IEA

IEA Research_
and Development_

An AI-Driven Approach to TIMSS Item Verification and Alignment

Ummugul Bezirhan[1]
Matthias von Davier[1]

[1]Boston College, Chestnut Hill, MA, USA

DOI: https://doi.org/10.58150/rd_4_1

Summary

This project explored how artificial intelligence (AI) can support the development and review of TIMSS assessment items by automating aspects of framework alignment. This project used a comprehensive, machine-readable item bank spanning mathematics and science items from 1995 to 2023, integrating item text, response options, metadata, psychometric information, and AI-generated captions of visual content. This resource was used for checking the ability of LLMs to align items to the TIMSS framework, enabling a systematic analysis of item characteristics across assessment cycles and has direct utility for TIMSS item development and review processes.

Using the pre-existing item bank, the project evaluated LLM and supervised machine learning approaches for classifying items by TIMSS content domains, cognitive domains, and difficulty levels. Results showed that AI methods can support content-domain alignment with strong agreement to expert classifications. Cognitive-domain classification was more challenging, reflecting overlap among framework categories, while difficulty prediction proved the most challenging task, suggesting that textual features alone may not fully capture the factors that determine empirical item difficulty.

Overall, the findings suggest that AI can strengthen TIMSS item development through decision-support for item screening, framework verification, and research use, while expert judgment remains essential for complex classifications and difficulty evaluation.

This work was supported in whole or part by the IEA Research and Development (R&D) Funds, which aim to advance and improve the science and methodology of IEA studies, ensuring that they remain at the forefront of international large-scale assessments in education. Funding is drawn from an overhead on participation fees across IEA studies. This does not constitute endorsement of the contents, which reflects the views only of the authors. IEA cannot be held responsible for any use which may be made of the information contained therein.

An AI-Driven Approach to TIMSS Item Verification and Alignment

Ummugul Bezirhan[1] and Matthias von Davier[1]

[1]Boston College, Chestnut Hill, MA, USA

Abstract

This study investigates the feasibility of using artificial intelligence (AI) to automate the alignment of assessment items with the Trends in International Mathematics and Science Study (TIMSS) framework. Specifically, it evaluates prompt-based large language models (LLMs) and supervised machine learning approaches across three classification dimensions: content domain, cognitive domain, and item difficulty. To support this analysis, a comprehensive, structured item bank was constructed from TIMSS assessment cycles spanning 1995 to 2023, including the integration of AI-generated image captions for non-textual elements. Results indicate that prompt-based LLMs achieve high accuracy in content-domain classification, demonstrating that semantically distinct constructs can be reliably recovered from item text when guided by framework definitions. Supervised models show comparable patterns, with performance declining as construct boundaries become less discrete. Cognitive-domain classification shows moderate performance across both approaches, reflecting inherent conceptual overlap between categories. Difficulty prediction remains the most challenging task, with models explaining only a limited proportion of variance, underscoring the dependence of difficulty on empirical psychometric data beyond textual features. Overall, findings suggest that AI-based alignment tools are most effective as decision-support instruments that augment expert judgment in large-scale assessment development, particularly for content alignment and first-pass screening tasks.

Keywords: TIMSS; large-scale assessments; item alignment; large language models; natural language processing

Introduction

International large-scale assessments (ILSAs) play a critical role in informing educational policy, curriculum development, and instructional practice. The Trends in International Mathematics and Science Study (TIMSS) is one such assessment, providing comparative data on student achievement across countries and over time. Conducted every four years since 1995 at the fourth and eighth grades, TIMSS provides data for over 70 countries seeking to monitor their education systems in a global context, benchmarking national curricula and informing policy and practice.

Central to TIMSS validity is the assessment framework, which is updated with each cycle through a collaborative, expert-driven process to ensure alignment with current educational objectives, practices, and priorities across participating countries. The framework provides comprehensive specifications that define what should be measured and how that measurement should occur. This includes detailed descriptions of content domains and their constituent topics, cognitive processes that students engage in when responding to items, target percentages for assessment coverage across domains, guidelines for item development and difficulty expectations, and integration of contemporary priorities such as computational thinking and environmental awareness.

The TIMSS Mathematics and Science Assessment Frameworks (von Davier & Kennedy, 2025) serve as the foundation for item development, ensuring that test items align with the constructs they aim to measure. Evidence-centered design (ECD; Mislevy et al., 2003) often referenced as a basis for principled test development to support validity arguments. ECD consists of three interconnected models: the domain model, which defines the constructs to be measured; the evidence model, which identifies the types of evidence needed to support inferences about these constructs; and the task model, which specifies how test items should be structured to generate that evidence. Consequently, item development requires systematic verification that each item aligns with the framework across specified content domains, cognitive domains, and difficulty levels.

Recent methodological developments have further increased the demand for items. With the transition to a country-level adaptive design (Yin & Foy, 2021), TIMSS now requires a larger pool of item blocks representing varying difficulty levels. This design necessitates pre-assessment classification of items as easy, medium, or difficult to enable appropriate test form assembly. Although each country administers the full assessment, booklet difficulty is balanced according to achievement distributions, increasing the importance of accurate difficulty prediction.

While new items are developed to reflect updated frameworks and evolving educational content, part of the assessment is composed of trend items to track changes in student performance over time. To estimate trend results, only a few items are released following each cycle, while the remaining items are reused for linking. These released items represent a small selection of the assessment, and countries often request additional examples to support learning and policy

decisions. Also, the released items are typically not trend items, so the use as examples for illustrating what is measured in TIMSS carries an inherent contradiction. To some extent, the example items show what was measured in past cycles but not what is common between current and future rounds. Ideally, independently developed new items could be judged to establish how well they align with the updated TIMSS framework. This would enable countries and item developers to assess how well their locally developed items align with the TIMSS framework, reducing dependency on a limited set of official items that often exemplify past frameworks.

Traditionally, item alignment is performed by SMEs through a manual review processes. While expert judgment remains essential, this approach is time-intensive, costly, and difficult to scale. From a technical perspective, automating TIMSS alignment presents substantial challenges. Assessment items often contain nuanced mathematical or scientific reasoning, contextual narratives, and implicit assumptions that are difficult to capture using surface-level text features. Additionally, available labeled data are limited in size, unevenly distributed across categories, and subject to annotation noise.

Recent advances in natural language processing (NLP) and large language models (LLMs) have generated interest in their potential application to complex classification tasks involving educational content. These models demonstrate capabilities in semantic understanding and reasoning over structured frameworks. However, whether these capabilities translate effectively to the specific demands of ILSAs remains an open empirical question.

This project investigates the extent to which AI methods, specifically prompt-based large language models and supervised machine learning techniques can contribute to automated alignment of mathematics and science assessment items with the TIMSS framework. The study evaluates these approaches across multiple dimensions, including content domain, cognitive domain, and difficulty level, examining their performance and limitations across assessment cycles.

To support empirical investigation, the project includes two additional development components. First, because TIMSS was paper-based until the 2019 cycle, an item bank was constructed by converting flat files from Word documents and PDFs into a structured digital repository suitable for computational modeling. Second, following the modeling phase, a research-grade web-based application was developed to enable experimentation with AI-assisted item analysis. This application is designed as a research-grade investigative platform rather than a validated operational tool. It generates structured classifications, allows export of results for further analysis, and facilitates systematic examination of when and how AI methods may supplement expert judgment in assessment development.

By documenting both the capabilities and limitations of current AI approaches for TIMSS framework alignment, this project aims to provide realistic insights into the feasibility of automated methods and identify directions for future methodological development in educational measurement.

Background and Related Work

Validity arguments for large-scale assessment are frequently grounded in ECD, which frames assessment development as an evidentiary reasoning process that links observations from tasks to claims about student proficiency. ECD formalizes the relationships among constructs (domain model), observable evidence (evidence model), and task features intended to elicit that evidence (task model). This approach is recognized as a framework for describing coherent test design because it makes assumptions explicit and supports traceability from test items to intended inferences (Mislevy & Riconscente, 2011). Within this framework, item alignment is not merely a content-tagging exercise; it is an essential component of a validity argument by showing that tasks operationalize the intended construct definitions. Moreover, misaligned items introduce construct-irrelevant variance and distort diagnostic interpretations (Karimi-Malekabadi et al., 2025). As item banks expand and frameworks evolve, manual alignment becomes increasingly burdensome, motivating the exploration of automated methods.

Early automated alignment research relied on surface-level linguistic features to represent item text. Bag-of-words representations, Term Frequency–Inverse Document Frequency (TF-IDF) weighting, and keyword overlap measures were applied to map assessment items to content standards using classifiers such as support vector machines (SVMs; Karlovcec et al., 2012) and latent Dirichlet allocation (Anderson et al., 2020). While these approaches demonstrated the feasibility of automated alignment, extracted features were shallow and failed to capture the contextual or sequential information inherent in well-constructed assessment items (Fu et al., 2025).

Subsequent work introduced distributed word representations, including Word2Vec, GloVe, and FastText embeddings. These methods mapped tokens into dense vector spaces and enabled cosine similarity-based matching between items and standards. Zhou and Ostrow (2022) demonstrated that averaging word-level embeddings and computing cosine similarity achieved approximately 35% accuracy for standard-to-item alignment, with contextual sentence embeddings yielding modest further gains. Tian et al. (2022) combined Word2Vec embeddings and XGBoost to align high school mathematics items, outperforming baseline models. However, static embeddings assign a single vector to each word regardless of context, limiting their ability to capture polysemy or nuanced reasoning structures embedded within assessment tasks.

Neural sequence models such as convolutional neural networks (CNNs) and bidirectional LSTMs (BiLSTM) introduced improved sensitivity to sequential structure. These models outperformed classical classifiers in question alignment tasks (Sun et al., 2018), confirming that sequence-aware architectures better capture the internal structure of assessment items. Nonetheless, even these models lacked the depth of contextual understanding required for fine-grained alignment to hierarchical content frameworks.

The emergence of the transformer architecture (Vaswani et al., 2017) and large (or small) language models fundamentally reshaped automated alignment research. Bidirectional pre-

trained models such as BERT and its variants encode contextual dependencies across entire token sequences, allowing polysemous terms and syntactic relationships to be represented dynamically.

Fine-tuned so-called small language models ([1]SLMs) consistently outperform embedding-based pipelines for item alignment. In large-scale mathematics alignment tasks, DeBERTa-v3-base achieved a weighted F1 score of 0.950 for domain classification and RoBERTa-large achieved 0.869 for topic alignment (Xu et al., 2025). Similar patterns were observed in reading and writing alignment tasks, where fine-tuned SLMs substantially outperformed classical supervised baselines (Fu et al., 2025). These studies emphasized that richer input representations matter more than simply increasing sample size. Incorporating answer options, keys, and rationale text yields larger performance gains than expanding the training dataset alone (Fu et al., 2025). Semantic overlap among fine-grained categories remains a persistent challenge. Xu et al. (2025) report extremely high cosine similarity among skill embeddings, making category boundaries difficult to separate even for advanced models.

LLMs introduce a complementary paradigm: prompt-based classification without task-specific fine-tuning. Rather than relying exclusively on labeled training data, LLMs leverage instruction following, zero-shot or few-shot prompting, and internal reasoning capabilities. Comprehensive evaluation of GPT-based models for item alignment demonstrates promising results. In large-scale K-5 mathematics and reading alignment tasks involving over 12,000 item-skill pairs, GPT-4o-mini achieved strong performance in binary misalignment detection (Karimi-Malekabadi et al., 2025). However, performance dropped substantially for subtle within-domain misalignments and open-set classification tasks involving large candidate pools. This decline mirrors findings in supervised transformer studies: when skill categories are highly similar, classification accuracy worsens. Similarly, evaluations of pre-trained transformer models in standards-to-standards alignment show that performance differences across models are often smaller than expected, suggesting that semantic representation quality and task framing may be more influential than parameter scale alone (Choi et al., 2025). Prompt engineering also meaningfully influences LLM performance. Structured reasoning prompts, chain-of-thought approaches and self-reflection prompts can improve classification robustness, though gains depend on task design and alignment granularity (Li et al., 2025). Overall, LLMs demonstrate strong flexibility and scalability but remain sensitive to category overlap.

Substantial attention has also been devoted to item difficulty prediction. Early approaches relied on lexical and syntactic features such as readability indices, sentence length, and word frequency to approximate difficulty (Freedle & Kostin, 1993; Perkins et al., 1995). With advances in NLP, embedding-based and machine learning methods were introduced to capture semantic features beyond surface statistics (Mikolov et al., 2013; Yaneva et al., 2019). More recently, transformer-

[1] It should be noted that today's small language models are, in terms of parameter count, much larger than what large language models were in 2020 or before.

based models such as BERT have demonstrated improved performance in difficulty prediction tasks compared to traditional methods (Yaneva et al., 2020; Gombert et al., 2024). Li et al. (2025) further showed that fine-tuned small language models outperform large language models for regression-based difficulty modeling, particularly when combined with data augmentation strategies to address imbalanced difficulty distributions for items in the medical domain.

Methodology

Item Bank Construction

The foundation of this project required constructing a structured, machine-readable item bank spanning all TIMSS assessment cycles from 1995 to 2023. Because TIMSS was administered exclusively on paper until the 2019 cycle when approximately half of participating countries transitioned to the eTIMSS digital format, previous assessment materials were stored in flat file formats such as Word, PDF, and RTF. The 2023 cycle marked the full transition to digital administration. The absence of a unified, computationally accessible repository for earlier cycles presented a critical barrier to systematic modeling, motivating a two-phase item bank development effort.

Phase I: Digitization and Structured Extraction

Building on an earlier digitization effort led by Dr. von Davier, Phase I extended the existing parsing pipeline to encompass all assessment cycles and conducted systematic quality checks to verify that content extraction into structured JSON preserved accuracy. The entire workflow was implemented in Python across the full item pool. Quality checks were conducted at two levels. First, automated validation procedures identified structural integrity by confirming that all required JSON fields were present, consistently named across cycles, and free of empty or null values; items failing these checks were flagged and reprocessed. Second, human review was conducted through random spot checking in which JSON files compared against the source material to confirm accurate capturing of the item text, response option and item metadata.

The primary extraction method relied on RTF processing using the pandoc library, with supplementary PDF parsing via PyPDF2 and pdfplumber for items not available in RTF format. RTF versions of the assessment materials were used directly where available, while PDF documents served only as a fallback source. An initial approach using PDF parsing was found to be insufficiently reliable due to inconsistent formatting, embedded images, and complex layouts across cycles, leading to RTF processing being adopted as the preferred extraction. The system navigated the TIMSS archive directory structure and automatically identified test items using unique identifier patterns embedded within each document.

The parsing extracted content blocks and linked them to their respective item identifiers. Multiple choice (MC) items required special handling to separate response option labels and corresponding text, while constructed response (CR) items focused on item stem content and associated tables. Where available, PNG image files from TIMSS publishing folders were

identified and their file paths appended to the corresponding JSON records for subsequent processing.

Parsed content was then merged with item metadata extracted from item information files, including content domain, cognitive domain, and item type. This integration used unique item identifiers as keys, producing a unified record for each item that maintained both content and classification metadata. The CTT and IRT parameters from separate sources were incorporated in the final phase, linking each item's content representation to its psychometric properties.

Phase II: AI-Generated Image Captioning and QC

Non-textual elements presented a persistent challenge for text-based computational models. Phase I addressed this partially by representing tables in LaTeX and generating image captions for a subset of items from the 2023 cycle, where the digital delivery format facilitated easier image extraction. However, the majority of items from earlier cycles, leaving visual content that is essential for mathematics and science items inaccessible to text-based models. Phase II addressed this gap through a comprehensive caption generation and validation pipeline covering the earlier cycles.

Caption generation was implemented in Python using standard data manipulation and file handling libraries and leveraged multiple OpenAI model variants including GPT-4.1, GPT-o3, and GPT-5. GPT-5 served as the primary captioning engine given its balance of accuracy, conciseness, and computational efficiency. GPT-o3 was reserved for targeted quality control given its precision in following structured instructions, though its higher computational cost prevented use at scale. Smaller model variants, such as GPT-5 nano and GPT-5 mini, were evaluated for efficiency but excluded from final production due to systematic errors in enumeration, geometric interpretation, and chart reading tasks. To prevent redundant or uninformative captions, the pipeline incorporated content-aware filtering logic embedded within the system prompt. For mathematics items, captions were generated only when an image conveyed substantive graphical or geometric information not already stated in the item stem; for science items, only when the image depicted a physical system or experimental configuration. Images that failed these criteria were assigned a "SKIP_BLANK" flag, which was converted to an empty caption field in the final JSON output. Figure A1 in the appendix provides representative examples of both included and excluded visual content, illustrating the operational distinction between substantive and non-substantive images.

All AI-generated captions were exported to validation spreadsheets for human review, with reviewer corrections taking precedence over automated outputs in the final integration step. Caption quality was formally evaluated for four cycles (2007 to 2019), covering 881 captioned items. Of these, 673 (76.4%) were judged accurate without modification, while 208 (23.6%) required correction following manual inspection. Reviewers compared AI-generated captions against the original images and revised inaccuracies as needed. The most common error types included incorrect object enumeration, misinterpretation of spatial orientation, inaccurate reading

of data visualizations (e.g., graphs and charts), and overly verbose descriptions of scientific diagrams that captured surface features without conveying underlying relationships. Figure A2 in the appendix presents an illustrative example of an AI-generated caption alongside its human-corrected version. A precedence hierarchy was created during integration: human-edited captions superseded all automated entries, and previously validated captions were preserved when reprocessing. In addition to the caption generation, Phase II also addressed schema inconsistencies identified across cycles, including mismatched field names and different data structures, to establish a uniform item bank schema suitable for the modeling phases that follow. The item bank encompasses all TIMSS items across cycles, including restricted-use, and unreleased items. All items are from the international version of the assessment in English. Table 1 shows the number of items processed from each TIMSS cycle, along with the corresponding item numbers.

Year	Items	Captioned Items	Text only Items
1995	512	177	335
1999	298	74	224
2003	722	146	576
2007	765	351	414
2011	735	117	618
2015	537	212	325
2019	518	201	317
2023	648	180	468
Total	4735	1458	3277

Table 1. Number of Items Processed for Captioning

Prompt Based Large Language Model Approach

The first modeling strategy employed in this study used a prompt-based LLM to automate the alignment of TIMSS items with the assessment framework. This approach treats the model as a zero-shot or few-shot classifier guided exclusively by structured prompting, without task-specific fine-tuning. The objective was to evaluate whether an LLM, when provided with explicit framework definitions and carefully designed instructions, could approximate expert classification decisions across content and cognitive domains, and difficulty levels.

The prompt-based approach is grounded in the assumption that language models, through pretraining, already encode sufficiently rich semantic and reasoning representations to support structured classification tasks, such that embedding construct definitions directly in the prompt is sufficient to elicit appropriate classification without requiring any task-specific fine-tuning. Rather than training a predictive model on labeled data, this method operationalizes framework alignment as an instruction-following task.

Each classification prompt was designed to mirror the reasoning process expected of SMEs. The model was instructed to interpret the item, evaluate its underlying content and required cognitive processes, and classify it according to the official TIMSS framework definitions. In this sense, the procedure parallels the domain-model reasoning described in evidence-centered design: the framework definitions define the constructs of interest, and the item text constitutes observable evidence from which inferences must be drawn. This design allows the classification mechanism to remain flexible across assessment cycles and subject areas without retraining, provided that the relevant framework definitions are programmed to be inserted into the prompt based on each item's metadata, specifically its subject, grade level, and assessment cycle.

Data

The prompt-based classification procedure was applied to newly developed TIMSS mathematics and science items spanning multiple assessment cycles. The initial prompt engineering was conducted using 217 items from the TIMSS 2019 Grade 4 and Grade 8 assessments, after removing multipart items. After the prompt structure was defined, the implementation was extended through 2015 and 2023. Each item in the structured item bank contained:

- The complete item stem text
- Response options (for multiple-choice items),
- Associated image captions when visual elements were present,
- Associated tables in the LaTeX format,
- Official framework labels for content and cognitive domains,
- Empirical difficulty classifications derived from operational data.

Multipart items, that is item sets consisting of multiple independently scored sub-questions sharing a common stimulus, were excluded from analysis. In TIMSS, such sets include a derived item that aggregates across sub-parts and cannot be treated as standalone classification unit. Including derived items alongside their sub-parts would have introduced redundancy into the unit of classification. Difficulty categories were operationalized using empirically derived percent-correct values and grouped into three levels: Easy (>60%), Medium (30-60%), and Hard (<30%). Ground truth labels for all three classification dimensions: content domain, cognitive domain and difficulty level, served as reference labels for evaluation only and were not disclosed to the model during classification. The model received only the item content and the framework definitions.

Prompt Engineering

A central feature of this implementation was the construction of a custom framework database derived from official TIMSS Assessment Framework documents. This database enabled the dynamic loading of full definitional text for content domains, cognitive domains, and difficulty levels according to subject, grade, and assessment cycle. Rather than supplying abbreviated labels or keyword summaries, the prompts embedded the complete construct definitions from the

framework. This design ensured that the model's classification decisions were conditioned on the same formal construct descriptions used by expert panels.

Each prompt presented: the complete item content, including captioned visual descriptions when applicable; the TIMSS subject, grade level, and assessment year; descriptions of all relevant content domains; descriptions of the three cognitive domains (Knowing, Applying, Reasoning); and structured difficulty guidance incorporating conceptual complexity, reasoning requirements, and distractor strength.

The model was instructed to assign exactly one label per dimension, using the framework's exact wording. To reduce superficial pattern matching and encourage construct-level reasoning, meta-prompting principles were embedded directly into the instructions. The model was explicitly guided to consider both an expert perspective (alignment with framework definitions) and a student perspective (cognitive processes required to solve the item) before committing to final classifications. This structure was intended to approximate the inferential reasoning process undertaken during human item review.

Four prompting configurations were implemented to examine the influence of prompt structure. These varied along two dimensions: the presence of worked examples and the inclusion of chain-of-thought (CoT) reasoning instructions.

- Zero-shot (ZS) prompting presented only the item and dynamically loaded framework definitions.
- Zero-shot CoT (ZS-CoT) added explicit instructions requiring step-by-step reasoning before final classification.
- Few-shot (FS) prompting supplemented the framework definitions with representative example items illustrating each cognitive domain and difficulty level combination, yielding approximately nine to ten structured examples.
- Few-shot CoT (FS-CoT) combined example-based guidance with structured reasoning instructions.

Figure A3 presents the full FS-CoT prompt structure, which represents the most complete prompting configuration used in this study. The remaining three conditions follow the same structure but differ in which components are included. The ZS condition omits both the step-by-step analysis instructions and the annotated examples, providing the model only with the item content and dynamically loaded framework definitions. The ZS-CoT condition retains the step-by-step reasoning instructions but omits the annotated examples. The FS condition includes the annotated examples but omits the reasoning instructions. In all configurations, placeholders in curly braces (e.g., {subject_name}, {grade}, {year}) are populated programmatically at inference time based on each item's metadata.

In the few-shot configurations, example items were selected to span the full decision space across classification dimensions. Specifically, examples were drawn to represent each

combination of cognitive domain (Knowing, Applying, Reasoning) and difficulty level (Easy, Medium, Hard), ensuring that the model was exposed to both prototypical instances and boundary cases. Items were selected randomly within each category combination, subject to the constraint that all combinations were represented in the final example set. This yielded approximately nine to ten examples per grade and subject. The inclusion of CoT reasoning was motivated by the inherently multi-step nature of cognitive domain and difficulty judgments. Unlike content-domain tagging, which may rely more heavily on topical cues, cognitive and difficulty classifications require the model to weigh competing interpretive signals, simulate problem-solving processes, and evaluate conceptual depth.

All model calls were executed with temperature set to zero, or reasoning effort set to none to maximize determinism and reproducibility. Each item was submitted independently to prevent contextual contamination across classifications. Following initial prompt development, a single structured prompting configuration was fixed and applied uniformly in subsequent analyses to ensure consistency across datasets and cycles.

Supervised Models

Unlike the prompt-based approach, which conditions each classification directly on embedded framework definitions at inference time, the supervised modeling component leverages historically labeled data to estimate parameters that generalize across assessment cycles. Supervised models were developed for cognitive domain classification and difficulty prediction, but not for content domain classification. This decision was motivated by two considerations. First, preliminary analyses indicated that content-domain alignment could be achieved reliably using the unsupervised, framework-aware prompting approach, thereby reducing the need for additional supervised modeling. Second, both cognitive domain classifications and difficulty parameters are structurally stable across TIMSS cycles. Cognitive domains are defined consistently over time, and difficulty is operationalized through comparable psychometric metrics (e.g., IRT b parameters or percent-correct values). This structural invariance permits pooling historical item data across cycles to train models that generalize to newly developed items.

Cognitive Domain Classification

Cognitive domain classification was modeled as a three-class supervised learning problem, corresponding directly to the three cognitive domains defined in the TIMSS framework (von Davier & Kennedy, 2025): Knowing (recalling and recognizing facts, concepts and procedures), Applying (using knowledge to solve problems in familiar concepts), and Reasoning (applying higher-order thinking to analyze, synthesize, or evaluate situations in novel or complex settings). Because the cognitive domain framework is invariant across cycles and grades, items from TIMSS cycles 2003, 2007, 2011, 2015, 2019, and 2023 were pooled to maximize training data and improve parameter stability. To reflect the operational structure of TIMSS item development, a temporally ordered training design was implemented. Items from the 2003-2019

cycles (N = 2,853) were used for model training, while items from the 2023 cycle (N = 509) were held out as an independent test set. This forward-chaining design mirrors the realistic scenario in which models trained on historical items are applied to newly developed items in a subsequent assessment cycle, thereby providing a more stringent test of generalizability.

Items from both Grade 4 and Grade 8 and from both Mathematics and Science were included. Pooling across grades is justified by the identical structural definition of the cognitive domains across grade levels. Similarly, cross-subject pooling is defensible because cognitive demand categories are defined in domain-general terms rather than subject-specific constructs. The prediction target was the official expert-assigned cognitive domain label for each item. Textual input included the full item stem, response options (for MC items), any associated item content table, and captioned visual descriptions when available.

Three model families were evaluated:

TF-IDF + Linear SVM (Baseline). Character-level n-grams (3–5) with TF-IDF weighting were used to capture subword and symbolic patterns. A linear Support Vector Classifier with class-balanced weights was employed. Probability calibration was performed using sigmoid scaling with cross-validation.

Enhanced TF-IDF Feature Ensemble. This model combined character n-grams, word n-grams, and boundary-aware features to capture both lexical and structural signals associated with cognitive demand.

Fine-Tuned Transformer Models. Two transformer models were also fine-tuned. RoBERTa (Liu et al., 2019), a 355M-parameter model, was evaluated both with and without oversampling to assess sensitivity to class imbalance. DeBERTa (He et al., 2021), fine-tuned with standard hyperparameters. Both transformer models were trained for up to 20 epochs using early stopping and a batch size of 16. All models were evaluated using accuracy, macro-F1, and Cohen's kappa. This supervised classification task allows direct comparison between parameter-optimized models and the framework-aware prompt-based approach described in Section 3.2.

To investigate whether classification difficulties reflected model limitations or inherent properties of the cognitive domain construct itself, a separability analysis was conducted using Sentence-BERT embeddings. For each item, dense vector representations were computed, and cosine similarity was calculated both within and across cognitive domains. Separability was operationalized as the difference between average within-domain and across-domain cosine similarity. This analysis was also applied to content domains to provide comparative reference.

Difficulty Prediction

The difficulty prediction used the same data spanning six TIMSS assessment cycles for both mathematics and science at Grade 4 and Grade 8 from 2003 to 2023. Starting from approximately 4,180 items, those lacking either IRT b-parameter estimates, percent-correct

values, or item text were removed; the final dataset comprised 3,362 unique items (51% mathematics, 49% science; 48% Grade 4, 52% Grade 8). Items were split into training (N=2,428), validation (N=429), and test (N=505) sets. We modeled item difficulty using both CTT-based percent-correct (p-value) and IRT-based b-parameters, because both are commonly used operational difficulty representations in ILSAs and appear frequently in prior automated difficulty prediction studies (Peters et al., 2025). Using identical splits ensures that performance differences across targets reflect properties of the outcome rather than sampling variation. All feature extraction procedures were fit exclusively on the training set and applied unchanged to validation and test sets. For categorical evaluation, b-parameter values were binned into Easy ($b < -0.847$), Medium ($-0.847 \leq b \leq 0.405$), and Hard ($b > 0.405$), and p-values into Easy ($p > 0.60$), Medium (0.30-0.60), and Hard ($p < 0.30$).

A comprehensive feature engineering pipeline was constructed to transform item content into structured numerical representations. The final feature matrix comprised 1,250 features combining three types. TF-IDF text features (1,000 features) were extracted using unigrams and bigrams fit exclusively on training data, with sparse matrices subsequently transformed for validation and test sets. Categorical metadata features (219 features, one-hot encoded – that is each categorical variable is converted into a set of binary indicators, one per category with a value of 1 indicating the presence of that category and 0 otherwise) captured item-level attributes including subject, grade, cognitive domain, content domain, cognitive area, and content area. Numeric metadata features (31 features) encompassed caption characteristics, readability indices (Flesch-Kincaid, Flesch Reading Ease, Gunning Fog, SMOG, Automated Readability Index), text complexity statistics, mathematics-specific features (number count, operator count, fraction and exponent detection, variable count), question type indicators, and LLM-extracted cognitive features. The LLM-extracted features were generated using GPT-4.1 with structured prompts and included continuous ratings of cognitive demand, multi-step reasoning, visual complexity, mathematical complexity, and predicted difficulty on standardized scales. A detailed description of all features is provided in Table A1 in the appendix.

Six models were evaluated for their ability to predict difficulty.

Baseline Models. Two mean-based baselines were implemented to establish lower-bound benchmarks. The Overall Mean baseline predicts the training set mean difficulty value for all items regardless of any item features. The Subject Mean baseline predicts separate training set means for Mathematics and Science items, representing the simplest possible use of subject-level information.

Tree-Based Ensemble Models. Two ensemble regressors were implemented: A Random Forest Regressor using bootstrap aggregation and limited tree depth to control overfitting, and a Gradient Boosting Regressor using sequential boosting with shrinkage and subsampling were trained on the full feature matrix. These models allow nonlinear interactions among lexical, structural, and metadata features.

Transformer-Based Regression. To assess the contribution of deep contextual representations, BERT-base-uncased was also fine-tuned as a regression model under two configurations: a text-only variant using item text as sole input, and a metadata-added variant combining text with item-level metadata features. Both BERT variants applied a regression head to the [CLS] token and were trained for up to 25 epochs using the AdamW optimizer with mean squared error loss.

Primary evaluation used mean absolute error (MAE) on both continuous targets, complemented by root mean square error (RMSE), proportion of variance explained by the model (R^2), and Pearson and Spearman correlations. It is important to note that baseline models (overall mean and subject mean) produce constant predictions within their respective groups; therefore, Pearson and Spearman correlations are undefined and omitted for these models. Categorical accuracy was also computed following bin assignment. Feature importance analysis was conducted for each ensemble model, separately for each target, to identify which feature categories contributed most to difficulty prediction.

Results

Prompt-Based Classification

The first step was to assess classification performance across prompting conditions. For this we utilized TIMSS 2019 items in addition to the framework. Content domain classification demonstrated consistently high performance, with all prompting conditions exceeding 94% accuracy and kappa values above 0.92, indicating substantial agreement beyond chance. FS-CoT achieved the highest accuracy (94.9%) and kappa (0.933), reflecting the model's strong ability to differentiate TIMSS content domains. In contrast, classification accuracy for the cognitive domain showed greater variation, ranging from 60.4% to 64.1%, with kappa values between 0.382 and 0.438. The FS-CoT condition yielded the highest accuracy, followed by the ZS baseline and FS. Kappa values across these conditions suggest fair to moderate agreement with expert labels, indicating that while the model captures meaningful cognitive distinctions, it does so with less precision than in the content domain.

Prompt	Content Domain		Cognitive Domain		Difficulty Level	
	Acc	κ	Acc	κ	Acc	κ
ZS	94.1	0.922	62.2	0.410	44.2	0.072
ZS CoT	94.2	0.923	60.4	0.382	49.8	0.134
FS	94.1	0.923	61.3	0.397	44.7	0.074
FS CoT	94.9	0.933	64.1	0.438	48.4	0.097

Table 2. Classification Performance Across Prompting Conditions

Difficulty classification, while the most challenging of the three dimensions, showed improvement over previous iterations. Accuracy scores ranged from 44.2% to 49.8%, and all conditions resulted in positive kappa values, indicating better-than-chance agreement. ZS-CoT

led in both accuracy and agreement, though overall performance remained modest, highlighting the inherent complexity of predicting empirically derived difficulty levels.

Grade-level analysis, as shown in Table 3, revealed consistently stronger model performance on Grade 4 items across all classification dimensions. For content domain classification, Grade 4 items achieved exceptional accuracy scores ranging from 96.9% to 97.7%. Grade 8 content domain performance, while lower, remained strong with accuracy scores from 91.0% to 92.3%. A similar pattern was also observed in cognitive domain classification. Grade 4 accuracy ranged from 62.6% to 65.0%, while Grade 8 performance varied from 57.5% to 62.8%. Notably, the FS-CoT condition achieved the smallest grade-level gap in cognitive domain performance (65.0% vs. 62.8%). For difficulty classification, Grade 4 items consistently outperformed Grade 8 items across all conditions. Grade 4 difficulty accuracy ranged from 50.4% to 57.7%, with ZS-CoT achieving the highest Grade 4 performance (57.7%). Grade 8 difficulty classification proved more challenging, with accuracy scores ranging from 33.0% to 40.4%, with FS-CoT achieving the best Grade 8 performance (40.4%).

Prompt	Grade	Content Domain		Cognitive Domain		Difficulty Level	
		Acc	κ	Acc	κ	Acc	κ
ZS	4	96.9	0.949	65.0	0.450	50.4	0.149
	8	91.0	0.876	58.5	0.351	36.2	-0.060
ZS CoT	4	97.6	0.956	62.6	0.413	57.7	0.229
	8	91.0	0.876	57.5	0.336	39.4	0.001
FS	4	96.9	0.948	63.4	0.426	53.7	0.189
	8	91.1	0.877	58.5	0.353	33.0	-0.085
FS CoT	4	97.7	0.959	65.0	0.450	54.5	0.174
	8	92.3	0.894	62.8	0.421	40.4	-0.025

Table 3. Classification Performance by Prompting Condition and Grade Level

Given its overall better performance across all three classification dimensions, the FS-CoT condition was selected for further analysis. Table 4 presents classification performance across all three classification dimensions (content domain, cognitive domain, difficulty level) for each cycle-subject combination. Results are consistent with the initial 2019 mathematics findings and reveal systematic patterns across subjects and cycles.

Cycle	Subject	Domain	Accuracy	Kappa
2015	Mathematics	Content	94.5	0.887
		Cognitive	68.7	0.500
		Difficulty	33.9	0.008
	Science	Content	94.9	0.939
		Cognitive	71.0	0.530

		Difficulty	45.9	0.056
2019	Mathematics	Content	94.4	0.930
		Cognitive	64.1	0.438
		Difficulty	48.4	0.097
	Science	Content	96.0	0.952
		Cognitive	56.6	0.325
		Difficulty	36.6	-0.011
2023	Mathematics	Content	93.1	0.910
		Cognitive	68.1	0.477
		Difficulty	35.2	0.005
	Science	Content	96.1	0.952
		Cognitive	61.8	0.395
		Difficulty	49.7	0.110

Table 4. FS-CoT Prompt-Based Classification Results

Content domain classification maintained consistently high performance across all cycle-subject combinations. This represents substantial to near-perfect agreement with expert classifications, indicating that content domain alignment is highly reliable when supported by comprehensive framework definitions embedded in the prompt. Science items demonstrated marginally stronger content domain classification (94.9% to 96.1%) compared to mathematics items (93.1% to 94.5%) across all cycles. This may reflect more distinct conceptual boundaries between science content domains (e.g., Biology, Chemistry, Physics, Earth Science) than between mathematics domains, where topics such as Number and Algebra often interrelate in complex ways, particularly at Grade 8, where many Number items include algebraic components. This pattern is also reflected in the confusion structure. In Mathematics, misclassification is concentrated at the boundary between Number and Algebra as shown in Figure 1, 21% of the Number items were misclassified as Algebra, whereas in science, similar boundary effects are observed between closely related domains such as Physics and Chemistry or Earth Science and Physical Science. These effects are more pronounced at higher grade levels, where domain distinctions become less discrete and items increasingly integrate multiple content areas. Across cycles, no clear temporal pattern emerged in content domain performance, suggesting that changes in framework language or item development practices across the 2015, 2019, and 2023 cycles did not substantially impact the model's ability to assign content classifications.

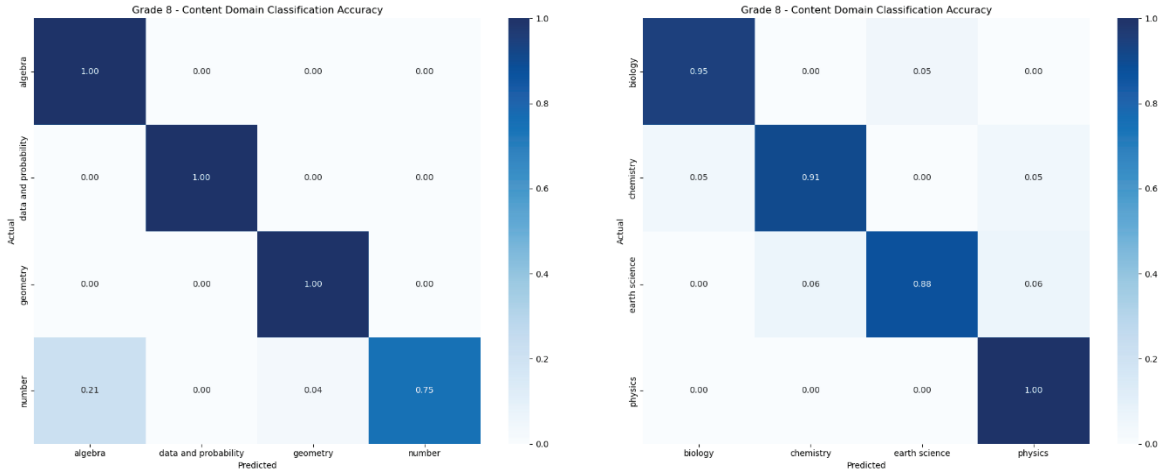


Figure 1. Confusion Matrices for Content Domain Classifications for Grade 8 TIMSS 2019 Math (left) and Science (right) items

Cognitive domain classification showed moderate performance with considerably more variation than content domain results. These results align closely with the 56.6% to 71% accuracy range observed in the initial evaluation, confirming that cognitive domain classification represents a consistently challenging task across the item bank. The 2015 confusion matrices for Mathematics and Science, shown in Figure 2, further clarify the nature of this performance. In both subjects, Applying was the most accurately classified category (Mathematics: 81%; Science: 82%), followed by Knowing (Mathematics: 67%; Science: 70%), while Reasoning showed the lowest accuracy (Mathematics: 47%; Science: 52%). A substantial proportion of Reasoning items were misclassified as Applying (Mathematics: 49%; Science: 43%), indicating difficulty distinguishing higher-order reasoning from procedural application. Additionally, nearly one-third of Knowing items were also classified as Applying in both subjects, suggesting that the boundary between recall and application tasks is similarly blurred. These consistent cross-subject patterns indicate that misclassification is not random but reflects systematic overlap among cognitive domains.

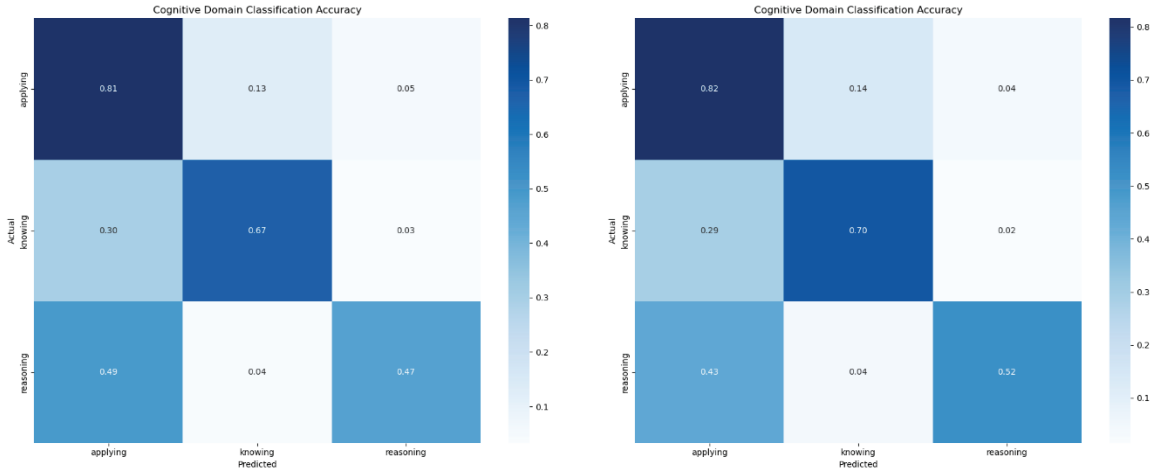


Figure 2. Confusion Matrices for Cognitive Domain Classifications for TIMSS 2015 Math (left) and Science (right) items

Difficulty classification proved most challenging across all cycles and subjects, with accuracy ranging from 33.9% to 49.7%. These results confirm that predicting empirically-derived difficulty categories from item content alone remains extremely challenging, even for advanced language models.

Supervised Cognitive Domain Classification

The supervised cognitive domain analysis was evaluated using two different train/test splits: a temporally ordered split that reflects cross-cycle generalization, and a random split that reflects within-distribution performance.

In the temporal split, items from the 2003-2019 cycles served as the training set (N = 2,853), and items from the 2023 cycle were held out as an independent test set (N = 509). The training distribution consisted of 1,197 Applying items, 1,031 Knowing items, and 625 Reasoning items. The test set contained 223 Applying, 171 Knowing, and 115 Reasoning items. This forward-chaining design mirrors the operational scenario in which models trained on historical TIMSS items are applied to newly developed items from a subsequent cycle. In the random split, items from all cycles (2003-2023) were randomly partitioned into training (N = 2,689) and test (N = 673) sets while preserving class proportions. The training distribution included 1,136 Applying items, 961 Knowing items, and 592 Reasoning items. The test set comprised 284 Applying, 241 Knowing, and 148 Reasoning items. This design evaluates model performance when training and test data are drawn from the same overall distribution. Together, these two evaluations allow comparison between cross-cycle generalization performance and within-distribution classification performance.

Table 5 presents classification results under both the temporal and random split evaluation designs. Under the temporal split, RoBERTa-large achieved the strongest overall performance, followed by DeBERTa-v3-base and enhanced TF-IDF. The performance gap between RoBERTa-large and the TF-IDF+SVM baseline was modest but meaningful, while DeBERTa-v3-base and enhanced TF-IDF offered only marginal improvements over the simpler baseline in kappa values (+0.013 and +0.007, respectively).

Training Split	Model	Accuracy	Macro-F1	Kappa
Temporal	RoBERTa-Large	0.688	0.672	0.507
	DeBERTa-v3-Base	0.668	0.651	0.468
	Enhanced TF-IDF	0.664	0.650	0.462
	TF-IDF + SVM	0.660	0.640	0.455

	RoBERTa-Large + Oversampling	0.642	0.614	0.419
Random	DeBERTa-v3-Base	0.636	0.627	0.427
	Enhanced TF-IDF	0.634	0.605	0.420
	TF-IDF + SVM	0.624	0.592	0.404
	RoBERTa-Large	0.602	0.597	0.378
	RoBERTa-Large + Oversampling	0.504	0.507	0.276

Table 5. Classification Results for Cognitive Domain

In terms of class specific recall, recall for the Reasoning category, defined here as the proportion of true Reasoning items correctly identified by the model, varied substantially across models under the temporal split. Reasoning recall is highlighted because it constitutes the minority class and is the category most susceptible to degradation under class imbalance, making it the most diagnostic indicator of model sensitivity. RoBERTa-large achieved the highest Reasoning recall (62.6%), followed by DeBERTa-v3-base (60.9%), while the TF-IDF-based models showed considerably lower Reasoning recall (TF-IDF+SVM: 46.1%; enhanced TF-IDF: 48.7%). This suggests that transformer architectures capture features relevant to identifying higher-order reasoning demand more effectively than lexical approaches, even when overall accuracy differences are small. Counter-intuitively, RoBERTa-large with oversampling performed worse than the standard RoBERTa-large on both accuracy (64.2% vs. 68.8%) and Reasoning recall (39.1% vs. 62.6%), indicating that oversampling degraded rather than improved cross-cycle generalization under the temporal design.

Confusion matrix analysis for the best-performing temporal model (RoBERTa-large) reveals systematic patterns. Applying items were classified most accurately (184/223, 82.5%), with errors distributed between Knowing and Reasoning. Knowing accuracy was moderate (94/171, 55.0%), with the majority of errors involving misclassification as Applying. Reasoning items proved most challenging (72/115, 62.6%), with errors split primarily between Applying (39 items) and very few misclassified as Knowing (4 items). This pattern, where Reasoning is confused with Applying but rarely with Knowing is consistent with the unsupervised results which showed the same directional confusion across both Mathematics and Science. Across both approaches, the primary classification challenge lies in distinguishing higher-order reasoning from procedural application rather than factual recall.

Under the random split design, which evaluates within-distribution performance when training and test data are drawn from the same overall item pool, the ranking of models shifted. DeBERTa-v3-base led accuracy of 63.6% ($\kappa = 0.427$), with enhanced TF-IDF (63.4%, $\kappa = 0.420$) and TF-IDF+SVM (62.4%, $\kappa = 0.404$) again remaining competitive. RoBERTa-large without oversampling declined markedly under the random split (60.2%, $\kappa = 0.378$) falling below the TF-IDF baseline, while RoBERTa-large with oversampling performed substantially

worse (50.4% with $\kappa = 0.276$). Overall accuracy levels were lower under the random split than the temporal split for most models, indicating that temporal cross-cycle generalization is not artificially easier than within-distribution evaluation; models trained on earlier cycles generalize to 2023 items at least as effectively as models evaluated on random holdouts from the same pooled data. The consistency of the TF-IDF approaches across both designs further supports the interpretation that a meaningful cognitive domain signal is present in lexical features and does not require deep contextual modeling to exploit.

Separability Analysis

To investigate whether classification difficulties reflected model limitations or inherent properties of the cognitive domain construct, a semantic separability analysis was conducted using Sentence-BERT embeddings across the full data (N = 3,362; cycles 2003-2023; both subjects and grades). Separability was defined as the difference between mean within-domain and across-domain cosine similarity. For content domains, mean within-domain similarity was 0.281 compared to inter-domain similarity of 0.196, yielding a separability gap of 0.085. This indicates that items within the same content domain are meaningfully more similar to one another than to items from other domains, confirming that content domain boundaries correspond to detectable structure in semantic embedding space. In contrast, cognitive domains exhibited minimal separability. Mean within-domain similarity (0.209) was nearly identical to across-domain similarity (0.202), producing a gap of only 0.007. This near-zero difference suggests that items labeled as Knowing, Applying, and Reasoning do not form clearly differentiated clusters in embedding space.

Disaggregated analyses by subject and grade reinforced this pattern. Content domain separability gaps ranged from 0.051 (Mathematics Grade 4) to 0.066 (Science Grade 8), remaining consistently positive across all subgroups. Cognitive domain gaps remained small throughout, ranging from 0.009 (Science Grade 4) to 0.018 (Mathematics and Science Grade 8). Although mathematics items exhibited slightly larger cognitive separability than science items, all subgroup values remained close to zero as shown in Figure 3.

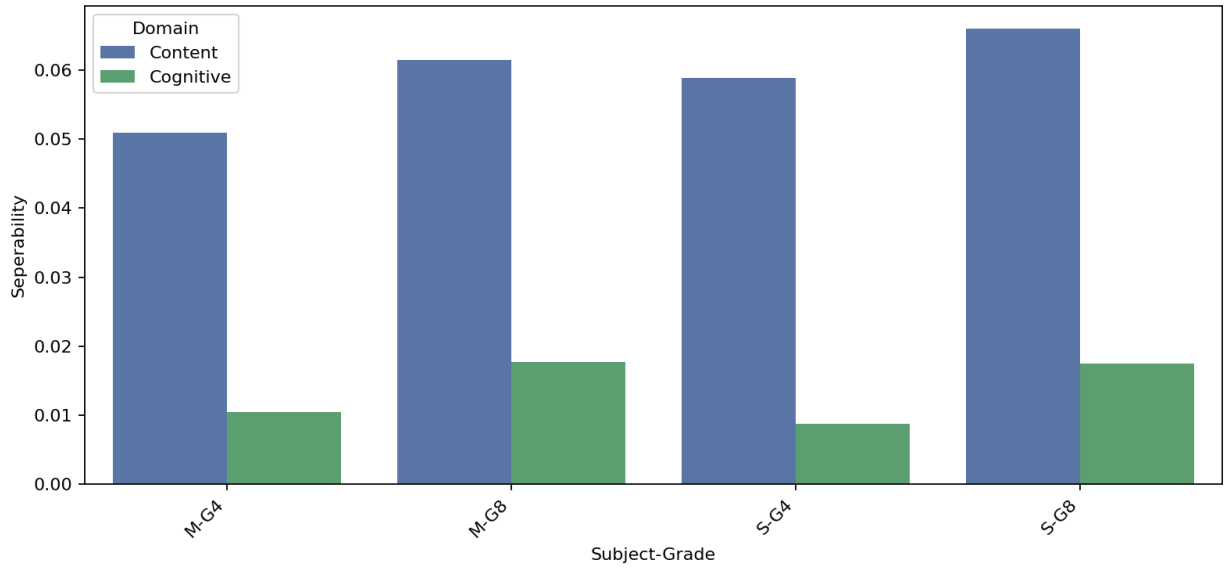


Figure 3. Separability of Content and Cognitive Domains by Subject and Grade

These findings indicate that items from different cognitive domains occupy largely overlapping regions of semantic space across subjects, grades, and assessment cycles, with textual features providing minimal basis for distinguishing Knowing, Applying, and Reasoning at the level of individual item representations.

Difficulty Prediction

Table 6 presents prediction performance across all models for both the b-parameter and p-value targets. The mean-based baselines produced categorical accuracy of approximately 47% for the b-parameter and 59% for p-value. This relatively high categorical accuracy of mean-based baselines reflects effects of class distribution rather than predictive power. Because a substantial proportion of items fall within the middle difficulty band, constant predictions centered near the mean achieve moderate apparent accuracy despite zero explanatory variance.

For the b-parameter, Random Forest achieved the strongest performance (MAE = 0.563, R^2 = 0.203) followed closely by Gradient Boosting (MAE = 0.569, R^2= 0.178). Both ensemble models substantially outperformed transformer-based approaches. The BERT text-only configuration achieved MAE = 0.598 and Pearson r = 0.335, while adding metadata further degraded performance (MAE = 0.611, Pearson r = 0.245), suggesting that simple concatenation of structured metadata with dense contextual embeddings does not effectively enhance the prediction of the difficulty parameter.

Target	Model	MAE	RMSE	R^2	Pearson r	Spearman ρ	Category Accuracy
b	Overall Mean	0.635	0.811	0.000			0.471
	Subject Mean	0.636	0.811	-0.001	-0.001	0.007	0.471
	Random Forest	0.563	0.724	0.203	0.467	0.443	0.582
	Grad. Boosting	0.569	0.735	0.178	0.434	0.394	0.564
	BERT Text-Only	0.598	0.781	0.074	0.335	0.313	0.535
	BERT + Metadata	0.611	0.789	0.054	0.245	0.252	0.542
p-value	Overall Mean	0.137	0.168	0.000			0.588
	Subject Mean	0.137	0.168	0.001	0.046	0.049	0.588
	Random Forest	0.116	0.144	0.272	0.523	0.513	0.608
	Grad. Boosting	0.114	0.141	0.303	0.554	0.536	0.615
	BERT Text-Only	0.120	0.155	0.148	0.508	0.486	0.617
	BERT + Metadata	0.141	0.177	-0.108	0.236	0.258	0.546

Table 6. Difficulty Prediction Results by Outcome Variable

The pattern shifted for the p-value target. Gradient Boosting achieved the strongest overall performance (MAE = 0.114, R^2= 0.303), slightly outperforming Random Forest (MAE = 0.116, R^2= 0.272). BERT text-only remained competitive for p-value (Pearson r = 0.508, categorical accuracy = 61.7%), approaching ensemble performance in rank correlation despite weaker R^2. In contrast to the b-parameter results, the BERT + metadata configuration exhibited a negative R^2 = -0.108, indicating unstable integration of heterogeneous feature types for this outcome.

Across both targets, correlations and categorical accuracy were consistently higher for the p-value than for the b-parameter. This suggests that proportion correct carries a more directly recoverable signal from item text and structural features than the IRT-calibrated latent difficulty parameter, which may be influenced by additional scaling processes.

Feature importance analysis from the Gradient Boosting models revealed distinct profiles for the two targets, as shown in Figure 4. For the b-parameter, grade level indicators dominated, with Grade_4 emerging as the single most important predictor, followed by Grade_8. LLM-derived indicators (likely_difficulty, cognitive_demand) ranked immediately below grade features, indicating that structured AI-extracted signals contribute meaningfully beyond surface text measures. Cognitive area indicators and readability metrics (e.g., type_token_ratio, text_length_words, and avg_sentence_length) also appeared among the top predictors.

For p-value, the importance structure is different from the b-parameter. Cognitive_demand became the dominant feature, followed by multi_step_reasoning, number of variables, and explanatory markers. Grade-level features were comparatively less influential, while lexical and question-type indicators (e.g., which, answer, for) increased in importance. This suggests that the b-parameter and p-value capture partially distinct aspects of item difficulty: latent trait difficulty

is more strongly associated with developmental level, whereas observed performance is more sensitive to cognitive and linguistic demands embedded in item text.

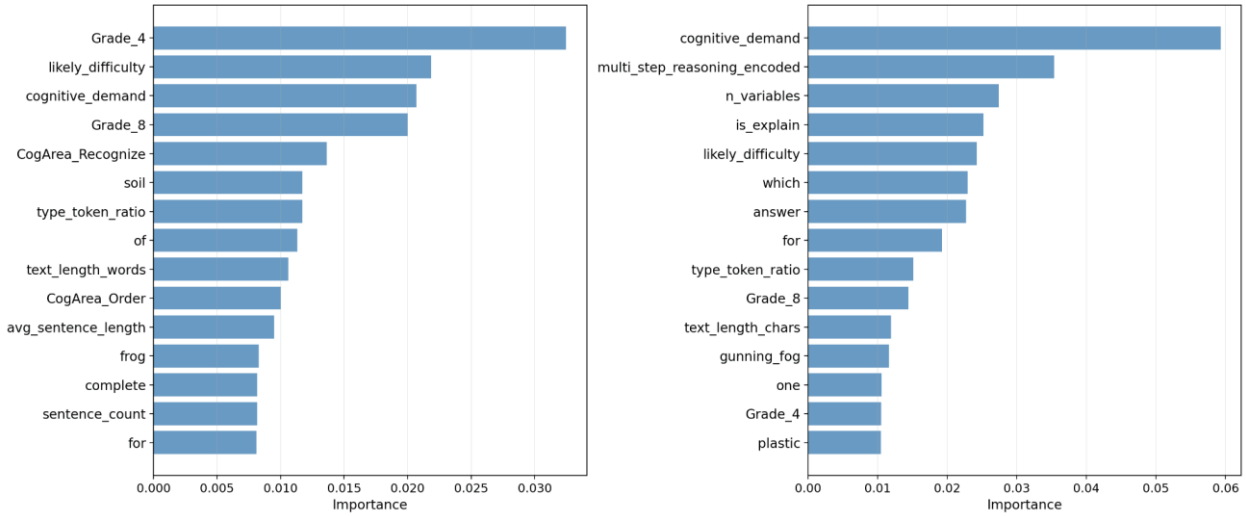


Figure 4. Relative feature importance derived from Gradient Boosting models for the b-parameter (left) and proportion correct (right).

Overall, difficulty prediction performance remained modest. The strongest models explained 20.3% of the variance in the b-parameter and 30.3% of the variance in p-value, indicating that item text and metadata account for only a limited share of empirical difficulty variance. Consistent with the prompt-based classification results, difficulty emerged as the most resistant framework dimension to automated prediction from item content alone.

Research Grade Application

To operationalize the modeling framework developed in this study, we implemented a web-based research application to automate the alignment of assessment items with the TIMSS framework. The implemented interface is presented in the appendix (Figure A4).

The application adopts an LLM architecture rather than supervised models. Although supervised cognitive domain models demonstrated measurable improvements relative to classical baselines, their gains over the framework-aware prompt-based approach were modest. Across evaluations, improvements in agreement were incremental, and confusion patterns remained structurally similar. Given the limited performance, architectural flexibility and framework adaptivity that the LLM-based version provides were prioritized.

The system supports four classification dimensions consistent with the hierarchical TIMSS framework structure:

- Content Domain (subject area)
- Content Area (specific topic within the domain)

- Cognitive Domain (Knowing, Applying, Reasoning)
- Cognitive Area (specific cognitive skill descriptors)

Although difficulty prediction models were developed and evaluated as part of the broader research investigation, automated difficulty classification was not incorporated into the application.

This decision was guided by both empirical and theoretical considerations. Empirically, while supervised regression models demonstrated outcomes comparable to the broader literature, predictive accuracy did not reach a level sufficient to justify deployment in an application. More importantly, item difficulty in TIMSS is fundamentally an empirical psychometric property derived from large-scale response data within a calibrated measurement model. Textual or metadata-based predictions, even when statistically informative, do not substitute for psychometric calibration. Given these considerations, we adopted a conservative design stance and excluded automated difficulty labeling from the application interface.

For each classification dimension, the model also produces both a categorical assignment and a confidence score expressed as a percentage. Confidence is computed through an explicit ranking procedure embedded within the prompt. Rather than selecting a category directly, the model is instructed to evaluate all possible categories, rank them according to fit, and quantify the distinction between the top-ranked category and competing alternatives. The reported confidence reflects the relative separation between the most plausible classification and the remaining options.

Confidence scores are presented as decision-support indicators rather than guarantees of correctness. LLMs may exhibit high confidence even in erroneous classifications; therefore, confidence is interpreted as a measure of decisiveness rather than validity. To support intuitive interpretation, confidence values are visualized using a color gradient ranging from red (low confidence) to green (high confidence), allowing users to quickly identify items that may warrant additional review.

The application was implemented as a Django-based web service with modular architecture consisting of four components:

User Input Layer: Users may submit individual items through a web interface or upload batches of items in CSV format.

Framework Retrieval Module: Subject-, grade-, and cycle-specific TIMSS framework definitions are dynamically retrieved from a structured internal repository. This ensures that classification decisions are explicitly conditioned on the appropriate version of the framework.

LLM Classification Engine: Submitted items are embedded within structured, framework-aware prompts that include full domain definitions and ranking-based confidence instructions.

The language model (GPT 5.2) returns standardized outputs specifying classifications and associated confidence scores.

Post-Processing and Visualization Layer: Model outputs are parsed, validated, and formatted for display. Confidence values are converted into visual indicators, and batch outputs can be downloaded for further analysis.

The application was intentionally developed as a research-grade investigational tool rather than a validated operational classification system. The tool is intended to augment expert review by providing structured preliminary classifications and transparent confidence indicators. It does not replace subject-matter expertise, particularly in cognitively ambiguous cases. Consistent with the empirical findings reported earlier, cognitive domain distinctions exhibit structural overlap in semantic space. The application therefore supports but does not automate expert judgment, providing a transparent decision-support system grounded in the current assessment framework.

Discussion

This study examined the feasibility of automated alignment of TIMSS assessment items with formal framework classifications using both prompt-based LLMs and supervised machine learning approaches. The findings reveal a differentiated pattern across construct dimensions, with implications for validity, construct representation, and responsible AI deployment in ILSA contexts.

Across all prompting conditions, content-domain classification demonstrated consistently high agreement with expert labels, with unsupervised prompt-based models achieving approximately 95% accuracy ($\kappa > 0.92$) in earlier analyses and comparably strong performance across cycles in the extended study. These findings align with prior supervised alignment studies. Studies using fine-tuned transformer models have similarly reported strong domain-level performance when content categories are semantically distinct (Xu et al., 2025). In large-scale mathematics alignment tasks, transformer architectures achieved weighted F1 values exceeding .90 for domain classification, suggesting that domain categories are highly recoverable from textual features (Xu et al., 2025). Likewise, investigations in reading and writing assessments demonstrate that domain alignment is substantially more stable than fine-grained skill classification (Fu et al., 2025).

What distinguishes the present findings is that comparable performance was achieved using a framework-aware few-shot chain-of-thought prompting strategy, without fine-tuning on large, labeled datasets. Whereas prior work relies heavily on supervised adaptation of pretrained models, the prompt-based approach embeds full framework definitions directly into the prompt, allowing the model to reason against formal domain descriptions. This suggests that when domain constructs are well-articulated and semantically coherent, prompt-based methods can approximate or outperform supervised performance.

The embedding-based separability analysis provides a structural explanation for this success. Content domains exhibited substantial semantic clustering across the full 2003-2023 corpus, indicating that items within a domain are meaningfully more similar to each other than to items from other domains. This aligns with findings from embedding-based alignment research showing that domain-level distinctions correspond to observable lexical regularities (Fu et al., 2025). The slightly stronger performance for science compared to mathematics further supports this structural interpretation. Science domains (Biology, Chemistry, Physics, Earth Science) are conceptually compartmentalized, whereas mathematics domains, particularly Number and Algebra, often intertwine procedurally and symbolically at Grade 8. The 21% cross-classification between Number and Algebra illustrates that even strong-performing models reveal construct-level overlap rather than purely random error.

Cognitive domain classification proved substantially more challenging. Across models, agreement clustered in the moderate range (e.g., $\kappa \approx 0.41$- 0.51 depending on method and split). These levels of agreement are not surprising when considering the broader literature. Nasström (2009) reported inter-rater reliability values of $\kappa \approx 0.41$- 0.47 among experts classifying items according to Bloom's taxonomy. Similarly, Karpen and Welch (2016) found only 46% agreement among faculty when categorizing exam questions by cognitive demand. These findings suggest that moderate agreement is characteristic of cognitive classification tasks, even among trained experts. The separability analysis reinforces this interpretation. Cognitive domains exhibited minimal semantic differentiation across the full item corpus (separability = 0.007), with near-identical within- and across-domain similarity. This pattern was consistent across subjects, grades, and cycles. The persistent confusion between Applying and Reasoning categories and the "middle-category bias" observed in both unsupervised and supervised models reflect this structural overlap.

These findings align with longstanding psychometric concerns regarding hierarchical cognitive taxonomies. Haberman and von Davier (2006) and von Davier and Haberman (2014) note that hierarchical attribute structures often collapse into Guttman-like or essentially unidimensional patterns in empirical data. From this perspective, attempts to distinguish finely graded cognitive levels may encounter inherent structural limits, as distinctions among categories are not strongly sustained in observable item features. Thus, the ceiling observed in cognitive classification performance reflects construct-level properties rather than purely model deficiencies.

As for the difficulty, prediction results exhibited a qualitatively different pattern from both content and cognitive alignment. Whereas content domains were highly recoverable and cognitive domains moderately recoverable, empirical difficulty proved substantially more resistant to automated inference from item content alone. This pattern is consistent with the broader literature on text-based item difficulty modeling (e.g., AlKhuzaey et al., 2023; Li et al., 2025; Peters et al., 2025; Yaneva et al., 2019), which characterizes difficulty as the most complex and least directly observable construct dimension.

Across both operationalizations of difficulty, ensemble tree-based models outperformed transformer-based approaches. For the b-parameter, Random Forest achieved the strongest performance ($R^2 = .203$, $r = .467$) and for p-value, Gradient Boosting produced the highest explanatory variance ($R^2 = .303$, $r = .554$), with Random Forest showing comparable performance. These results indicate that approximately one-fifth of latent difficulty variance and nearly one-third of observed performance variance can be recovered from item text, structured metadata, and LLM-derived cognitive features.

The consistently higher predictive accuracy of p-values compared to b-parameters suggests that observed performance is more linguistically grounded than IRT-calibrated difficulty estimates. As AlKhuzaey et al. (2023) highlighted in their systematic review, text-based features are excellent proxies for the cognitive load presented by the item's surface features. Our feature importance pattern supports this claim that p-values were sensitive to cognitive demand and multi-step reasoning markers. In contrast, the b-parameter was more strongly associated with developmental (grade-level) indicators. This indicates that while LLMs can identify the complexity of a task, the latent difficulty involves scaling processes that go beyond the text itself.

Transformer models did not surpass ensemble methods on either target, consistent with findings from a recent systematic review showing that classical machine learning approaches remain competitive with and often outperform state-of-the-art language models for difficulty prediction, particularly when structured metadata features are available, and dataset size is constrained (Peters et al., 2025). Moreover, the degradation in performance when adding metadata to BERT suggests a feature-integration gap, in which the dense contextual embeddings of transformers do not intuitively merge with discrete structural indicators such as grade level.

Taken together, these results support three broader conclusions. First, difficulty is fundamentally less recoverable from item text, reflecting its dependence on both cognitive and psychometric processes. Second, the choice of difficulty parameter materially affects model performance and interpretation; p-value and IRT b capture overlapping but distinct aspects of difficulty. Third, hybrid pipelines that combine structured metadata, cognitive modeling signals, and ensemble learning currently provide the most stable performance in large-scale STEM contexts.

Overall, item difficulty is shaped not only by linguistic and cognitive features but also by curriculum alignment, prior instruction, ability distribution, and scaling transformations (AlKhuzaey et al., 2023). These influences constrain the theoretical upper bound of content-only prediction. Within these structural limits, explaining 20-30% of empirical variance represents a meaningful, though bounded, contribution. Automated methods can therefore support preliminary estimation, screening, and expert review prioritization, but they cannot substitute for psychometric calibration. This conclusion aligns with the broader validity-oriented perspective adopted throughout this study: AI methods are most appropriately viewed as augmentative tools

within an evidentiary assessment framework, rather than as replacements for human judgment and empirical field testing.

Several limitations should be considered when interpreting these findings. First, all models rely on textual representations of items supplemented by AI-generated image captions. While captions convey structural and relational information about visual elements, they are an imperfect substitute for direct visual processing. Mathematics and science items frequently contain diagrams and graphs whose difficulty-relevant properties may not be fully captured in text descriptions. Multimodal modeling approaches that incorporate raw image features alongside text could potentially improve both cognitive domain and difficulty prediction, particularly for items where visual complexity is a primary source of cognitive demand.

Second, the prompt-based classification approach was evaluated using a single model family (GPT). Comparative evaluation across different LLMs, including open-access models, would strengthen conclusions about the generalizability of the framework-aware prompting strategy and clarify the extent to which performance is model-specific rather than method-specific.

Future research should examine data augmentation strategies to improve both cognitive domain and difficulty predictions. Techniques such as back-translation, paraphrase generation, or LLM-based item synthesis could help balance the distributions. Comparative evaluation of the prompt-based approach across different LLM families would further clarify the generalizability of the framework-aware prompting strategy.

Building on these directions, the findings also suggest several practical recommendations. Content-domain classification can be used as an efficient first-pass screening mechanism to verify alignment of newly developed or AI-generated items with framework definitions. Cognitive-domain classification is more appropriately applied as a triage tool, helping to identify clearly classifiable items while flagging ambiguous cases for expert review. Automated difficulty estimates should be interpreted cautiously and used only for preliminary screening or rough ordering of items, rather than for calibration purposes. Across all applications, automated outputs should complement, not replace, psychometric field testing and subject-matter expert judgment. When integrated in this way, automated alignment methods can reduce reviewer burden, support systematic checks across large item pools, and facilitate more iterative and scalable item development processes.

References

AlKhuzaey, S., Grasso, F., Payne, T. R., & Tamma, V. (2023). Text-based question difficulty prediction: A systematic review of automatic approaches. International Journal of Artificial Intelligence in Education, 1-53.

Anderson, D., Rowley, B., Stegenga, S., Irvin, P. S., & Rosenberg, J. M. (2020). Evaluating content-related validity evidence using a text-based machine learning procedure. *Educational Measurement: Issues and Practice*, *39*(4), 53-64.

Choi, H. J., Butterfuss, R., & Fan, M. (2025, October). Pre-trained Transformer Models for Standard-to-Standard Alignment Study. In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Full Papers* (pp. 306-311).

Fu, Y., Jiao, H., Zhou, T., Zhang, N., Li, M., Xu, Q., ... & Lissitz, R. W. (2025, October). Text-Based Approaches to Item Alignment to Content Standards in Large-Scale Reading & Writing Tests. In Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Coordinated Session Papers (pp. 19-36).

Freedle, R., & Kostin, I. (1993). The prediction of TOEFL reading comprehension item difficulty for expository prose passages for three item types: Main idea, inference, and supporting idea items. ETS Research Report Series, 1993(1), i-48.

Gombert, S., Menzel, L., Di Mitri, D., & Drachsler, H. (2024, June). Predicting Item Difficulty and Item Response Time with Scalar-mixed Transformer Encoder Models and Rational Network Regression Heads. In Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024) (pp. 483-492).

Haberman, S. J., & von Davier, M. (2006). 31b some notes on models for cognitively based skills diagnosis. Handbook of statistics, 26, 1031-1038.

He, P., Gao, J., & Chen, W. (2021). Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. arXiv preprint arXiv:2111.09543.

Karimi-Malekabadi, F., Razavi, P., & Powers, S. (2025). Scaling Item-to-Standard Alignment with Large Language Models: Accuracy, Limits, and Solutions. arXiv preprint arXiv:2511.19749.

Karlovčec, M., Córdova-Sánchez, M., & Pardos, Z. A. (2012, June). Knowledge component suggestion for untagged content in an intelligent tutoring system. In International Conference on Intelligent Tutoring Systems (pp. 195-200). Berlin, Heidelberg: Springer Berlin Heidelberg.

Karpen, S. C., & Welch, A. C. (2016). Assessing the inter-rater reliability and accuracy of pharmacy faculty's Bloom's Taxonomy classifications. Currents in Pharmacy Teaching and Learning, 8(6), 885-888.

Li, M., Jiao, H., Zhou, T., Zhang, N., Peters, S., & Lissitz, R. W. (2025). Item difficulty modeling using fine-tuned small and large language models. *Educational and Psychological Measurement*, *85*(6), 1065-1090.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. https://arxiv.org/abs/1907.11692

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. ETS Research Report Series, 2003(1), i-29.

Mislevy, R. J., & Riconscente, M. M. (2011). Evidence-centered assessment design. In Handbook of test development (pp. 75-104). Routledge.

Näsström, G. (2009). Interpretation of standards with Bloom's revised taxonomy: a comparison of teachers and assessment experts. International Journal of Research & Method in Education, 32(1), 39-51.

Perkins, K., Gupta, L., & Tammana, R. (1995). Predicting item difficulty in a reading comprehension test with an artificial neural network. Language testing, 12(1), 34-53.

Peters, S., Zhang, N., Jiao, H., Li, M., Zhou, T., & Lissitz, R. (2025). Text-Based Approaches to Item Difficulty Modeling in Large-Scale Assessments: A Systematic Review. *arXiv preprint arXiv:2509.23486*.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information processing & management, 24(5), 513-523.

Tian, Z., Flanagan, B., Dai, Y., & Ogata, H. (2022). Automated Matching of Exercises with Knowledge components. In *ICCE*.

von Davier, M., & Haberman, S. J. (2014). Hierarchical diagnostic classification models morphing into unidimensional 'diagnostic'classification models—a commentary. Psychometrika, 79(2), 340-346.

von Davier, M. & Kennedy, A. M. (2025). TIMSS 2027 Assessment Frameworks. Boston College, TIMSS & PIRLS International Study Center. https://doi.org/10.6017/lse.tpisc.timss.vp1245

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Wang, T., Stelter, K., Floyd, J., O'Neill, T., Hendrix, N., Bazemore, A., Rode, K., & Newton, W. (2023). Blueprinting the future: Automatic item categorization using hierarchical zero-shot and few-shot classifiers. arXiv. https://arxiv.org/abs/2312.03561

Xu, Q., Jiao, H., Zhou, T., Li, M., Zhang, N., Peters, S., & Fu, Y. (2025). Automated Alignment of Math Items to Content Standards in Large-Scale Assessments Using Language Models. *arXiv preprint arXiv:2510.05129*.

Yaneva, V., Baldwin, P., & Mee, J. (2019, August). Predicting the difficulty of multiple choice questions in a high-stakes medical exam. In Proceedings of the fourteenth workshop on innovative use of NLP for building educational applications (pp. 11-20).

Yaneva, V., Baldwin, P., & Mee, J. (2020, May). Predicting item survival for multiple choice questions in a high-stakes medical exam. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 6812-6818).

Yin, L., & Foy, P. (2021). TIMSS 2023 Assessment Design. TIMSS 2023 Assessment Frameworks, 71.Zhou, Z., & Ostrow, K. S. (2022, June). Transformer-based automated content-standards alignment: A pilot study. In International Conference on Human-Computer Interaction (pp. 525-542). Cham: Springer Nature Switzerland.

Appendix

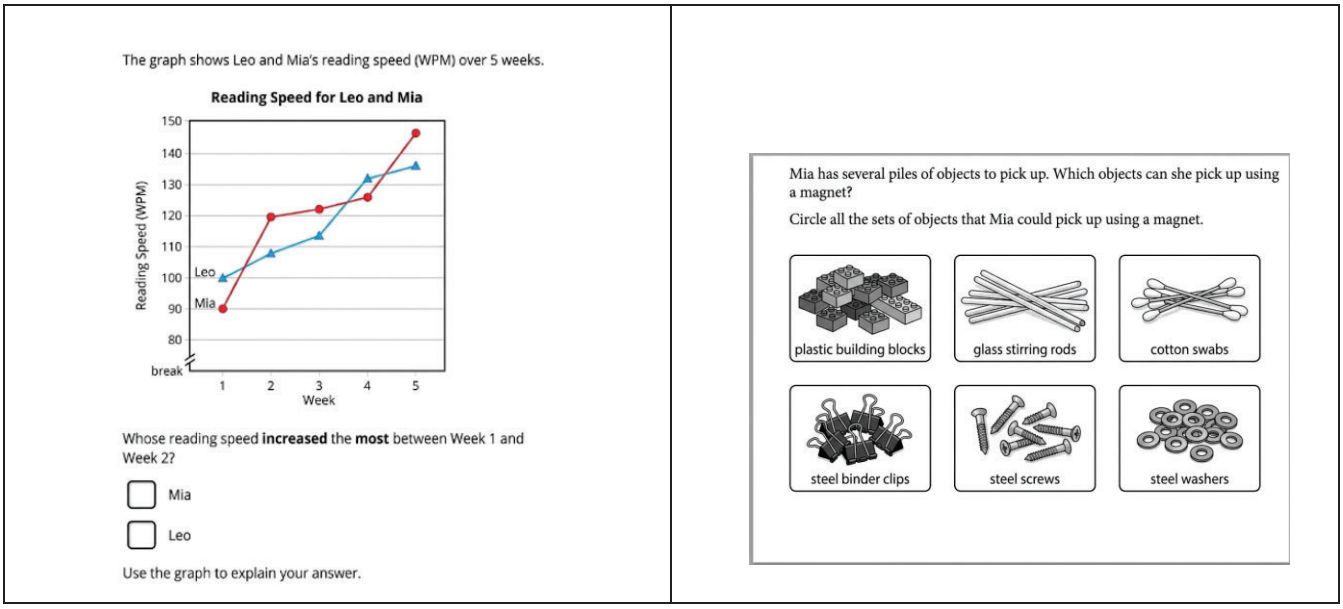
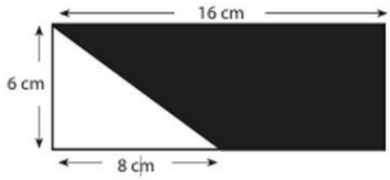


Figure A1. Examples of substantive and non-substantive visual content in TIMSS items. The left panel presents a data visualization requiring interpretation and is therefore captioned. The right panel presents visual elements that duplicate information in the options and are excluded from captioning (SKIP_BLANK).



In the figure above, what is the area of the shaded region in cm^2?

A. 24

B. 44

C. 48

D. 72

AI-Generated Caption: A rectangle with a width of 16 cm and a height of 6 cm is shown. A diagonal line extends from the top-left corner to a point on the bottom edge, dividing the rectangle into two regions. The left region is shaded, while the right region is unshaded.

Human-Corrected Caption: A rectangle with a width of 16 cm and a height of 6 cm is shown. A diagonal line extends from the top-left corner to a point on the bottom edge located 8 cm from the left side. This line divides the rectangle into two regions. The region to the right of the diagonal is shaded, while the triangular region on the left is unshaded.

Figure A2. Example of AI-generated and human-corrected image captions for a TIMSS item.

```
Act as an expert specializing in the TIMSS assessment framework.
Your task is to simulate how students interact with a {subject_name} item, diagnose
its cognitive demand, and judge its difficulty level from both an expert and a
student perspective. Analyze the given TIMSS Grade {grade} {subject_name} assessment
item.
Classify this item according to the TIMSS {year} {subject_name} Framework. Use these
three categories:
1. **Content Domain**: Select the main content domain from this list (use the exact
name):
{content_domains_text}

2. **Cognitive Domain**: Identify the main cognitive domain (choose exactly one:
Knowing, Applying, Reasoning):
{cognitive_domains_text}

3. **Difficulty Level**: Indicate the item's difficulty (Easy / Medium / Hard),
based not only on typical student success rates (Easy: >60%, Medium: 30–60%, Hard:
<30%) but also on complexity, required reasoning, potential misconceptions,
distractor strength, and student accessibility:
{difficulty_text}

---
## ANALYSIS (THINK STEP-BY-STEP)
### Step 1: **Student Simulation**
- Solve the item *as a typical Grade {grade} student*.
- Show your reasoning aloud, step by step.
- Point out where students may get confused, make errors, or fall for distractors.
### Step 2: **Content Domain Classification**
- Use exact names from the TIMSS content domains list.
- Choose based on the core concept tested, not superficial features.
### Step 3: **Cognitive Domain Classification**
Use this flowchart:
1. Does the item only require recall, recognition, or direct computation?
   - YES → **Knowing**
2. Does it require applying known procedures, using formulas, or following multi-
step operations?
   - YES → **Applying**
3. Does it require unfamiliar reasoning, making inferences, connecting ideas, or
justifying solutions?
   - YES → **Reasoning**
### Step 4: **Difficulty Judgment**
Consider:
- Multi-step vs. single-step processing
- Familiar vs. novel context
- Surface-level vs. conceptual thinking
- Strength and plausibility of distractors
- Language complexity and accessibility

### Examples
The following annotated examples demonstrate how to use the TIMSS framework for
accurate classification. Each example includes the item and final judgment.
Use these to guide your classification of the next item.
{examples_text}

Classify the following item considering all information available to you and
calibrate using the provided examples.
ITEM:
"""{item_text}"""

Provide only your classification in strict JSON format as below:
```json
   {{
      "content_domain": "{content_domain_names}",
      "cognitive_domain": "Knowing / Applying / Reasoning",
      "difficulty": "Easy / Medium / Hard"
   }}```
```

Figure A3. Full FS-CoT prompt structure
```

Table A1. Features Used for Difficulty Prediction

Feature Group	Feature	Description	Type
TF-IDF Text Features	Unigrams and bigrams	Term frequency–inverse document frequency weights extracted from item text	Continuous
Categorical Metadata	Subject	Mathematics or Science	Binary
	Grade	Grade 4 or Grade 8	Binary
	Cognitive Domain	Knowing, Applying, or Reasoning	Binary
	Content Domain	Subject-specific content area	Binary
	Cognitive Area	Fine-grained cognitive skill descriptor	Binary
	Content Area	Fine-grained content topic	Binary
Caption Features	Has caption	Whether item contains an image caption	Binary
	Caption length	Number of characters in caption	Continuous
	Caption word count	Number of words in caption	Continuous
Readability Indices	Flesch-Kincaid Grade Level	Text readability grade estimate	Continuous
	Flesch Reading Ease	Readability ease score	Continuous
	Gunning Fog Index	Reading complexity estimate	Continuous
	SMOG Index	Grade level readability estimate	Continuous
	Automated Readability Index	Character-based readability estimate	Continuous
Text Complexity	Text length (characters)	Total character count of item stem	Continuous
	Text length (words)	Total word count of item stem	Continuous
	Average word length	Mean number of characters per word	Continuous
	Type-token ratio	Lexical diversity measure	Continuous
	Sentence count	Number of sentences in item stem	Continuous

	Average sentence length	Mean words per sentence	Continuous
Mathematics-Specific	Contains number	Whether item stem contains digits	Binary
	Number count	Count of numeric values in item	Continuous
	Operator count	Count of mathematical operators	Continuous
	Has fraction	Whether item contains a fraction	Binary
	Has exponent	Whether item contains an exponent	Binary
	Variable count	Count of single-letter variables	Continuous
Question Type	What question	Item stem begins with "what"	Binary
	How question	Item stem begins with "how"	Binary
	Why question	Item stem begins with "why"	Binary
	Which question	Item stem begins with "which"	Binary
	Calculate/compute/find	Item requests a calculation	Binary
	Explain/describe/why	Item requests an explanation	Binary
LLM-Extracted	Cognitive demand	Cognitive load	Continuous
	Multi-step reasoning	Whether item requires multi-step reasoning	Binary
	Visual complexity	Visual complexity	Continuous
	Mathematical complexity	Mathematical complexity	Continuous
	Likely difficulty	Predicted difficulty	Continuous

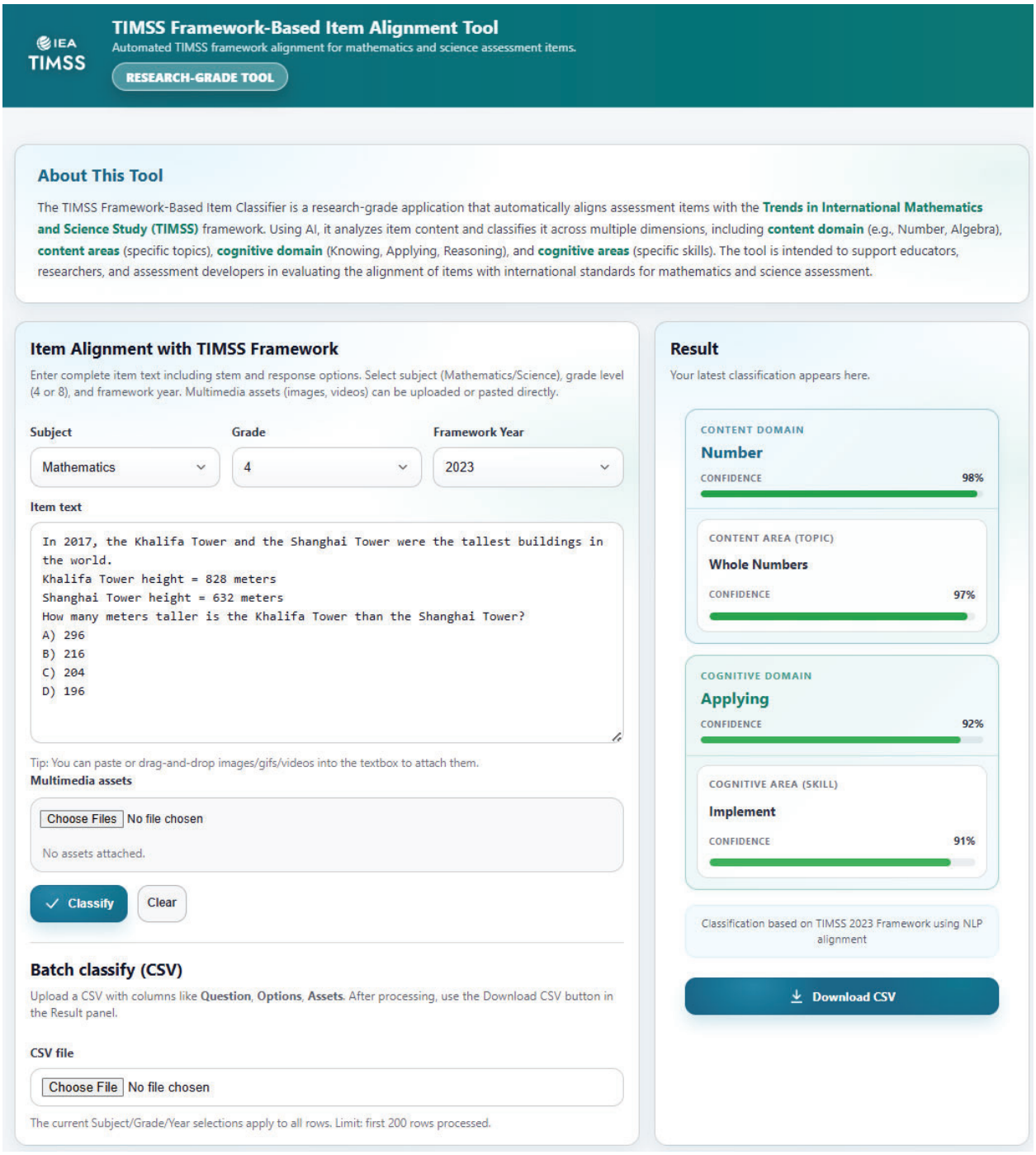


Figure A4. Web-Based Research-Grade Prototype for Automated Framework Alignment