APPENDIX A: ERROR CATEGORIES, SAMPLE METADATA AND LLM PROMPTS

Item Writing Flaw	Description of desired question attribute
Unclear information	Questions and all options should be written in clear, unambiguous language.
Implausible Distractors	Make all distractors plausible as good items depend on having effective distractors.
None of the above	Avoid none of the above as it only really measures students ability to detect incorrect answers.
Longest option correct	Avoid having the correct option be longer and include more detailed information, since this may attract students to this option.
Gratuitous information	Avoid unnecessary information in the stem that is not required to answer the question.
True/False question	The options should not be a series of true/false statements.
Convergence cues	Avoid convergence cues in options where there are different combinations of multiple components to the answer.
Logical cues	Avoid clues in the stem and the correct option that can help test-savvy students identify the correct option.
All of the above	Avoid all of the above options as students can guess correct responses based on partial information.
Fill-in-the-blank	Avoid omitting words in the middle of the stem that students must insert from the options provided.
Absolute terms	Avoid the use of extreme absolute terms (e.g. never, always, all) in the options, as students are aware that these are almost always false.
Word repeats	Avoid similarly-worded stems and correct responses, or words repeated in the stem and correct response.
Unfocused stem	The stem should present a clear and focused question that can be understood and answered without looking at the options.
Complex or K-type	Avoid questions that have a range of correct responses, so that students need to select from a number of possible combinations of the responses.
Grammatical cues	All options should be grammatically consistent with the stem and should be parallel in style and form.
Lost sequence	All options should be arranged in chronological or numerical order.
Vague terms	Avoid the use of vague terms (e.g. frequently, occasionally) in the options as there is seldom agreement on their actual meaning.
Negative worded	Negatively worded stems are less likely to measure important learning outcomes and can confuse students.
More than one correct	In single best-answer form, questions should have exactly one best answer.

Table 1: Description of the IWF categories used in this study for automated evaluation (rule-based and LLM-based).

Extracted metadata

Sample additional non-question metadata extracted from a lecture on machine learning is given below in Table 7.

LLM Prompts

The prompts we used to generate/extract the data are below:

- **Summary**: Following is the transcript of a lecture: {lecture_text}. Please summarize this in a way that it can be used by a student to review the lecture and attempt questions.
- **Lecture Segmentation**: Please split the transcript of the lecture into a minimum of 3 and a maximum of 4 segments. A segment is a group of consecutive lines in the text where topics within the group are semantically similar to topics across all the sentences. Return the full text for each segment from the lecture. Return the segments in JSON format. The Segment number for each segment will be the key, and the text for the associated segment will be the value. Try to keep the size of each segment the same/similar. Every line in the lecture must fall in some segment.
- **Key Topics**: Please get me the five most important Machine Learning topics (in the form of very short phrases) that are discussed in {lecture_segment}. Only include topics from {wikipedia_set} Return the topics in the form of a JSON, with the key being "Concepts" and the list of topics being the value. Please do not include any extra text in the response
- **Key Definitions**: Please give me the key definitions, if any, that are discussed in {lecture_segment}. Give the definitions in JSON format, with the words as the keys and the definitions as the values. Please do not include any extra text or characters like the next line character in the response
- **Key Examples**: Please give me the explanation of the key examples (if any), used to explain a concept, in this text: {lecture_segment}. Give the response in the form of a JSON format. The key will be the name of the example - i.e., the text being referenced from the lecture, and the value will be the explanation. Please do not include any extra text like "Example <number>" or characters like the next line character in the response

Metadata	Example
Procedural Knowledge	**How to create an ensemble model**: An ensemble model is created by combining multiple individual learning models to produce an aggregate model that is more powerful than any of its individual learning models alone. This is effective because different learning models, although each of them might perform well individually, they'll tend to make different kinds of mistakes on a data set. Typically, this happens because each individual model might overfit to a different part of the data. By combining different individual models into an ensemble, we can average out their individual mistakes to reduce the risk of overfitting while maintaining strong prediction performance.
Key concepts	Ensembles Bagging Boosting Random forest, Decision trees, Overfitting, Supervised Learning, Regression
Key definitions	**Ensembles**: A method in machine learning that involves creating learning models by combining multiple individual learning models to produce an aggregate model that is more powerful than any of its individual learning models alone. **Overfitting**: A modeling error in machine learning occurs when a function is too closely fit to a limited set of data points.
Key Examples	**Random Forests**: Random forests are given as an example of the ensemble idea applied to decision trees. They are widely used in practice and achieve very good results on a wide variety of problems. Random forests can be used as classifiers via the scikit learn random forest classifier class or for regression using the random forest regressor class both in the sklearn ensemble module. The use of random forests helps to overcome the disadvantage of using a single decision tree, which is prone to overfitting the training data.

Table 2: Sample (non-question) metadata from a machine learning lecture

- **Procedural Knowledge**: Please give me the "how to" explanations that are given, if any, in {lecture_segment}. Please directly start the response with the "how to" explanations and do not include any extra text in the response. Return the explanations in a JSON form The keys of the JSON should be "How to <the procedure>" and the value should be the explanation. Keep the explanation in the form of a paragraph.
- **Questions**: Please write a minimum of 10 and maximum of 15 unique multiple choice questions, with four choices each, from the text: {lecture_segment}. The questions will be used to test the knowledge of the students regarding the different concepts and examples covered in the text, hence the questions need to cover them. The questions should be good enough to be given in a technical exam. The questions need to be returned in a JSON format, with the keys being "Question <question number>" and values being another JSON. The sub-JSON containing the question data needs to be in this format:
  "Question": <The question statement>
  "A": <Option 1>
  "B": <Option 2>
  "C": <Option 3>
  "D": <Option 4>
  "Correct answer": <The correct answer A, B, C or D>
  "Explanation": <Explanation for the correct answer>

We gave specific formatting instructions to facilitate automated parsing and storage of the data. To make sure GPT was consistent with its output, we used the JSON key-value format for our data.

Sample evaluated questions

- Sample question that passed all human evaluation metrics:
  - Question: What is the main drawback of overfitting?
  - A: The model doesn't capture the trends in the data
  - B: The model captures both the general trend and the noise in the data
  - C: The model focuses too much on local variations
  - D: The model generalizes well to test data
- Sample question that failed most human evaluation metrics:
  - Question: How many data set samples are present in the regression problem?
  - A: 10
  - B: 50
  - C: 100
  - D: 200
- Sample question that passed all automated evaluation metrics:

  - Question: What are the two main types of data leakage?
  - A: Leakage in the training data and leakage in features
  - B: Leakage in the testing data and leakage in labels
  - C: Leakage in the prediction data and leakage in algorithms
  - D: Leakage in the validation data and leakage in models
- Sample question that failed 6 important automated evaluation metrics:
  - Question: What is a common issue when fixing data leakage problems?
  - A: Data leakage problems are usually easy to fix and do not require much effort
  - B: Fixing one leaking feature can reveal the existence of a second one
  - C: Fixing data leakage problems often leads to a decrease in model performance
  - D: Data leakage problems are typically isolated and do not affect other features