

# AMELIA - Argument Mining Evaluation on Legal documents in ItAlian: A CALAMITA Challenge

Giulia Grundler<sup>2,\*</sup>, Andrea Galassi<sup>1,\*</sup>, Piera Santin<sup>3</sup>, Alessia Fidelangeli<sup>2</sup>, Federico Galli<sup>2</sup>, Elena Palmieri<sup>1</sup>, Francesca Lagioia<sup>2,3</sup>, Giovanni Sartor<sup>2,3</sup> and Paolo Torroni<sup>1</sup>

<sup>1</sup>DISI, Alma-AI, University of Bologna, Italy

<sup>2</sup>CIRSFID Alma-AI, Faculty of Law, University of Bologna, Italy

<sup>3</sup>European University Institute, Law Department, Italy

## Abstract

This challenge consists of three classification tasks, in the context of argument mining in the legal domain. The tasks are based on a dataset of 225 Italian decisions on Value Added Tax, annotated to identify and categorize argumentative text. The objective of the first task is to classify each argumentative component as premise or conclusion, while the second and third tasks aim at classifying the type of premise: legal vs factual, and its corresponding argumentation scheme. The classes are highly unbalanced, hence evaluation is based on the macro F1 score.

## Keywords

LLM, Argument Mining, Legal Analytics, VAT, CALAMITA, CLiC-it

## 1. Challenge: Introduction and Motivation

To what extent are Large Language Models (LLMs) capable of reasoning, as opposed to simply recognizing patterns from vast amounts of data is an open research question and the subject of a lively ongoing debate [1]. A way to describe human reasoning is through its ability to understand, evaluate, and invent arguments composed by claims, evidence, and conclusions meaningfully connected with one another [2]. For this reason, the ability to recognize arguments could be considered as a first step in a sequence of reasoning tasks of increasing complexity, that goes from the detection and classification of argumentative discourse units or *argument components*, through argument structure prediction, reconstruction, evaluation, down to argument generation. Automating these tasks is the object of argument mining [3, 4, 5]. We believe that gauging the ability of LLMs to address even basic argument mining tasks would provide meaningful cues as to these models' ability to process and understand logical relations expressed in natural language.

While several datasets for argument mining in English have been developed over the last decade [6, 7, 8, 9, 10],

resources for other languages remain scarce. To the best of our knowledge, only a few works exist for Italian. In [11], the authors use the CorEA corpus of user comments to online newspaper articles, to assign the correct relation (support or attack), to pairs of arguments. In [12], the authors propose a new model for stance detection, trained and evaluated on a corpus of Italian tweets where users were discussing on a highly polarized political debate.

Among the many domains of interest for argument mining, our focus is on the legal domain, where argumentation is fundamental for the decision-making process. Legal reasoning relies heavily on well-structured arguments, as legal professionals must construct and deconstruct arguments within formal documents, providing a challenging setting for assessing an LLMs' ability to engage in complex reasoning tasks. Despite its relevance, little attention has been given to argument mining in the legal domain in Italian. Most existing work in legal NLP for Italian has focused on tasks such as law article retrieval [13, 14], outcome prediction [15], analysis of contracts [16, 17], and summarization [18, 19].

Our challenge for CALAMITA [20] consists of three classification tasks over argumentative texts. We mostly follow the setting used in Demosthenes [21, 22], a corpus for argument mining on legal documents in English. Since we leverage real legal documents, not synthetic or artificially constructed case studies, our dataset reflects the real complexity and nuances of legal argumentation. It is therefore particularly relevant for a robust assessment of LLMs' abilities in real-world applications. To the best of our knowledge, we are the first to propose a challenge of argument mining over legal documents in Italian.

The challenge requires understanding not only the

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 – 06, 2024, Pisa, Italy

\*Corresponding authors.

✉ giulia.grundler2@unibo.it (G. Grundler); a.galassi@unibo.it (A. Galassi)

🆔 0000-0002-7255-9343 (G. Grundler); 0000-0001-9711-7042

(A. Galassi); 0000-0002-0734-9657 (P. Santin); 0000-0003-3739-5387

(F. Galli); 0000-0001-5176-8843 (E. Palmieri); 0000-0001-7083-3487

(F. Lagioia); 0000-0003-2210-0398 (G. Sartor); 0000-0002-9253-8638

(P. Torroni)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Italian language but domain-specific technical language. Such a language uses complex syntactic structures, and a specialized terminology. Besides language, the challenge tests LLMs' ability to recognize and interpret legal arguments by recognizing typical argumentation *schemes* [23], e.g., patterns of reasoning used in human discourse, offering a principled approach to argument analysis and evaluation. Identifying schemes is challenging as there are many possible schemes, and arguments are often only partially laid out in the text, leaving many important parts implicit for brevity or because they are considered common knowledge. Nonetheless, this task lends itself to generalization beyond the legal domain, making the insights transferable to other fields where structured reasoning plays a critical role.

## 2. Challenge: Description

We consider an *argument* as a set of interconnected portions of texts called *argument components*. The connections between components form a specific pattern of relationships that represents a reasoning paradigm.

The following tasks presume that argument components have already been identified from the source documents. Argument components can therefore be classified according to their role in the connections (such as Premises or Conclusions), according to their content (such as Legal or Factual), and according to the relationship pattern they contribute to (the Argumentative Scheme).

This challenge proposes three classification tasks, in the context of argument mining in the legal domain:

- **Argument Component classification:** given an argumentative component, classify it as premise or conclusion.
- **Premise Type classification:** given a premise, classify it as factual or legal.
- **Argument Scheme classification:** given a pre-determined set of argument schemes, classify a legal premise as belonging to one or more such schemes.

The following paragraphs contain a definition of each class, along with an example extracted from the dataset. The translated version of the examples is available in Appendix A.

**Argument Component classification.** Binary classification: given an argumentative component, classify it as premise or conclusion.

- *Argument premise:* a proposition that provides a reason or support for the argument.

*Si osserva poi che ritenere che la mancata possibilità di detrazione a favore di soggetti come il ricorrente comporti un aiuto di Stato in favore degli ospedali pubblici, in quanto le perdite degli stessi vengono ripianate dalle USL e dalla Regioni trascura di considerare l'accessibilità, indiscriminata, ai servizi dei nosocomi pubblici da parte dei soggetti iscritti al SSN, rispetto a quella ad un libero professionista sanitario che, in quanto tale, ben potrebbe rifiutarsi di prestare i propri servizi al pare di un normale contraente.*

- *Argument conclusion:* the statement that follows logically from the premise(s) and represents the final point being argued for.

*Dunque, l'ufficio ha riconosciuto la non imponibilità IVA delle cessioni all'esportazione, così cessando sul punto la materia del contendere.*

Argument components can be involved in more than one relationship, therefore a component may be the conclusions of other premises, as well as a premise of other arguments. In that case, the component is to be classified as a premise.

**Premise Type classification.** Multi-label classification: classify an argumentative premise as factual or legal (or both).

- *Factual premise:* a premise that describes factual situations and events, pertaining to the substance or the procedure of the case.

*Indubbiamente, la contribuente ha impugnato la sentenza di prime cure, rappresentando nuovamente di non aver potuto proporre appello avverso la pronuncia di condanna di primo grado, per causa di forza maggiore.*

- *Legal premise:* a premise that specifies the legal content (legal rules, precedents, interpretation of applicable laws and principles).

*La giurisprudenza citata, alla motivazione della quale si fa rinvio, ha tra l'altro preso posizione espressamente e positivamente sulla conformità della normativa italiana rispetto a quella dell'Unione Europea, risultando così confutata anche la doglianza della difesa sul punto che ha chiesto la sospensione del procedimento, con investitura della Corte di Giustizia Europea della questione.*

Since a premise could be both factual and legal, this task is framed as multi-label binary classification.

**Argument Scheme classification.** Legal premises determine the nature of the legal reasoning they support, hence they are labeled with the corresponding reasoning pattern, called argument scheme. We define five schemes relevant for tax law. Each legal premise may be assigned multiple schemes, therefore we frame this task as multi-label multi-class classification.

Given a legal premise, classify it as belonging to one or more of the following schemes: (*established*) rule, precedent, classification, interpretative, or principle.

- *Rule (or established rule) scheme*: it is used whenever an explicit reference to codified law is present. This reference can be the reference to a certain article or the quotation of the text of a certain article.

*Infatti, è ben vero che, ai sensi del combinato disposto dagli articoli 54 e 23 D.Lgs. n. 546/1992, il convenuto in appello deve costituirsi entro 60 giorni dal giorno in cui ricorso è stato notificato.*

- *Precedent scheme*: it is used whenever there is an explicit reference to a previous decision. In the dataset we considered only the references to a decision of both the Court of Cassation or the European Court of Justice.

*L'Amministrazione "ha l'onere di provare ed allegare gli elementi probatori su cui si fonda la contestazione, tra i quali possono rilevare, in via indiziaria, quali elementi sintomatici della mancata esecuzione della prestazione dal fatturante, l'assenza della minima dotazione personale e strumentale, l'immediatezza dei rapporti (cedente/prestatore fatturante interposto e cessionario/committente), una conclamata inidoneità allo svolgimento dell'attività economica e la non corrispondenza tra i cedenti e la società coinvolta nell'operazione".*

- *Classification scheme*: it is used whenever a legal concept is defined, its properties are listed, and a certain fact or legal deed must be qualified as having those properties.

*In conclusione, per quanto fin qui esposto, i "compro oro" possono essere definiti come "esercizi commerciali che acquistano, commerciano o rivendono oggetti d'oro, di metalli preziosi o recanti pietre preziose usati e li cedono nella forma di materiale, di rottami d'oro o di metalli preziosi alle fonderie o ad altre aziende specializzate nel recupero di materiali preziosi". Trattano esclusivamente prodotti finiti e non possono, congiuntamente, acquistare oro da gioielleria usato, fonderlo (per proprio conto o*

*con incarico a terzi) e cedere il prodotto finito ottenuto.*

- *Interpretative scheme*: it is used whenever the Court expresses new interpretative assertions (that may depend on previous case law) thereby creating new precedents.

*Si vuole dire, in sostanza, che la finalità del contraddittorio anticipato è quella di mettere il contribuente nella condizione di potere fare valere le proprie osservazioni prima che la decisione sia adottata e, quindi, di far sì che l'Amministrazione possa tener conto di tutti gli elementi del caso nell'adottare (o non adottare) il provvedimento ovvero nel dare a questo un contenuto piuttosto che un altro.*

- *Principle scheme*: it is used whenever the Court explicitly refers to a principle of law (e.g. the Principle of proportionality).

*Nell'ordinamento unionale, pertanto, il principio del contraddittorio in ambito tributario prescinde dalla natura del tributo e deve trovare applicazione ogni qualvolta l'amministrazione sulla base della documentazione esibita ritenga dovere dare alla stessa documentazione interpretazione diversa da quella data dal contribuente invitandolo, come detto, a fornire nel corso del contraddittorio le ragioni della propria scelta.*

### 3. Data description

#### 3.1. Origin of data

The data consists of argumentative portions of text extracted from 225 Italian decisions on Value Added Tax (VAT) by the Regional Tax Commissions from various judicial districts. The decisions were downloaded partially from the open Giustizia Tributaria database<sup>1</sup> and from other judicial databases accessed through university licensing agreements. The decisions range from 2010 to 2022 and concern taxable transactions, exemptions, out-of-scope transactions, and the right to obtain a deduction. The argumentative components were extracted from the sections "Motivi della decisione", "Diritto" or "Fatto e diritto", depending on the format of each decision.

The collected data were anonymised modifying any identification data of natural or legal persons involved in the proceedings. In particular, the names of the parties in the proceeding and, to provide the highest privacy standards, also the names of the companies have been replaced with initials (e.g., Mario Rossi in "MR", Company

<sup>1</sup>Tax Justice database accessible at: <https://www.giustizia-tributaria.it/>.

s.r.l in “C s.r.l.”). The names of the judges composing the judicial panel have been replaced by “giu1, giu2, [...] giuN”. Also, addresses and places were replaced with ‘XXX’, and dates were changed to show only the year in the following format: DD/MM/2015.

### 3.2. Annotation details

The dataset was annotated by four tax law experts. Annotation guidelines are significantly based on our previous work on the Demosthenes corpus [21], a dataset with English documents from the Court of Justice of the European Union. The guidelines were adapted to the Italian decisions, and refined through an iterative process of validation and discussion, to solve conflicts between annotators. In particular, the annotation is based on the same classes used in Demosthenes. However, the structure of the decisions is different: while in the English corpus the annotation is done at the sentence level, it is not always possible to meet this criterion in Italian decisions. Therefore, the constraint has been relaxed, allowing a single annotation to cover multiple sentences and a single sentence to contain multiple annotations. The tagged decisions are available in our GitHub repository.<sup>2</sup>

### 3.3. Data format

Data are available as a Hugging Face Dataset,<sup>3</sup> divided in three splits: train, val and test. Each row represents an argumentative component, with the following columns:

- Text: the text of the component
- Document: the document it belongs to
- Component: if it is a premise (prem) or a conclusion (conc)
- Type: a list value representing the type of a premise; the list contains F for a Factual premise and L for a Legal one.
- Scheme: a list value representing the argumentative schemes of a legal premise. The values are: Rule, Prec, Class, Itpr and Princ.
- Chain\_id: univocal for each document, it specifies the argumentative chain the component belongs to (e.g. A1, A2,..., B1, B2,...)
- Id: an univocal numerical id

### 3.4. Example of prompts used for zero and few shots

For each task, we propose both a zero-shot and a few-shot prompt. For the few-shot version, we have selected some particularly representative examples from the training

<sup>2</sup><https://github.com/adele-project/AMELIA/>

<sup>3</sup><https://huggingface.co/datasets/nlp-unibo/AMELIA>

set, some of which are included in Section 2. Here we report the zero-shot version. The translation of the zero-shot prompts is available in Appendix B. The few-shot version is available in Appendix C.

**Argument Component classification:** given an argumentative text, classify it as premise or conclusion.

Prompt: “*Classifica il seguente testo argomentativo come premessa ‘prem’ o conclusione ‘conc’. Per premessa (prem) si intende una proposizione che fornisce una ragione o un supporto per l’argomentazione. Per conclusione (conc) si intende l’affermazione che segue logicamente dalle premesse e rappresenta il punto finale che viene argomentato. Testo:*”

**Premise Type classification:** given a premise, classify it as factual, legal or both.

Prompt: “*Classifica la seguente premessa come di fatto ‘F’, legale ‘L’ o entrambe. Le premesse di fatto (F) descrivono situazioni ed eventi fattuali relativi al caso di specie. Le premesse legali (L) specificano il contenuto giuridico (norme giuridiche, precedenti, interpretazione delle leggi e dei principi applicabili). L’output atteso è una lista con tutte le label applicabili. Ad esempio: [‘F’, ‘L’]. Testo:*”

**Argument Scheme classification:** given a legal premise, classify it as one or more of the following argumentative schemes: Rule, Prec, Class, Itpr, Princ.

Prompt: “*Classifica la seguente premessa legale in uno o più dei seguenti schemi argomentativi: Rule, Prec, Class, Itpr, Princ. Rule: se esiste un riferimento esplicito o implicito a un articolo di legge o la citazione del testo di una norma. Prec: se esiste un riferimento ad una precedente pronuncia della Corte di Cassazione o della Corte di Giustizia dell’Unione Europea. Class: se c’è la definizione di un concetto giuridico o degli elementi costitutivi dello stesso. Itpr: se c’è il riferimento a uno dei criteri interpretativi contenuti all’art. 12 delle preleggi (letterale, teleologica, psicologica, sistematica) al codice civile. Princ: se c’è un riferimento espresso a un principio generale del diritto (es. principio di proporzionalità). L’output atteso è una lista con tutte le label applicabili. Ad esempio: [‘Prec’, ‘Princ’, ‘Rule’]. Testo:*”

### 3.5. Detailed data statistics

The composition of the dataset is summarized in Table 1. The splitting between train, validation, and test data was done at the document level so that components of the same document belong to the same split. It was performed manually, with a ratio of approximately 60:20:20, and the aim of balancing the Scheme classes as much as possible. We adopt the train/val/test format to make the results comparable with as many methods as possible, such as fine-tuned transformer-based models.

Split	N docs	Component		Premise Type		Argument Scheme				
		Prem	Conc	Factual	Legal	Rule	Prec	ltpr	Princ	Class
Train	135	1866	242	1254	812	350	264	224	92	51
Validation	44	528	81	315	266	107	82	83	21	22
Test	46	516	78	323	260	118	73	67	31	27
Total	225	2910	401	1892	1338	575	419	374	144	100

**Table 1**  
Composition of the dataset

## 4. Metrics

Due to the heavy unbalance between the classes, we evaluate the results using the macro F1 score. Additionally, we evaluate the F1 score of each class to provide further insights.

As a reference, in Demosthenes [21] the best macro F1 results for the three tasks are 0.88 for Argument Component classification, 0.85 for Premise Type classification, and 0.75 for Argument Scheme classification. It is important to specify that these scores are not directly comparable and we provide them only as a reference of the difficulty of the tasks.

## 5. Limitations

The original documents, along with the argument mining annotation, are already available as part of the Adele tool.<sup>4</sup> The original documents, annotated according to the task of outcome prediction instead of argument mining, are also published in [15].

The dataset is limited in size, consisting of only 225 legal decisions on Value Added Tax (VAT). While this provides a valuable resource for testing argument mining models in the Italian tax legal domain, the relatively small dataset may not capture the full diversity of argumentative structures present in the broader Italian tax legal system or other legal domains. This could limit the scalability of models trained on this dataset. Also, given that the legal decisions are from a specific time frame (2010-2022), the dataset may not reflect more recent developments or changes in legal reasoning or tax law.

Secondly, the dataset has been anonymised to protect the privacy of individuals and legal entities. While this is necessary to comply with data protection regulations, the anonymisation process may have removed certain contextual details (e.g., names of places or entities) that could be relevant for understanding the nuances of certain legal arguments. As a result, models may not fully

capture or leverage such contextual details that would otherwise aid in more accurate argument classification.

Another limitation is the manual annotation process, which, despite efforts to ensure consistency through expert annotators and conflict resolution, may still be subject to human bias or interpretation inconsistencies. These subjective elements could affect the quality and reproducibility of the tasks.

## 6. Ethical issues

The dataset comprises legal decisions that have been anonymised to protect the privacy of the individuals. However, it is important to acknowledge the potential risks related to re-identification, even with anonymisation efforts, especially in legal contexts where case details could be cross-referenced with external sources. Care was taken to remove any personal identifiers, such as names, addresses, and dates, but residual risks may remain.

Additionally, the use of this dataset raises questions regarding the deployment of AI systems in legal contexts. AI used by a judicial authority in researching and interpreting facts and the law are considered high-risk by the AI Act.<sup>5</sup> Those systems must conform to the essential requirements (e.g. data governance, user transparency, human oversight, etc.) and the conformity must be documented.

Finally, a critical aspect is the transparency and accountability of AI systems when applied in sensitive domains like law. Users of the models should understand their limitations, especially in tasks involving nuanced reasoning like legal argumentation. Furthermore, ensuring that legal professionals and stakeholders have the ability to audit and interpret the decisions made by AI models is crucial to avoid undermining trust in legal institutions.

<sup>4</sup><https://adele-tool.eu/>

<sup>5</sup><https://eur-lex.europa.eu/eli/reg/2024/1689/oj>.



## 7. Data license and copyright issues

The dataset used in this challenge consists of legal decisions on Value Added Tax (VAT) made by the Regional Tax Commissions in Italy, available and downloaded from the Giustizia Tributaria and other judicial databases accessed through university licensing agreements. These legal texts, being official public documents, are generally not subject to copyright restrictions. The dataset consists of a non-substantial part of the respective databases. Moreover, the use of data is compliant with the text and data mining exception under the EU Copyright Directive and implementing national law.<sup>6</sup>

Since the data has been processed and annotated, the annotations and derived data are subject to copyright by the authors of this challenge. To promote transparency and further research, the dataset is released under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. This license allows others to share, use, and adapt the data, as long as appropriate credit is given to the creators, and any modifications are explicitly indicated.

## Acknowledgments

This work was partially supported by the following projects: “ADELE – Analytics for DEcision of LEgal cases” (Justice Programme, GA. No. 101007420); PRIN2022 PRIMA - PRivacy Infringements Machine-Advice (Ref. Prot. n.: 20224TPEYC - CUP J53D23005130001); “FAIR - Future Artificial Intelligence Research” – Spoke 8 “Pervasive AI”, under the European Commission’s NextGeneration EU programme, PNRR – M4C2 – Investimento 1.3, Partenariato Esteso (PE00000013).

## References

- [1] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big?, in: FAccT, ACM, 2021, pp. 610–623.
- [2] D. Walton, Argumentation Theory: A Very Short Introduction, Springer US, Boston, MA, 2009, pp. 1–22. doi:10.1007/978-0-387-98197-0\_1.
- [3] M. Lippi, P. Torroni, Argumentation mining: State of the art and emerging trends, ACM Trans. Internet Techn. 16 (2016) 10:1–10:25. doi:10.1145/2850417.
- [4] E. Cabrio, S. Villata, Five years of argument mining: a data-driven analysis, in: IJCAL, ijcai.org, 2018, pp. 5427–5433.
- [5] J. Lawrence, C. Reed, Argument Mining: A Survey, Computational Linguistics 45 (2020) 765–818. doi:10.1162/coli\_a\_00364.
- [6] I. Habernal, D. Faber, N. Recchia, S. Bretthauer, I. Gurevych, I. S. genannt Döhmann, C. Burchard, Mining legal arguments in court decisions, Artif. Intell. Law 32 (2024) 1–38.
- [7] V. Niculae, J. Park, C. Cardie, Argument mining with structured svms and rnns, in: ACL (1), Association for Computational Linguistics, 2017, pp. 985–995.
- [8] P. Poudyal, J. Savelka, A. Ieven, M. F. Moens, T. Goncalves, P. Quaresma, ECHR: Legal corpus for argument mining, in: E. Cabrio, S. Villata (Eds.), Proceedings of the 7th Workshop on Argument Mining, Association for Computational Linguistics, Online, 2020, pp. 67–75. URL: <https://aclanthology.org/2020.argmining-1.8>.
- [9] T. Mayer, S. Marro, E. Cabrio, S. Villata, Enhancing evidence-based medicine with natural language argumentative analysis of clinical trials, Artif. Intell. Medicine 118 (2021) 102098.
- [10] P. Accuosto, H. Saggion, Mining arguments in scientific abstracts with discourse-level embeddings, Data Knowl. Eng. 129 (2020) 101840.
- [11] P. Basile, V. Basile, E. Cabrio, S. Villata, Argument Mining on Italian News Blogs, volume 1749 of CEUR Workshop Proceedings, CEUR-WS.org, 2016. URL: <https://ceur-ws.org/Vol-1749/paper8.pdf>.
- [12] M. Lai, V. Patti, G. Ruffo, P. Rosso, Stance evolution and twitter interactions in an italian political debate, in: M. Silberztein, F. Atigui, E. Kornysheva, E. Métais, F. Meziane (Eds.), Natural Language Processing and Information Systems, Springer International Publishing, Cham, 2018, pp. 15–27.
- [13] A. Tagarelli, A. Simeri, Unsupervised law article mining based on deep pre-trained language representation models with application to the italian civil code, Artificial Intelligence and Law 30 (2021) 417–473. doi:10.1007/s10506-021-09301-8.
- [14] V. Bellandi, S. Castano, P. Ceravolo, E. Damiani, A. Ferrara, S. Montanelli, S. Picascia, A. Polimeno, D. Riva, Knowledge-based legal document retrieval: A case study on italian civil court decisions, in: D. Symeonidou, R. Yu, D. Ceolin, M. Poveda-Villalón, D. Audrito, L. D. Caro, F. Grasso, R. Nai, E. Sulis, F. J. Ekaputra, O. Kutz, N. Troquard (Eds.), Companion Proceedings of the 23rd International Conference on Knowledge Engineering and Knowledge Management, Bozen-Bolzano, Italy, September 26–29, 2022, volume 3256 of CEUR Workshop Proceedings, CEUR-WS.org, 2022. URL: <https://ceur-ws.org/Vol-3256/km4law2.pdf>.
- [15] F. Galli, G. Grundler, A. Fidelangeli, A. Galassi, F. Laggioia, E. Palmieri, F. Ruggeri, G. Sartor, P. Torroni,

<sup>6</sup><https://eur-lex.europa.eu/eli/dir/2019/790/oj>.

- Predicting outcomes of italian VAT decisions, in: JURIX, volume 362 of *Frontiers in Artificial Intelligence and Applications*, IOS Press, 2022, pp. 188–193. doi:10.3233/FAIA220465.
- [16] A. Galassi, F. Lagioia, A. Jablonowska, M. Lippi, Unfair clause detection in terms of service across multiple languages, *Artificial Intelligence and Law* (2024) 1–49. doi:10.1007/s10506-024-09398-7.
- [17] K. Drawzeski, A. Galassi, A. Jablonowska, F. Lagioia, M. Lippi, H. Micklitz, G. Sartor, G. Tagiuri, P. Torroni, A corpus for multilingual analysis of online terms of service, in: *NLLP@EMNLP*, Association for Computational Linguistics, 2021, pp. 1–8. doi:10.18653/v1/2021.nllp-1.1.
- [18] L. Ragazzi, G. Moro, S. Guidi, G. Frisoni, Law-suit: a large expert-written summarization dataset of italian constitutional court verdicts, *Artificial Intelligence and Law* (2024) 1–37. doi:10.1007/s10506-024-09414-w.
- [19] D. Licari, P. Bushipaka, G. Marino, G. Comandé, T. Cucinotta, Legal holding extraction from italian case documents using italian-legal-bert text summarization, in: *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law, ICAIL '23*, Association for Computing Machinery, New York, NY, USA, 2023, p. 148–156. doi:10.1145/3594536.3595177.
- [20] G. Attanasio, P. Basile, F. Borazio, D. Croce, M. Francis, J. Gili, E. Musacchio, M. Nissim, V. Patti, M. Rinaldi, D. Scalena, CALAMITA: Challenge the Abilities of LAnguage Models in ITALian, in: *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, Pisa, Italy, December 4 - December 6, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024.
- [21] G. Grundler, P. Santin, A. Galassi, F. Galli, F. Godano, F. Lagioia, E. Palmieri, F. Ruggeri, G. Sartor, P. Torroni, Detecting arguments in CJEU decisions on fiscal state aid, in: G. Lapesa, J. Schneider, Y. Jo, S. Saha (Eds.), *Proceedings of the 9th Workshop on Argument Mining*, International Conference on Computational Linguistics, Online and in Gyeongju, Republic of Korea, 2022, pp. 143–157. URL: <https://aclanthology.org/2022.argmining-1.14>.
- [22] P. Santin, G. Grundler, A. Galassi, F. Galli, F. Lagioia, E. Palmieri, F. Ruggeri, G. Sartor, P. Torroni, Argumentation structure prediction in CJEU decisions on fiscal state aid, in: *ICAIL, ACM*, 2023, pp. 247–256.
- [23] D. Walton, C. Reed, F. Macagno, *Argumentation schemes*, Cambridge University Press, 2008.

## A. Translated Examples

### Argument Component classification.

Argument premise:

*It should be noted that viewing the inability to deduct expenses for individuals such as the plaintiff as state aid to public hospitals overlooks the indiscriminate accessibility of public hospital services for individuals registered with the National Health Service (SSN). In contrast, a self-employed healthcare professional may refuse to provide services as an ordinary contractor.*

Argument conclusion:

*Thus, the office recognized the VAT non-taxable nature of the exportation, thus considering there is no longer any grounds to proceed on the matter.*

### Premise Type classification.

Factual premise:

*Undoubtedly, the taxpayer appealed the first instance ruling, again representing that she could not appeal against the first instance decision due to force majeure.*

Legal premise:

*The cited case law, to which reference is made for its reasoning, has explicitly and positively addressed the conformity of Italian legislation with that of the European Union. This effectively refutes the defense's objection on this point, which requested the suspension of the proceedings and the referral of the issue to the European Court of Justice.*

### Argument Scheme classification.

Rule Scheme:

*In fact, it is true that under Articles 54 and 23 of Legislative Decree No. 546/1992, the defendant on appeal must come up for trial within 60 days from the day on which appeal was served.*

Precedent Scheme:

*The Administration “has the burden of proving and attaching the evidence on which the dispute is based, among which the absence of the minimum personal and instrumental equipment, the immediacy of the relationships (transferor/interposed invoicing provider and transferee/buyer), an overt unsuitability to carry out the economic activity and the mismatch between the transferors and the company involved in the transaction may be circumstantial.”*

Classification Scheme:

*In conclusion, given what has been said so far, “gold shop” can be defined as “business establishments that buy, trade or resell used objects of gold, precious metals or bearing precious stones and dispose of them in the form of material, scrap gold or precious metals to foundries or other companies specialising in the recovery of precious materials”. They deal only in finished products and may not purchase used jewelery gold, melt it down (for their account or by commissioning a third party) and dispose of the resulting finished product.*

Interpretative Scheme:

*It means that, in essence, the purpose of the right to be heard is to put the taxpayer in the position of being able to make his or her observations before the decision is made and, therefore, to ensure that the administration can take into account all the elements of the case in adopting (or not adopting) the measure or in giving this one content rather than another.*

Principle Scheme:

*In the European Union system, therefore, the right to be heard in tax matters is independent of the nature of the tax and must be applied whenever the administration on the basis of the documentation exhibited deems it necessary to give the same documentation an interpretation that differs from that given by the taxpayer, inviting him, as mentioned, to provide in the exercise of the right to be heard the reasons for his choice.*

## B. Translated Prompts

### Argument Component classification.

“Classify the following argumentative text as premise ‘prem’ or conclusion ‘conc’. A premise (prem) is a proposition that provides a reason or support for the argument. A conclusion (conc) is the statement that follows logically from the premise(s) and represents the final point being argued for. Text:”

### Premise Type classification.

“Classify the following premise as factual ‘F’, legal ‘L’ or both. Factual premises (F) describe factual situations and events, pertaining to the substance or the procedure of the case. Legal premises (L) specify the legal content (legal rules, precedents, interpretation of applicable laws and principles). The expected output is a list with all applicable labels. For example: [‘F’, ‘L’]. Text:”

### Argument Scheme classification.

“Classify the following legal premise as one or more of the following argumentative schemes: Rule, Prec, Class, Itpr, Princ. Rule: whether there is an explicit or implicit reference to an article of law or citation of the text of a certain article. Prec: whether there is a reference to a previous ruling of the Supreme Court or the Court of Justice of the European Union. Class: if there is a definition of a legal concept or its constituent elements. Itpr: if there is reference to one of the interpretative criteria contained in Article 12 of the prelegislations (literal, teleological, psychological, systematic) to the Civil Code. Princ: if there is a reference to a general principle of law (e.g. principle of proportionality). The expected output is a list with all applicable labels. For example: [‘Prec’, ‘Princ’, ‘Rule’]. Text:”

## C. Few-shot prompts

### Argument Component classification.

“Classifica il seguente testo argomentativo come premessa ‘prem’ o conclusione ‘conc’. Per premessa (prem) si intende una proposizione che fornisce una ragione o un supporto per l’argomentazione. Per conclusione (conc) si intende l’affermazione che segue logicamente dalle premesse e rappresenta il punto finale che viene argomentato.

Esempi:

Testo: Si osserva poi che ritenere che la mancata possibilità di detrazione a favore di soggetti come il ricorrente comporti un aiuto di Stato in favore degli ospedali pubblici, in quanto le perdite degli stessi vengono ripianate dalle USL e dalla Regioni trascura di considerare l’accessibilità, indiscriminata, ai servizi dei nosocomi pubblici da parte dei soggetti iscritti al SSN, rispetto a quella ad un libero professionista sanitario che, in quanto tale, ben potrebbe rifiutarsi di prestare i propri servizi al pare di un normale contraente

Risposta: prem

Testo: L’appello è infondato e va respinto

Risposta: conc

Testo: Va osservato che la motivazione dell’atto di accertamento non può esaurirsi nel rilievo dello scostamento, ma deve essere integrata con la dimostrazione dell’applicabilità in concreto dello ‘standard’ prescelto e con le ragioni per le quali sono state disattese le contestazioni sollevate dal contribuente. (cfr. Cass. S.U. 26635/2009, Cass. 12558/2010, Cass. 12428/2012, Cass. 23070/2012)



Risposta: prem

Testo: Dunque, l'ufficio ha riconosciuto la non imponibilità IVA delle cessioni all'esportazione, così cessando sul punto la materia del contendere

Risposta: conc

Testo: Risulta d'altronde dalle osservazioni scritte del governo spagnolo che quest'ultimo non riesce a discernere tale differenza ad un esame delle pertinenti norme dell'ordinamento spagnolo.

Risposta: prem

Testo: Il Collegio, esaminata l'eccezione preliminare svolta nel suo appello dall'Ufficio e relativa alla richiesta nullità della sentenza per mancata instaurazione del contraddittorio, la respinge

Risposta: conc

Testo: ”

### Premise Type classification.

“Classifica la seguente premessa come di fatto ‘F’, legale ‘L’ o entrambe. Le premesse di fatto (F) descrivono situazioni ed eventi fattuali relativi al caso di specie. Le premesse legali (L) specificano il contenuto giuridico (norme giuridiche, precedenti, interpretazione delle leggi e dei principi applicabili). L'output atteso è una lista con tutte le label applicabili. Ad esempio: ['F', 'L'].

Esempi:

Testo: Per i primi giudici nel caso di specie questa esenzione non poteva essere applicata perché la complessiva attività di ‘A’ srl era un'attività commerciale svolta in concorrenza con altre imprese operanti nel settore

Risposta: ['F']

Testo: In assenza di siffatti elementi, che in via presuntiva avrebbero potuto fare giungere questo giudice a conclusioni diverse in via logica, si deve confermare l'esito cui è giunta la commissione provinciale

Risposta: ['F']

Testo: Su questo si osserva che si deve condividere la circostanza dedotta dal giudice di prime cure per cui deve essere il contribuente, ove sia contestata la inerenza e verità della rappresentazione ricavabile dal documento contabile, a dare la dimostrazione della fondatezza e della correttezza del comportamento tenuto

Risposta: ['L']

Testo: L'Ufficio non potrà impedire ad un imprenditore, per esempio, di cedere immobili con prezzi

bassi onullti per ricavare liquidità a fronte di nuovi impegni, ma dovrà rilevare la condotta antieconomica dello stesso sulla base dell'utile di esercizio

Risposta: ['L']

Testo: Invero l'avviso di accertamento è fondato sul mancato rispetto, da parte del contribuente, nel calcolo del ROL, delle disposizioni dell'articolo 96, secondo comma, del TUIR, che ne definisce le modalità

Risposta: ['F', 'L']

Testo: La società ‘A’, per quanto previsto dall'art. 4, comma 18 del Regolamento CEE n. 2913/1992, riveste il ruolo di ‘dichiarante in Dogana’, soggetto passivo della obbligazione

Risposta: ['F', 'L']

Testo: ”

### Argument Scheme classification.

“Classifica la seguente premessa legale in uno o più dei seguenti schemi argomentativi: Rule, Prec, Class, Iptr, Princ. Rule: se esiste un riferimento esplicito o implicito a un articolo di legge o la citazione del testo di una norma. Prec: se esiste un riferimento ad una precedente pronuncia della Corte di Cassazione o della Corte di Giustizia dell'Unione Europea. Class: se c'è la definizione di un concetto giuridico o degli elementi costitutivi dello stesso. Iptr: se c'è il riferimento a uno dei criteri interpretativi contenuti all'art. 12 delle preleggi (letterale, teleologica, psicologica, sistematica) al codice civile. Princ: se c'è un riferimento espresso a un principio generale del diritto (es. principio di proporzionalità). L'output atteso è una lista con tutte le label applicabili. Ad esempio: ['Prec', 'Princ', 'Rule'].

Esempi:

Testo: Infatti, è ben vero che, ai sensi del combinato disposto dagli articoli 54 e 23 D.Lgs. n. 546/1992, il convenuto in appello deve costituirsi entro 60 giorni dal giorno in cui ricorso è stato notificato.

Risposta: ['Rule']

Testo: L'Amministrazione “ha l'onere di provare ed allegare gli elementi probatori su cui si fonda la contestazione, tra i quali possono rilevare, in via indiziaria, quali elementi sintomatici della mancata esecuzione della prestazione dal fatturante, l'assenza della minima dotazione personale e strumentale, l'immediatezza dei rapporti (cedente/prestatore fatturante interposto e cessionario/committente), una conclamata inidoneità allo svolgimento dell'attività economica e la non corrispondenza tra i cedenti e la società coinvolta nell'operazione”

Risposta: ['Prec']

Testo: In conclusione, per quanto fin qui esposto, i "compro oro" possono essere definiti come "esercizi commerciali che acquistano, commerciano o rivendono oggetti d'oro, di metalli preziosi o recanti pietre preziose usati e li cedono nella forma di materiale, di rottami d'oro o di metalli preziosi alle fonderie o ad altre aziende specializzate nel recupero di materiali preziosi". Trattano esclusivamente prodotti finiti e non possono, congiuntamente, acquistare oro da gioielleria usata, fonderlo (per proprio conto o con incarico a terzi) e cedere il prodotto finito ottenuto

Risposta: ['Class']

Testo: Si vuole dire, in sostanza, che la finalità del contraddittorio anticipato è quella di mettere il contribuente nella condizione di potere fare valere le proprie osservazioni prima che la decisione sia adottata e, quindi, di far sì che l'Amministrazione possa tener conto di tutti gli elementi del caso nell'adottare (o non adottare) il provvedimento ovvero nel dare a questo un contenuto piuttosto che un altro.

Risposta: ['Itpr']

Testo: Nell'ordinamento unionale, pertanto, il principio del contraddittorio in ambito tributario prescinde dalla natura del tributo e deve trovare applicazione ogni qualvolta l'amministrazione sulla base della documentazione esibita ritenga dovere dare alla stessa documentazione interpretazione diversa da quella data dal contribuente invitandolo, come detto fornire nel corso del contraddittorio le ragioni della propria scelta

Risposta: ['Princ']

Testo: In sintesi per esteroestensione si intende la fittizia localizzazione della residenza fiscale di un soggetto all'estero, in particolare in un Paese con un trattamento fiscale più vantaggioso di quello nazionale, che la giurisprudenza configura in termini di abuso del diritto riconosciuto, in via tendenziale, come principio generale anche nel diritto dei singoli Stati membri (v. Cass., Sez. Un., n. 30055 del 2008, secondo la quale il divieto di abuso del diritto si traduce in un principio generale antielusivo che trova fondamento, in tema di tributi non armonizzati, nei principi costituzionali di capacità contributiva e di progressività dell'imposizione).

Risposta: ['Prec', 'Class', 'Princ']

Testo: La denuncia, infatti, non codificata nel codice di procedura penale (a differenza della notizia di reato di cui all'articolo 347 c.p.p.), può definirsi come qualunque atto con il quale chiunque abbia notizia di un reato perseguibile d'ufficio ne informa il pubblico ministero o un ufficiale di polizia giudiziaria.

Risposta: ['Rule', 'Itpr', 'Class']

Testo: "