Koh, E., et al. (Eds.) (2026). Proceedings of the 2nd International Conference on Learning Evidence and Analytics. Asia-Pacific Society for Computers in Education.

Automated Classification of Bloom's Taxonomy Levels for Japan's Information Technology Engineers Examinations

Toshiyo SETO[a*], Brendan FLANAGAN[a,b] , Yiling DAI[c], Masako OKAMOTO[b] & Hiroaki Ogata[a]

[a]*Graduate School of Informatics, Kyoto University, Japan*
[b]*Institute for Liberal Arts and Sciences, Kyoto University, Japan*
[c]*Graduate School of Advanced Science and Engineering, Hiroshima University, Japan*
*toshiyo.saitoh.27p@st.kyoto-u.ac.jp

Abstract: Understanding the cognitive demands of assessment items is essential for providing evidence-based learning support. This study analyzes Japan's Information Technology Engineers Examinations (ITEE), a national qualification exam system whose hierarchical proficiency levels align with the cognitive stages of the revised Bloom's Taxonomy. We first conducted a manual classification of exam questions across all four levels and found that the distribution of cognitive process levels remains largely consistent regardless of proficiency level, suggesting that exam difficulty is differentiated by factors other than cognitive complexity, such as the depth of domain-specific knowledge. To alleviate manual classification burdens, we compared four automated approaches: a lexical baseline (TF-IDF), a fine-tuned transformer (DeBERTa), LLM prompting (Gemini), and a multi-agent deliberation framework (Claude). Results show that while four-class classification remains challenging, binary classification into lower-order and higher-order thinking skills achieves high accuracy across all four baseline methods. Among the four-class approaches, the multi-agent method outperformed LLM prompting single-agent classification, demonstrating the value of structured deliberation for nuanced cognitive level assessment. These findings contribute to the development of metrics for characterizing exam content at scale, which can inform adaptive learning systems that optimize practice based on learners' cognitive skill profiles.

Keywords: Bloom's taxonomy, automated classification, ITEE, multi-agent LLM, learning analytics

1. Introduction

The Information Technology Engineers Examinations (ITEE) assesses IT knowledge across four proficiency levels. With over 500,000 annual applicants (IPA, 2024), developing personalized learning support is a pressing challenge. Such support should move beyond exam preparation to foster learning explicitly aligned with cognitive frameworks like Bloom's Taxonomy.

To provide AI-driven adaptive learning, it is necessary to understand what cognitive processes each exam question demands. A question requiring recall of a definition differs fundamentally from one requiring application of a concept to a novel scenario, even if both address the same topic. Characterizing this cognitive demand at the item level would provide a valuable metric for designing practice sequences that systematically develop learners' cognitive skills.

The revised Bloom's Taxonomy (Anderson et al., 2001) offers a well-established framework for classifying cognitive processes into six hierarchical levels. However, applying it to large-scale exam databases requires substantial expert effort, making manual classification impractical for the thousands of questions accumulated over decades of ITEE administration.

This study addresses this challenge in two stages. First, we manually classified 290 ITEE questions across all four proficiency levels for two primary purposes: 1) identifying the distribution of cognitive levels, showcasing a possible application of the classification results, and 2) generating the ground truth for the evaluation of the automated classification methods. Second, we developed and compared multiple automated classification approaches—a lexical baseline, a fine-tuned transformer model, single-agent LLM prompting, and a multi-agent deliberation framework—to enable scalable cognitive level annotation.

2. Related Work

The revised Bloom's Taxonomy (Anderson et al., 2001) classifies cognitive processes into six hierarchical levels—remember, understand, apply, analyze, evaluate, and create—and has been widely used in educational assessment to evaluate the cognitive complexity of test items (Krathwohl, 2002). While this framework has been applied across various educational contexts, its application to large-scale professional IT certification exams remains unexplored.

The automatic classification of exam questions based on Bloom's Taxonomy has attracted growing research interest. Approaches range from rule-based keyword matching (Mohammed & Omar, 2020) and TF-IDF with SVM classifiers (Yahya & Osama, 2011) to transformer-based models and LLMs. Kumar and Gulwani (2025) compared traditional machine learning and LLMs, finding that SVMs outperformed zero-shot LLMs on an English dataset. While most existing studies focus on English-language assessments, there is a significant opportunity to explore the potential of LLM approaches and their strategic combination with traditional machine learning methods in this domain. This study addresses this gap by investigating these advanced techniques specifically for Japanese exam questions, aiming to determine how structured AI collaboration can enhance classification performance.

In the broader area of educational content labeling, Flanagan et al. (2024) demonstrated the effectiveness of combining automated labeling with human-in-the-loop verification to improve both accuracy and efficiency. Building on this foundation, Yamauchi et al. (2025) demonstrated that lightweight text classification methods can effectively label Japanese mathematics exercises with curriculum-aligned knowledge components, even with incomplete textual input. These evolving methodologies, particularly those optimized for Japanese educational contexts, provide the direct methodological foundation for our current work. We extend these approaches by shifting the focus from subject-matter labeling to the more nuanced task of classifying cognitive process levels within the ITEE framework.

Recently, the paradigm of agentic AI and multi-agent systems has emerged as a promising approach for complex reasoning tasks. By assigning specialized personas to multiple LLM agents and facilitating structured debate, these systems can mitigate individual hallucinations and achieve higher reasoning accuracy. While such multi-agent frameworks have shown success in general problem-solving, their specific application to evaluating educational assessments—particularly mapping items to nuanced cognitive frameworks like Bloom's Taxonomy—remains largely unexplored, presenting a critical area for investigation.

3. The ITEE and Cognitive Level Analysis

3.1 Overview of the ITEE

The system comprises 13 examination categories organized into four proficiency levels: Level 1 (IT Passport) for basic literacy, Level 2 (Fundamental) for foundational knowledge, Level 3 (Applied) for practical application, and Level 4 (Specialist, e.g., Information Security and Database) for advanced expertise. Exam formats scale in complexity across these levels, progressing from simple multiple-choice questions at Level 1 to complex, scenario-based, and extended-response written items at Levels 3 and 4.

3.2 Manual Classification Using Bloom's Taxonomy

We manually classified 290 questions sampled across all four levels using the cognitive process dimension of the revised Bloom's Taxonomy. Two domain experts independently classified each question, with disagreements resolved through discussion. Only four of the six levels were observed—Remember (91 questions, 31.4%), Understand (112 questions, 38.6%), Apply (75 questions, 25.9%), and Analyze (12 questions, 4.1%)—as the exam formats do not elicit Evaluate or Create processes.

Table 1. Cognitive level distribution across ITEE proficiency levels (%)

Cognitive Level	IP (Level 1)	FE (Level 2)	AP (Level 3)	SC+DB (Level 4)
Remember	32.0	31.7	33.8	26.0
Understand	38.0	33.3	37.5	48.0
Apply	28.0	30.0	23.8	20.0
Analyze	2.0	5.0	5.0	6.0

As shown in Table 1, the most striking finding is that the cognitive level distributions are largely consistent across all four proficiency levels. Understand is the most frequent category at every level (33–48%), followed by Remember (26–34%) and Apply (20–30%), while Analyze remains rare (2–6%). This suggests that the differentiation of exam difficulty across proficiency levels is not primarily driven by shifts in cognitive complexity. Instead, the increasing difficulty may be attributed to the specialized nature, breadth, or depth of domain-specific knowledge required at each level, an area that warrants further investigation.

4. Automated Classification Approaches

We developed and compared four automated classification approaches using the 290 manually labeled questions. To ensure reproducibility and transparency, the full text of the prompts, inference settings, and model configurations used in this study are made publicly available on GitHub at: https://github.com/toshiyo-seto/blooms-taxonomy-classifier.All experiments used 5-fold cross-validation with Accuracy, Cohen's Kappa, and Macro F1 as evaluation metrics. We calculated Cohen’s Kappa from the contingency matrix to account for chance agreement. Macro F1-score was averaged across the four levels to prevent bias toward the dominant "Understand" and "Remember" classes.

4.1 Logistic regression, Transformer, and LLM Prompting Approaches

We evaluated three approaches as baselines: (1) a lexical approach using TF-IDF features with logistic regression; (2) a fine-tuned DeBERTa (He et al., 2021) model, a Japanese pre-trained transformer model whose disentangled attention mechanism captures nuanced linguistic cues; and (3) a prompting-based approach using Google's Gemini (gemini-3.1-pro-preview), evaluated with both zero-shot and few-shot settings. In the few-shot setting, two example questions per cognitive level (eight examples total) were provided as in-context demonstrations.

4.2 Multi-Agent Deliberation Framework

This framework was implemented using Claude (claude-sonnet-4.6), chosen for its superior performance in structured reasoning and the ease of building specialized agentic workflows. Inspired by the observation that human experts often disagree on borderline cases in Bloom's Taxonomy classification, we designed a multi-agent framework in which three specialized AI analysts independently evaluate each question and then reach a consensus through deliberation. Each analyst is assigned a distinct analytical perspective:

- **Analyst A (Cognitive Process):** Maps cognitive verbs to Bloom's 19 sub-categories.
- **Analyst B (Knowledge Dimension):** Infers the cognitive level via the Taxonomy Table based on knowledge type.
- **Analyst C (Difficulty Gradient):** Evaluates required processing depth using the exam level as a weak prior.

Each analyst independently assigns a Bloom's level with a justification. The three assessments are then combined through a structured deliberation process: if all three agree, that level is assigned; otherwise, the analysts exchange their reasoning and a final majority vote determines the classification. This design mirrors the multi-perspective approach recommended for reliable Bloom's classification (Anderson et al., 2001) and leverages the complementary strengths of different analytical lenses.

We compared this multi-agent approach against a single-agent baseline in which one LLM performs the classification without role specialization or deliberation.

5. Results and Discussion

5.1 Four-class and Binary Classification

Table 2 presents classification results for both four-class and binary (LOTS: Remember/Understand vs. HOTS: Apply/Analyze) settings. For conciseness, Table 2 reports only Accuracy; the full set of metrics, including Cohen's Kappa and Macro F1, is provided in Table 3 for the multi-agent comparison. In the four-class task, Gemini with few-shot prompting achieved the highest accuracy (0.689) among the single-agent approaches, outperforming fine-tuned DeBERTa (0.672) and the TF-IDF baseline (0.517). Error analysis revealed that most misclassifications occur between Remember and Understand. This is consistent with the known difficulty of distinguishing these adjacent levels even among human raters (Krathwohl, 2002).

Table 2. Classification Results for Four-class and Binary Settings

Method	4-class Accuracy	Binary Accuracy
TF-IDF + Logistic Regression	0.517	0.827
DeBERTa	0.672	0.827
Gemini (Zero-shot)	0.586	**0.879**
Gemini (Few-shot)	**0.689**	0.844

Simplifying to binary LOTS/HOTS classification substantially improved the performance of all methods, with Gemini's zero-shot prompting reaching the highest binary accuracy of 0.879. Interestingly, while few-shot examples were necessary to navigate the nuances of the four-class task, zero-shot prompting proved more effective for the broader binary distinction. This indicates that the LOTS/HOTS boundary is linguistically more salient and practically sufficient for identifying whether a question demands lower-order or higher-order thinking.

5.2 Multi-Agent vs. Single-Agent Classification

Table 3 compares the multi-agent deliberation framework against a single-agent baseline on the four-class task. The multi-agent approach consistently outperformed the single-agent method across all metrics.

Table 3. Multi-Agent vs. Single-Agent Comparison (Four-class)

Metric	Multi-Agent	Single-Agent (Claude)	Difference
Accuracy	0.710	0.679	+0.031
Cohen's Kappa	0.570	0.543	+0.027

Macro F1	0.648	0.622	+0.026
Inter-Analyst Agreement	0.807	—	—

As shown in Table 3, the multi-agent Claude approach achieved the highest accuracy of 0.710, outperforming both its single-agent counterpart and the best-performing single-agent Gemini model from the previous experiment. The inter-analyst agreement rate of 0.807 indicates convergence in most cases, while the consistent improvement over the single-agent baseline across all metrics (+0.026–0.031) demonstrates the value of decomposing cognitive assessment into complementary analytical perspectives.

Table 4. Per-class F1-scores for LLM-based approaches

Method	Remember	Understand	Apply	Analyze
Gemini (Few-shot)	0.810	0.620	0.610	0.670
Single-Agent (Claude)	0.773	0.558	0.571	0.500
Multi-Agent (Claude)	0.733	0.702	0.667	0.400

To address the severe class imbalance, particularly for the rare "Analyze" class (which constitutes only 4.1% of the dataset), Table 4 presents the per-class F1-scores for the LLM-based approaches. We focused this fine-grained analysis on the LLM methods as they demonstrated the highest overall accuracy in the four-class setting. While the minority "Analyze" class remains inherently challenging, the LLMs showed capability in identifying these rare instances. Notably, while Gemini excelled at identifying the rare class, the Multi-Agent framework delivered the most balanced and consistent performance across the dominant instructional categories ("Understand" and "Apply"), highlighting the value of structured deliberation for resolving the core cognitive processes.

This finding aligns with recent advancements in agentic AI, demonstrating that while single LLMs are susceptible to prompt sensitivity and hallucinations in subjective assessment tasks, a multi-agent deliberation framework effectively regularizes these outputs. By externalizing the internal reasoning steps into a transparent debate among specialized roles, the system mirrors the human consensus-building process, ultimately leading to more reliable educational labeling.

5.3 Summary of Findings

These results yield three key insights. First, four-class Bloom's Taxonomy classification of Japanese IT exam questions is inherently challenging, with even the best method achieving 0.710 accuracy. Second, simplifying the task to a LOTS/HOTS binary classification makes high-accuracy automated annotation feasible (0.827–0.879), providing a practical tool for large-scale learning analytics. Third, multi-agent deliberation with specialized analytical roles outperforms all other approaches for fine-grained classification, demonstrating the value of structured reasoning over both statistical patterns matching and general-purpose LLM inference.

6. Conclusion

This study investigated the cognitive demand of Japan's ITEE using the revised Bloom's Taxonomy and developed automated classification models. Manual analysis of 290 questions revealed consistent cognitive level distributions across all four proficiency levels, with Understand being most prevalent (33–48%) and Analyze remaining rare (2–6%). This finding indicates that ITEE difficulty is primarily differentiated by domain-specific knowledge rather than cognitive complexity, which has direct implications for the design of adaptive learning systems.

For automated classification, we compared traditional machine learning, deep learning, single-agent LLM prompting, and a multi-agent deliberation framework. While four-class

classification remains challenging, the few-shot Gemini approach (0.689) outperformed the fine-tuned DeBERTa model (0.672). The multi-agent Claude framework achieved the highest overall performance (Accuracy: 0.710), demonstrating the effectiveness of decomposing cognitive assessment into complementary analytical perspectives. Binary classification into LOTS and HOTS achieved high accuracy across all baseline methods (up to 0.879 with zero-shot prompting), offering a practical and reliable metric for large-scale exam content characterization.

Several limitations exist. First, the 290-question dataset is relatively small for fine-tuning. Second, the ground truth relies on subjective expert consensus, and formal inter-rater reliability (IRR) metrics were not calculated; addressing this rigor remains for future research. Additionally, the current analysis is limited to multiple-choice and scenario-based questions; the framework has not yet been applied to long-form written questions at higher levels. The multi-agent approach, while achieving the best accuracy, incurs higher computational cost due to multiple LLM calls per question. Finally, the current evaluation of the LLM-based approaches relies on a single inference run. Assessing performance variability across multiple runs to firmly establish the robustness of the results will be addressed in future extended research.

Future work will pursue two directions. First, we plan to expand the labeled dataset and refine the multi-agent framework to further improve four-class classification accuracy. Second, we intend to integrate the automated cognitive level annotations into an AI-driven adaptive learning system that optimizes practice question selection based on both topic coverage and cognitive demand profiles, thereby providing personalized and evidence-based learning support for ITEE candidates.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers JP23H01001, JP23H00505, JP24K20902, JP26K06443, and the MEXT Project for Training Experts in Statistical Sciences.

References

Anderson, L. W., et al. (2001). A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives. Longman.

Flanagan, B., Tian, Z., Yamauchi, T., Dai, Y., & Ogata, H. (2024). A human-in-the-loop system for labeling knowledge components in Japanese mathematics exercises. *Research and Practice in Technology Enhanced Learning*, *19*, 28. https://doi.org/10.58459/rptel.2024.19028

He, P., Liu, X., Gao, J., & Chen, W. (2021). DeBERTa: Decoding-enhanced BERT with disentangled attention. In *Proceedings of the International Conference on Learning Representations (ICLR 2021)*. https://doi.org/10.48550/arXiv.2006.03654

Information-technology Promotion Agency, Japan. (2024). *The Information Technology Engineers Examination*. https://www.ipa.go.jp/en/it-examinations/jitec.html

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory Into Practice*, *41*(4), 212–218. https://doi.org/10.1207/s15430421tip4104_2

Kumar, R., & Gulwani, D. (2025). Automated analysis of learning outcomes and exam questions based on Bloom's taxonomy. *arXiv preprint arXiv:2511.10903*. https://doi.org/10.48550/arXiv.2511.10903

Mohammed, M., & Omar, N. (2020). Question classification based on Bloom's taxonomy cognitive domain using modified TF-IDF and word2vec. *PLoS ONE*, *15*(3), e0230442. https://doi.org/10.1371/journal.pone.0230442

Yahya, A. A., & Osama, A. (2011). Automatic classification of questions into Bloom's cognitive levels using support vector machines. In *Proceedings of the International Arab Conference on Information Technology*. Naif Arab University for Security Sciences.

Yamauchi, T., Nakamoto, R., Flanagan, B., Dai, Y., Wijerathne, I., & Ogata, H. (2025). Augmentation of learning content with knowledge components: Automatic unit labeling for various forms of Japanese math materials. *IEEE Transactions on Learning Technologies*. https://doi.org/10.1109/TLT.2025.3584038