INTERNATIONAL MONETARY FUND

Enhancing IMF Economics Training: AI-Powered Analysis of Qualitative Learner Feedback

Andras Komaromi, Xiaomin Wu, Ran Pan, Yang Liu, Pablo Cisneros, Anchal Manocha, Hiba El Oirghi

WP/24/166

2024
AUG

WORKING PAPER

 WP/24/166

IMF Working Paper
Institute for Capacity Development

Enhancing IMF Economics Training: AI-Powered Analysis of Qualitative Learner Feedback

Prepared by Andras Komaromi, Xiaomin Wu, Ran Pan, Yang Liu, Pablo Cisneros, Anchal Manocha, and Hiba El Oirghi*

Authorized for distribution by Oussama Kanaan
August 2024

ABSTRACT: The International Monetary Fund (IMF) has expanded its online learning program, offering over 100 Massive Open Online Courses (MOOCs) to support economic and financial policymaking worldwide. This paper explores the application of Artificial Intelligence (AI), specifically Large Language Models (LLMs), to analyze qualitative feedback from participants in these courses. By fine-tuning a pre-trained LLM on expert-annotated text data, we develop models that efficiently classify open-ended survey responses with accuracy comparable to human coders. The models' robust performance across multiple languages, including English, French, and Spanish, demonstrates its versatility. Key insights from the analysis include a preference for shorter, modular content, with variations across genders, and the significant impact of language barriers on learning outcomes. These and other findtpeings from unstructured learner feedback inform the continuous improvement of the IMF's online courses, aligning with its capacity development goals to enhance economic and financial expertise globally.

RECOMMENDED CITATION: Komaromi, A., Wu, X., Pan, R., Liu, Y., Cisneros, P., Manocha, A., Oirghi, H. (2024). Enhancing IMF Economics Training: AI-Powered Analysis of Qualitative Learner Feedback, IMF Working Papers, WP/24/166.

JEL Classification Numbers:	A29, C55, I21
Keywords:	capacity development; online learning; artificial intelligence; large language models; qualitative feedback
Authors' E-Mail Address:	akomaromi@imf.org; fred.xiaomin.wu@gmail.com; rpan@imf.org; yliu10@imf.org; pcisneros@imf.org; amanocha@imf.org; heloirghi@imf.org

* The authors are grateful to Daniel Maldonado, Rebeca Hassan, and Amy Lee for developing the learner feedback evaluation framework and for providing the data that underlies this research.

WORKING PAPERS

Enhancing IMF Economics Training: AI-Powered Analysis of Qualitative Learner Feedback

Prepared by Xiaomin Wu, Andras Komaromi, Ran Pan, Yang Liu, Pablo Cisneros, Anchal Manocha, and Hiba El Oirghi[1]

[1] The authors are grateful to Daniel Maldonado, Rebeca Hassan, and Amy Lee for developing the learner feedback evaluation framework and for providing the data that underlies this research.

Contents

1. Introduction

As part of its broader capacity development (CD) mandate, the International Monetary Fund (IMF) provides policy-oriented economics and finance training to member countries through a variety of modalities. For a long time, IMF training courses and workshops have been delivered in traditional classroom settings. Since 1981, IMF economists have taught over 7,000 courses worldwide and trained over 200,000 country officials in face-to-face and live virtual classrooms. As the demand for training grew and technology-supported self-paced learning became mainstream, the IMF started to experiment with online courses: not only allowing for the scaling up of training delivery but also expanding its reach.

The IMF Online Learning (OL) Program has expanded exponentially since its inception in 2013. As of May 2024, the IMF offered over 100 Massive Open Online Courses (MOOCs) in six languages on the edX platform. The program has registered more than 210,000 active participants from across the globe, with more than one-third of participating government officials from Sub-Saharan Africa, highlighting its value in increasing the reach of CD delivery.

The IMF learning team gathers participant feedback on the vast portfolio of courses (hereafter referred to as IMFx courses) to understand how training is received and what areas need improvement. The feedback survey includes quantitative and closed-form responses as well as open-ended responses to capture qualitative information on learners' experience and on the perceived strengths and weaknesses of the online training. This paper describes how learning experts at the IMF started to use Artificial Intelligence (AI), in particular Large Language Models (LLM), to gain insights from a growing amount of unstructured qualitative data.

To efficiently process open-ended questions, we combined a pre-trained LLM and human-labelled text responses in a supervised classification task. First, learning experts created a framework to classify text responses and manually labelled thousands of data points. Second, we appended an encoder-only LLM with a simple classification layer. The classifier receives the vector representation (or embedding) of the text responses and learns to predict the categories defined by the experts. Finally, we trained this deep neural network on the manually labelled data.

Our model outperforms several simple benchmarks, and its reliability is on par with human coders. For example, the model has significantly higher accuracy than searching for carefully selected keywords. Interestingly, the model's alignment with human coders matches closely the level of agreement between humans when asked to code the same responses following the same guidelines. We also show that the LLM's pre-training on a multi-lingual corpus allows it to deliver similar performance across languages despite the considerably lower amount of non-English training data for the classification task.

We also examine and improve the calibration of the model to better assess the uncertainty around its predictions. Without further adjustments, neural network classifiers are not well-suited for a probabilistic interpretation of their outputs. In particular, we find that our classifier tends to be underconfident in its final predicted class. We apply a post-processing step to map the model's raw confidence scores to the probability that the prediction is correct. This adjusted probability can be useful, for example, to design a decision rule for human review of the survey response.

Although this paper focuses on describing the technical details and performance of the developed models, we illustrate the value of feedback evaluation system with two examples. First, we show that shorter and modular learning content caters better to the IMF online learners, providing the flexibility which is particularly appreciated by female participants. Second, we provide evidence from the qualitative survey responses that the IMF's efforts to adapt courses to non-English languages increases the efficiency and reach of its training.

Ultimately, our language models enable large-scale analysis of qualitative learner feedback. The manual annotation and processing of learners' written comments is impractical due to the large and growing volume of data. However, with AI language models we can gain interesting insights into the preferences and perceptions of MOOC participants, aiding in the improvement and better targeting of online training.

The rest of the paper is structured as follows. Sections 2 and 3 present the framework and tools for impact measurement and the manual process of feedback evaluation that was in place before developing AI-based methods. Section 4 describes the development of an automatic classification model based on finetuning a pre-trained LLM on expert-classified text responses. In Section 5 we discuss the model's performance and the uncertainty around its predictions. Finally, Section 6 provides initial insights derived with the help of the model.

2. Measuring Training Impact and Learner Experience

The broad framework for monitoring and evaluating the impact of IMFx courses follows the Kirkpatrick (1976) model, which is a well-established standard for training evaluation at the workplace. This model distinguishes four levels of measuring and understanding training impact: participant reaction, learning, behavioral change, and results (Figure 1).

Figure 1. The Kirkpatrick Model

Level 1 and Level 2 evaluation of IMFx training takes place during and immediately after completing the online courses. Participant's acquired knowledge and skills are measured through an identical pre-course and post-course test that provides information on learning gains (Level 2). Participants' reaction to the course, such as perception of usefulness, strengths, weaknesses, and applicability, is measured through a post-course survey (Level 1). Behavior surveys are conducted every two years to gauge the application of acquired knowledge in day-to-day work and to identify enablers and barriers for application on the job (Level 3). In addition, a survey is conducted for participants who sign up for online courses but do not complete them. This non-completing participant survey provides insights into the reasons for dropping out.[1]

The post-course (Level 1) survey captures a large amount of quantitative and qualitative information on participants' self-reported satisfaction together with demographic information. The questionnaire is administered using Cvent's survey functionality, and includes both close-ended (fixed-alternative) questions and open-ended questions where participants can enter any text (Appendix I). The survey's response rate is over 20 percent, which compares favorably to similar MOOC feedback surveys (Tzeng, Lee, Huang, Huang, & Lai, 2022). This yields responses from over 7,000 participants per trimester in six languages. Most text responses are in English, but a significant portion come in French (10 percent) and Spanish (6 percent), while Portuguese, Arabic, and Russian account for about 1 percent.

While the analysis of the quantitative responses can be automated using standard statistical packages, making sense of the qualitative data previously required laborious manual review and coding. The next section provides an overview of the manual coding process and lays out the argument for applying natural language modelling techniques to increase efficiency.

[1] Level 4 evaluation is currently not undertaken for IMFx courses.

3. Manual Review of Qualitative Responses

Open-ended questions in the survey seek learners' views on the strengths and weaknesses of the course, on gaps in the content, and on the usefulness of the training for respondents' job responsibilities. Participants can also share any other comments or suggestions for improvements. These completely open-form questions do not have an effective word limit, but most responses are short with an average length of 13 words. Table 1 provides some representative responses to each of the five qualitative questions.

Table 1. Examples of open-ended responses in the feedback survey

Survey Questions	Sample Feedback Comments
What were the strengths of the course? (Strengths)	*"The courses were given in a very simple and practical way that makes it easy to understand, especially for a non-English speaking person like me."* *"Richness of the course in relation to the different aspects of public finance management. Country examples and experiences. Very complete documentation and many tools provided."*
What were the weaknesses of the course? (Weaknesses)	*"The course may not keep up with emerging issues and trends in the PFM field, such as the increasing importance of digitalization and technology in PFM."* *"It is too lengthy so duration can be cut short and graphical presentation can be applied in the course."*
Are there topics that were not covered or not explained well enough that you think should be included? Please describe. (Missing topics)	*"It would be good to expose in detail the various indirect methods of compiling QNA (mathematical and econometric methods)."* *"I think all the proposed topics were covered thoroughly and in detail."*
How will you apply the learning from this course to your job? (Application)	*"As I am also a member in Governance and Audit committee of the nodal Environmental Conservation agency (BHTFEC), the knowledge I gained from the course will be an immense benefit for assessing the area of investments and its long-term impacts."* *"As someone who works in the socio-economic planning body in our country, the learnings from this course will help me contribute inputs in the crafting of regional plans and policies in relation to climate change."*
Please share any additional comments you may have around this course. (Other comments)	*"Course was well thought and thorough in covering the content."* *"This course should be kept open for all throughout the year as it's insightful and can help every stakeholder of PFM."*

Note: For brevity, the phrases in parenthesis will be used to refer to each question throughout the paper.

To rigorously incorporate qualitative feedback into course evaluation, a manual coding process was developed. Reading comments can provide detailed insights into learners' views, but

without systematically categorizing this information, it is hard to identify patterns that could be useful for course evaluation and improvement. To tackle this challenge, a group of six learning experts read through the comments of six courses (totaling 185 surveys and over 700 individual comments) and identified, through iterative discussions, a set of frequently mentioned themes for each question.

Table 2 summarizes the coding guidelines for the Strengths question with identified themes, descriptions, and examples of representative responses. When describing the strengths of the course, participants often referred to the quality or organization of the material, the relevance or applicability of the content, the expertise of the instructors, and the convenience of taking IMF online training. The complete codebook for all questions is included in Appendix II.

Table 2. Question-specific codebook for response categorization

Question-specific Codebooks		
What were the strengths of the course?		
Codes/Themes	**Description**	**Example quotes**
Strength - Material	Positive comments pertaining to the quality or organization of material, course concepts, explanations, etc., as well as to the visual display or delivery of content (e.g., video, activity, exercises, etc.)	*It is very well structured and organized. The information is very clear and easy to understand.*
Strength - Practicality	Positive comments pertaining to the practicality and relevance of content; and its real-life applicability in country/work context	*Examples from several jurisdictions gave a holistic view of the subject and the best practices followed by different countries can be taken as a model for implementation by others.*
Strength - Expertise	Positive comments pertaining to the expertise, experience, and aptitude of the instructors/trainers	*The Lecturers, they are well versed in topics discussed throughout the module and very effective delivering the main learning objective of the course.*
Strength - Convenience	Positive comments related to ease, affordability, accessibility of taking a course	*- I was able to study at the comfort of my home and at a convenient time.* *- I did not have pay to study the course.* *- Wider data study material.*
Strength - Other	Other comments not captured in the codes/descriptions above. Vague or unintelligible comments were also merged into this category.	

Each comment was assigned relevant codes (typically one or two, and no more than three) from the codebook, with a mechanism to build consensus in borderline cases. Sometimes learners addressed more than one theme in a single response. Although coders were allowed to use multiple codes, if necessary, in practice less than 5 percent of comments received more than

one label. Despite following the same coding guidelines, human classification inevitably contains a subjective element and different coders may assign different labels to the same response. To ensure consistency, whenever the assigned coder had doubts about classifying a response, it was escalated to a group of at least three team members who made a consensus decision on the final label. About 4 percent of comments went through this group audit.

This elaborate manual labeling process required considerable resources, highlighting the potential benefits of automation via natural language processing techniques. A group of learning experts collectively spent an estimated 90 hours to classify over 15,000 comments from 89 courses between June 2022 and Feb 2023. This large initial investment produced a vast amount of human-labelled text data, which made it feasible to experiment with automation by fine-tuning a pre-trained large language model.

4. Automation with an LLM

Pre-trained Large Language Models (LLMs) have been shown to dramatically reduce the cost of processing qualitative user feedback. For example, LLMs are routinely used to automatically extract key themes and sentiments from lengthy customer reviews, saving substantial human labor. Since pre-trained LLMs already encode syntactic and semantic language features, they can be efficiently fine-tuned using only a limited amount of human-labelled data.

A popular choice for many natural language processing (NLP) tasks is the Bidirectional Encoder Representations from Transformers (BERT) model (Devlin, Chang, Lee, & Toutanova, 2018).[2] BERT is a deep learning model that encodes words (tokens) in a sentence (sequence) considering their context (surrounding words) from both directions. During pre-training, the encodings are optimized on a large corpus of text to achieve two objectives: masked word prediction and next sentence prediction. In masked word prediction, the model is trying to guess randomly hidden (masked) words in the sentence. In next sentence prediction, the model is trying to determine whether one sentence follows another one in the training corpus. It has been shown that optimizing for these two simple tasks produces encodings that capture important linguistic features.

Building on the general BERT architecture, models can be efficiently trained for many downstream NLP tasks. The last hidden layer of BERT produces contextualized embeddings (vector representations) of each token and the whole sequence, which can be fed into additional layers for further transformations. By adding this additional block (head) on top of the pre-trained model, BERT can be fine-tuned to perform specific tasks such as sentiment analysis or topic classification. During fine-tuning, the weights of the pre-trained BERT model are

[2] Even after the appearance of powerful generative AI models (e.g., OpenAI's GPT series), BERT remains the preferred choice for NLP tasks in many applications. First, the bidirectional nature of the encoder enables good understanding of the context of full paragraphs. Second, BERT is open source and has an extensive community of developers who have created easily re-usable code bases. Third, its relatively smaller size allows for efficient fine-tuning for specific tasks and quick inference.

updated together with the additional layers based on the task-specific labelled training data. The resulting model inherits BERT's general language understanding, but its performance is optimized for the given task.

In our application, we added a classification head on top of BERT's base architecture to enable fine-tuning on the manually coded survey responses.[3] The classification head simply involves projecting the sequence embedding into a space with dimensionality equal to the number of different classes. Taking the maximum value of these class scores (logits) yields the predicted class. To facilitate further interpretation and analysis, we also apply the SoftMax function on the logits to get normalized confidence scores that are akin to a probability distribution over the classes. In the next section we revisit the question on whether the raw confidence score can be interpreted as true probabilities.

Before fine-tuning, the labelled data is preprocessed. We drop entries indicating an empty response (e.g., blank, na, none) and nonsensical inputs that contain only one unique character (0.4 percent of non-empty responses). For survey responses with more than one code (<5 percent of observations), only the first code was retained. We also remove any characters that are not words, white spaces, or digits from the response sentences. Additionally, all response sentences are converted to lowercase.

For each survey question we trained a separate model, randomly assigning 90 percent of the expert-coded data for model training and 10 percent for testing metrics. The size of the resulting training data is reported in Table 3. We run the training for a maximum of 30 epochs, with early stopping based on the best F1 score within the first 5 epochs to prevent overfitting to the training samples. We use a mini-batch size of 16 samples and the Adam optimizer (Kingma & Ba, 2014) with a decreasing learning rate schedule to facilitate quicker convergence.

Table 3. Size of training data after pre-processing responses

Question	Observations
Strengths	3,517
Weaknesses	3,794
Missing Topics	1,509
Application	3,661
Other Comments	1,370
Total	**13,851**

After iterative training and model evaluation, a simple front end was developed for the BERT classifier. This user interface takes the data files downloaded from edX and outputs the classified responses and associated confidence scores (Appendix III). This AI solution yields tremendous efficiency gains. The process that took about 30 hours of manual effort every trimester is executed in 5 minutes. Combined with closed-form survey questions and

[3] We use the BertForSequenceClassification implementation from the Hugging Face Transformers library with "bert-large-uncased" as the pre-trained language model.

course-level information, the model allows detailed analysis of learner experience and its driving factors, including the topic, language, length, difficulty of the course, and participant characteristics.

5. Model Performance

This section evaluates the text classifier's performance along two dimensions. First, we present evidence on the accuracy of the model's final prediction using various benchmarks. We also analyze differences in performance between languages. Second, we investigate the uncertainty around the final prediction: that is, how confident the model is when identifying the most likely label. This type of analysis helps conduct selective human review and can inform the design of further empirical analysis that uses machine-classified data.

5.1 Predictive Accuracy

The model's performance metrics on the test dataset, together with various benchmarks, are reported in Table 4. We use the standard measures of accuracy (share of correct matches), precision (Type I error) and recall (Type II error). The F1 score is the harmonic mean of precision and recall, better suited to unbalanced test dataset, as it balances the two types of errors. For all metrics, we calculate the weighted average according to class size in the training data. To put the model's performance in context, we evaluate it relative to a naïve baseline, an alternative keyword-based classification method, and the reliability of experienced human coders.

The naïve baseline simply uses the most frequent (mode) class in the training data as a prediction for all observations in the test data. For highly unbalanced datasets, the mode may be a useful benchmark when interpreting the accuracy measure. For example, for the Strengths question, 78 percent of responses in the training data were labelled Material. In this situation, even the naïve mode baseline achieves a high accuracy score, so measures of precision and recall become more important. The F1 scores for this experiment are reported in Table 4 under the Mode baseline.

Table 4. Model performance and various benchmarks

Survey Question	Model	Accuracy	F1	Precision	Recall
Strengths (n = 389)	BERT	0.905	0.823	0.827	0.844
	Baselines:				
	Mode	0.781	0.686	0.611	0.781
	Keyword		0.189		
	Human reliability		0.822		
Weaknesses (n = 422)	BERT	0.903	0.775	0.774	0.795
	Baselines:				
	Mode	0.359	0.189	0.129	0.359
	Keyword		0.223		
	Human reliability		0.762		
Missing Topics (n = 168)	BERT	0.929	0.811	0.820	0.805
	Baselines:				
	Mode	0.604	0.455	0.365	0.604
	Keyword		0.215		
	Human reliability		-		
Application (n= 407)	BERT	0.811	0.652	0.658	0.657
	Baselines:				
	Mode	0.442	0.271	0.195	0.442
	Keyword		0.095		
	Human reliability		-		
Other Comments (n = 153)	BERT	0.948	0.824	0.827	0.846
	Baselines:				
	Mode	0.537	0.375	0.288	0.537
	Keyword		0.328		
	Human reliability		-		

Note: The sample size of the test data is reported in parenthesis for each question.

Keyword-based methods can also serve as a useful benchmark. Our human experts provided carefully selected keywords for each label in all five question categories. For example, complaints about the length of the course typically include words like “length,” “time,” “hours,” “long,” “shorter,” etc. We performed stem processing to all responses and keywords to increase the likelihood of finding relevant matches. Then, we applied two approaches to evaluate the potential effectiveness of keyword search.

First, we calculated the share of responses that included any of the keywords associated with the human-coded label (Table 5). This strategy is biased towards finding matches because responses can contain keywords from multiple classes. Nevertheless, the match rate is far below the model’s accuracy for each question. This significant difference demonstrates that the finetuned BERT model learns more complicated patterns than simple word matching.

Table 5. Share of responses that contain at least one relevant keyword

Survey Question	Keyword Matches (in percent)
Strengths	32.7
Weaknesses	28.8
Missing Topics	14.4
Application	16.6
Other Comments	29.7

Second, we assigned a predicted label to each response based on the highest number of matching keywords from the expert-provided list. If there was no matching keyword or if equal matches were found for several labels, we randomly picked a class based on the distribution of classes in the training data. The resulting F1 scores are reported in Table 4 under the Keywords baseline, and they indicate clearly inferior performance.

As a final benchmark, we compare the model's classification accuracy with the inter-coder reliability of two human experts. Since the class definitions can be interpreted differently by coders in specific cases, it is important to acknowledge that our ground truth data has subjective elements. To get a sense of this uncertainty, we assigned a pair of human experts to independently code 100 responses for the Strengths question and another pair to code 100 responses for the Weaknesses question. We compared the alignment between the coders by calculating the average F1 score treating each of them as ground truth.[4] We report this agreement measure in Table 4 under the Human reliability baseline.

The results confirm that the model significantly outperforms simpler benchmarks, and its accuracy is on par with the reliability of human coders. All accuracy measures of the BERT model are above the Mode or Keyword benchmarks by a wide margin. More importantly, the degree of agreement between the model predictions and the collective coding of human experts is indistinguishable from the degree of agreement between two independent human coders.

Model Accuracy Across Languages

Most of our survey responses come from courses in English, but the training data contain comments from Spanish (7 percent) and French (5 percent) courses as well.[5] Since the number of training examples for non-English languages is much smaller, one could expect that the machine learning model will perform relatively poorly on these responses. However, the underlying LLM was pre-trained on a multilingual corpus, which means that the vector representations (embeddings) of semantically similar sentences in different languages are

[4]Because we use class size-weighted averages, the two F1 scores are slightly different. However, in practice the difference was negligible.

[5] We dropped a very low number of Portuguese responses from the analysis (below 0.2 percent). Occasionally, participants provide comments in a language different from the course language, but manual examination showed that this is rather rare.

expected to be close, especially for languages with similar vocabulary and linguistic patterns. This suggests that the mappings learned from English survey responses may be transferable to other languages. We find some evidence that this is the case.

For most survey questions the variation in model performance across languages is not statistically significant. Table 6 reports the share of correct predictions by question type and language, and the p-value of Pearson's Chi-squared test of equal accuracy. Although the number of non-English observations in the test sample is low, we observe similar accuracy levels. The exception is the Strength question category where the classification of French and Spanish responses is less successful than those in English. The statistical test confirms that this is the only significant difference between languages, while pooling all survey questions together yields almost identical predictive accuracy (p=0.64).

Table 6. Prediction accuracy across languages

Survey Question	Sample								
	Training				Test				
	En	Fr	Sp	All	En	Fr	Sp	All	P(χ^2)
Strengths									
Accuracy	0.91	0.92	0.81	0.90	0.92	0.75	0.79	0.91	0.01
# of responses	3,056	187	264	3,507	342	16	29	387	
Weaknesses									
Accuracy	0.88	0.81	0.83	0.87	0.91	0.90	0.78	0.90	0.13
# of responses	3,344	180	259	3,783	379	20	23	422	
Missing Topics									
Accuracy	0.90	0.92	0.93	0.91	0.93	0.92	1.00	0.93	0.69
# of responses	1,279	107	121	1,507	147	12	9	168	
Application									
Accuracy	0.81	0.84	0.80	0.81	0.81	0.83	0.86	0.81	0.73
# of responses	3,215	180	256	3,651	360	18	29	407	
Other Comments									
Accuracy	0.94	0.89	0.96	0.94	0.93	1.00	1.00	0.95	0.28
# of responses	1,161	85	121	1,367	117	18	17	152	
All									
Accuracy	0.88	0.87	0.85	0.87	0.89	0.88	0.86	0.89	0.64
# of responses	12,055	739	1,021	13,815	1,345	84	107	1,536	

Note: The last column reports p-values for Pearson's chi-squared test for the hypothesis that the share of correct responses and the course language is independent.

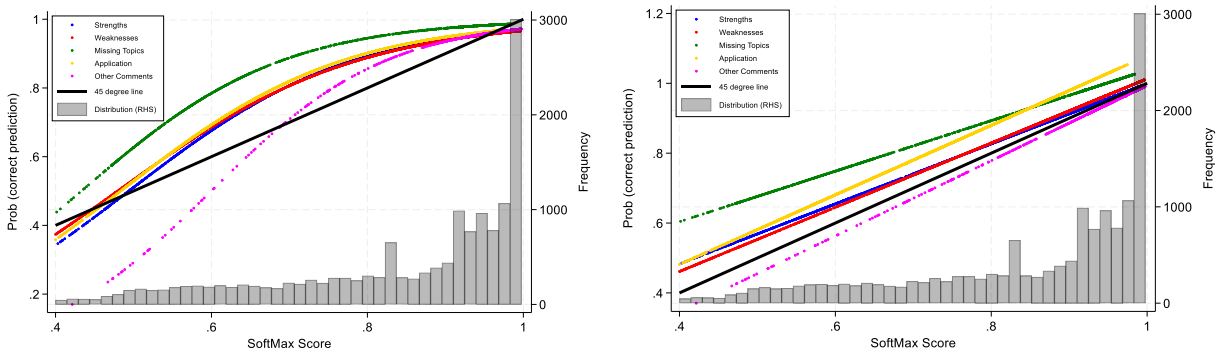
5.2 Uncertainty Around Model Predictions

In addition to the final prediction in a classification task, it is often useful to get a meaningful assessment of the relative probability of possible labels. For example, we may decide to validate some model predictions through human review. In this case, it makes sense to review the predictions with low confidence or when another label has only marginally smaller probability. Scrutinizing these cases can help identify and understand edge cases which is useful for improving the model. If the classification results are used in further empirical analysis or in decision-making, it could also be critical to have a sense of the uncertainty around model predictions.

To interpret the output of our BERT classifier as probabilities, we need to *calibrate* the model. By default, the output of most machine learning models cannot be interpreted in terms of probabilities. For example, assume that the final SoftMax layer outputs 0.9 for the most likely class of a survey response. Ideally, we would like this value to represent that if we were to take 100 responses with this predicted score, then in reality 90 of those 100 responses would be correctly classified. However, this interpretation is only valid in a *calibrated* model where the confidence scores reflect the true probability of the prediction being correct.

We found that our trained classifiers tend to be "underconfident" in predicting the most likely class. On Figure 2 we fitted logistical regressions (left panel) and linear probability models (right panel) using the BERT classifiers' highest SoftMax score to explain a binary variable that indicates whether the prediction was correct or not. The resulting curves can be interpreted as a mapping between the model's confidence and the true probabilities, assuming specific functional forms for this relationship. The 45-degree line represents a well-calibrated model where the SoftMax scores reflect true probabilities. Apart from the Other Comments category, the curves lie above the 45-degree line indicating under-confidence in the predicted class.

Figure 2. Model calibration: SoftMax scores vs. estimated probabilities



Note: The graphs present the predicted values from logit regressions (left) and linear probability models (right) with a binary dependent variable indicating whether the model prediction was correct. The explanatory variables are a constant and the SoftMax score for the predicted class. The grey histogram provides the frequency of top SoftMax scores pooled across all survey questions.

We can gain additional insights on the model's confidence and its calibration by examining the margin between the top two classes. This margin captures how distinctly the model distinguishes between its top choices, providing another angle to uncertainty. For instance, in automated systems where fallback to human decision-making is costly, such as manual review of survey responses, a smaller margin can serve as a heuristic for the usefulness of human intervention.

Table 7 demonstrates that for some question types the margin of the predicted class has information on whether the prediction is correct or not. The table reports results from logistic regressions that include a measure of the distance from the second most likely class. The

estimated coefficient on this margin is consistently positive and it is highly significant in three out of the five cases. This indicates that a more “distinctive” prediction is more likely to be correct. It is interesting to note that in a perfectly calibrated model the margin should not carry any extra information for the correctness of the predicted class, because the model’s confidence scores should already reflect this probability.

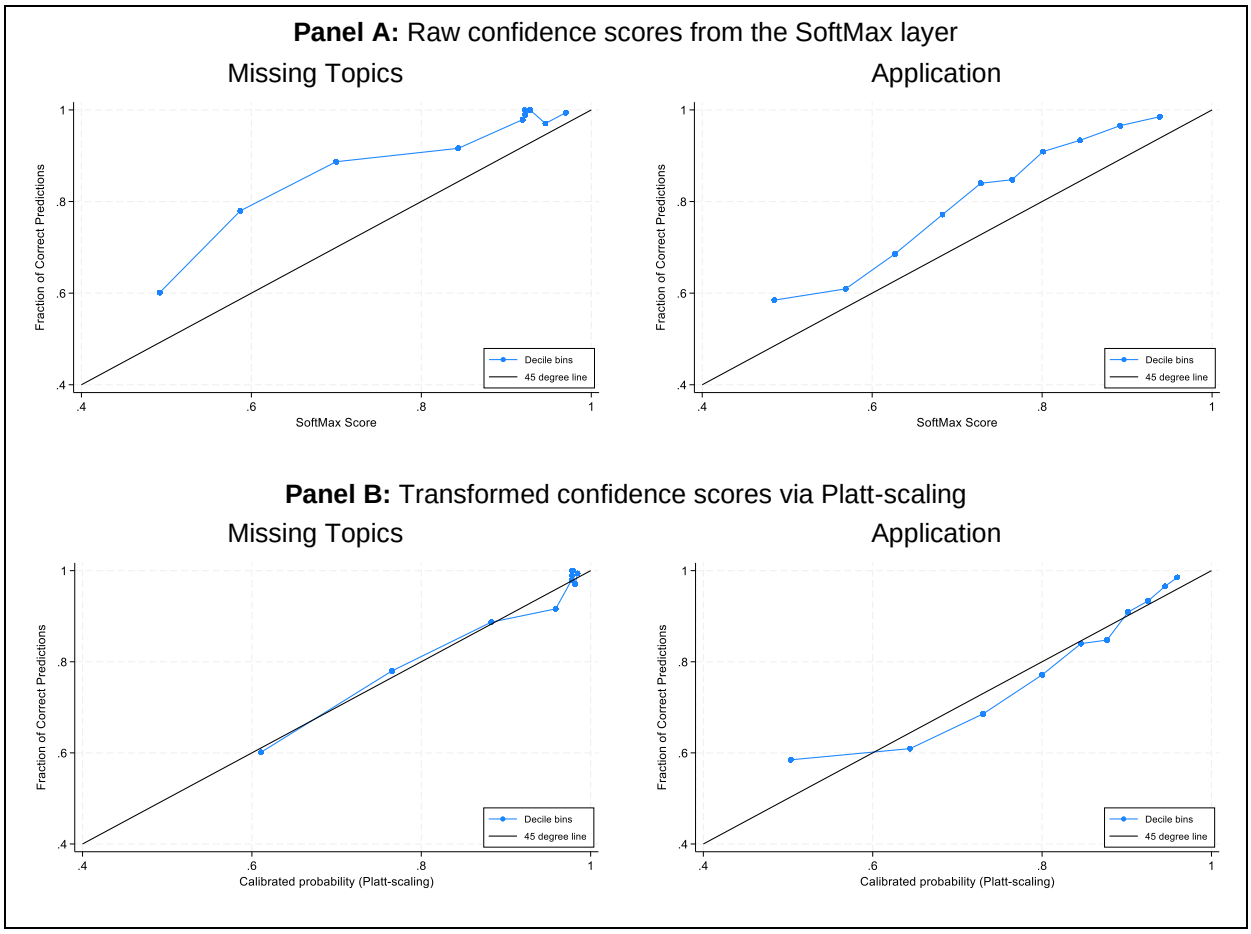
Table 7. Prediction confidence and distinctiveness

Prob(Correct Class)	Strengths	Weaknesses	Missing Topics	Application	Other Comments
Confidence	6.61***	6.64***	7.32***	7.05***	8.85***
	(0.32)	(0.27)	(0.64)	(0.34)	(0.92)
Margin	0.94***	2.45***	0.64	0.56***	0.06
	(0.29)	(0.32)	(0.42)	(0.20)	(0.73)
Constant	-3.65***	-4.67***	-3.33***	-3.54***	-5.34***
	(0.25)	(0.29)	(0.38)	(0.23)	(0.71)

Note: The table presents the results of logit regressions on a binary variable indicating whether the model prediction was correct. The explanatory variables are the SoftMax score for the predicted class (Confidence) and the difference between the highest and second highest SoftMax scores, normalized by the maximum possible range (Margin). Standard errors in parentheses. ***p < 0.01, **p < 0.05, *p < 0.1.

The reliability diagrams confirm the model’s bias for under-confidence. Reliability diagrams, or calibration curves, are a visual method to inspect whether a classification model is well-calibrated. It can be viewed as a non-parametric version of the probability models fitted in Figure 2. Panel A of Figure 3 shows the reliability diagrams for two question types. The graphs group the predicted scores into bins and for each bin they plot the average of the predicted scores on the x-axis and the empirical probabilities (fraction of correctly classified responses) on the y-axis. Here we used deciles to create the cutoff points for the 10 bins. As before, the resulting line can be compared to the 45-degree line to assess the model’s calibration. The dots tend to be above this line, indicating the model is under-predicting the true probabilities.

Figure 3. Reliability diagrams before and after Platt-scaling



Note: The figure illustrates model calibration for two survey questions. Panel A plots average SoftMax scores (x-axis) against the fraction of correct predictions (y-axis) in each decile bin. Panel B illustrates the calibration improvement from Platt-scaling by replacing the raw SoftMax scores on the x-axis with predicted probabilities from a logit regression (see notes under Figure 2).

We use Platt-scaling to improve the calibration of our machine learning models. The most common approach to calibration is to apply post-processing methods to the output of any classifier without re-training the machine learning model itself. These methods create a link between the raw confidence scores and the empirical probabilities in the data, with an aim to better align the two. Platt scaling (Platt, 1999) is a straightforward post-processing method which assumes a logistic relationship between the model predictions and the true probabilities, as in Figure 2. After applying this transformation to the model outputs, the reliability diagrams in Panel B of Figure 3 are much better aligned with the 45-degree line, indicating a better calibrated classifier.

Finally, we confirm that the calibrated prediction probabilities are, on average, higher for correct predictions than for incorrect predictions (Table 8). For correctly classified responses the models tend to have very high confidence with a 90 percent average probability across all

question types. In the case of incorrect predictions, the classifiers were much less confident in their output, indicating only 70 percent probability on average. This result provides reassurance that the models do not systematically misclassify responses with high confidence.

Table 8. Average calibrated probabilities for correct and incorrect predictions

Survey Question	Correct Predictions	Incorrect Predictions	Difference
Strengths	0.92	0.72	0.20
Weaknesses	0.90	0.68	0.22
Missing Topics	0.93	0.75	0.18
Application	0.84	0.71	0.13
Other Comments	0.95	0.81	0.14
All	**0.90**	**0.71**	**0.19**

Note: The table reports the average probability (after Platt-scaling) of the most likely class.

6. Illustrative Insights

Insights from qualitative learner feedback can help to improve and tailor online training to meet participant needs. The audience for the IMF's MOOCs is diverse: it includes officials from government agencies (e.g., ministries, central banks, statistical offices), students, private sector professionals, and other members of the public. While catering to all learners is challenging, targeted enhancements can be made by analyzing survey feedback. This section provides two examples using model-classified text feedback from over 10,000 survey respondents participating in one of the 205 IMFx courses offered between January 2022 and December 2023.

6.1 Course Length and Learner Satisfaction

We investigated what qualitative questions reveal about attitudes toward the length of online courses. Learning experts recommend shorter, focused, and modular content to provide flexibility and sustain engagement.[6] However, subject matter experts often believe that longer courses are necessary to provide sufficient substance to develop applicable skills. One solution is to break information into smaller chunks and design training as a series of self-contained modules. These design choices should be informed by an analysis of participant feedback.

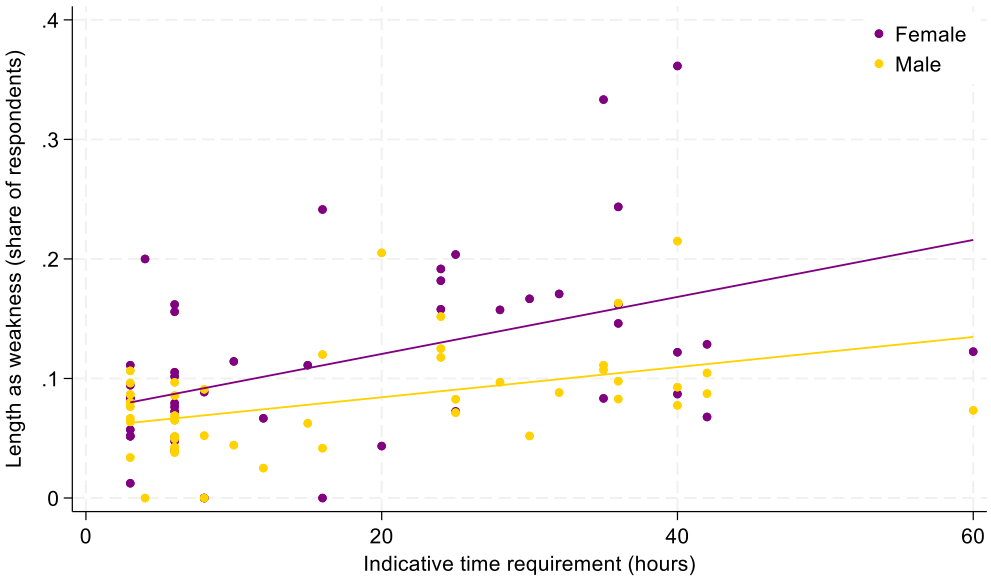
Survey responses indicate that course length influences learner satisfaction, especially among female participants. Figure 4 shows the relationship between courses' indicative time requirements[7] and the prevalence of critical comments about course length, brevity, content amount, or time requirements. There is a clear positive association, with longer courses

[6] For example, the IMF's Learning Channel on YouTube features more than 300 micro-learning videos that provide "bite-sized" (3-5 minutes) content for busy learners.

[7] Indicative total learning hours are estimated for the average participant based on the amount of text, videos, exercises, and assessments in the course.

receiving more complaints about length. Furthermore, female participants are generally more sensitive to time requirements when asked about the salient weaknesses of a MOOC.

Figure 4. Course length as a perceived weakness



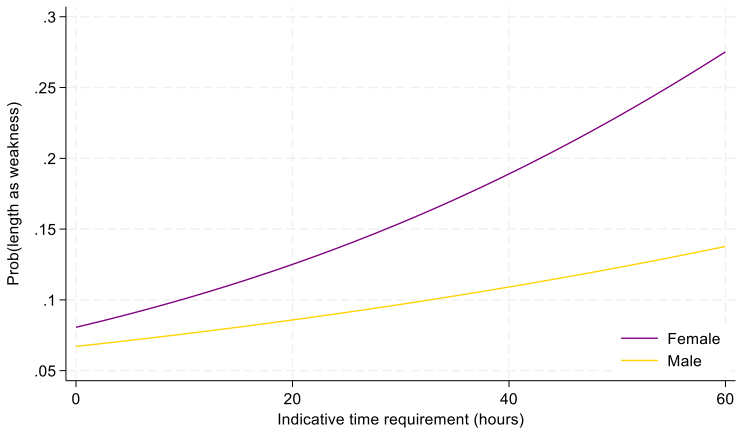
Note: The figure plots the share of survey respondents indicating "length" as a weakness (y-axis) against the estimated total learning hours of the course (x-axis). Respondents are grouped by gender, and each dot represents a course-gender pair. Learning hours are estimated by the course developers based on the amount of text, videos, exercises, and assessments.

These gender differences among courses of different length are also statistically significant according to a logit regression estimated at the level of individual survey respondents (Table 9). Overall, the results suggest that shorter and modular courses can increase learner satisfaction and facilitate more equal access to training across genders.

Table 9. Probability of negative comment on course length: Logit regression

Variable	Length as Weakness
Course Length	0.0133***
	(0.0028)
Female	0.1980*
	(0.1116)
Female × Course Length	0.0111***
	(0.0042)
Constant	-2.6318***
	(0.0749)
Observations	9,780

Implied probability estimates by gender

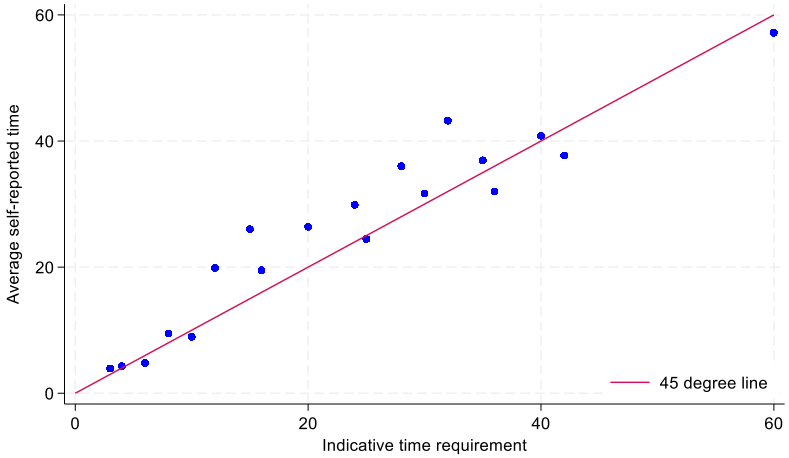


Note: The table presents logit regression results on a binary variable that indicates whether the respondent mentioned overly long learning content as a weakness. The explanatory variables are the estimated course length (in hours), a female dummy, and the interaction term. Standard errors in parentheses. ***p < 0.01, **p < 0.05, *p < 0.1. The graph on the right plots the corresponding probability curves by gender.

6.2 Self-Reported Learning Time and Language Barriers

We can also gain insights from analyzing self-reported learning time and its correlates. The indicative time requirements mentioned above are estimated for the "representative" learner. Accordingly, Figure 5 shows that these estimates are positively related to the average self-reported learning times across survey respondents. However, the actual time each participant spends on completing a course can vary significantly based on individual factors such as educational background, perceived usefulness of the course, or language barriers.

Figure 5. Estimated and self-reported time for course completion



Note: The figure plots the relationship between the indicative time requirement estimated by the course developers (x-axis) and the average

self-reported time spent on the same courses (y-axis). Numerical self-reported times are imputed using the top of the time brackets in the survey, using 8 hours for the uncapped 6+ hours bracket.

There is suggestive evidence that language can be a constraining factor, disproportionately impacting learners from certain regions. For example, participants who spend more time completing a course tend to flag language issues more frequently in their written comments. This pattern is more pronounced for courses delivered in English. Participants citing language constraints are more likely to come from Sub-Saharan Africa and the Western Hemisphere, where dominant regional languages facilitate communication. (Table 10).

Table 10. Critical comments about language issues (percent of respondents)

Course Language	Respondent's Self-reported Study Time (hours/week)				Region of Respondent's Country					Total
	< 2	3-4	5-6	6+	AFR	APD	EUR	MCD	WHD	
All languages	1.92	3.07	3.71	4.34	4.38	1.88	1.66	2.18	3.76	3.18
English	2.85	4.26	4.68	6.03	5.38	3.00	2.27	3.74	5.96	4.30

Note: The table shows the share of answers indicating 'language issues' as a major weakness of the course, among respondents that identified any weakness. The data are broken down by self-reported study time and region. Regions follow the IMF's organizational structure of Area Departments: Sub-Saharan Africa (AFR), Asia and Pacific (APD), Europe (EUR), Middle East and Central Asia (MCD), and Western Hemisphere (WHD).

Additional survey results support the IMF's efforts to develop language adaptations for its training courses, ensuring equal support for all member countries. Many participants reporting language difficulties also note that they intend to apply their new knowledge and skills in their jobs, expecting it to contribute to their institutions' key mandates and their own career prospects. Such responses appear in 51 percent of surveys from Africa and 56 percent from Latin America, compared to 35 percent in Europe. This highlights the importance of offering IMF economics and finance training in languages other than English.

These results on attitudes toward course length by gender and language constraints by region illustrate how qualitative learner feedback can inform evidence-based improvements and tailoring of the IMF's online learning program. The evaluation framework and AI-assisted approach presented in this paper greatly facilitate a detailed analysis of learner perceptions and their relation to learning outcomes across various dimensions, such as age, educational and employment background, and work experience.

7. Conclusion

This paper demonstrated that AI solutions can drastically improve the efficiency of learning analytics, especially the processing of qualitative feedback. The IMF Online Learning Program collects feedback from every active MOOC participant on their perception of the strengths, weaknesses, and applicability of IMF training. A large portion of the survey comprises open-ended questions, resulting in over 7,000 responses every trimester. It used to require

substantial human labor to read and manually categorize each comment following established coding guidelines. After finetuning a pre-trained open-source LLM (BERT) on labelled data, the classification process was reduced to a matter of minutes.

We demonstrated that the model's accuracy is on par with human annotators. Moreover, the classifier trained mostly on English responses was able to deliver largely consistent performance on other languages. This suggests that pre-trained language models can be effectively fine-tuned to multilingual use cases even if task-specific training data for certain languages is sparse. We also analyzed the uncertainty around the model predictions and adjusted the raw confidence scores to better reflect the likelihood of a correctly predicted class.

The ability to process a vast amount of qualitative feedback helps gain deeper understanding of the learner experience. For example, preliminary analysis shows that IMF learners generally prefer shorter learning content, but the disutility associated with lengthy courses varies significantly by gender. Survey responses also indicate that language barriers are an important consideration, and course adaptations to languages other than English help reach more learners from all 190 members of the IMF. These results illustrate how unstructured feedback can be useful for tailoring the IMF's capacity development programs to better meet the diverse needs of its global audience, thus enhancing the effectiveness of economic policy training.

Of course, classifying qualitative responses into pre-determined categories significantly compresses the information content of the underlying data. In ongoing work, IMF learning experts utilize state-of-the-art generative AI models to allow the unstructured text data to speak more freely. For example, LLMs help transform, in an unsupervised manner, large volumes of learner feedback into summaries that can be read by course developers. Likewise, sentiment analysis and topic clustering are now viable avenues to identify specific aspects of each course that are relevant to learners. AI-assisted analysis of learner feedback opens a new chapter in learner-centered design.

References

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kirkpatrick, D. L. (1976). Evaluation of training. *In R. L. Craig (Ed.), Training and development handbook: A guide to human resource development. New York: McGraw-Hill.*, (2nd ed., pp. 301–319).

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3), 61-74.

Tzeng, J.-W., Lee, C., Huang, N., Huang, H., & Lai, C. (2022). MOOC Evaluation System Based on Deep Learning. *The International Review of Research in Open and Distributed Learning*, 23 (1):21-40.

Appendix I. Post-Course Survey Questions

1. What country are you from? Countries are listed alphabetically. If you do not see your country in the list (Type box).
    - List of Countries
2. Which of the following best characterizes your occupation?
    - Student
    - Government Official (working at the Ministry of Finance, Central Bank, or other government agency)
    - Academic/Professor
    - Economist/Analyst (non-government)
    - Researcher
    - Engineer
    - Journalist/Media professional
3. Please indicate your number of years of relevant work experience.
    - Less than a year
    - 1–5 years
    - 6–10 years
    - More than 10 years
4. Please select your gender.
    - Male
    - Female
    - Other
    - I prefer not to answer
5. Please select your age group.
    - Under 25
    - 25–30
    - 31–35
    - 36–45
    - Over 45
6. What was your main goal(s) for taking this course? Select all that apply.
    - Personal challenge
    - Increase knowledge and skills
    - Social community of the course and networking
    - Interest in topic
    - Employment/job advancement opportunities
    - To apply newly gained knowledge/skills to my work
    - To access learning opportunities and materials not otherwise available to me

- Curiosity about the topics
- To earn a certificate

6.1 Did the course meet your expectations? If no, why do you feel your expectations were not met?

- Yes
- No, because _____

7. Did you face any challenges completing parts of this course?
    - Yes
    - No

    7.1 What were your barriers to completion?
    - I had problems with internet connectivity and/or video streaming.
    - I did not have enough time due to work or other commitments.
    - The course was too difficult.
    - There was too much material in the course.
    - The course was not offered in my preferred language.

    7.2 What problems with internet connectivity and/or video streaming did you face?
    - I had internet connectivity issues.
    - I could not stream the videos and had to download them.
    - I could not stream or download the videos and had to rely on transcripts.
8. What were the strengths of the course?
   (Open-ended answer)
9. What were the weaknesses of the course?
   (Open-ended answer)
10. On average, how much time did you spend on this course per week?
    - Less than 2 hours
    - 3–4 hours
    - 5–6 hours
    - More than 6 hours
11. The following course elements were essential for my learning:

	Strongly agree	Agree	Neutral	Disagree	Strongly disagree	Not applicable
Videos and transcripts						
Graded assessments						
Ungraded activities / exercises						

Text and reading materials (handouts and resources)						
Discussion forums						
Visuals and interactives (such as infographics, charts, illustrations.						

12. Please indicate the extent to which you agree with the following statements:

	Strongly agree	Agree	Neutral	Disagree	Strongly disagree	Not applicable
The module's learning objectives are clearly defined (i.e., you knew at the beginning of the module what you would learn in the module).						
Module content sufficiently covered the module objectives.						
Module content was relevant and easy to understand.						
Module content was well-structured and presented in a						

clear and logical way.						

13. Are there topics that were not covered or not explained well enough that you think should be included? Please describe. (Open-ended answer)
14. Please indicate the extent to which you agree with the following statements:

	Strongly agree	Agree	Neutral	Disagree	Strongly disagree	Not applicable
The instructions for the graded assessments were easy to understand.						
The graded assessments adequately covered and tested the module content (i.e. you could answer the questions based on the content).						
The graded assessments provide quality feedback where required.						
The graded assessments helped me better grasp the module content.						

15. Please indicate the extent to which you agree with the following statements:

	Strongly agree	Agree	Neutral	Disagree	Strongly disagree	Not applicable
The platform is well organized,						

easy to navigate and easy to use.						
The platform is visually appealing and functionally consistent.						
The course information is communicated clearly on the platform.						
The platform has enough information about technical support.						

16. Please indicate the extent to which you agree with the following statements:

	Strongly agree	Agree	Neutral	Disagree	Strongly disagree	Not applicable
The IMF course team adequately responded to your questions and supported discussions in an effective manner.						
The IMF course team provided support in technical matters related to the course (for example, ease of registration).						

17. Please indicate the extent to which you agree with the following statements:

	Strongly agree	Agree	Neutral	Disagree	Strongly disagree	Not applicable
I learned new knowledge and skills from this course.						
The content of this course was relevant to my job.						
The investment (time and other resources) in attending this course was worthwhile.						
I would recommend this course to others.						
Overall, I was satisfied with this course.						

17.1 How will you apply the learning from this course to your job? (Open-ended answer)

18. After taking this course, I have a better understanding of the International Monetary Fund (IMF) and its work.
    - 5 (Strongly agree)
    - 4 (Agree)
    - 3 (Neutral)
    - 2 (Disagree)
    - 1 (Strongly disagree)
19. After taking this course, I have a better understanding of the economic policy issues in my country and/or globally.
    - 5 (Strongly agree)
    - 4 (Agree)
    - 3 (Neutral)
    - 2 (Disagree)

- 1 (Strongly disagree)

20. Please share any additional comments you may have around this course. (Open-ended answer)

Appendix II. Question-Specific Codebook for Response Categorization

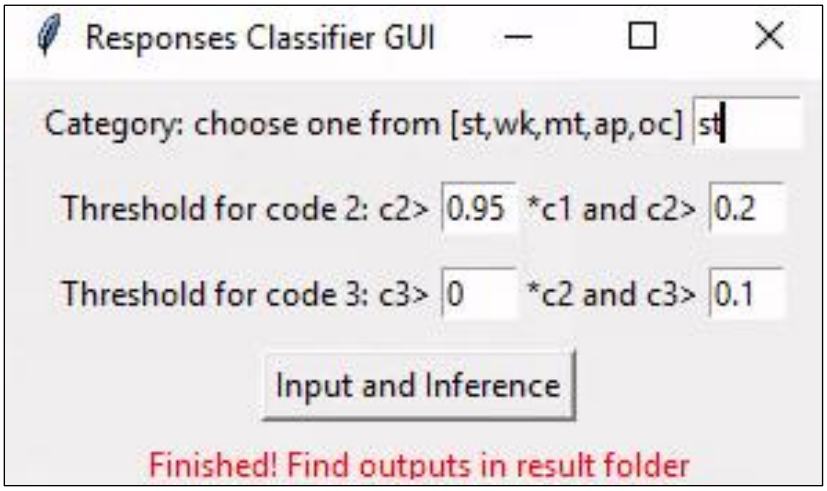
Question-specific Codebooks		
What were the strengths of the course?		
Codes/Themes	**Description**	***Example quotes***
Strength - Material	Positive comments pertaining to the quality or organization of material, course concepts, explanations, etc., to the visual display or delivery of content (e.g., video, activity, exercises, etc.)	*It is very well structured and organized. The information is very clear and easy to understand.*
Strength - Practicality	Positive comments pertaining to the practicality and relevancy of content; and real-life applicability in country/work context	*Examples from several jurisdictions gave a holistic view of the subject and the best practices followed by different countries can be taken as a model for implementation by others.*
Strength - Expertise	Positive comments pertaining to the expertise, experience, and aptitude of the instructors/trainers	*The Lecturers, they are well versed in topics discussed throughout the module and very effective delivering the main learning objective of the course.*
Strength - Convenience	Positive comments related to ease, affordability, accessibility of taking course.	*- I was able to study at the comfort of my home and at a convenient time.* *- I did not have pay to study the course.* *- Wider data study material.*
Strength - Other	Other comments not captured in the codes/descriptions above	
Strength - Indecipherable	Comments that were vague, unintelligible or could not easily be interpreted	

What were the weaknesses of the course?		
Codes/Themes	**Description**	***Example quotes***
Weakness - Material	Critical comments pertaining to the quality or organization of material, course concepts, explanations, etc. AND/OR any inaccuracies or issues with content or grades.	*Strategic management not covered in broader perspective.*
Weakness - Practicality	Critical comments pertaining to the lack of practicality of content; and real-life applicability in country/work context	*maybe more practical examples would be useful*
Weakness - Technical	Critical comments pertaining to any technical issues (e.g., issues with downloading content, reviewing videos, internet connectivity, etc.)	*The video scripts could not be easily downloaded*
Weakness - Length	Critical comments pertaining to course length, brevity, amount of content, or time requirements	*Some modules are long, therefore it is not easy to keep focusing.*
Weakness - Communication	Critical comments pertaining to communication or interaction with participants and/or experts	*The weak point in the course is the lack of direct communication with the lecturer who is presenting the course, and thus the inability to make use of it quickly and ask questions in a timely manner.*
Weakness - Language and Culture	Critical comments pertaining to language issues, language offerings and/or lack of representation.	*Not too many videos with speakers from the Asian region.*
Weakness - Certificate	Critical comments pertaining to the difficulty in obtaining a certificate, requirement to pay for the certificate, and/or any other comments specifically related to the course certificate.	*I wish we could all get a certificate regardless of whether we paid the or not.*
Weakness - None	Comments stating there being no weaknesses/improvements needed for the course	*There are no weaknesses, I was completely satisfied with the course.*

Weakness - Other	Other comments not captured in the codes/descriptions above	
Weakness-Indecipherable	Comments that were vague, unintelligible or could not easily be interpreted	
Are there topics that were not covered or not explained well enough that you think should be included? Please describe.		
Codes/Themes	**Description**	***Example quotes***
Missing Topics - Specific	Comments with specific topic suggestions	*Review of strategic failures of some Tax Administration strategies*
Missing Topics - None	Comments stating there being no missing nor ill-explained topics	*All of the topics were well explained.*
Missing Topics - Other	Comments not relating to topic suggestion	*I would add more lectures*
How will you apply the learning from this course to your job?		
Codes/Themes	**Description**	***Example quotes***
Application - General	Broad or general comments that learning from course can/will be applied in job and/or is relevant to job	*It will be applied to the everyday learning involved with my job*
Application - Specific topics	Comments that refer to specific topics, without explicitly mentioning how it'll be applied in their job.	*To better understand how macroeconomics work.*
Application - Work	Comments pertaining to how learning contributes towards key organizational mandate or strategy, or how learning contributes to job support, improvement, or performance	*As a planner in the Ministry of Youth and Culture, I will integrate in the Ministry's plan how the young generation can be skilled about financial inclusion and literacy.*
Application - Opportunities	Comments pertaining to having better work opportunities or future career and educational prospects	*As a business student the course generally increased my knowledge in my field of study.*
Application - Unknown	Comments pertaining to not knowing yet how learning will be applied	*I'm not sure yet that I'll apply this to my job but it helps me somehow*
Application - Other	Other comments not captured in the codes/descriptions above	

Application - Indecipherable	Comments that were vague, unintelligible or could not easily be interpreted	
Please share any additional comments you may have around this course.		
Codes/Themes	**Description**	***Example quotes***
Comments - Praise	Comments expressing gratitude or positive praise for course	*Congratulations to the IMF team, you are wonderful teachers and visual designers.*
Comments - Technical Issues	Comments pertaining to any technical issues (e.g., issues with downloading content, reviewing videos, incorrect grading, etc.)	*Question 3 on Unit 10 assessment questions, failed to understand what it required. My choice was marked wrong, and I'm not sure of the correct answer. Please explain what it required, may you send response directly to my email.*
Comments - Suggestions	Comments pertaining to specific suggestions (e.g., topics, target audience, course structure, etc.)	*Include developments on the derivative markets.*
Comments - None	Any comments indicating they have no comment or nothing else to express	*Congratulations to the IMF team, you are wonderful teachers and visual designers.*
Comments - Other	Other comments not captured in the codes/descriptions above	
Comments - Indecipherable	Comments that were vague, unintelligible or could not easily be interpreted	

Appendix III. User Interface of the BERT Classifier



Note: The five qualitative survey questions are identified with two-letter codes (st-Strengths, wk-Weaknesses, mt-Missing Topics, ap-Application, oc-Other Comments). To derive secondary or tertiary classifications, relative and absolute threshold rules can be set on the model's SoftMax confidence scores.

PUBLICATIONS

Enhancing IMF Economics Training: AI-Powered Analysis of Qualitative Learner Feedback

Working Paper No. WP/2024/166