# Large Language Model-based Pipeline for Item Difficulty and Response Time Estimation for Educational Assessments

**Hariram Veeramani** [1]
hariram@ucla.edu

**Surendrabikram Thapa** [2]
surendrabikram@vt.edu

**Natarajan Balaji Shankar** [1]
balaji1312@ucla.edu

**Abeer Alwan** [1]
alwan@ee.ucla.edu

[1] **University of California, Los Angeles**

[2] **Virginia Tech**

## Abstract

This work presents a novel framework for the automated prediction of item difficulty and response time within educational assessments. Utilizing data from the BEA 2024 Shared Task, we integrate Named Entity Recognition, Semantic Role Labeling, and linguistic features to prompt a Large Language Model (LLM). Our best approach achieves an RMSE of 0.308 for item difficulty and 27.474 for response time prediction, improving on the provided baseline. The framework's adaptability is demonstrated on audio recordings of 3rd-8th graders from the Atlanta, Georgia area responding to the Test of Narrative Language. These results highlight the framework's potential to enhance test development efficiency.

## 1 Introduction

Standardized tests are essential tools for evaluating knowledge and ability for academic and professional purposes, and thus must be rigorously designed and meet stringent criteria. Key aspects include diverse item difficulty for comprehensive skill evaluation and appropriate response time allocation – insufficient time compromises fairness, while excessive time leads to inefficiencies (Huggins-Manley et al., 2022). Traditionally, item difficulty and response time optimization have relied on *pretesting*, where new items are embedded in live exams. However, this labor-intensive process limits the number of new items and introduces security risks through potential overuse (Settles et al., 2020). In high-stakes examinations like the United States Medical Licensing Examination (USMLE) [1], these challenges necessitate the exploration of alternative approaches for more secure and efficient test design.

In response to these challenges, recent research explores automated prediction using the text of items themselves. This approach promises to streamline test development, enhance exam fairness, and mitigate security risks associated with item overexposure. The automated prediction of *item difficulty* and *item response time* shared task at the 19th BEA Workshop aims to address this gap (Yaneva et al., 2024). Advancements in Large Language Models (LLMs), trained on massive text corpora, hold significant potential for discerning language patterns indicative of item difficulty and response time. This paper outlines our methodology for automated prediction of these characteristics, leveraging named entity recognition, semantic role labeling, and LLMs. We further evaluate the framework's validity across modalities by applying it to a dataset of children's oral responses to the Test of Narrative Language. Our approach integrates these technologies to analyze the complexities of test item texts, aiming to accurately predict both difficulty level and response time.

## 2 Related Works

In recent years, the prediction of item difficulty and response time has garnered significant attention in the field of educational assessment research. Prior work in this field employed techniques rooted in classical test theory and item response theory. More recently, the advent of sophisticated machine learning approaches has enabled novel methods for modeling these parameters (Yaneva et al., 2020, 2021).

In Lin et al. (2019) an LSTM-based method for Chinese reading comprehension tests was proposed. It achieved high accuracy utilizing word embeddings and text correlation networks. Similarly, Hochreiter and Schmidhuber (1997) employed word embeddings within a semantic space to analyze relationships between multiple-choice test components, finding correlations between semantic similarity and item difficulty. Research on item difficulty prediction in medical exams has also advanced significantly with Qiu et al. (2019) introducing the Document enhanced Attention based

[1] https://www.usmle.org/

neural Network (DAN) framework using semantic relevance and similarity for difficulty assessment. Ha et al. (2019) further demonstrated that embeddings and linguistic features extracted from test documents outperform simple text complexity measures in predicting construct-relevant difficulty in MCQs. Baldwin et al. (2021) incorporated item response time prediction, emphasizing the importance of understanding how test-takers interact with items. In a similar vein, Xue et al. (2020) found transfer learning beneficial for USMLE item difficulty prediction, suggesting stems alone are optimal for difficulty, while the entire question benefits response time prediction. Despite these advancements, the joint prediction of item difficulty and response time remains under-explored, motivating our proposed technique designed to address this gap.

## 3 Data

We evaluate our framework primarily on the 2024 BEA shared task dataset constructed from the USMLE. As an auxiliary task, we also test its validity on the Test of Narrative Language.

### 3.1 Shared Task Description

The BEA 2024 Shared Task focuses on the automated prediction of item difficulty and item response time for standardized exams, with an emphasis on the USMLE. This task seeks to enhance the fairness and validity of standardized exams by streamlining the estimation of item characteristics, reducing the reliance on extensive pretesting. The shared task comprises two tracks:

- **Track 1: Item Difficulty Prediction** predicts the difficulty level of test items using item text and relevant metadata.
- **Track 2: Item Response Time Prediction** predicts the average time required by test-takers to answer an item utilizing item text and metadata.

#### 3.1.1 Dataset

This task utilizes a dataset of 667 retired questions from USMLE Steps 1, 2 CK (Clinical Knowledge), and 3. These items cover a range of medical knowledge and were authored by experts. The dataset includes the following components for each item:

- **Item Text (Stem)**: Clinical scenario/question presented.
- **Answer Options**: Response choices (A-J, some items may have fewer options).
- **Correct Answer (Key)**: Correct response letter.
- **Item Type**: Indicates text-only or image-based (images not provided).
- **Exam Step**: Which USMLE step the item belongs to.
- **Item Difficulty**: Numerical difficulty value (higher=more difficult).
- **Response Time**: Average response time (seconds) from live exam data.

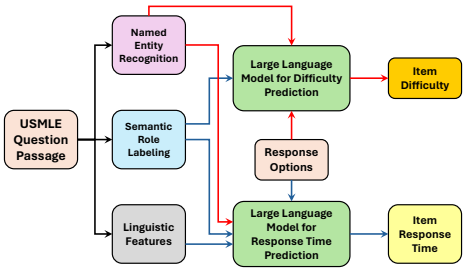


Figure 1: The proposed framework for item difficulty and response time prediction

### 3.2 Test of Narrative Language (TNL)

This work also uses audio recordings of 185 3rd-8th grade students from the Atlanta, Georgia, area as they perform the “Test of Narrative Language (TNL)" assessment (data collected in Fisher et al. (2019)). In "Task 2 - Picture Description" in the TNL, the children were shown an image containing a character and several elements to describe. The students were then asked to tell a story about the image, making their story as complete as possible. Each child’s response to the prompt was recorded, and each child, on average, took about 3 minutes to complete their story. Each child’s assessment was administered and audio recorded by a trained member of the project staff according to the TNL protocols. Recordings were then independently scored by two speech-language pathologists. If disagreements occurred in scoring, the two scorers reviewed the audio and discussed differences to reach a consensus. Each child’s score was an integer value between 0 and the total number of test keywords. Recordings were taken at the child’s school. Audio was recorded in stereo at a sampling rate of 48kHz. All recordings were resampled to mono audio with a sampling rate of 16kHz for experimentation.

### 3.3 Evaluation

The evaluation for both tracks of the shared task, and the Test of Narrative Language, is based on the Root Mean Squared Error (RMSE) metric, offering an objective measure of the accuracy of predictions made by the proposed pipeline.

## 4 Methodology

### 4.1 Item Difficulty Prediction

Our item difficulty prediction methodology integrates multiple advanced NLP techniques to enhance the precision of our predictions. We outline our approach in three main steps: Named Entity Recognition (NER), Semantic Role Labeling (SRL), and the final difficulty prediction.

#### 4.1.1 Named Entity Recognition

For Named Entity Recognition (NER), we employ a dual-model strategy using both Longformer (Beltagy et al., 2020) and a choice between three Large Language Models (LLMs), Mistral-7B (Jiang et al., 2023), Llama-7B (Touvron et al., 2023), or Gemma-7B (Team et al., 2023) to extract named entities from the entire question text. For LLMs, we provide input as the question and specifically prompt them as follows: `Understand the input sentence and annotate the named entities from the Input Context`. This process can be represented as follows:

$$NER_{longformer} = Extract_{longformer}(Question)$$

$$NER_{LLM} = Extract_{LLM}(Question)$$

$$NER_{Union} = NER_{longformer} + NER_{LLM}$$

This process yields three combinations of NER outputs, one for each LLM, by taking the union of NERs extracted from Longformer and the selected LLM. This approach ensures a more comprehensive and accurate set of named entities by leveraging the strengths of each model.

#### 4.1.2 Semantic Role Labeling

Following Named Entity Recognition (NER), we employ Semantic Role Labeling (SRL) utilizing both AllenNLP SRL Model (BERT Variant) (Gardner et al., 2018) and the selected LLM. SRL functions to identify semantic relationships within the sentence, attributing roles to entities according to their contextual significance. For SRL, the process is analogous to that of NER, employing both AllenNLP SRL and LLM to analyze the text. This process can be represented as:

$$SRL_{BERT} = Analyze_{BERT}(Question)$$

$$SRL_{LLM} = Analyze_{LLM}(Question)$$

$$SRL_{Union} = SRL_{BERT} + SRL_{LLM}$$

For LLMs to generate SRL, we provide the question and specifically prompt them as follows: `Understand the input context, which consists of the input sentence and the associated named entities, then annotate the semantic role labels of the input context`. This step deepens our pipeline's comprehension of the question's structure and content, thus facilitating more precise predictions of item difficulty.

#### 4.1.3 Difficulty Prediction

Finally, we integrate NER and SRL outputs to predict item difficulty. The LLM is prompted to estimate difficulty based on the complexity of relating the correct answer to the identified entities and their semantic roles.

$$Difficulty = Predict_{LLM}(NER_{union}, SRL_{output})$$

We prompt the LLMs by providing input as the question, NER, SRL, answer, and the prompt as: `For answer option set, understand the input context consisting of an input sentence, a collection of named entities and semantic role information, summarize the association with the` $i$`th answer option. Depending on the difficulty level of the linkages between input context and [answer options], assign the input context a score in the range of 0 to 1.4.` This approach leverages the LLM's language understanding capabilities, enriched by the detailed insights from NER and SRL, enabling a more informed prediction of item difficulty.

### 4.2 Item Response Time Prediction

For item response time prediction, as shown in Fig. 1, we use linguistic features in addition to the NER and SRL features. For NER and SRL features, we follow the same steps as for the difficulty prediction subtask.

#### 4.2.1 Linguistic Features from Question

For item response time prediction, we begin by extracting a subset of the 255 hand-crafted linguistic features from LingFeat (Lee et al., 2021). Among all features, we only take numerical and syntactic features. The LLM is then prompted to estimate response time using the question, NER, SRL, answer, linguistic features and the following prompt: `For answer option set, understand the input context consisting of an input sentence, a collection of named entities, semantic role information, Concatenate lingfeat numerical and syntactic features to summarize the association with the` $i$`th answer option. Depending on the exhaustiveness of the linkages`

`demonstrated with input context and [answer options], assign the entire input context a response time in the range of 25.0 to 230.00. Higher value would indicate longer response time and higher exhaustiveness.`

Both item difficulty and response time predictions are performed utilizing the Langchain library (LangChain, 2024) for chaining API calls to the LLM models in different stages, as well as to post-process the outputs after each stage.

### 4.3 Difficulty and Item Response Time for Oral Assessments

For recordings from the Test of Narrative Language, we first generate Automatic Speech Recognition (ASR) transcripts using the Whisper model (Radford et al., 2023) as in Veeramani et al. (2023). Prior studies on literacy development (MEIERS and MENDELOVITS, 2016), highlight the role played by item response theory in measuring narrative proficiency and literacy among school children. Item difficulty is assessed utilizing two metrics: 1) Transcription Word Accuracy: Calculated as described in Oliveira et al. (2022). 2) Proportion of Correct Responses: We measure the percentage of children who correctly answer a test item, providing an additional indicator of item difficulty. To model item response time, we analyze the time taken by disfluencies exhibited by speakers during the assessment. These disfluencies, classified as filled pauses (FP), partial words (PW), repetitions (RP), revisions (RV), and restarts (RS), are extracted using models pretrained on the Switchboard corpus (Godfrey et al., 1992) following the methodology outlined in Romana et al. (2023).

### 4.4 System Design

As per the BEA 2024 Shared Task guidelines, we attempt the item difficulty and response time prediction task with three separate pipelines. The runs use identical pipelines and differ only in the choice of the LLM, with Run 1 using Llama2-7B, Run 2 Mistral-7B, and Run 3 using Gemma-7B.

## 5 Results and Discussion

### 5.1 BEA 2024 Shared Task

We first report our results on the BEA 2024 shared task, comparing the baseline with three variants of our proposed pipeline. Our findings (Table 1) demonstrate that prompting Llama2-7B (Run 1) for simultaneous prediction of response time and item difficulty outperforms the DummyRegressor baseline and other LLMs. Similarly, Gemma-7B (Run 3) also exceeds the baseline. We did not perform any ablation studies. However, these results align with prior research on LLM reasoning capabilities (Johnson et al., 2023), supporting the value of our chosen handcrafted features as supplementary input.

Table 1: RMSE values of different runs on the BEA 2024 Shared Task. Numbers in bold represent best results

| Method | Item Difficulty | Response Time |
|---|---|---|
| Baseline | 0.311 | 31.68 |
| Run 1 | **0.308** | **27.474** |
| Run 2 | 0.329 | 31.962 |
| Run 3 | **0.308** | 28.191 |

### 5.2 Test of Narrative Language

Table 2: RMSE values from different runs on the TNL - Task 2 data

| Method | Item Difficulty | Response Time |
|---|---|---|
| Baseline | 4.043 | 4.941 |
| Run 1 | 2.162 | 2.038 |
| Run 2 | 2.0578 | 2.0237 |
| Run 3 | **2.007** | **2.022** |

As shown in Table 2, Gemma-7B (Run 3) demonstrates superior performance in predicting both response time and item difficulty, exceeding the baseline and other LLMs. Similar to the results seen in the BEA 2024 Shared Task, the inclusion of numerical, lexical, and linguistic features likely aides in understanding the complex interplay of within the input and the syntactic/semantic relationships needed to correctly identify the answer.

### 5.3 Conclusion

This paper introduces a novel framework for automating the prediction of item difficulty and response time, a crucial aspect of educational assessment design. Our system, utilizing Named Entity Recognition, Semantic Role Labeling, and linguistic features in conjunction with a Large Language Model, demonstrates promising performance on the BEA 2024 Shared Task data, achieving RMSE values of 0.308 (item difficulty), and 27.474 (item response time). The framework's adaptability was further evidenced by its successful application to audio recordings from the Test of Narrative Language, highlighting the potential of this approach to streamline test development.

## Limitations

While promising, our framework has limitations: **Model Interpretability:** The LLM's decision-making process lacks transparency. Future research should explore methods for increasing interpretability and providing human-understandable explanations.

**Linguistic Feature Scope:** Our current implementation analyzes a specific set of linguistic features for item response time prediction. It is possible that additional features, such as specific domain-related vocabulary, could further enhance prediction accuracy.

**Domain Specificity:** While our framework shows promise for both written and oral assessments, its performance may vary across different domains and test formats. Further research is needed to evaluate and potentially adapt the framework for optimal performance in specific testing contexts.

Addressing these limitations will improve the framework's accuracy, efficiency, and fairness in educational assessments.

## Ethics Statement

We offer a brief discussion of the licensing requirements for the models and datasets used in our submission.

**Datasets:** The USMLE dataset employed for item difficulty and response time prediction is provided by the BEA Shared Task. The auxiliary data from the Test of Narrative Language is derived from a copyrighted assessment. Data from individual participants is not publicly released to protect test-taker anonymity.

**Pretrained Models:** The Longformer, BERT, Mistral, and Whisper models utilized in this work are released under an Apache-2.0 license. Gemma and Llama2 are available for use with a custom license permitting non-commercial use.

## Acknowledgements

We would like to thank Prof. Robin Morris for collecting and labeling the audio recording from the Test of Narrative Language used as part of this study. This work was also funded in part by the National Science Foundation.

## References

Peter Baldwin, Victoria Yaneva, Janet Mee, Brian E Clauser, and Le An Ha. 2021. Using natural language processing to predict item response times and improve test construction. *Journal of Educational Measurement*, 58(1):4–30.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Evelyn L Fisher et al. 2019. Executive Functioning and Narrative Language in Children with Dyslexia. *American Journal of Speech-Language Pathology*, 28(3):1127–1138.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, speech, and signal processing, ieee international conference on*, volume 1, pages 517–520. IEEE Computer Society.

Le An Ha, Victoria Yaneva, Peter Baldwin, and Janet Mee. 2019. Predicting the difficulty of multiple choice questions in a high-stakes medical exam. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 11–20, Florence, Italy. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

A Corinne Huggins-Manley, Brandon M Booth, and Sidney K D'Mello. 2022. Toward argument-based fairness with an application to ai-enhanced educational assessments. *Journal of Educational Measurement*, 59(3):362–388.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Alexander Johnson, Hariram Veeramani, Natarajan Balaji Shankar, and Abeer Alwan. 2023. An Equitable Framework for Automatically Assessing Children's Oral Narrative Language Abilities. In *Proc. INTERSPEECH 2023*, pages 4608–4612.

LangChain. 2024. [link].

Bruce W Lee, Yoo Sung Jang, and Jason Hyung-Jong Lee. 2021. Pushing on text readability assessment: A transformer meets handcrafted linguistic features. *arXiv preprint arXiv:2109.12258*.

Li-Huai Lin, Tao-Hsing Chang, and Fu-Yuan Hsu. 2019. Automated prediction of item difficulty in reading comprehension using long short-term memory. In *2019 international conference on asian language processing (ialp)*, pages 132–135. IEEE.

MARION MEIERS and JULIETTE MENDELOVITS. 2016. A longitudinal study of literacy development in the early years of school. *UNDERSTANDING WHAT WORKS IN ORAL READING ASSESSMENTS*, page 118.

Chaina S Oliveira, João VC Moraes, Telmo Silva Filho, and Ricardo BC Prudêncio. 2022. A two-level item response theory model to evaluate speech synthesis and recognition. *Speech Communication*, 137:19–34.

Zhaopeng Qiu, Xian Wu, and Wei Fan. 2019. Question difficulty prediction for multiple choice problems in medical exams. In *Proceedings of the 28th acm international conference on information and knowledge management*, pages 139–148.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning, ICML 2023*, volume 202, pages 28492–28518. PMLR.

Amrit Romana, Kazuhito Koishida, and Emily Mower Provost. 2023. Automatic disfluency detection from untranscribed speech. *arXiv preprint arXiv:2311.00867*.

Burr Settles, Geoffrey T. LaFlair, and Masato Hagiwara. 2020. Machine learning–driven language assessment. *Transactions of the Association for Computational Linguistics*, 8:247–263.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hariram Veeramani, Natarajan Balaji Shankar, Alexander Johnson, and Abeer Alwan. 2023. Towards Automatically Assessing Children's Oral Picture Description Tasks. In *Proc. 9th Workshop on Speech and Language Technology in Education (SLaTE)*, pages 119–120.

Kang Xue, Victoria Yaneva, Christopher Runyon, and Peter Baldwin. 2020. Predicting the difficulty and response time of multiple choice questions using transfer learning. In *Proceedings of the fifteenth workshop on innovative use of NLP for building educational applications*, pages 193–197.

Victoria Yaneva, Le An Ha, Peter Baldwin, and Janet Mee. 2020. Predicting item survival for multiple choice questions in a high-stakes medical exam. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6812–6818, Marseille, France. European Language Resources Association.

Victoria Yaneva, Daniel Jurich, Le An Ha, and Peter Baldwin. 2021. Using linguistic features to predict the response process complexity associated with answering clinical MCQs. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 223–232, Online. Association for Computational Linguistics.

Victoria Yaneva, Kai North, Peter Baldwin, An Ha Le, Saed Rezayi, Yiyun Zhou, Sagnik Ray Choudhury, Polina Harik, and Brian Clauser. 2024. Findings from the First Shared Task on Automated Prediction of Difficulty and Response Time for Multiple Choice Questions. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, Mexico City, Mexico. Association for Computational Linguistics.