Mistral at SemEval-2024 Task 5: Mistral 7B for Argument Reasoning in Civil Procedure

Marco Siino
Department of Electrical, Electronic
and Computer Engineering
University of Catania
Italy
marco.siino@unipa.it

Abstract

At the SemEval-2024 Task 5, the organizers introduce a novel natural language processing challenge and corpus within the realm of the United States civil procedure. Every datum within the corpus comprises a comprehensive overview of a legal case, a specific inquiry associated with it, and a potential argument in support of a solution, supplemented with an in-depth rationale elucidating the applicability of the argument within the given context. Derived from a text designed for legal education purposes, this dataset presents a multifaceted benchmarking task for contemporary legal language models. Our manuscript delineates the approach we adopted for participation in this competition. Specifically, we detail the use of a Mistral 7B model to answer the questions provided. Our only and best submission reaches an F1-score equal to 0.5597 and an Accuracy of 0.5714, outperforming the task's baseline.

1 Introduction

The content of the Task 5 hosted at SemEval-2024 (Held and Habernal, 2024), was originally introduced in (Bongard et al., 2022).

Asserting a legal argument represents a fundamental proficiency necessary for aspiring legal professionals to acquire. This proficiency demands not only a comprehension of pertinent legal domains but also advanced reasoning skills, including the utilization of analogy-based arguments and the identification of implicit contradictions. Despite recent strides in establishing objective metrics for contemporary natural language processing (NLP) models across diverse facets of legal language comprehension, the absence of a sophisticated task addressing argumentative reasoning within legal contexts persists.

In this article, is discussed a novel task alongside a corresponding benchmark dataset. The introduction of a genuinely challenging task, sourced from legal educational resources, will serve to elucidate strengths and weaknesses inherent in contemporary legal transformer models, including but not limited to Legal-BERT (Chalkidis et al., 2020). Specifically, at the SemEval-2024 Task 5 is unveiled a novel, openly accessible legal dataset tailored for the binary text classification of issues within U.S. civil procedure. The primary objective is to ascertain whether a proposed solution to a given inquiry is deemed accurate or erroneous. The corpus draws inspiration from "The Glannon Guide To Civil Procedure" authored by Joseph Glannon (Glannon, 2023), which caters to law students by offering a comprehensive examination of fundamental U.S. civil procedure topics, inclusive of multiple-choice queries designed to assess reader comprehension.

Through the inception of this freshly minted corpus, the intent extends to scrutinizing the efficacy of various methodological approaches while establishing performance benchmarks.

To address these objectives, there is an ongoing demand for automated tools capable of extracting and categorizing data, facilitating the classification with recent NLP models. Recent advancements in the area of the machine and deep learning architectures have spurred heightened interest in Natural Language Processing (NLP). Substantial endeavours have been directed towards devising techniques for the automated identification and categorization of textual content accessible on the internet today. In the literature, to perform text classification tasks, several strategies have already been proposed. In the last fifteen years, some of the most successful strategies have been based on SVM (Colas and Brazdil, 2006; Croce et al., 2022), on Convolutional Neural Network (CNN) (Kim, 2014; Siino et al., 2021), on Graph Neural Network (GNN) (Lomonaco et al., 2022), on ensemble models (Miri et al., 2022; Siino et al., 2022) and, recently, on Transformers (Vaswani et al., 2017; Siino et al., 2022b).

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 155–162

Participants in SemEval-2024 Task 5 were tasked as follows. The task at hand involves evaluating the accuracy of an answer candidate provided in response to a question, accompanied by a brief introductory passage pertaining to the subject of the question. The objective is to ascertain whether the candidate answer is indeed incorrect or correct. To face with the task, we propose a Transformer-based approach which made use of Mistral 7B (Jiang et al., 2023). We used the model in a zero-shot setup described in the rest of this paper. Specifically, we prompted the latest pretrained version of Mistral with each sample in the dataset. Specifically, we provided a *candidate answer* to a *question*, asking the model if the answer to the legal question was correct or not. The model replied with a yes or no, eventually providing some further explanation.

The subsequent sections of this work are structured as follows: Section 2 offers background information on Task 5, held at SemEval-2024. In Section 3, we outline the approach introduced in this study. Section 4 delves into the specifics of the experimental setup employed to reproduce our findings. The outcomes of the official task and relevant discussions are presented in Section 5. Finally, Section 6 concludes our study and suggests avenues for future research.

We make all the code publicly available and reusable on GitHub[1].

2 Background

For the Task 5 at SemEval-2024 is proposed a legal corpus, publicly accessible for binary text classification tasks focusing on issues within U.S. civil procedure. The primary objective is to determine the correctness of solutions provided in response to specific questions. This corpus draws its content from "The Glannon Guide To Civil Procedure" authored by Joseph Glannon (Glannon, 2023), tailored for law students. The book encompasses fundamental U.S. civil procedure topics and includes multiple-choice questions aimed at evaluating reader comprehension.

Through collaboration with the author and publisher, task organizers secured permission to utilize the content of "The Glannon Guide To Civil Procedure" for constructing this dataset, which is freely available to the research community. The book comprises 25 chapters, each containing multiple-choice questions pertaining to a particular topic, prefaced by an introduction. Every question is followed by 3 to 5 answer candidates, among which one is deemed correct. These answer candidates serve as hypotheses, necessitating an examination of their respective prerequisites for accuracy. The correctness or incorrectness of an answer is subsequently expounded upon in the accompanying analysis.

The dataset construction process involved automated parsing of the book's content, leveraging its structured format to extract individual components of each instance (i.e., introduction, question, answers, and analysis). Additional parsing rules were employed to detect anomalies in the structure, such as instances where the same introduction was shared across multiple questions. However, certain sections of the book required manual extraction, particularly regarding the correctness of answer candidates, as this information was typically embedded within the free-text analysis section. The analysis segments were organized to address each answer candidate separately, classifying them as true or false. To achieve this, the organizers adopted a strategy of isolating the relevant aspects for each answer, despite the absence of explicit keywords or structural indicators guiding the segmentation process. Despite efforts to maintain consistency, some structural inconsistencies were noted throughout the dataset.

Two samples from provided datasets are available online[2] and reported in the Table 3 in the Appendix sectionA. In this case, the two samples contain the same introduction and the same question while providing different answers. Given the Introduction and the Question, the first answer (first row) is wrong, while the second one (second row) is correct.

The organizers adhere to the schedule for SemEval24, which means the following dates:

- Tasks announced (with sample data available): 17 July 2023
- Training data ready 4 September 2023
- Evaluation start 10 January 2024
- Evaluation end by by 31 January 2024
- Paper submission due 19 February 2024
- Notification to authors 18 March 2024

[1] https://github.com/marco-siino/SemEval2024/

[2] https://github.com/trusthlt/semeval24

- Camera ready due 01 April 2024
- SemEval workshop: June 16–21, 2024 (co-located with NAACL 2024 in Mexico City, Mexico)

3 System Overview

Even if it has already been proved that the Transformers are not necessarily the best option for any text classification task (Siino et al., 2022a), depending on the goal, some strategies like domain-specific fine-tuning (Sun et al., 2019; Van Thin et al., 2023), or data augmentation (Lomonaco et al., 2023; Mangione et al., 2022; Siino et al., 2024a) can be beneficial for the considered task.

So far, several Large Language Models (LLMs) have proved to be able to address a plethora of different NLP tasks. For example, in the recent literature, there has been mention of LLaMA, as presented by (Touvron et al., 2023). LLaMA stands out as a collection of publicly available Large Language Models (LLMs) that rival the capabilities of closed-source counterparts like GPT-3.

However, to address the Task 5 hosted at SemEval-2024 we made use of a zero-shot learning approach (Chen et al., 2023; Wahidur et al., 2024), making use of Mistral 7B (Jiang et al., 2023). Mistral 7B, a language model boasting 7 billion parameters, is engineered to excel in both performance and efficiency. In comparison to the leading open 13B model (Llama 2), Mistral 7B demonstrates superior performance across all assessed benchmarks. Moreover, it outperforms the leading publicly available 34B model (LLaMA 1) across various tasks involving code generation, mathematical operations, and reasoning. The model capitalizes on grouped-query attention (GQA) to expedite inference, complemented by sliding window attention (SWA) to effectively process sequences of varying lengths while minimizing inference costs. Additionally, a fine-tuned variant, Mistral 7B – Instruct, is tailored for adhering to instructions. This version, outperforms Llama 2 13B – chat model across both automated and human benchmarks.

The introduction of Mistral 7B Instruct underscores the ease with which the base model can be fine-tuned to achieve notable performance enhancements. Notably, this variant lacks any moderation mechanisms.

Our approach is few-shot (Littenberg-Tobias et al., 2022) and make use of the above-mentioned Mistral 7B. More specifically, given the task hosted at SemEval-2024, we asked the model: *"Is the Answer to the Question above True or False? Answer using ONLY True or False:"*. To this request, the model replied with one or more words - usually starting with a *true* or *false* - that we parsed to extract one of the two labels (i.e., 0 for false and 1 for true). For example, given the introduction:

> *"Defendant in denial. Cardozo is in an accident on Main Street with two other cars, driven by Hooper and Lopes. Cardozo brings a suit in federal court against Hooper and Lopes for his damages. Paragraph 21 of Cardozo's complaint alleges that Hooper had signaled before he turned onto Main Street. The police report on the accident states that, according to a bystander, Hooper had signaled before turning onto Main Street. Lopes, who was coming from Hooper's left, had no view of the right side of Hooper's car, and did not see whether he signaled or not. At the time an answer is due, Lopes's counsel has seen the police report, but has not yet been able to locate other witnesses to obtain their testimony. The most appropriate response for Lopes to Paragraph 21 of Cardozo's complaint would be to."*

The answer:

> *"state that he is without sufficient information to form a belief about the truth of the allegation."*

And our question:

> *Is the Answer to the Question above True or False? Answer using ONLY True or False:*

The model replied with:

> *true. lopes' answer could state that he lacks sufficient information to admit*

that we mapped into the binary label *1* corresponding to *true*.

We did not find any inconsistency in the outputs generated by Mistral along all the provided prompts. Specifically, we did not notice any variation in the behaviours of the model at different times of prompting. This leads us to the conclusion that given always the same input context (i.e.,

few-shot samples) during the prompt, the output provided is always consistent disregarding the time and the previous prompts provided. Finally, we collected all the predictions provided on the test set to into a JSON file with the required format to submit our predictions.

As noted in the recent study by (Siino et al., 2024b), the contribution of preprocessing for text classification tasks is generally not impactful when using Transformers. More specifically, the best combination of preprocessing strategies is not very different from doing no preprocessing at all in the case of Transformers. For these reasons, and to keep our system highly fast and computationally light, we have not performed any preprocessing on the text.

4 Experimental Setup

We implemented our model on Google Colab. The library we used come from HuggingFace and as already mentioned is Mistral 7B[3]. We employed the v0.2 iteration of Mistral 7B, which represents an enhanced version of the Mistral-7B-Instruct-v0.1 model. To harness the capabilities of instruction fine-tuning, prompts must be enclosed within [INST] and [/INST] tokens. Additionally, the initial instruction should commence with a sentence identifier. The next instructions should not. The assistant generation will be ended by the end-of-sentence token ID. We also imported the Llama library (Touvron et al., 2023) from *llama_cpp*. The library is fully described on GitHub[4]. The dataset provided for all the phases are available on the official competition page. We did not perform any additional fine-tuning on the model. To run the experiment, a T4 GPU from Google has been used. After the generation of predictions, we exported the results on the format required by the organizers. As already mentioned, all of our code is available on GitHub.

5 Results

Given the binary nature of the classification task, the organizers proposed F1 score and Accuracy as the two evaluation metrics to be considered for the final ranking. The F1 score is defined in the Equation 1. Where TP stands for the number of correctly predicted right answers, FP stands for the number of wrongly predicted right answers, and FN stands for the right answers wrongly predicted as wrong answers.

	F1	Accuracy
Mistral 7B	0.5597	0.5714

Table 1: The method's performance on the test set. In the table, the results obtained and shown on the official GitHub page are reported.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \quad (1)$$

Given the previous definitions, the accuracy is defined as stated in the Equation 2.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

In Table 1, we present the outcomes derived from our methodology. They are the same results publicly availble on the official final ranking shown on the official task page[5] and on CodaLab[6].

In the Table 2, the results obtained by the first three teams and by the last one, as showed on the official task page, are reported. Compared to the best performing models, our simple approach exhibits some room for improvements. However, it is worth notice that required no further pre-training and the computational cost to address the task is manageable with the free online resources offered by Google Colab. Finally, the proposed approach is able to outperform the baseline provided.

6 Conclusion

This paper presents the application of a Mistral 7B-model for addressing the Task 5 at SemEval-2024. For our submission, we decided to follow a zero-shot learning approach, employing as-is, an in-domain pre-trained Transformer. After several experiments, we found beneficial to build a prompt containing the question for the model. Then we provide as a prompt: the introduction, the question and an answer candidate. The model is asked to decide whether the candidate answer is correct or not. The task is challenging, and there is still opportunity for improvement, as can be noted looking at the final ranking. Possible alternative approaches include

[3] https://huggingface.co/TheBloke/Mistral-7B-Instruct-v0.2-GGUF

[4] https://github.com/ggerganov/llama.cpp

[5] https://github.com/trusthlt/semeval24

[6] https://codalab.lisn.upsaclay.fr/competitions/14817

TEAM NAME	F1	Accuracy
HW-TSC (1)	0.8231	0.8673
PoliToHFI (2)	0.7747	0.8265
SU-FMI (3)	0.7728	0.8367
lena.held (21)	0.4269	0.7449

Table 2: Comparing performance on the test set. In the table are shown the results obtained by the first three teams and by the last one. In parentheses is reported the position in the official final ranking.

utilizing the few-shot capabilities or also the use of other models like GPT and T5, eventually using further data, or directly integrating other samples from the training and from the development sets. Further improvements could be obtained with a fine-tuning and modelling the problem as a text classification task. Furthermore, given the interesting results recently provided on a plethora of tasks, also other few-shot learning (Wang et al., 2023; Maia et al., 2024; Siino et al., 2023; Meng et al., 2024) or data augmentation strategies (Muftie and Haris, 2023; Tapia-Téllez and Escalante, 2020; Siino and Tinnirello, 2023) could be employed to improve the results. Looking at the final ranking, our simple approach exhibits some room for improvements. However, it is worth notice that required no further pre-training and the computational cost to address the task is manageable with the free online resources offered by Google Colab.

Acknowledgments

We extend our gratitude to the anonymous reviewers for their insightful comments and valuable suggestions, which have significantly enhanced the clarity and presentation of this paper.

References

Leonard Bongard, Lena Held, and Ivan Habernal. 2022. The Legal Argument Reasoning Task in Civil Procedure. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 194–207, Abu Dhabi, UAE. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Shiming Chen, Ziming Hong, Wenjin Hou, Guo-Sen Xie, Yibing Song, Jian Zhao, Xinge You, Shuicheng Yan, and Ling Shao. 2023. Transzero++: Cross attribute-guided transformer for zero-shot learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):12844 – 12861.

Fabrice Colas and Pavel Brazdil. 2006. Comparison of svm and some older classification algorithms in text classification tasks. In *IFIP International Conference on Artificial Intelligence in Theory and Practice*, pages 169–178. Springer.

Daniele Croce, Domenico Garlisi, and Marco Siino. 2022. An SVM ensemble approach to detect irony and stereotype spreaders on twitter. In *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022*, volume 3180 of *CEUR Workshop Proceedings*, pages 2426–2432. CEUR-WS.org.

Joseph W Glannon. 2023. *Glannon guide to civil procedure: learning civil procedure through multiple-choice questions and analysis*. Aspen Publishing.

Lena Held and Ivan Habernal. 2024. SemEval-2024 Task 5: Argument Reasoning in Civil Procedure. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, Mexico City, Mexico. Association for Computational Linguistics.

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*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1746–1751. ACL.

Joshua Littenberg-Tobias, G. R. Marvez, Garron Hillaire, and Justin Reich. 2022. Comparing few-shot learning with GPT-3 to traditional machine learning approaches for classifying teacher simulation responses. In *AIED (2)*, volume 13356 of *Lecture Notes in Computer Science*, pages 471–474. Springer.

Francesco Lomonaco, Gregor Donabauer, and Marco Siino. 2022. COURAGE at checkthat!-2022: Harmful tweet detection using graph neural networks and ELECTRA. In *Proceedings of the Working Notes of*

CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022, volume 3180 of *CEUR Workshop Proceedings*, pages 573–583. CEUR-WS.org.

Francesco Lomonaco, Marco Siino, and Maurizio Tesconi. 2023. Text enrichment with japanese language to profile cryptocurrency influencers. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023*, volume 3497 of *CEUR Workshop Proceedings*, pages 2708–2716. CEUR-WS.org.

Beatriz Matias Santana Maia, Maria Clara Falcão Ribeiro de Assis, Leandro Muniz de Lima, Matheus Becali Rocha, Humberto Giuri Calente, Maria Luiza Armini Correa, Danielle Resende Camisasca, and Renato Antonio Krohling. 2024. Transformers, convolutional neural networks, and few-shot learning for classification of histopathological images of oral cancer. *Expert Systems with Applications*, 241:122418.

Stefano Mangione, Marco Siino, and Giovanni Garbo. 2022. Improving irony and stereotype spreaders detection using data augmentation and convolutional neural network. In *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022*, volume 3180 of *CEUR Workshop Proceedings*, pages 2585–2593. CEUR-WS.org.

Zong Meng, Zhaohui Zhang, Yang Guan, Jimeng Li, Lixiao Cao, Meng Zhu, Jingjing Fan, and Fengjie Fan. 2024. A hierarchical transformer-based adaptive metric and joint-learning network for few-shot rolling bearing fault diagnosis. *Measurement Science and Technology*, 35(3).

Mohsen Miri, Mohammad Bagher Dowlatshahi, Amin Hashemi, Marjan Kuchaki Rafsanjani, Brij B Gupta, and W Alhalabi. 2022. Ensemble feature selection for multi-label text classification: An intelligent order statistics approach. *International Journal of Intelligent Systems*, 37(12):11319–11341.

Fuad Muftie and Muhammad Haris. 2023. Indobert based data augmentation for indonesian text classification. In *2023 International Conference on Information Technology Research and Innovation, ICITRI 2023*, page 128 – 132.

Marco Siino, Elisa Di Nuovo, Ilenia Tinnirello, and Marco La Cascia. 2021. Detection of hate speech spreaders using convolutional neural networks. In *Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021*, volume 2936 of *CEUR Workshop Proceedings*, pages 2126–2136. CEUR-WS.org.

Marco Siino, Elisa Di Nuovo, Ilenia Tinnirello, and Marco La Cascia. 2022a. Fake news spreaders detection: Sometimes attention is not all you need. *Information*, 13(9):426.

Marco Siino, Marco La Cascia, and Ilenia Tinnirello. 2022b. Mcrock at semeval-2022 task 4: Patronizing and condescending language detection using multi-channel cnn, hybrid lstm, distilbert and xlnet. In *Proceedings of the 16th International Workshop on Semantic Evaluation, SemEval@NAACL 2022, Seattle, Washington, United States, July 14-15, 2022*, pages 409–417. Association for Computational Linguistics.

Marco Siino, Francesco Lomonaco, and Paolo Rosso. 2024a. Backtranslate what you are saying and i will tell who you are. *Expert Systems*, n/a(n/a):e13568.

Marco Siino, Maurizio Tesconi, and Ilenia Tinnirello. 2023. Profiling cryptocurrency influencers with few-shot learning using data augmentation and ELECTRA. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023*, volume 3497 of *CEUR Workshop Proceedings*, pages 2772–2781. CEUR-WS.org.

Marco Siino and Ilenia Tinnirello. 2023. Xlnet with data augmentation to profile cryptocurrency influencers. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023*, volume 3497 of *CEUR Workshop Proceedings*, pages 2763–2771. CEUR-WS.org.

Marco Siino, Ilenia Tinnirello, and Marco La Cascia. 2022. T100: A modern classic ensemble to profile irony and stereotype spreaders. In *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022*, volume 3180 of *CEUR Workshop Proceedings*, pages 2666–2674. CEUR-WS.org.

Marco Siino, Ilenia Tinnirello, and Marco La Cascia. 2024b. Is text preprocessing still worth the time? a comparative survey on the influence of popular preprocessing methods on transformers and traditional classifiers. *Information Systems*, 121:102342.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18*, pages 194–206. Springer.

José Medardo Tapia-Téllez and Hugo Jair Escalante. 2020. Data augmentation with transformers for text classification. In *Advances in Computational Intelligence*, pages 247–259, Cham. Springer International Publishing.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Dang Van Thin, Duong Ngoc Hao, and Ngan Luu-Thuy Nguyen. 2023. Vietnamese sentiment analysis:

An overview and comparative study of fine-tuning pretrained language models. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(6):1–27.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Rahman S. M. Wahidur, Ishmam Tashdeed, Manjit Kaur, and Heung-No Lee. 2024. Enhancing zero-shot crypto sentiment with fine-tuned language model and prompt engineering. *IEEE Access*, 12:10146 – 10159.

Xixi Wang, Xiao Wang, Bo Jiang, and Bin Luo. 2023. Few-shot learning meets transformer: Unified query-support transformers for few-shot classification. *IEEE Trans. Circuits Syst. Video Technol.*, 33(12):7789–7802.

A Appendix

As stated in the background section, in this appendix are shown two samples from the provided datasets. The two samples in the Table 3 give an example of a wrong answer candidate (first row in the table) and an example of a correct answer candidate (second row in the table).

Introduction	Question	Answer Candidate	Label
"My students always get confused about the relationship between removal to federal court and personal jurisdiction. Suppose that a defendant is sued in Arizona and believes that she is not subject to personal jurisdiction there. Naturally, she should object to personal jurisdiction. [...] But generally the scope of personal jurisdiction in the federal court will be the same as that of the state court, because the Federal Rules require the federal court in most cases to conform to state limits on personal jurisdiction. Fed. R. Civ. P. 4(k)(1)(A). I've stumped a multitude of students on this point. Consider the following two cases to clarify the point."	*"7. A switch in time. Yasuda, from Oregon, sues Boyle, from Idaho, on a state law unfair competition claim, seeking $250,000 in damages. He sues in state court in Oregon. Ten days later (before an answer is due in state court), Boyle files a notice of removal in federal court. Five days after removing, Boyle answers the complaint, including in her answer an objection to personal jurisdiction. Boyle's objection to personal jurisdiction is"*	not waived by removal, but will be denied because the federal courts have power to exercise broader personal jurisdiction than the state courts.	0
"My students always get confused about the relationship between removal to federal court and personal jurisdiction. Suppose that a defendant is sued in Arizona and believes that she is not subject to personal jurisdiction there. Naturally, she should object to personal jurisdiction. [...] But generally the scope of personal jurisdiction in the federal court will be the same as that of the state court, because the Federal Rules require the federal court in most cases to conform to state limits on personal jurisdiction. Fed. R. Civ. P. 4(k)(1)(A). I've stumped a multitude of students on this point. Consider the following two cases to clarify the point."	*"7. A switch in time. Yasuda, from Oregon, sues Boyle, from Idaho, on a state law unfair competition claim, seeking $250,000 in damages. He sues in state court in Oregon. Ten days later (before an answer is due in state court), Boyle files a notice of removal in federal court. Five days after removing, Boyle answers the complaint, including in her answer an objection to personal jurisdiction. Boyle's objection to personal jurisdiction is"*	not waived by removal. The court should dismiss if there is no personal jurisdiction over Boyle in Oregon, even though the case was properly removed.	1

Table 3: Two different samples from the official dataset are provided. Together with the introduction, a question and a candidate answer the label is provided (i.e., 0 if the answer is incorrect, 1 if the answer is correct)