparation should include careful inspection of question stems for ambiguous verb usage. When possible, additional contextual cues such as task descriptions, procedural requirements, or domain-specific terminology should be retained during preprocessing rather than removed.

Second, feature design should extend beyond isolated action verbs. Signals related to task structure can provide stronger indicators of cognitive demand. Examples include the presence of conditional reasoning, requirement to trace algorithmic steps, comparison between alternatives, or identification of relationships among components. These features often reflect the reasoning process required to answer the question. Third, evaluation procedures should examine error patterns between adjacent cognitive levels rather than relying only on overall accuracy. Confusion matrices and qualitative inspection of borderline cases can help identify systematic misclassification tendencies. Such analysis allows developers to detect cases where questions intended for higher cognitive levels are consistently interpreted as lower level tasks. Finally, when Bloom level classifiers are integrated into educational software systems, their outputs should be interpreted cautiously. Automated labels may serve as preliminary indicators rather than definitive judgments of cognitive demand. Manual review may remain necessary for question stems located near cognitive boundaries, particularly in datasets containing short prompts or strong domain-specific terminology. Relying less on action verbs introduces a trade-off between simplicity and interpretability. Verbs are easy to extract and align with instructional conventions, but they provide weak signals when used in isolation. Alternative features should therefore emphasize task structure, such as the presence of procedural steps, comparison requirements, or decomposition of components. Contextual cues drawn from surrounding nouns, operators, and domain-specific constructs can further support cognitive inference. While these features increase analytical complexity, they offer more stable indicators of cognitive demand than surface-level verbs alone. The key findings of the study are summarized in Figure 9.

Overall, the findings suggest that improvements in automated Bloom-level classification are unlikely to result solely from increasing model complexity or expanding training data. Instead, greater emphasis is required on dataset design, contextual representation, and careful interpretation of classification outputs. Practical approaches may involve richer task descriptions or selective human oversight for ambiguous cases. Finally, the data-centric perspective adopted in this work underscores the importance of diagnostic analysis prior to model optimization. Without understanding the structural sources of

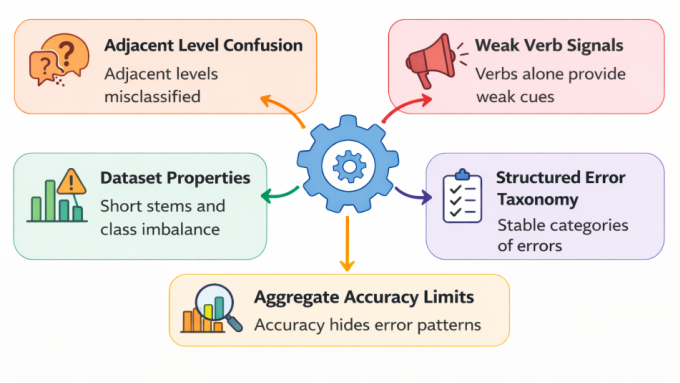

Figure 9. Key findings highlighting factors that contribute to misclassification in automated Bloom-level classification of MCQ stems.

ambiguity present in the data, apparent performance gains may conceal persistent weaknesses rather than resolve them. The insights presented here therefore provide a basis for more cautious and context-aware use of automated Bloom-level classification systems.

The key findings of the study are summarized as follows:

- Automated Bloom-level classification shows systematic confusion between adjacent cognitive levels, particularly when action verbs are used as primary cues.
- Dataset properties such as class imbalance, short question stems, and domain-specific semantics strongly influence misclassification behavior.
- Action verbs alone provide weak signals of cognitive demand when task structure and reasoning requirements are not considered.
- Recurring error patterns can be organized into a structured taxonomy that remains stable across annotation and inspection stages.
- Aggregate accuracy obscures meaningful misclassification tendencies, making error-pattern analysis necessary for reliable interpretation.

8 Limitations and future work

Several limitations of the present study should be noted. Bloom-level annotation involves interpretive judgment, and even when standard definitions are followed, some degree of variability is difficult to avoid. This issue becomes more apparent at higher cognitive levels, where distinctions often depend on subtle differences in task intent and depth of reasoning. The dataset used in this study has several limitations that should be acknowledged. Its size is modest, which constrains the ability to observe rare misclassification patterns with high frequency. The single-domain focus on computer science MCQs further limits generalization to other disciplines where linguistic structure and task formulation may differ. These constraints mean that the observed error patterns should be interpreted as illustrative rather than exhaustive. Despite this, the dataset remains suitable for theory-driven inspection, as it reflects realistic assessment conditions under which automated Bloom-level classification is commonly applied.

A further limitation relates to the scope of the dataset, which is restricted to a single academic domain. While this focus allows for controlled and detailed diagnostic analysis, it limits the extent to which the observed error patterns can be generalized to other disciplines. Different subject areas may employ distinct linguistic conventions and assessment styles, which could influence the cognitive demand to be expressed. The dataset also exhibits a naturally imbalanced class distribution, with comparatively fewer instances representing higher-order Bloom levels. Although this imbalance was intentionally retained to reflect typical assessment practices, it may accentuate error tendencies for minority classes. The findings should therefore be interpreted within the context of realistic, rather than idealized, data conditions.

Future research should evaluate the generality of the proposed misclassification taxonomy by applying it to a multiple-choice question dataset from a different academic domain, such as biology. A comparative analysis across domains would allow assessment of whether the same error types and adjacent-level confusions persist under different semantic and task structures. Additional work may also examine the interaction between data-centric diagnostics and limited modeling interventions to test whether identified error patterns remain stable across classification approaches. These directions extend beyond the present study, which is focused on explaining structural sources of Bloom-level misclassification within a single-domain dataset. Future work should also address annotation subjectivity more explicitly. This may involve incorporating multiple annotators with domain expertise and examining areas of disagreement between closely related Bloom levels. Structured adjudication protocols and clearer task-oriented annotation guidelines could further reduce ambiguity.

Such steps would support more consistent labeling while preserving the realistic uncertainty inherent in Bloom-level interpretation.

9 Conclusion

This paper examined Bloom-level classification from a data-centric perspective, with emphasis on dataset structure, linguistic ambiguity, and recurring misclassification patterns. Through dataset diagnostics and systematic error analysis, the analysis suggests that many classification difficulties arise before any modeling decisions are made. Action verbs, while informative, function as incomplete proxies for cognitive demand, as similar phrasing can support different forms of reasoning and distinctions between adjacent Bloom categories are often linguistically under-specified. As a result, misclassification is often observed to follow predictable patterns rather than occurring at random. These observations indicate that improvements in automated Bloom-level classification depend not only on modeling choices, but also on careful consideration of dataset design and task interpretation. By highlighting structural sources of ambiguity in MCQ stems, this study contributes a diagnostic framework for interpreting classifier behavior and for guiding future work on Bloom taxonomy automation. Together, these contributions shift attention from model-centric performance reporting to data-centric explanation, enabling systematic analysis of recurring error patterns driven by linguistic cues, task structure, and dataset properties.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

AI Use Statement

The authors declare that no generative AI was used in the preparation of this manuscript.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Sobral, S. R. (2021). Bloom's taxonomy to improve teaching-learning in introduction to programming. *International Journal of Information and Education Technology, 11*(3), 148–153. [CrossRef]

[2] Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into practice, 41*(4), 212-218. [CrossRef]

[3] Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives: complete edition*. Addison Wesley Longman, Inc.

[4] Sucipto, S., Prasetya, D. D., & Widiyaningtyas, T. (2024). A review questions classification based on Bloom taxonomy using a data mining approach. *ITEGAM-JETIA, 10*(48), 161-170. [CrossRef]

[5] Crowe, A., Dirks, C., & Wenderoth, M. P. (2008). Biology in bloom: implementing Bloom's taxonomy to enhance student learning in biology. *CBE—Life Sciences Education, 7*(4), 368-381. [CrossRef]

[6] Santos, O. C., & Boticario, J. G. (2015). Practical guidelines for designing and evaluating educationally oriented recommendations. *Computers & Education, 81*, 354-374. [CrossRef]

[7] Ullah, Z., Lajis, A., Jamjoom, M., Altalhi, A., & Saleem, F. (2020). Bloom's taxonomy: A beneficial tool for learning and assessing students' competency levels in computer programming using empirical analysis. *Computer Applications in Engineering Education, 28*(6), 1628-1640. [CrossRef]

[8] Zhang, J., Wong, C., Giacaman, N., & Luxton-Reilly, A. (2021, February). Automated classification of computing education questions using Bloom's taxonomy. In *Proceedings of the 23rd Australasian computing education conference* (pp. 58-65). [CrossRef]

[9] Li, Y., Rakovic, M., Poh, B. X., Gaševic, D., & Chen, G. (2022). Automatic Classification of Learning Objectives Based on Bloom's Taxonomy. *International Educational Data Mining Society*. [CrossRef]

[10] Mohammed, M., & Omar, N. (2020). Question classification based on Bloom's taxonomy cognitive domain using modified TF-IDF and word2vec. *PloS one, 15*(3), e0230442. [CrossRef]

[11] Sebbaq, H., & El Faddouli, N. E. (2022). Fine-tuned BERT model for large scale and cognitive classification of MOOCs. *International Review of Research in Open and Distributed Learning, 23*(2), 170-190. [CrossRef]

[12] Mustafidah, H., Suwarsito, S., & Pinandita, T. (2022). Natural language processing for mapping exam questions to the cognitive process dimension. *International Journal of Emerging Technologies in Learning (iJET), 17*(13), 4-16. [CrossRef]

[13] Rawat, A., Kumar, S., & Samant, S. S. (2023, July). A systematic review of question classification techniques based on bloom's taxonomy. In *2023 14th international*

conference on computing communication and networking technologies (*ICCCNT*) (pp. 1-7). IEEE. [CrossRef]

[14] Almatrafi, O., & Johri, A. (2025). Leveraging generative AI for course learning outcome categorization using Bloom's taxonomy. *Computers and Education: Artificial Intelligence, 8*, 100404. [CrossRef]

[15] Alammary, A., & Masoud, S. (2025). Towards Smarter Assessments: Enhancing Bloom's Taxonomy Classification with a Bayesian-Optimized Ensemble Model Using Deep Learning and TF-IDF Features. *Electronics, 14*(12), 2312. [CrossRef]

[16] Talha. (2025). AI-Ready Large-Scale Student Performance Dataset: Quiz Assessments, Bloom's Taxonomy & ML Applications [Data set]. Zenodo. [CrossRef]

[17] Patil, P. M., Bhavsar, R. P., & Pawar, B. V. (2024). Bloom's Taxonomy based automatic Marathi question generation. *International Journal of Advanced Technology and Engineering Exploration, 11*(113), 575. [CrossRef]

[18] Prasetya, D. D., & Widiyaningtyas, T. (2024, December). An Evaluation of the Impact of Dataset Size on Classification Performance in the Cognitive Bloom's Taxonomy. In *2024 Beyond Technology Summit on Informatics International Conference* (*BTS-I2C*) (pp. 131-136). IEEE. [CrossRef]

[19] Mazza, A., El Makkaoui, K., Ouahbi, I., & Maleh, Y. (2025, June). Automating Educational Assessment with AI: Leveraging Bloom's Taxonomy and Transformer Models for Question Classification. In *2025 International Conference on Circuit, Systems and Communication* (*ICCSC*) (pp. 1-5). IEEE. [CrossRef]

[20] Banujan, K., Kumara, S., Prasanth, S., & Ravikumar, N. (2023). Revolutionising Educational Assessment: Automated Question Classification Using Bloom's Taxonomy and Deep Learning Techniques–A Case Study on Undergraduate Examination Questions. *International Journal of Education and Development using Information and Communication Technology, 19*(3), 259-278.

[21] Shaikh, S., Daudpotta, S. M., & Imran, A. S. (2021). Bloom's learning outcomes' automatic classification using lstm and pretrained word embeddings. *IEEE Access, 9*, 117887-117909. [CrossRef]

[22] Gavhane, J. M., & Pagare, R. (2025, April). Revolutionizing Academic Evaluation: Bloom's Taxonomy Meets Deep Learning and NLP. In *2025 IEEE Global Engineering Education Conference* (*EDUCON*) (pp. 1-10). IEEE. [CrossRef]

[23] Lee, K. W., Ang, Y. S., & Lai, J. W. (2025). Learning Analytics with Scalable Bloom's Taxonomy Labeling of Socratic Chatbot Dialogues. *Computers, 14*(12), 555. [CrossRef]

[24] Azzi, A., Erdős, F., Németh, R., Varadarajan, V., & Afrifa, S. (2025). Comparative analysis of NLP-driven MCQ generators from text sources. *Computers and Education: Artificial Intelligence, 9*, 100440. [CrossRef]

[25] Nevid, J. S., & McClelland, N. (2013). Using Action Verbs as Learning Outcomes: Applying Bloom's Taxonomy in Measuring Instructional Objectives in Introductory Psychology. *Journal of Education and Training Studies, 1*(2), 19-24. [CrossRef]

[26] Cammies, C., Cunningham, J. A., & Pike, R. K. (2024). Not all Bloom and gloom: Assessing constructive alignment, higher order cognitive skills, and their influence on students' perceived learning within the practical components of an undergraduate biology course. *Journal of Biological Education, 58*(3), 588-608. [CrossRef]

[27] Gani, M. O., Ayyasamy, R. K., Sangodiah, A., & Fui, Y. T. (2023). Bloom's Taxonomy-based exam question classification: The outcome of CNN and optimal pre-trained word embedding technique. *Education and Information Technologies, 28*(12), 15893-15914. [CrossRef]

[28] Das, S., Das Mandal, S. K., & Basu, A. (2022). Classification of action verbs of Bloom's taxonomy cognitive domain: An empirical study. *Journal of Education, 202*(4), 554-566. [CrossRef]

Maleeha Mahboob received her MS degree in Computer Science from Virtual University of Pakistan. Her research interests include educational data analytics, learning assessment systems, and the application of intelligent techniques in education. She is particularly interested in the structuring, evaluation, and cognitive alignment of educational content to enhance learning effectiveness. Her work primarily focuses on data-driven approaches to the design and evaluation of educational software systems. (Email: maleeha.mahboob2@gmail.com)