

# 2025 Activity Report

RESEARCH CENTRE: Inria Centre at Université Côte d'Azur  
IN PARTNERSHIP WITH: Université Côte d'Azur, CNRS

---

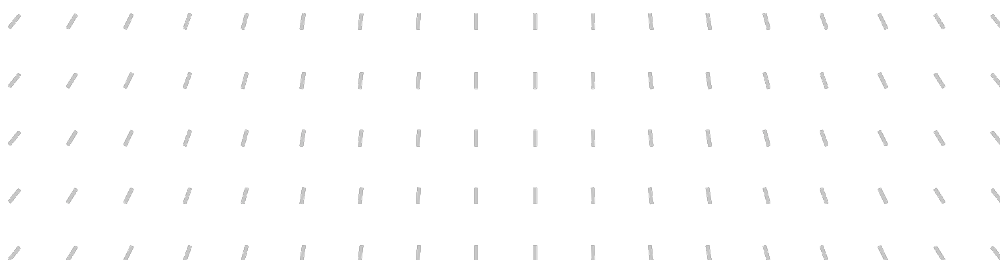
Project-Team

## MARIANNE

Models and data for computational argumentation in  
natural language

---

*In collaboration with* Laboratoire informatique, signaux systèmes de Sophia Antipolis  
(I3S)



## **Project-Team MARIANNE**

*Creation of the Project-Team: 2025 February 01*

Each year, Inria research teams publish an Activity Report presenting their work and results over the reporting period. These reports follow a common structure, with some optional sections depending on the specific team. They typically begin by outlining the overall objectives and research programme, including the main research themes, goals, and methodological approaches. They also describe the application domains targeted by the team, highlighting the scientific or societal contexts in which their work is situated. The reports then present the highlights of the year, covering major scientific achievements, software developments, or teaching contributions. When relevant, they include sections on software, platforms, and open data, detailing the tools developed and how they are shared. A substantial part is dedicated to new results, where scientific contributions are described in detail, often with subsections specifying participants and associated keywords. Finally, the Activity Report addresses funding, contracts, partnerships, and collaborations at various levels, from industrial agreements to international cooperations. It also covers dissemination and teaching activities, such as participation in scientific events, outreach, and supervision. The document concludes with a presentation of scientific production, including major publications and those produced during the year.

## Keywords

### Computer sciences and digital sciences

- A9.2.1. – Supervised learning
- A9.2.8. – Deep learning
- A9.4. – Natural language processing
- A9.6. – Decision support
- A9.8. – Reasoning
- A9.11. – Generative AI
- A9.14. – Evaluation of AI models
- A9.15. – Symbolic AI
- A9.16. – Societal impact of AI
- A9.17. – Cybersecurity and AI

### Other research topics and application domains

- B9.6.8. – Linguistics
- B9.6.10. – Digital humanities
- B9.9. – Ethics

## Contents

<b>Project-Team MARIANNE</b>	<b>1</b>
<b>1 Team members, visitors, external collaborators</b>	<b>5</b>
<b>2 Overall objectives</b>	<b>6</b>
2.1 Context: the need for automating natural language argument analysis . . . . .	6
2.2 Scientific objectives: mining, assessing and generating arguments . . . . .	6
<b>3 Research program</b>	<b>7</b>
<b>4 Application domains</b>	<b>12</b>
<b>5 Highlights of the year</b>	<b>13</b>
5.1 Awards . . . . .	13
<b>6 Latest software developments, platforms, open data</b>	<b>13</b>
6.1 Latest software developments . . . . .	13
6.1.1 HERACLES . . . . .	13
6.1.2 StreamETM . . . . .	14
6.1.3 ACTA . . . . .	14
6.1.4 PEACE . . . . .	14
6.1.5 DispuTool 2.0 . . . . .	15
6.1.6 MARIANNE-SAFE . . . . .	15
6.2 Open data . . . . .	15
<b>7 New results</b>	<b>16</b>
7.1 Argument Mining . . . . .	16
7.1.1 FALCON: A multi-label graph-based dataset for fallacy classification in the COVID-19 infodemic . . . . .	16
7.1.2 AM4DSP: Argumentation Mining in Structured Decentralized Discussion Platforms for Deliberative Democracy . . . . .	16
7.1.3 RooseBERT: A New Deal For Political Language Modeling . . . . .	17
7.1.4 Leveraging Graph Structural Knowledge to Improve Argument Relation Prediction in Political Debates . . . . .	17
7.1.5 Merging Embedded Topics with Optimal Transport for Online Topic Modeling on Data Streams . . . . .	18
7.1.6 Stick-Breaking Embedded Topic Model with Continuous Optimal Transport for Online Analysis of Document Streams . . . . .	18
7.1.7 A Topicality-Driven QUD Model for Discourse Processing . . . . .	19
7.2 Argumentation quality assessment and reasoning . . . . .	19
7.2.1 Fast Computing of Dung Semantics in Acyclic Probabilistic Argumentation Frameworks . . . . .	19
7.2.2 A Logic-based Framework for Decoding Enthymemes in Argument Maps involving Implicitness in Premises and Claims . . . . .	20
7.2.3 An Axiomatic Study of a Modular Evaluation of Enthymeme Decoding in Weighted Structured Argumentation . . . . .	20
7.2.4 Similarity Measures for First-Order Logical Arguments . . . . .	20
7.2.5 DataLens: Enhancing Dataset Discovery via Network Topologies . . . . .	21
7.2.6 Mining Implicit Arguments for Reasoning : A Survey . . . . .	21
7.2.7 Before the Outrage: Challenges and Advances in Predicting Online Antisocial Behavior . . . . .	22
7.2.8 Addressing Antisocial Behavior in Multi-Party Dialogs Through Multimodal Representation Learning . . . . .	22
7.2.9 Effectiveness of Counter-Speech against Abusive Content: A Multidimensional Annotation and Classification Study . . . . .	22

7.2.10	'Detectors Lead, LLMs Follow': Integrating LLMs and traditional models on implicit hate speech detection to generate faithful and plausible explanations . . . . .	23
7.2.11	DISPUTool 3.0: Fallacy Detection and Repairing in Argumentative Political Debates . . . . .	23
7.2.12	Repairing Fallacious Argumentation in Political Debates . . . . .	24
7.2.13	Contextualizing Toxicity: An Annotation Framework for Unveiling Pragmatics in Conversations of Online Discussion Forums . . . . .	25
7.3	Natural Language Argument Generation . . . . .	25
7.3.1	Beating Harmful Stereotypes Through Facts: RAG-based Counter-speech Generation . . . . .	25
7.3.2	Overview of the Critical Questions Generation Shared Task . . . . .	25
7.3.3	Argument generation for fact-checking . . . . .	26
7.4	Other contributions . . . . .	26
7.4.1	On Estimating the Strength of Differentially Private Mechanisms in a Black-Box Setting . . . . .	26
<b>8</b>	<b>Partnerships and cooperations</b>	<b>27</b>
8.1	International initiatives . . . . .	27
8.1.1	Associate Teams in the framework of an Inria International Lab or in the framework of an Inria International Program . . . . .	27
8.1.2	Participation in other International Programs . . . . .	27
8.2	International initiatives . . . . .	27
8.2.1	Visits of international scientists . . . . .	28
8.2.2	Visits to international teams . . . . .	28
8.3	European initiatives . . . . .	28
8.3.1	Horizon Europe . . . . .	28
8.4	National initiatives . . . . .	28
<b>9</b>	<b>Dissemination</b>	<b>29</b>
9.1	Promoting scientific activities . . . . .	29
9.1.1	Scientific events: organization . . . . .	29
9.1.2	Scientific events: selection . . . . .	30
9.1.3	Journal . . . . .	31
9.1.4	Invited talks . . . . .	32
9.1.5	Leadership within the scientific community . . . . .	32
9.1.6	Scientific expertise . . . . .	33
9.1.7	Research administration . . . . .	33
9.2	Teaching - Supervision - Juries - Educational and pedagogical outreach . . . . .	33
9.2.1	Teaching . . . . .	34
9.2.2	Supervision . . . . .	35
9.2.3	Juries . . . . .	35
9.3	Popularization . . . . .	36
9.3.1	Productions (articles, videos, podcasts, serious games, ...) . . . . .	36
9.3.2	Participation in Live events . . . . .	36
<b>10</b>	<b>Scientific production</b>	<b>36</b>
10.1	Major publications . . . . .	36
10.2	Publications of the year . . . . .	37

# 1 Team members, visitors, external collaborators

## Research Scientists

- Serena Villata [Team leader, CNRS, Senior Researcher, from Feb 2025, HDR]
- Victor David [INRIA, ISFP, from Feb 2025]
- Federica Granese [INRIA, Starting Research Position, from Feb 2025]

## Faculty Members

- Elena Cabrio [UNIV COTE D'AZUR, Professor, from Feb 2025]
- Anaïs Ollagnier [UNIV COTE D'AZUR, Associate Professor, from Feb 2025]

## Post-Doctoral Fellows

- Sofiane Elguendouze [CNRS, Post-Doctoral Fellow, from Feb 2025]
- Yingxue Fu [I3S, Post-Doctoral Fellow, from Feb 2025]

## PhD Students

- Greta Damo [UNIV COTE D'AZUR, from Feb 2025]
- Deborah Dore [UNIV COTE D'AZUR, from Feb 2025]
- Thi Thao Ha [I3S, from Nov 2025]
- Cyprien Michel-Deletie [ENS DE LYON, from Feb 2025]
- Nicolas Ocampo [UNIV COTE D'AZUR, from Feb 2025 until Apr 2025]
- Ekaterina Sviridova [UNIV COTE D'AZUR, from Feb 2025]
- Xiaoou Wang [CNRS, from Feb 2025 until Nov 2025]

## Technical Staff

- Mariana Eugenia Chaves Espinoza [UNIV COTE D'AZUR, Engineer, from Feb 2025 until Aug 2025]
- Theo Alkibiades Collias [UNIV COTE D'AZUR, from Feb 2025]

## Interns and Apprentices

- Hajar Bakarou [UNIV COTE D'AZUR, Intern, from Apr 2025 until Sep 2025]
- Loup Doinel [INRIA, Intern, from Mar 2025 until Sep 2025]
- Mohamed Sinane El Messoussi [UNIV COTE D'AZUR, Intern, from Apr 2025 until Sep 2025]
- Erwan Hain [INRIA, Apprentice, from Feb 2025]
- Nino Pireaud [UNIV COTE D'AZUR, Intern, from Apr 2025 until Oct 2025]
- Diego Zuniga Blassio [UNIV COTE D'AZUR, Intern, from Sep 2025]

## Administrative Assistant

- Delphine Robache [INRIA]

## 2 Overall objectives

### 2.1 Context: the need for automating natural language argument analysis

The field of computational argumentation is emerging as an important aspect of Artificial Intelligence (AI) research. The reason is based on the recognition that, if we are to develop robust intelligent systems, then it is imperative that they can handle incomplete and inconsistent information in a way that somehow emulates the way humans tackle such a complex task. Humans handle incomplete and inconsistent information by using argumentation either internally, by evaluating arguments and counterarguments, or externally, by for instance entering into a discussion or in a debate where arguments are exchanged.

Argumentation is an effective approach for solving various theoretical and practical problems, like explaining and justifying the decision making outcomes. The idea of “argumentation” as the process of creating arguments for and against competing claims, has been a subject of interest to philosophers and lawyers. In Computer Science, argumentation offers a novel framework shedding light on classical forms of reasoning, such as logical deduction, induction and abduction, interactive explanations, discussion and negotiation in computer-supported cooperative work, and learning.

Argumentation is the process by which arguments are constructed and handled: this means that arguments are compared, evaluated in some respect and judged to establish whether any of them is warranted. In computational models of argument, each argument is a set of assumptions that, together with a conclusion, is obtained by a reasoning process. In monological argumentation, a single agent reasons by constructing arguments to support and attack a conclusion to reach a decision, while, in dialogical argumentation, a group of arguers interacts to construct arguments supporting or attacking a particular claim and finally deliberate. An argumentation framework is essentially a directed graph in which the arguments are represented as nodes and the attack relation is represented by the arrows. This basic definition can be extended with different kinds of relations, e.g., support. A major step concerns the assessment of a set of arguments and their conclusions to establish their justification status, and therefore compute their acceptability degree. The assessment of the justification status of the arguments allows the agent to decide what to believe and what to do. Argumentation semantics provide formal criteria to determine which sets of arguments (i.e., extensions) can be regarded as collectively acceptable. Both qualitative and quantitative approaches have been proposed in the literature to assess the acceptance of an argument.

Argument(ation) mining (AM) is the research field in artificial argumentation aiming at automatically processing natural language arguments and reasoning upon them. It aims at extracting natural language arguments and their relations from text, with the final goal of providing machine-processable structured data for computational models of argument.

MARIANNE intends to carry out research at the interface of Natural Language Processing (NLP) and Argumentation, for modeling, mining, generating and reasoning upon argument structures from complex, uncertain, genre-specific, and multilingual textual data. One major goal of MARIANNE is to validate the scaling of the models produced by validating the main proposals through experimental implementations, in collaboration with researchers from other disciplines like Linguistics, Philosophy, Sociology and Law, to pursue high-impact research in AI.

### 2.2 Scientific objectives: mining, assessing and generating arguments

The MARIANNE project-team proposes and investigates NLP methods and algorithms for natural language argumentation to address the following three axes:

**Axis A - Argument mining:** The first research axis is the development of models and algorithms designed for mining natural language arguments from text. The Argument Mining (AM) scientific community has focused till now on the two main tasks constituting the AM pipeline, i.e., the detection of argumentative components (namely, evidence and claims), and the prediction of the relations of support and attack holding among them. They represent an obligatory starting step, but the resulting argumentation frameworks (i.e., the graph structure composed by the arguments as the nodes of the graph and the relations representing the links) are still quite simple with respect to the needs raised in the team application scenarios. Our goal is to enhance the extraction of machine-processable natural language argument structures to allow reasoning over complex real world natural arguments, with a focus on the English, French, Italian, Spanish and German languages.

**Axis B - Natural language argument quality assessment:** The second research axis is focused on the definition of computational methods to automatically assess the quality of natural language arguments. Despite a few existing approaches, the issue of automatically assessing the quality of an argumentation remains largely unexplored. On the one side, it consists in assessing the quality of the mined arguments to decide, for instance, whether a certain argument has to be selected for synthesizing a debate, or whether the overall debate is of good quality. On the other side, it consists in ensuring that the newly generated arguments satisfy the defined quality criteria in order to assess them from the qualitative point of view, i.e., a counter-argument to attack a fake news needs to be concise and without repetitions. The quality of the arguments is also characterized by formal properties of the argumentation graph, e.g., the argument strength, argument preferences, and argument acceptability.

**Axis C - Generation of natural language arguments:** In addition to the definition of more effective methods to mine (fine-grained) argumentative structures from text, and to automatically assess their quality, the third research axis of the team consists in the definition of new (generative and not generative) methods to generate natural language arguments, with a focus on English and French initially. This process is incremental and starts with the generation of single argumentative components towards the generation of arguments in the context of interactive dialogues with users. These dialogues are then employed in different use cases with different goals, i.e., explanation, counter-argumentation. These arguments are firstly generated starting from the mined arguments, and they rely on abductive reasoning schema based on the set of critical questions and reasoned responses necessary to reach the user’s understanding.

**Ethics.** Like all AI algorithms, the results of the technological solutions developed in MARIANNE may potentially be used in a misleading way. This potential danger concerns mainly the argumentation generation part, e.g., the generation of hateful and fallacious arguments. This is why a significant part of the research activity of MARIANNE is dedicated to the definition of computational solutions to “ensure” that the argumentation generated by our algorithms in human-machine conversations will not contain harmful or unfair content. This means to go beyond the pure inclusion of language guardrails, as in current Large Language Model (LLM) solutions, as the ethical foundation of an argumentation must be “controlled” on the discourse level too. The team members are already working on this topic, as witnessed by the recent publication at EMNLP 2024. In addition, the awareness of the ethical concerns around AI is high, thanks also to the participation of team members (Serena Villata) to national instances like the National Pilot Committee of Digital Ethics (CNPEN).

### 3 Research program

The MARIANNE project-team aims at developing natural language processing (NLP) methods and algorithms for argumentation in natural language. The research program of the team is organized around the following three axes:

**Axis A - Argument mining:** The first research axis will be the development of models and algorithms designed for mining natural language arguments from text.

**Axis B - Natural language argument quality assessment:** The second research axis will be focused on the definition of computational methods to automatically assess the quality of natural language arguments.

**Axis C - Generation of natural language arguments:** The third research axis of the team will consist in the definition of new (generative and not generative) methods to generate natural language arguments.

#### Axis A: Argument mining

##### Long term objectives

Current models are shown to learn linguistic patterns or cues rather than a real semantic understanding of the arguments. For instance, a model could predict a relation correctly, but it cannot explain the underlying reasoning of why the two linked components are in a relationship, since the required warrants are not explicitly mentioned in the text. Current classification algorithms can therefore be effective until a certain point, where the relation can be inferred from explicit mentions in the text. The team will therefore investigate methods to

inject both common sense knowledge and expert domain knowledge that humans exploit to carry out such kind of reasoning, in the form of knowledge bases or ontologies, to boost the performance of current neural methods applied to Argument Mining (AM) tasks. This long-term objective takes different forms depending on the application scenario: for the healthcare and the legal ones it consists in precise domain knowledge pointed out by domain experts, whilst in the disinformation, hate speech and the politics ones, it consists in a subtle mixture of precise domain knowledge (e.g., if we talk about vaccines or a new tax law) and common sense knowledge. This step is mandatory to boost the performances of argument mining models.

Thus far, we concentrated on the task of predicting intra-argument relations, i.e., those relations holding between the evidence and claims of a single argument. However, a necessary further step consists in moving from intra-argument relation prediction to inter-argument relation prediction. Establishing an argumentative link between two (or more) different arguments is mandatory to obtain a full argumentation framework (e.g., a graph representation of all the evidence and claims extracted from the clinical essays about Covid-19, where arguments are linked through argumentative relations) able to capture possible inconsistencies (e.g., self-attacking arguments, rebuttal on points supporting the candidate's viewpoint in political debates) and fallacious arguments (e.g., ad hominem attacks), that are at the basis of the research challenge on argument quality discussed below. We plan to tackle this issue through neural models like Graph Neural Networks (GNNs), which represent a natural choice given the graph-based structure of the argumentation, with a particular focus on the use of GNNs as a model of neural-symbolic computing, ensuring explainability.

Whilst a high-level classification of the argument components using the classes *evidence* and *claim* and of the relations with *attack* and *support* is required to identify the main argumentative elements present in the text, as soon as we move to precise application scenarios this general classification becomes insufficient and needs to be refined with more precise classes. Both for argumentative components and relations, the finer-grained classes strongly depend on the application domain (e.g., in the political domain we may have *critical discussion* and *strategic maneuvering* as argument classes, and *rebut* and *undercut* as relation classes). For argumentative relations, the idea is to start considering the two classical inter-argument relations in formal computational models of argument, namely rebut (i.e., the claim of an argument attacks the claim of another argument) and undercut (i.e., the claim of an argument attacks the premises of another argument).

Overall, in this first axis, the team will extend and empower neural AM models and algorithms to detect even more precisely both argument components and relations in the application scenarios of the team (i.e., medicine, politics, legal cases, hate speech, disinformation), with a focus on the English, French, Italian, Spanish and German languages. All the above mentioned objectives require to pass through (i) the definition of detailed annotation guidelines of high quality linguistic resources, (ii) the annotation of inter-argument relations and finer-grained classes of components/relations, and (iii) the definition of neural AM algorithms empowered with knowledge bases and domain ontologies. The team will also release a number of annotated linguistic resources for these tasks that will be made available to contribute to the advancements of the research community.

### Medium-term projects

In order to reach the long term objectives, we work on the following projects:

- *Mining argument misalignments and conflicts*: Another mid-term objective of MARIANNE will be the identification and classification of misalignments and overt conflicts that can hinder argumentative discussion and block democratic deliberation. We will design technology-enhanced dialogue spaces based on dispute mediation, that can be applied also in the context of social media discussions to detect and counter the use of abusive language. This innovative method, which scales up mediation techniques for large-scale dialogues, foresees two steps: (i) Conflict analysis: analysis of the main sources and the pattern of escalation of conflicting episodes in the use cases, based on argument mining methods (e.g., frequent debated issues, conflicting arguments) and argument-based qualitative analysis of discourses (which allows to see the dynamics of conflict escalation). The argument conflict analysis will uncover the real issues, recurring types of conflicting arguments that are based on different starting points, but also uncover real issues on which the discussion can proceed and points of commonalities. To achieve this goal, our objective will be to leverage the structural information present in the debates to enhance the performance of relation prediction between argument components. We will tackle this task by integrating Knowledge Graph Embedding models with Transformer-based models based on Pre-trained Large Language Models.

- *Mining implicit arguments*: In argumentation, it is crucial to comprehend the main arguments being put forward, the underlying reasoning, and how these arguments interrelate within a specific context. In this mid-term objective, we aim to develop systems capable of automatically extracting arguments and discerning their relationships. However, human argumentation is more intricate than mere argument extraction and lexical or semantic analysis. Argumentative units may not always be explicitly stated in discourse; instead, they are often connected through implicit inferences. These inferences depend on various factors, including background general knowledge, domain-specific expertise (as medical or juridical ones), subjective reasoning, or other patterns, which complicate argument comprehension and comprehensive analysis in the different domains. Therefore, our objectives are to investigate the role of external knowledge and other extra-linguistic features in elucidating implicit reasoning chains in argumentation, and to utilize these implicit elements to enhance existing methods for extracting arguments and predicting their relationships.
- *Mining arguments over time*: The way argumentation is carried out in public discourse, political debates and scientific communication changes over time. Another mid-term goal of the team will be to automatically explore the dynamics of inter and intra-argument structures over time. On the one side, the task will be to investigate through semi-supervised and unsupervised learning methods how the argumentation evolved over time. For instance, for political debates, we plan to start with the USElecDeb60To20 dataset we built, containing all the US presidential debates since 1960 to 2020. The objective is to study the temporal evolution trends of the argumentation, to see how the structure of the arguments evolved (e.g., number and fine-grade degree of the premises, presence of major claims, employment of rhetorical elements, choice of news events, change points). On the other side, the task will consist in the investigation of the dynamics of the argumentation in terms of attacks and supports among the candidates' arguments (i.e., graph level analysis of the argumentation). The final goal will be to assess if and how the dynamics of the argumentation impacted the outcome of the decision making process. For instance, this would allow us to learn from past argumentation dynamics to predict the results of the future elections in a country.

## Axis B: Natural language argument quality assessment

### Long term objectives

To automatically assess the quality of argument structures, we will consider first the following three main high-level dimensions to characterize arguments' quality: (i) Cogency: an argument should be seen as cogent if it has individually acceptable premises that are relevant to the argument's conclusion and that are sufficient to draw the conclusion; (ii) Rhetorical effectiveness: an argumentation should be seen as effective if it persuades the audience of the author's stance on the discussed issue. Besides the logical grounding of the actual arguments (*logos*), it also takes into consideration the credibility of the argument source (*ethos*) and the emotional force of the argumentation (*pathos*); (iii) Reasonableness: an argumentation should be seen as reasonable if it contributes to the resolution of the given issue in a sufficient way that is acceptable to everyone from the expected target audience. These three dimensions of argument quality will then be specified in further sub-dimensions, like the fact that a premise of an argument should be seen as relevant if it contributes to the acceptance or rejection of the argument's conclusion (i.e., if it is worthy considering it as a reason, evidence, or similar regarding the conclusion), or that an argumentation is successful in creating credibility if it conveys arguments in a way that makes the author trustworthy (e.g., by indicating the honesty of the author or by revealing the author's expertise regarding the discussed issue), and it should be seen as not successful if the opposite holds.

In addition to these three main argument quality dimensions, we will consider also two other elements, namely conceptual notions, and influence factors. Conceptual notions are the notions of maximal quality (based on arguing goals such as agreement and deliberation or on preferences between different arguments), and the notions of minimal quality, which represent what makes an argument valuable or appropriate to be stated as well as how to avoid fallacies. Influence factors are factors that influence the perception of quality beyond the content, structure, and style of the argument itself like argument-related factors (such as the argument's length, its structure in terms of relations between units, and revisions applied to it), and context-related factors (such as the domain of the discussion, the audience addressed, and the debaters involved).

These three dimensions takes into account both the content of the argument itself (cogency and rhetorical effectiveness) and the argumentation structure (reasonableness, considering both counter-arguments and their rebuttals), whilst maximal/minimal quality and the influence factors concerns more the context of the arguments to be assessed. To address the automatic quality assessment of the arguments, we will need to jointly consider the content of the textual arguments (i.e., by fine-tuning generalist or domain-specific Large Language Models) and the graph structure of the whole argumentation (i.e., through the generation of graph embeddings or the usage of Graph Neural Network models). These node-level features will be used to create graph level statistics useful for the reasonableness dimension in particular.

Finally, the same criteria proposed for assessing the quality of the mined arguments will be used to assess, as a long-term objective, the quality of the generated arguments. In addition, the argument generation will also be evaluated with diversity, which is of paramount importance, since verbatim repetition of arguments can become detrimental in explanation and counter-argumentation scenarios. Both *lexical diversity* and *semantic diversity* can be considered at this stage, where lexical diversity focuses on the variability of the generated arguments and can be captured by word overlapping metrics, and semantic diversity focuses on meaning and is harder to be captured, as in the case of generated arguments with similar meaning but different wordings. The model to generate arguments will be also evaluated along with standard metrics, i.e., BLEU and BertScore (computing a similarity score, using contextual embeddings, for each token in the candidate sentence with each token in the reference sentence) concerning the lexical and semantic generation performances.

### Medium-term projects

In order to reach the long term objectives, we work on the following projects:

- *Counter-narrative Quality Assessment*: This first mid-term objective aims to perform a quality assessment of Counter Narratives (CN). In particular, it focuses on the evaluation of the effectiveness of counter narratives against hate speech. Furthermore, it delves into fairness evaluation, proposing the development of metrics to assess representation, and inclusivity, along with employing bias detection techniques to mitigate biases in counter narratives. The research will be conducted on publicly available CN datasets, created both by experts from Non-Governmental organisations (NGOs) or collected directly from user-generated messages on social media. To assess effectiveness, various dimensions have to be considered. These include: the importance of emotional appeal, the analysis of argumentative structures and identification of argumentative language. Additionally, we will consider the need for soundness, target specificity, and complexity analysis in counter narratives. To assess fairness, we will define metrics for evaluating counter narratives against hate speech, focusing on avoidance of harmful stereotypes or biases, and appropriate language. The automatic quality assessment of CN will be addressed through a novel neural model including word embeddings, sentiment and emotion features, and pre-trained LLMs. Furthermore, this task addresses the differences in style between expert-generated and user-generated counter narratives, emphasizing the need for research to understand their impact on effectiveness and fairness.
- *Fallacious and Fake Argument Assessment*: Fallacies are arguments that employ faulty reasoning. Given their persuasive and seemingly valid nature, fallacious arguments are often used in political debates. Employing these misleading arguments in politics can have detrimental consequences for society, since they can lead to inaccurate conclusions and invalid inferences from the public opinion and policymakers. Automatically detecting and classifying fallacious arguments represents therefore a crucial challenge to limit the spread of misleading or manipulative claims, and promote a more informed and healthier political discourse. In this mid-term objective, we aim to go beyond the six categories of fallacious arguments we considered thus far (i.e., Ad Hominem, Appeal to Authority, Appeal to Emotion, False Cause, Slippery Slope, and Slogans), in order to address the two-fold task of fallacious argument detection and classification through defining neural network architectures based on Transformers models, combining text, argumentative features, and engineered features. More precisely, we will tackle challenging fallacy categories like causal ones (e.g., Common Cause, False Cause, Reversing Causation, Post-Hoc), where reasoning and knowledge-based features are jointly required to identify the fallacy. Finally, we also target the usage of AM methods to provide an argumentative

representation of fact-checkers' justifications to improve the precision and explainability of fake news classification systems.

## Axis C: Generation of Natural Language Arguments

### Long term objectives

The two last long term objectives are related to the generation of natural language arguments.

More precisely, we plan to start with the *extractive* argument generation, following the example of extractive summarization approaches. The idea is to select the main arguments identified in the text during the mining phase (Axis A), and generating these arguments verbatim producing a kind of argument-based summary of the original text. To select the main arguments to generate, we will exploit the quality criteria defined in Axis B. We will incrementally work starting from the generation of claims to the generation of whole argumentative structures composed of evidence and claims linked by attack and support relations. Secondly, we plan to move to the *abstractive* argument generation. The argument generation models we target will employ a text planning module to conduct content selection over the mined arguments and specify the suitable language style at sentence-level in order to select the talking points to cover for each sentence to be generated, combined with a content generation module to produce the final fluent argument on a certain topic. For this task, we will fine-tune Large Language Models for general application scenarios and for our use cases (i.e., counter-argument generation to fight online hate speech and disinformation, argument-based natural language explanations). The generated arguments need to meet two criteria, namely quality (e.g., variability of the arguments, no repetitiveness) and quantity. The evaluation of the generated arguments will be addressed by employing standard metrics for natural language generation, i.e., ROUGE (a recall-oriented metric), BLEU (based on n-gram precision) and METEOR (measuring unigram precision and recall by considering synonyms, paraphrases, and stemming).

The second long-term objective is to propose and test a more concrete argument-based dialogue, based on the results of the previous step. First, the argument-based dialogue neural model selects the mined arguments (i.e., components and relations) supporting or attacking a given (stance about a) topic and generates new arguments using the mined ones. Then, a task-oriented dialogue model provides the generated arguments about a certain topic and allows the user to ask clarification questions. It is in charge of generating appropriate queries to the argumentation model, and of generating an answer. The argumentation can be evaluated along with three axes: 1) validity of the used inferences; 2) quality of the generated interactions; 3) epistemic quality of conclusions in a given domain. Measurable factors such as the time spent by each user interacting with the system or the number of repeated questions are taken as indicators of this model. A quantitative analysis of the dialogues will be also conducted, based on different configurations of the dialogue simulator.

This research axis can be questioned with respect to the hype on generative AI the scientific community is facing these days. In the team, in the three scientific axes we propose, we will make use of Large Language Models which we will fine tune for our precise tasks, and we will employ generative AI models like GPT, Gemini, Claude or Mistral for zero-shot or few-shot evaluations. As this issue concerns more an engineering (i.e., prompt engineering) contribution than a scientific one, in this research program, we did not develop further this part, even though we are obliged to address it to compare with state-of-the-art systems.

### Medium-term projects

In order to reach the long term objectives, we work on the following projects:

- *Argument-based XAI*: Explainable Artificial Intelligence (XAI) has the goal to explain the decisions made by an intelligent system. This topic has recently raised a lot of interest in the AI research community due to the need to explain machine decisions, in particular in sensitive application domains like medicine. AI models must be able to provide interpretable and robust explanations for their decisions, as well as to learn the best way to explain their predictions to humans. This popular issue has already been investigated by different communities in the field, and several approaches have been proposed to get a better understanding of AI models from a mathematical point of view. Current results emphasize that there is the need to generate explanations coping with human reasoning methods, such as argumentation, and that natural language explanations is one of the most challenging ways to materialize them. Our objective is to connect argumentation to XAI through the interactive generation

of arguments to help humans to get a better understanding of machine predictions. Arguments need not only to be rational, but “manifestly” rational, so that arguers can see for themselves the rationale behind the inferential steps taken. We will propose novel argument generation methods to generate natural language explanations targeting different domains (i.e., medicine, fake news, hate speech, law, policy making). The important feature we target for generating these natural language explanations is two-fold: on the one side, they must unveil the reasons behind a decision through manifestly rational arguments, and they need to consider the human feedback to improve the firstly generated explanation (i.e., adding more details, broaden the discussion, provide reference to institutional sources). The generated argument-based explanations will benefit the user in terms of justification (exposing the reasoning behind a decision, thus aiding the user to decide how much credence to give in it), user involvement (allowing the user to add her knowledge and inference skills to the overall decision process), and system acceptance (in that the system’s functionality is fully transparent and its suggestions are adequately justified).

- *Counter-Argument Dialogue Generation*: Existing resources highlight the proficiency of large language models in generating counter-arguments, for instance, against hate speech or fake news. However, they primarily focus on simple two-turn exchanges, whereas real-life interactions often involve multiple, complex dialogues. Furthermore, among those approaches, it still remains unclear which arguments are crucial for steering conversations toward positive shifts. To address these issues, our research aims to delve deeper into the structure and dynamics of counter-argumentation in human and machine-generated resources extended dialogues. Our objectives are four-fold: (i) explore the use and impact of argumentative structures (premise, conclusion, major claims, and attack/support relations) and rhetorical strategies (logos, pathos, and ethos) in longer conversations; (ii) examine how the generated argument structures may evolve across multiple turns; (iii) assess how well existing counter-arguments taxonomies align with effective argument generation methods; and (iv) develop novel methods for generating counter-arguments that guarantee these conversation shifts, through reinforcement learning with human feedback (RLHF).

## 4 Application domains

The team mainly focuses on the application domains in healthcare, politics, law, hate speech and disinformation, given that argumentation-based decision making is becoming increasingly prominent in such contexts, and the team members already have experience in these application scenarios.

In the **healthcare domain**, clinicians need frameworks supporting them in extracting relevant information from textual documents (in particular from clinical trials), that is subsequently presented in a structured way. Given the aptness of Argument Mining methods to automatically detect in text those argumentative structures that are at the basis of evidence-based reasoning applications, AM represents a valuable contribution in the healthcare domain.

**Political debates** and political programs offer a rare opportunity for citizens to compare the candidates’ positions on the most controversial topics of the campaign. Given their innate argumentative features, they represent another natural playground for Argument Mining methods. The ability of identifying argumentative components and predicting their relations in such a kind of texts opens the door to cutting-edge tasks like fallacy detection and counter-argumentation generation.

Given its intrinsic argumentative features, **legal text** represents another natural but challenging use case scenario. The idea is to define novel methods to automatically identify those arguments from such texts, connect them by semantic relations and enrich them through the matching with the concepts present in knowledge graphs, to aid the comparative analysis of jurists like in case law.

Two application scenarios are linked to defense, and more precisely, to information warfare, i.e., **disinformation** and **hate speech**. First, content moderation is not enough to limit the diffusion of misleading or fake information, as successfully recognizing online disinformation depends not only on understanding whether factual statements are true, but also on interpreting and critically assessing the reasoning and arguments provided in support of conclusions.

Second, there are already tools that use AI and allow the automatic detection of violent and harmful speech, but most of these tools do not integrate cultural and contextual dimensions, and do not go beyond the

identification of explicit hate speech, as we focus on in the team. Using natural language argumentation to constitute a counter-argumentation would also be beneficial.

## 5 Highlights of the year

- Serena Villata has been granted with an **ERC Consolidator grant 2025** for the PANDORA project. The project will start in June 2026.
- Serena Villata has been nominated member of the French Council for Artificial Intelligence and Digital Technology in July 2025.
- Victor David coordinated the new Inria Associate Team with UCL, in collaboration with Prof. Tony Hunter.
- Elena Cabrio has been nominated member of the AI Advisory Council of the company ENGIE Energy.
- Serena Villata has been nominated member of the French Advisory Committee on Digital Ethics in September 2025.
- Serena Villata has been nominated Scientific Director of the Interdisciplinary Institute of Artificial Intelligence (3IA Côte d'Azur) in November 2025.

### 5.1 Awards

- Greta Damo, Benjamin Ocampo and Serena Villata were awarded with the "Prix d'excellence" of the University Côte d'Azur in December 2025.

## 6 Latest software developments, platforms, open data

In MARIANNE, we identified the following privileged application fields, based on existing research activities of team members and on the local eco-system: Defense, Digital Humanities, Medicine.

The main objective of the software implementations, i.e., proofs of concept, developed in the team is to experimentally validate the results obtained and ease the transfer of the developed methodologies to industry. Most of these proofs of concept are released as Python packages or Web platforms.

MARIANNE team members conceived and designed the above mentioned tools, identified the requirements and provided the specification design, as well as the formal definition of the algorithms involved. The main proofs of concept developed in the team and that we aim to pursue are the following.

### 6.1 Latest software developments

#### 6.1.1 HERACLES

**Name:** Online Topic Change Point Detection

**Keywords:** Incremental clustering, Continual Learning

**Functional Description:** This script performs topic modeling on chunks of documents using BERTopic and detects change points in topic distributions over time using Online Change Point Detection (OCPD). It processes document chunks iteratively, updating the topic model with each new chunk, and identifies significant changes in topics over time.

**Contact:** Serena Villata

### 6.1.2 StreamETM

**Keywords:** Machine learning, Artificial intelligence, Natural language processing

**Scientific Description:** StreamETM is an application designed for dynamic topic modeling using the Embedded Topic Model (ETM). It processes streaming text data, merges topic models over time, and detects change points in topic distributions.

Features: - Dynamic Topic Modeling: Continuously update topic models with new data chunks. - Topic Merging: Merge new topic models with existing ones to maintain a coherent topic structure. - Change Point Detection: Detect significant changes in topic distributions over time using Online Change Point Detection (OCPD). - Preprocessing: Preprocess text data, including lemmatization, stopword removal, and frequency-based filtering.

**Functional Description:** StreamETM is an application designed for online topic modeling using the Embedded Topic Model (ETM). It processes streaming text data, merges topic models over time, and detects change points in topic distributions.

**News of the Year:** The software has been developed as part of the paper, Merging Embedded Topics with Optimal Transport for Online Topic Modeling on Data Streams (<https://arxiv.org/pdf/2504.07711>)

**Publication:** [hal-05468591](https://hal.archives-ouvertes.fr/hal-05468591)

**Contact:** Federica Granese

**Participant:** 4 anonymous participants

### 6.1.3 ACTA

**Name:** A Tool for Argumentative Clinical Trial Analysis

**Keywords:** Artificial intelligence, Natural language processing, Argument mining

**Functional Description:** Argumentative analysis of textual documents of various nature (e.g., persuasive essays, online discussion blogs, scientific articles) allows to detect the main argumentative components (i.e., premises and claims) present in the text and to predict whether these components are connected to each other by argumentative relations (e.g., support and attack), leading to the identification of (possibly complex) argumentative structures. Given the importance of argument-based decision making in medicine, ACTA is a tool for automating the argumentative analysis of clinical trials. The tool is designed to support doctors and clinicians in identifying the document(s) of interest about a certain disease, and in analyzing the main argumentative content and PICO elements.

**Release Contributions:** In 2024, ACTA has been integrated in a suite called ANTIDOTE, which collects the software results from the ANTIDOTE project (all partners). In 2025, ACTA has been optimised to improve the mining and analysis of natural language arguments from clinical texts.

**URL:** <http://antidote.i3s.unice.fr/acta/>

**Contact:** Serena Villata

### 6.1.4 PEACE

**Name:** Providing Explanations and Analysis for Combating Hate Expressions

**Keywords:** Hate Speech Detection, Generating Explanations, Implicit Hate Speech, Subtle Hate Speech

**Functional Description:** PEACE is a web tool conceived to support content moderators in exploring and evaluating implicit and subtle hate speech on social media. It comprises three main functionalities: i) the exploratory analysis of hate speech messages characteristics (exploration), ii) the prediction of hatefulness (detection), and iii) the explanation of system predictions (explanation).

These functionalities incorporate not only a binary classification of whether a message is hateful (including explicit, implicit, and subtle messages), with a detailed explanation in natural language that clarifies why a message is considered hateful and an exploratory analysis of the message characteristics.

**URL:** <https://3ia-demos.inria.fr/peace/>

**Contact:** Elena Cabrio

### 6.1.5 DispuTool 2.0

**Keywords:** Argument mining, NLP, Web API, Relation Extraction, Component Detection

**Functional Description:** DISPUTool 2.0 is an automated tool which relies on Argument Mining methods to analyze the political debates from the US presidential campaigns to extract argument components (i.e., premise and claim) and relations (i.e., support and attack), and highlight fallacious arguments. DISPUTool 2.0 allows also for the automatic analysis of a piece of a debate proposed by the user to identify and classify the arguments contained in the text. A REST API is provided to exploit the tool's functionalities.

**Release Contributions:** The new version of DispuTool of 2025 allows to "repair" a fallacious argument with its non-fallacious version.

**URL:** <https://3ia-demos.inria.fr/disputool/>

**Contact:** Serena Villata

### 6.1.6 MARIANNE-SAFE

**Name:** Structured Argumentation for Fact-checking with Explanations

**Keywords:** Artificial intelligence, Disinformation detection

**Functional Description:** - Generate argument-structured summaries based on a fact-checking article or a list of evidence. - Retrieve evidence and generate argument-structured summaries when the fact-checking article is not available. - Assign a truthfulness label to a news claim which is enriched with a summary.

**Release Contributions:** First version published in 2025.

**URL:** <https://3ia-demos.inria.fr/safe/>

**Contact:** Serena Villata

## 6.2 Open data

MARIANNE team members also released some curated datasets (i.e., manually annotated linguistic resources) which are freely released for the scientific advancement of the research community:

- USElecDeb60To20 ([annotation guidelines and data](#)): an annotated dataset collected from the website of the Commission on Presidential Debates, including all the transcripts of the presidential debates from Kennedy and Nixon in 1960 until Trump and Biden in 2020. Dataset statistics: 29521 argument components (16087 claims and 13434 premises), 25012 relations (3723 attacks and 21289 supports), and 2745 fallacious arguments. USElecDeb60To20: <https://github.com/pierpaologoffredo/ElecDeb60to20>
- AbstRCT: an annotated dataset of randomized controlled trials retrieved from the MEDLINE database via PubMed search. Dataset statistics: 4073 argument components (2808 evidence, 1265 claims), and 2601 argument relations (2259 supports, 342 attacks). Topics: neoplasm, glaucoma, hepatitis, diabetes, hypertension. AbstRCT: <https://gitlab.com/tomaye/abstrct>
- DART: a dataset consisting of tweets annotated with argumentation, over the topics Brexit and Grexit. Restricted access as dataset developed in the context of an industrial collaboration.

- NoDE: a benchmark of natural language argumentation from heterogeneous online content. Dataset statistics: Debatepedia/ProCon dataset (260 pairs divided into 140 supports and 120 attacks), Twelve Angry Men dataset (80 pairs divided into 25 supports and 55 attacks), and Wikipedia dataset (452 pairs divided into 215 supports and 237 attacks). NoDE: <http://www-sop.inria.fr/NoDE/>

## 7 New results

In 2025, MARIANNE team members achieved the following results.

### 7.1 Argument Mining

In this section, we present the main new results of the team on the first research axis.

#### 7.1.1 FALCON: A multi-label graph-based dataset for fallacy classification in the COVID-19 infodemic

**Participants:** Mariana Chaves, Elena Cabrio, Serena Villata.

**Keywords:** Argument mining, fallacy detection, disinformation

Fallacies are arguments that seem valid but contain logical flaws. During the COVID-19 pandemic, they played a role in spreading misinformation, causing confusion and eroding public trust in health measures. Therefore, there is a critical need for automated tools to identify fallacies in media, which can help mitigate harmful narratives in future health crises. We present two key contributions to address this task. First, we introduce FALCON [15], a multi-label, graph-based dataset containing COVID-19-related tweets. This dataset includes expert annotations for six fallacy types—loaded language, appeal to fear, appeal to ridicule, hasty generalization, ad hominem, and false dilemma—and allows for the detection of multiple fallacies in a single tweet. The dataset’s graph structure enables analysis of the relationships between fallacies and their progression in conversations. Second, we evaluate the performance of language models on this dataset and propose a dual-transformer architecture that integrates engineered features. Beyond model ranking, we conduct statistical analyses to assess the impact of individual features on model performance.

#### 7.1.2 AM4DSP: Argumentation Mining in Structured Decentralized Discussion Platforms for Deliberative Democracy

**Participants:** Sofiane Elguendouze, Erwan Hain, Elena Cabrio, Serena Villata.

**Keywords:** Argument mining, e-Democracy, Deliberation Platforms

**Collaborations:** Lucas Anastasiou (The Open University (UK)), Anna de Liddo (The Open University (UK))

Argument(ation) mining (AM) is the automated process of identification and extraction of argumentative structures in natural language. This field has seen rapid advancements, offering powerful tools to analyze and interpret complex and large discourse in diverse domains (political debates, medical reports, etc.). In this paper [20], we introduce an AM-boosted version of BCause, a large-scale deliberation platform. Figure 1 visualizes the overall architecture of the platform. The system enables the extraction and analysis of arguments from online discussions in the context of deliberative democracy, which aims to enhance the understanding and accessibility of structured argumentation in large-scale deliberation processes.

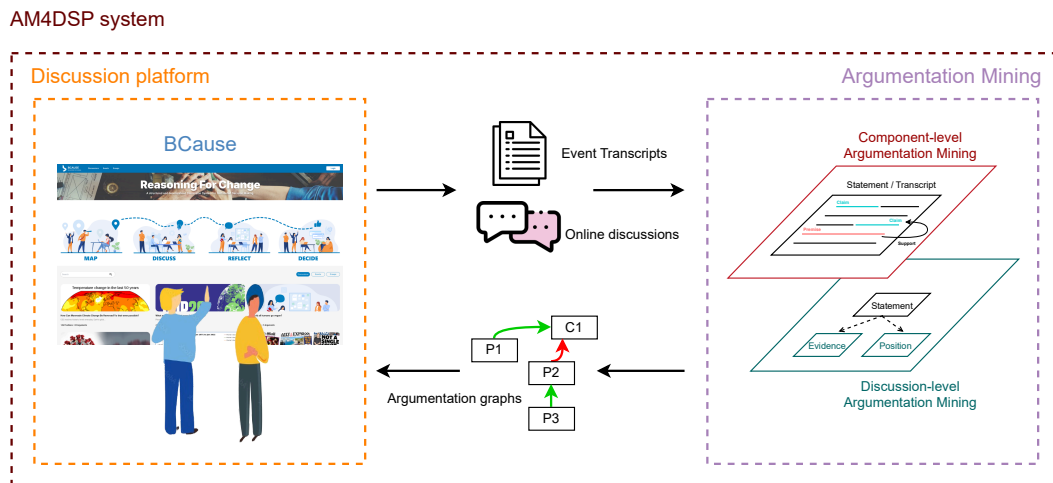


Figure 1: demonstrates the AM4DSP system. AM4DSP performs argumentation analysis on structured discussions in BCause, a large scale deliberation platform. See Subsection 7.1.2 for more details.

### 7.1.3 RooseBERT: A New Deal For Political Language Modeling

**Participants:** Deborah Dore, Elena Cabrio, Serena Villata.

**Keywords:** Argument mining, language modeling, political debates, sentiment analysis, stance detection

The increasing amount of political debates and politics-related discussions calls for the definition of novel computational methods to automatically analyze such content with the final goal of lightening up political deliberation to citizens. However, the specificity of the political language and the argumentative form of these debates (employing hidden communication strategies and leveraging implicit arguments) make this task very challenging, even for current general-purpose pre-trained Language Models. To address this issue, we introduce a novel pre-trained Language Model for political discourse language called RooseBERT. Pre-training a language model on a specialized domain presents different technical and linguistic challenges, requiring extensive computational resources and large-scale data. RooseBERT has been trained on large political debate and speech corpora (8K debates, each composed of several sub-debates on different topics) in English. To evaluate its performances, we fine-tuned it on four downstream tasks related to political debate analysis, i.e., stance detection, sentiment analysis, argument component detection and classification, and argument relation prediction and classification. Our results demonstrate significant improvements over general-purpose Language Models on these four tasks, highlighting how domain-specific pre-training enhances performance in political debate analysis. We release RooseBERT [34] for the research community.

### 7.1.4 Leveraging Graph Structural Knowledge to Improve Argument Relation Prediction in Political Debates

**Participants:** Deborah Dore, Serena Villata.

**Keywords:** Argument mining, political debates, knowledge graphs

**Collaborations:** Stefano Faralli (Sapienza University of Rome)

Argument Mining (AM) aims at detecting argumentation structures (i.e., premises and claims linked by attack and support relations) in text. A natural application domain is political debates, where uncovering the hidden dynamics of a politician’s argumentation strategies can help the public to identify fallacious and propagandist arguments. Despite the few approaches proposed in the literature to apply AM to political debates, this application scenario is still challenging, and, more precisely, concerning the task of predicting the relation holding between two argument components. Most of AM relation prediction approaches only consider the textual content of the argument component to identify and classify the argumentative relation holding among them (i.e., support, attack), and they mostly ignore the structural knowledge that arises from the overall argumentation graph. In this paper [19], we propose to address the relation prediction task in AM by combining the structural knowledge provided by a Knowledge Graph Embedding Model with the contextual knowledge provided by a fine-tuned Language Model (see Figure 2). Our experimental setting is grounded on a standard AM benchmark of televised political debates of the US presidential campaigns from 1960 to 2020. Our extensive experimental setting demonstrates that integrating these two distinct forms of knowledge (i.e., the textual content of the argument component and the structural knowledge of the argumentation graph) leads to novel pathways that outperform existing approaches in the literature on this benchmark and enhance the accuracy of the predictions.

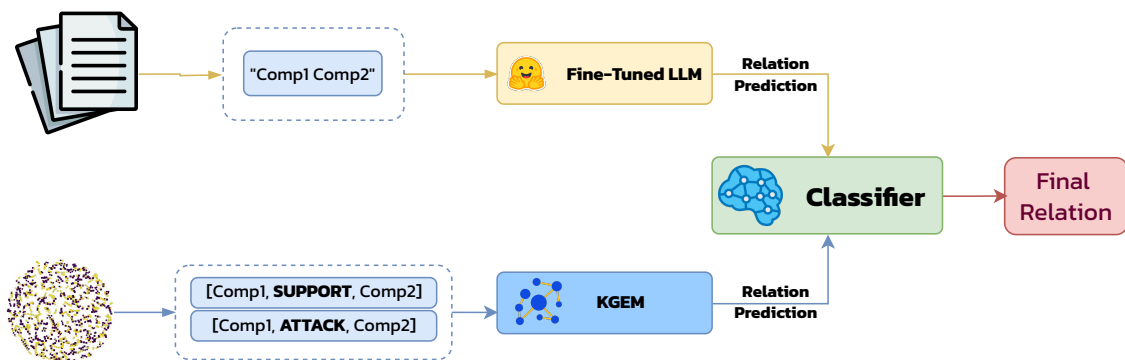


Figure 2: Architecture integrating Knowledge Graphs and Language Models for Argument Component Relation Prediction and Classification. For more details, see Subsection 7.1.4.

### 7.1.5 Merging Embedded Topics with Optimal Transport for Online Topic Modeling on Data Streams

**Participants:** Federica Granese, Serena Villata.

**Keywords:** Online Topic Modeling, Optimal Transport, NLP

**Collaborations:** Benjamin Navet (Université Côte d’Azur), Charles Bouveyron (Université Côte d’Azur)

Online topic models are unsupervised algorithms to identify latent topics in data streams that continuously evolve over time. Although these methods naturally align with real-world scenarios, they have received considerably less attention from the community compared to their offline counterparts. In this work, we introduce a novel approach to online topic modeling named StreamETM. This approach builds on the Embedded Topic Model (ETM) to handle data streams by merging models learned on consecutive partial document batches using unbalanced optimal transport. Additionally, an online change point detection algorithm is employed to identify shifts in topics over time, allowing for the detection of significant changes in the dynamics of text streams [24].

### 7.1.6 Stick-Breaking Embedded Topic Model with Continuous Optimal Transport for Online Analysis of Document Streams

**Participants:** Federica Granese, Serena Villata.

**Keywords:** Online Topic Modeling, Optimal Transport, NLP

**Collaborations:** Charles Bouveyron (Université Côte d’Azur)

In this work, we propose the Stick-Breaking Stream Embedded Topic Model (SB-SETM), an improved version of the StreamETM model presented in [24], for the online topic modeling setting. SB-SETM (i) leverages a truncated stick-breaking construction for the topic-per-document distribution, enabling the model to automatically infer from the data the appropriate number of active topics at each timestep; and (ii) introduces a merging strategy for topic embeddings based on a continuous formulation of optimal transport adapted to the high dimensionality of the latent topic space. Numerical experiments show that SB-SETM outperforms baselines on simulated scenarios. We extensively test it on a real-world corpus of news articles covering the Russian–Ukrainian war from 2022 to 2023 [30].

### 7.1.7 A Topicality-Driven QUD Model for Discourse Processing

**Participants:** Yingxue Fu, Anaïs Ollagnier.

**Keywords:** QUD, discourse modeling, discourse parsing, Rhetorical Structure Theory, Penn Discourse Treebank

**Collaborations:** Mark-Jan Nederhof (University of St Andrews)

Question Under Discussion (QUD) is a discourse framework that has attracted growing interest in NLP in recent years. Among existing QUD models, the QUD tree approach focuses on reconstructing QUDs and their hierarchical relationships, using a single tree to represent discourse structure. Prior implementation shows moderate inter-annotator agreement, highlighting the challenging nature of this task. In this paper, we propose a new QUD model for annotating hierarchical discourse structure [21]. Our annotation achieves high inter-annotator agreement: 81.45% for short files and 79.53% for long files of Wall Street Journal articles. We show preliminary results on using GPT-4 for automatic annotation, which suggests that one of the best-performing LLMs still struggles with capturing hierarchical discourse structure. Moreover, we compare the annotations with RST annotations. Lastly, we present an approach for integrating hierarchical and local discourse relation annotations with the proposed model.

## 7.2 Argumentation quality assessment and reasoning

In this section, we present the main new results of the team on the second research axis.

### 7.2.1 Fast Computing of Dung Semantics in Acyclic Probabilistic Argumentation Frameworks

**Participants:** Victor David.

**Keywords:** Probabilistic Argumentation, Algorithm, Complexity

**Collaborations:** Stefano Bistarelli (University of Perugia), Pierre Monnin (Inria), Francesco Santini (University of Perugia), Carlo Taticchi (University of Perugia)

This paper published in [14] presents fast and exact methods for computing the probability of an argument’s acceptance using Dung’s semantics in the Constellation paradigm of Abstract Argumentation. For (directed) Singly-Connected Graphs (SCGs), the problem can now be solved in linearithmic time instead of being exponential in the number of attacks, as reported in the literature. Moreover, in the more general

case of Directed Acyclic Graphs (DAGs), we provide an algorithm whose time complexity is linearithmic in the product of the out-degree of dependent arguments, i.e., arguments reaching the argument considered for acceptance through multiple paths in the graph. We theoretically show that this complexity is lower than the lower bound of the (exact) Constellation method, which is also supported by empirical results. Our approach to DAGs is also compared with the (approximate) Monte-Carlo method, which is stopped when exact results are obtained. Within this time constraint, Monte-Carlo still outputs significant errors, underlying the fast computation of our approach.

### 7.2.2 A Logic-based Framework for Decoding Enthymemes in Argument Maps involving Implicitness in Premises and Claims

**Participants:** Victor David.

**Keywords:** Enthymeme, Logical Argumentation, Argument mining

**Collaborations:** Anthony Hunter (UCL)

Argument mining is a natural language processing technology aimed at identifying the explicit premises and claims of arguments in text, and the support and attack relationships between them. To better understand and automatically analyze the argument maps that are output from argument mining, it would be desirable to instantiate the arguments in the argument map with logical arguments. However, most real-world arguments are enthymemes (i.e. some of the premises and/or claims are implicit), which need to be decoded (i.e. the implicit aspects need to be identified). A key challenge is to decode enthymemes so as to respect the support and attack relationships in the argument map. To address this, we present a novel framework [18], based on default logic, for representing arguments including enthymemes. We show how decoding an enthymeme means identifying the default rules that are implicit in the premises and claims. We then show how choosing a decoding of the enthymemes in an argument map can be formalized as an optimization problem, and that a solution can be obtained using MaxSAT solvers.

### 7.2.3 An Axiomatic Study of a Modular Evaluation of Enthymeme Decoding in Weighted Structured Argumentation

**Participants:** Victor David.

**Keywords:** Enthymeme, Logical Argumentation, Axiomatization

**Collaborations:** Jonathan Ben-Naim (CNRS), Anthony Hunter (UCL)

An argument can be seen as a pair of premises and a claim they support. Human arguments are often approximate, with some premises left implicit, leading to an implicit inference of the claim, i.e., forming enthymemes. To better understand and use them, we must decode these approximate enthymemes, typically by identifying missing premises to make the inference explicit, and, as we propose, by also removing irrelevant content to improve argument quality in specific contexts. Often, multiple decodings of an enthymeme are possible. However, no formal method has yet been proposed for identifying higher-quality decodings. To pave the way, we introduce [13] six types of criteria for evaluating aspects of decodings. Then, we introduce the concept of a criterion measure, designed to evaluate decodings based on a specific criterion. In parallel, we define desirable properties for criterion measures, referred to as axioms, and we systematically evaluate our criterion measures with respect to them. Finally, we introduce the notion of quality measure that combine specific criterion measures to give an overall evaluation of the quality of decodings.

### 7.2.4 Similarity Measures for First-Order Logical Arguments

**Participants:** Victor David.

**Keywords:** Similarity Measure, First Order Logic, Argumentation

**Collaborations:** Jérôme Delobelle (Université Paris Cité), Jean-Guy Mailly (Université de Toulouse)

Similarity in formal argumentation has recently gained attention due to its significance in problems such as argument aggregation in semantics and enthymeme decoding. While prior work has focused on propositional logic arguments, we extend these approaches to First-Order Logic (FOL) arguments, enabling reasoning based on the similarity of arguments in more complex and realistic contexts. We present a comprehensive framework [17] for FOL argument similarity, including: 1. An extended axiomatic foundation for similarity measures, 2. A parametric model decomposed into four levels to efficiently evaluate structured knowledge, 3. A Tversky-based family of measures to instantiate these concepts, 4. A set of constraints ensuring well-behaved models that satisfy axioms, and 5. We introduce and analyze non-symmetric similarity measures in formal argumentation for the first time.

### 7.2.5 DataLens: Enhancing Dataset Discovery via Network Topologies

**Participants:** Anaïs Ollagnier.

**Keywords:** Dataset search, Network visualization, Faceted search

**Collaborations:** Aline Menin (Université Côte d’Azur)

The rapid growth of publicly available textual resources, such as lexicons and domain-specific corpora, presents challenges in efficiently identifying relevant resources. While repositories are emerging, they often lack advanced search and exploration features. Most search methods rely on keyword queries and metadata filtering, which require prior knowledge and fail to reveal connections between resources. To address this, we present DataLens, a web-based platform that combines faceted search with advanced visualization techniques to enhance resource discovery. DataLens offers network-based visualizations, where the network structure can be adapted to suit the specific analysis task. It also supports a chained views approach, enabling users to explore data from multiple perspectives. A formative user study involving six data practitioners revealed that users highly value visualization tools—especially network-based exploration—and offered insights to help refine our approach to better support dataset search.

### 7.2.6 Mining Implicit Arguments for Reasoning : A Survey

**Participants:** Ekaterina Sviridova, Elena Cabrio, Serena Villata.

**Keywords:** Argument mining, Implicitness

Argumentation is the process of creating arguments for and against competing claims. Computational argumentation involves different ways of analyzing and reasoning upon arguments and their relations. More precisely, Argument Mining is the research field aiming at automatically identifying and classifying argument structures in text. The research field is mainly focused on the extraction of explicit argument structures (i.e., claims and premises connected by support and attack relations). However, an even more challenging task consists in extracting implicit argument structures in text (e.g., enthymemes). These structures are particularly valuable to then address argument reasoning, e.g., on incomplete and uncertain information, to finally compute the set of acceptable arguments, i.e., argument justification and skepticism. In this paper [11], we present and compare current approaches and available datasets for the novel task of Implicit Argument Mining. Future work perspectives are discussed to pave the way to further studies in this direction.

### 7.2.7 Before the Outrage: Challenges and Advances in Predicting Online Antisocial Behavior

**Participants:** Anaïs Ollagnier.

**Keywords:** Antisocial behavior prediction, Systematic literature review, Abusive behavior

Antisocial behavior (ASB) on social media—including hate speech, harassment, and trolling—poses escalating challenges for platform safety and societal well-being. While prior research has largely focused on detecting harmful content after its appearance, predictive approaches seek to anticipate harmful behaviors—such as hate speech propagation, conversation derailment, or user recidivism—before they fully unfold. Despite growing scholarly attention, the field of ASB prediction remains structurally uncoordinated and conceptually diffuse, lacking a unified taxonomy or integrative synthesis. This paper [36] addresses this gap through a systematic review of 49 machine learning studies on ASB prediction published between 2010 and 2025. It introduces a coherent framework that organizes existing work into five core task types: early harm detection, harm emergence prediction, harm propagation prediction, behavioral risk prediction, and proactive moderation support. The analysis compares tasks by temporal framing, prediction granularity, and operational goals, while tracing methodological trends from classical machine learning to pre-trained language models. By consolidating fragmented research and clarifying conceptual boundaries, this review establishes a structured foundation for future work. It identifies key challenges—including dataset scarcity, limited benchmarks, and temporal drift—and calls for ASB-specific evaluation protocols, multimodal and pragmatically informed representations, and explainable human-in-the-loop systems to advance predictive moderation.

### 7.2.8 Addressing Antisocial Behavior in Multi-Party Dialogs Through Multimodal Representation Learning

**Participants:** Hajar Bakarou, Mohamed Sinane El Messoussi, Anaïs Ollagnier.

**Keywords:** Social networks, antisocial behavior, multi-party dialogs, multimodal representation learning

Antisocial behavior (ASB) on social media, including hate speech, harassment, and cyberbullying, poses growing risks to platform safety and societal well-being. Prior research has focused largely on networks such as X and Reddit, while multi-party conversational settings remain underexplored due to limited data. To address this gap, we use CyberAgressionAdo-Large [32], a French open-access dataset simulating ASB in multi-party conversations, and evaluate three tasks: abuse detection, bullying behavior analysis, and bullying peer-group identification. We benchmark six text-based and eight graph-based representation-learning methods, analyzing lexical cues, interactional dynamics, and their multimodal fusion. Results show that multimodal models outperform unimodal baselines. The late fusion model mBERT + WD-SGCN achieves the best overall results, with top performance on abuse detection (0.718) and competitive scores on peer-group identification (0.286) and bullying analysis (0.606). Error analysis highlights its effectiveness in handling nuanced ASB phenomena such as implicit aggression, role transitions, and context-dependent hostility.

### 7.2.9 Effectiveness of Counter-Speech against Abusive Content: A Multidimensional Annotation and Classification Study

**Participants:** Greta Damo, Elena Cabrio, Serena Villata.

**Keywords:** Hate speech, counter speech, social media

The research presents a novel computational framework for evaluating the effectiveness of counter-speech (CS) in mitigating online hate speech [16]. Grounded in linguistics, communication, and argumentation theories, the framework defines six core dimensions – Clarity, Evidence, Emotional Appeal, Rebuttal, Audience Adaptation, and Fairness – used to annotate 4,214 counter-speech instances from two benchmark datasets. A human-annotated resource was created and released to the community, enabling further research in this area. The study also proposes two classification strategies, multi-task and dependency-based models, which outperform standard baselines, achieving average F1 scores of 0.94 and 0.96, respectively, across expert- and user-written CS. Results highlight the interdependence of effectiveness dimensions and provide a structured approach for assessing counter-speech, with potential societal impact in reducing online hostility and supporting safer digital discourse.

### 7.2.10 ‘Detectors Lead, LLMs Follow’: Integrating LLMs and traditional models on implicit hate speech detection to generate faithful and plausible explanations

**Participants:** Nicolás Benjamín Ocampo, Greta Damo, Elena Cabrio, Serena Villata.

**Keywords:** Hate speech, explainability, Large Language Models

The journal paper addresses the challenge of detecting and explaining implicit hate speech (HS) on social media [9]. It proposes a novel framework that enhances Large Language Models (LLMs) for both classification and natural language explanation of hateful content, particularly targeting nuanced and implicit instances often missed by traditional methods. The approach combines binary classification (HS vs. Non-HS) with generated explanations, and further investigates the impact of incorporating information from BERT-based models on detection performance. Comprehensive evaluation on widely used datasets, the HateCheck and Implicit Hate Corpus datasets, demonstrates that LLMs coupled with explanations outperform classical detectors, improving both accuracy and interpretability. The work provides insights into how predictive and explanatory systems can support automated moderation and enhance understanding of implicit hateful messages.

### 7.2.11 DISPUTool 3.0: Fallacy Detection and Repairing in Argumentative Political Debates

**Participants:** Deborah Dore, Elena Cabrio, Serena Villata.

**Keywords:** Argument mining, political debates, fallacy detection, fallacy repairing

**Collaborations:** Pierpaolo Goffredo (ALTEN)

This paper introduces and evaluates a novel web-based application designed to identify and repair fallacious arguments in political debates. DISPUTool 3.0 [22] offers a comprehensive tool for argumentation analysis of political debate, integrating state-of-the-art natural language processing techniques to mine and classify argument components and relations (see Figure 3). DISPUTool 3.0 builds on the ElecDeb60to20 dataset, covering US presidential debates from 1960 to 2020. In this paper, we introduce a novel task which is integrated as a new module in DISPUTool, i.e., the automatic detection and classification of fallacious arguments, and the automatic repairing of such misleading arguments. The goal is to show to the user a tool which not only identifies fallacies in political debates, but also shows how the argument looks like once the veil of fallacy falls down. An extensive evaluation of the module is addressed employing both automated metrics and human assessments. With the inclusion of this module, DISPUTool 3.0 advances even more user critical thinking in front of the augmenting spread of such nefarious kind of content in political debates and beyond.

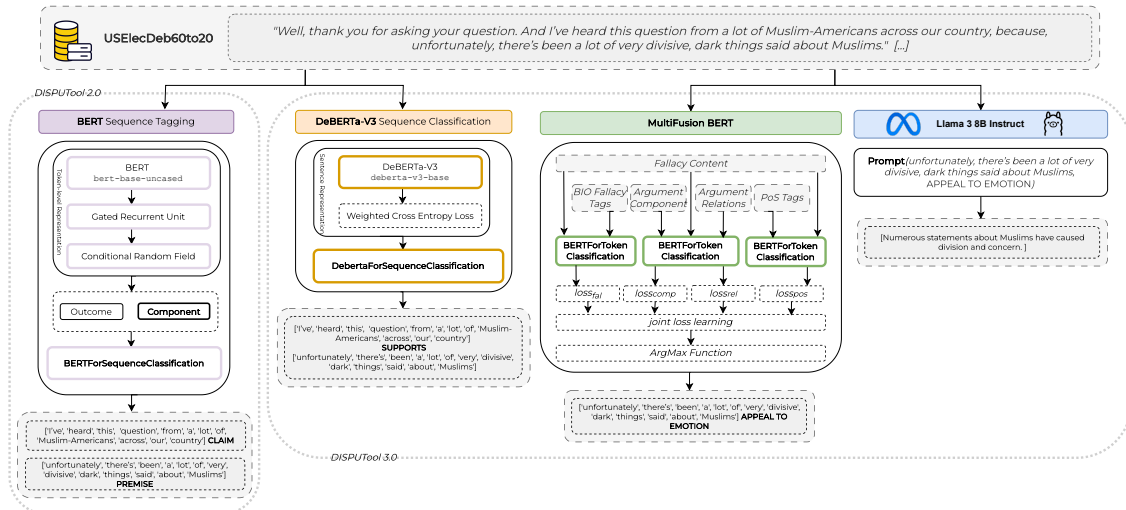


Figure 3: DispuTOOL 3.0’s Architecture. For mode details, see Subsection 7.2.11.

### 7.2.12 Repairing Fallacious Argumentation in Political Debates

**Participants:** Deborah Dore, Elena Cabrio, Serena Villata.

**Keywords:** Argument mining, political debates, fallacy detection, fallacy repairing

**Collaborations:** Pierpaolo Goffredo (ALTEN)

Fallacious arguments are defined as “invalid” arguments (e.g., the conclusion does not follow from the premises) or wrong moves in argumentative discourse. This kind of argumentation is therefore misleading or deceptive, in particular when employed in political debates. As the spreading of this nefarious content severely impacts the society and the decision-making of both citizens and policymakers, it is vital to prevent fallacious and propagandist arguments to circulate. To address this challenging task, several approaches proposing to identify fallacious argumentation in text have been presented in the literature. However, merely identifying this content is insufficient to ensure the audience realizes the impact of the fallacious argument on its deliberation process and to support the development of critical thinking skills. To tackle this challenging goal, it is necessary to unveil why a particular argument is fallacious and to demonstrate how it could be repaired as a valid, non-fallacious argument. In this paper [23], we address this key challenge by proposing a new task called repairing fallacious argumentation. The goal of this task is to modify statements that contain fallacious arguments into versions that are clearer, fairer, and free from any technique that could negatively persuade listeners. We carry out this task on political debates, where the need for this kind of solution is urgent. Our contribution in addressing this task is manifold: i) a novel dataset, FallacyFix, comprising repaired examples across various fallacy categories (Appeal to Fear, Appeal to Pity, Appeal to Popular Opinion, Flag Waving, and Loaded Language) based on the ElecDeb60to20-fallacy dataset; ii) modular prompt techniques for generating non-fallacious arguments, both dependent and independent of the specific fallacy label being addressed. Through an extensive evaluation, we assess these techniques using the most widely used Large Language Models (in Zero-Shot, Few-Shot, and Fine-Tuning settings) and a standard baseline model (BART); iii) a rigorous evaluation framework to assess the accuracy of the generated non-fallacious argument repairing the fallacy in the original argument, with respect to the manually annotated benchmark of non-fallacious arguments we built from the ElecDeb60to20 dataset; iv) a human evaluation of the generated non-fallacious arguments to assess the acceptability of these arguments across three dimensions, i.e., Relevance, Suitability, and Cogency. Future research will focus on integrating domain-specific knowledge to address complex fallacy categories, further analyzing language models’ behavior in countering fallacies, and exploring real-time fallacy repair methodologies. These efforts aim to enhance our ability to address fallacies dynamically in

various argumentation contexts, potentially improving the quality of public discourse and decision-making.

### 7.2.13 Contextualizing Toxicity: An Annotation Framework for Unveiling Pragmatics in Conversations of Online Discussion Forums

**Participants:** Yingxue Fu, Anaïs Ollagnier.

**Keywords:** pragmatics ; toxicity ; annotation ; Reddit conversations

The role of context has attracted increasing attention in research on toxicity detection. Interpreting toxic language remains a complex and multifaceted challenge, shaped by numerous linguistic, contextual, and social factors. However, current approaches often define “context” narrowly, focusing primarily on surface lexical cues such as hate lexicons, profanity markers, or sentiment polarity. These features, while useful, are insufficient to capture the interactional dynamics, user behaviors, and intentionality that shape such phenomena. To address this gap, this paper introduces a novel and systematic annotation framework [35], grounded in Speech Act Theory, aimed at deciphering the illocutionary and perlocutionary dimensions of conversation, which are unexplored in existing studies. We apply this framework to a new dataset of complete Reddit conversation threads, sampled to include discussions that turn toxic (124 conversations, 1990 messages). We evaluate the performance of GPT models (GPT-3, GPT-4, and GPT-5) on this challenging annotation task, providing insights into how large language models capture pragmatic and contextual dimensions of online toxicity.

## 7.3 Natural Language Argument Generation

In this section, we present the main new results of the team on the third research axis.

### 7.3.1 Beating Harmful Stereotypes Through Facts: RAG-based Counter-speech Generation

**Participants:** Greta Damo, Elena Cabrio, Serena Villata.

**Keywords:** Hate speech, counter speech, Retrieval-Augmented Generation

The study introduces a novel knowledge-grounded framework for automatic counter-speech (CS) generation [33]. The framework leverages advanced Retrieval-Augmented Generation (RAG) pipelines to produce trustworthy and factual counter-speech for eight target groups commonly affected by hate speech, including women, people of color, persons with disabilities, migrants, Muslims, Jews, LGBT persons, and others. A knowledge base of 32,792 texts from the United Nations Digital Library, EUR-Lex, and the EU Agency for Fundamental Rights supports the generation process. Evaluation on the MultiTarget-CONAN dataset, using both automated and LLM-based metrics, and human assessments, demonstrates that the framework outperforms standard large language model baselines and competitive approaches. The resulting system and knowledge base provide a reusable resource for research on effective and reliable counter-speech generation.

### 7.3.2 Overview of the Critical Questions Generation Shared Task

**Participants:** Ekaterina Sviridova, Elena Cabrio, Serena Villata.

**Collaborations:** Blanca Calvo Figueras (University of the Basque Country UPV/EHU), Jaione Bengoetxea (University of the Basque Country UPV/EHU), Maite Heredia (University of the Basque Country UPV/EHU), Rodrigo Agerri (University of the Basque Country UPV/EHU)

The proliferation of AI technologies has reinforced the importance of developing critical thinking skills. One promising direction involves leveraging Large Language Models (LLMs) to support the generation of critical questions: inquiries aimed at identifying weaknesses or fallacies in argumentative texts. The Critical Questions Generation (CQs-Gen) shared task [29] offers the first benchmark for this task. Thirteen participating teams explored diverse strategies for generating such questions, with the highest-performing system achieving 67.6 % accuracy, highlighting the challenge and complexity of the task. Notably, three of the four top-ranked submissions used argumentation scheme annotations to enhance performance. While most systems were based on open-weight LLMs, the two leading teams employed proprietary models.

### 7.3.3 Argument generation for fact-checking

**Participants:** Xiaou Wang, Elena Cabrio, Serena Villata.

**Keywords:** Argument generation, RAG, fact-checking, disinformation detection.

The widespread availability of the internet and social media platforms has fundamentally transformed how people consume and share information. This transformation has brought with it a significant challenge: online disinformation, i.e., namely, the intentional spread of false or misleading information. The speed and scale at which disinformation circulates present two major hurdles: (1) the development of automated fact-checking algorithms, and (2) the generation of explanations to accompany these systems. Explanations are crucial for several reasons: they help convince readers of the interpretation of evidence; they enable a feedback loop to correct judgment errors; and they reduce the risk of the “backfire effect,” where opaque predictions from black-box models actually strengthen belief in false claims. Currently, explanations are typically derived either from human-written fact-checking articles or from lists of evidence extracted from textual corpora. Meanwhile, the field of Argument Mining offers promising tools to enhance fact-checking. In the context of the ANR ATTENTION project, we enriched a subset of the LIAR-PLUS dataset with argumentative annotations, creating LIARArg [12] the first fake news classification dataset enhanced with argumentative layers. It includes 2,832 news titles with justifications, comprising 3,956 claims, 7,130 premises, and 8,205 argument relations. Moreover, we introduced a novel summarization method [27] that generates argument-structured explanations, significantly boosting the performance of top fact-checking algorithms across multiple datasets—including those with human-written fact-checks (LIAR-PLUS, FNC-1, Check-Covid) and those requiring automatic evidence retrieval (ExClaim). This retrieval-summarization framework, called SAFE [28], is also made available via a user-facing web interface. Our work bridges Argument Mining and automated fact-checking to create systems that are not only more effective, but also more transparent and persuasive. We showed that argumentative structures in evidence can improve both classification and explanation.

## 7.4 Other contributions

In this section, we summarize the new results obtained by team members on topics which are not related with the three main axes of MARIANNE.

### 7.4.1 On Estimating the Strength of Differentially Private Mechanisms in a Black-Box Setting

**Participants:** Federica Granese.

**Keywords:** Differential Privacy , Histogram-based Sampling, Impossibility of Provable Guarantees

**Collaborations:** Daniele Gorla (Sapienza University of Rome), Louis Jalouzet (CEA), Catuscia Palamidessi (Inria), Pablo Piantanida (CNRS)

We analyze to what extent final users can infer information about the level of protection of their data when the data obfuscation mechanism is a priori unknown to them (the so-called “black-box” scenario). In particular, we explore four notions of differential privacy. On the one hand, we prove that, without any assumptions on the underlying distributions, it is not possible to have an algorithm that can infer the level of data protection with provable guarantees. On the other hand, we demonstrate that, under reasonable assumptions (namely Lipschitzness of the involved densities on a closed interval), such guarantees exist for the local versions and can be achieved by a simple histogram-based estimator. We validate our results experimentally and note that, in two particularly well-behaved distributions (namely the Laplace and the Gaussian noise), our method performs better than expected, in the sense that in practice the number of samples needed to achieve the desired confidence is smaller than the theoretical bound, and the estimate of  $\epsilon$  is more precise than predicted [10].

## 8 Partnerships and cooperations

In this section, we summarize the main cooperation activities carried out within MARIANNE in 2025.

### 8.1 International initiatives

**Participants:** Victor David, Elena Cabrio, Serena Villata.

#### 8.1.1 Associate Teams in the framework of an Inria International Lab or in the framework of an Inria International Program

Victor David coordinated the activity of the Inria–UCL associated team EXPLAINER, which focuses on argumentation, natural language processing, logic, and neuro-symbolic approaches involving large language models. The team addresses the challenge of implicit arguments (enthymemes), which are pervasive in human communication but poorly handled by current argument mining and learning-based methods. By combining NLP with logical representations and automated reasoning, the goal is to enable principled analysis of argumentative structures, including consistency, validity, and relations between arguments. A central objective is the decoding and evaluation of enthymemes. Other team members working within the EXPLAINER team are Elena Cabrio and Serena Villata.

In 2025, the collaboration has already yielded strong outcomes, including two papers accepted at top-tier A\* conferences. Through the EXPLAINER team, funding was secured for a two-year postdoctoral position starting in January 2026, dedicated to enthymeme decoding using LLMs and argumentation schemes. In parallel, additional projects enabled the recruitment of a six-month intern followed by a six-month research engineer position, with a view toward a PhD on the persuasiveness of enthymemes. Further support was obtained through a one-year postdoctoral position funded by the IDEX AMI-Idées call of Université Côte d’Azur, focused on evaluating enthymeme decoding quality.

#### 8.1.2 Participation in other International Programs

Serena Villata co-led with Jaime Sichman (USP, Brasil) the UNBIAS Team (argUmentation and Norm Based Intelligent AgentS) within the IRC CNRS-USP ([UNBIAS website](#)). The scientific goal of the UNBIAS team is to incorporate AI argumentation and normative agents techniques to provide better interactions in social-technical systems.

### 8.2 International initiatives

**Participants:** Federica Granese.

### 8.2.1 Visits of international scientists

The collaboration with UCL within the EXPLAINER associated team was also strengthened by a one-month visit from Anthony Hunter in June-July 2025 (funded by the I3S lab), and by the submission of a JCJC ANR proposal directly linked to the EXPLAINER project.

### 8.2.2 Visits to international teams

Federica Granese participated in the first edition of the IVADO–Inria Initiative: Franco–Quebec Exchange Semester on Artificial Intelligence, a joint program designed to strengthen collaboration between AI research communities in France and Québec. The program provided financial support for a five-week research stay (mid-November to mid-December 2025) at the CNRS International Laboratory on Learning Systems (ILLS) and Mila – Quebec AI Institute (Montréal, Québec, Canada). The research project addressed emerging safety risks in multi-agent AI systems, with a particular emphasis on modeling manipulative interactions between AI agents. The project involved Federica Granese and Serena Villata (MARIANNE, Inria-I3S) and Pablo Piantanida (ILLS, Mila).

## 8.3 European initiatives

**Participants:** Elena Cabrio, Sofiane Elguendouze, Erwan Hain, Serena Villata.

### 8.3.1 Horizon Europe

The team participated to one European project in 2025.

- *ORBIS project (2022-2025)* ORBIS addresses the disconnects between ambitious ideas and collective actions at a large socio-technical scale. It responds to the profound lack of dialogue between citizenship and policy making institutions by providing a theoretically sound and highly pragmatic socio-technical solution to enable the transition to a more inclusive, transparent and trustful Deliberative Democracy in Europe. The project shapes and supports new democratic models that are developed through deliberative democracy processes; it follows a socio-constructive approach in which deliberative democracy is not a theory which prescribes new democratic practices and models, but rather the process through which we can collectively imagine and realize them. ORBIS provides new ways to understand and facilitate the emergence of new participatory democracy models, together with the mechanisms to scale them up and consolidate them at institutional level. It delivers: (i) a sound methodology for deliberative participation and co-creation at scale; (ii) novel AI-enhanced tools for deliberative participation across diverse settings; (iii) a novel socio-technical approach that augments the articulation between deliberative processes and representative institutions in liberal democracies; (iv) new evidence-based democratic models that emerge from the application of citizen deliberation processes; (v) demonstrated measurable impact of such innovations in real-world settings. The project builds on cutting-edge AI tools and technologies to develop a sustainable digital solution to e-deliberation. In this project, the MARIANNE team contributes with the development of argumentation-based solutions to analyze in a transparent manner the discussions and enhance decision making. The MARIANNE members involved in the project are: Sofiane Elguendouze, Erwan Hain, Elena Cabrio and Serena Villata. [Link to the project webpage](#)

## 8.4 National initiatives

**Participants:** Elena Cabrio, Mariana Chaves, Sofiane Elguendouze, Serena Villata, Xiaou Wang

The team participated to two ANR projects in 2025, ensuring the coordination of one of them.

- *ANR PRCE, Call 2021, "Artificial Intelligence" – Generating Counter Arguments to Fight Disinformation on the Web (ATTENTION) – (January 2022 - June 2026)*, total budget 522k€ – principal investigator: Serena Villata. The national research industrial project PRCE ATTENTION started on January 2022. The project includes the following partners, in addition to the CNRS - Laboratoire I3S (coordinator): CNRS - Centre Maurice Halbwachs, EURECOM, Université Paris 1 Sorbonne, the French start-up Buster.Ai. Objectives of the project: online disinformation is not a new phenomenon, but it has taken on an unprecedented scale, particularly during the acute health crisis. Social media limit the virality of disinformation mainly through content moderation. However, identifying disinformation and reporting its status is not enough to counter it. The ATTENTION project addresses this issue by designing intelligent ways to identify disinformation online and generate counter-arguments to fight the spread of such information online. The idea is to avoid the undesired effects of content moderation, e.g., overblocking, and to directly intervene in the discussion by engaging with people spreading incorrect information, through textual arguments that counter the fake content, and prevent it from spreading. A multidisciplinary perspective is adopted to ensure as a result AI solutions compliant with the ethical and sociological challenges of online disinformation. The PhD thesis of Xiaou Wang was funded on this project. Xiaou defended his PhD thesis on December 2025. The MARIANNE members involved in the project are: Xiaou Wang, Elena Cabrio and Serena Villata. [Link to the project webpage](#)
- *ANR ASTRID, Call 2022, "Artificial Intelligence" – Controversy and influence in the Ukraine war: a study of argumentation and counter-argumentation through Artificial Intelligence (CIGAIA) – (December 2022 - October 2026)*. Serena Villata is the **co-principal investigator** of the national research project ANR CIGAIA, which started on December 2022. The project includes the following partners, in addition to the CNRS - Laboratoire I3S: CREA Centre de Recherche de l'École de l'Air (coordinator). The proposed research is organized into two Axes. To understand the impact of controversies on armed conflict, the first axis studies the controversies in English and French taking place in the conflict between Russia and Ukraine. The second axis focuses on designing and implementing an Artificial Intelligence algorithm for the automatic analysis of argumentation in these controversies. This work is based on the joint creation of an annotated dataset from an initial mapping, allowing automatic argumentation analysis to identify and classify controversies. Building on the practical field of the conflict between Russia and Ukraine, this project addresses to the creation of a decision support tool, capable of mapping the actors through the identification of controversies and their characterizations by means of argumentation and counter-argumentative strategies. The MARIANNE members involved in the project are: Elena Cabrio and Serena Villata. [Link to the project webpage](#)

## 9 Dissemination

In this section, we summarize the main dissemination and popularization activities carried out by MARIANNE's members in 2025.

### 9.1 Promoting scientific activities

#### 9.1.1 Scientific events: organization

##### General chair, scientific chair

- Elena Cabrio was part of the Scientific Committee. Soph.IA Summit, 2025.

- Serena Villata was co-chair (together with Mauro Vallati) of the 14th Conference on Prestigious Applications of Intelligent Systems (PAIS-2025) which held in close association with the 28th European Conference of Artificial Intelligence (ECAI-2025) in Bologna during October 25-30, 2025.

#### Member of the organizing committees

- Victor David was a member of the organizing committee of the Workshop on Argumentation and Online Debates (ANR Project AGGREEY).
- Serena Villata co-organized the "Hybrid Argumentation and Responsible AI" seminar at Leibnitz Center (The Netherlands) on March 31 - April 4, 2025. This workshop proposed Hybrid Argumentation as an approach to responsible interaction between humans and AI systems. [Link to the workshop webpage](#)

#### Member of scientific boards and steering committees

- Elena Cabrio is an elected member of the Board of the scholarly association for Natural Language Processing (ATALA).
- Serena Villata is vice-president of the Steering Committee of COMMA (Computational Models of Arguments).
- Serena Villata is member of the Steering Committee of ECA (European Conference on Argumentation).
- Serena Villata is member of the IAAIL (International Association for Artificial Intelligence and Law) Executive Committee.

#### 9.1.2 Scientific events: selection

##### Chair of conference program committees

- Serena Villata was co-chair (together with Mauro Vallati) of the program committee of the 14th Conference on Prestigious Applications of Intelligent Systems (PAIS-2025).

##### Member of the conference program committees

- Elena Cabrio and Serena Villata have been members of the following program committees as:
  - Senior action editors, ACL Rolling Review (ARR)
  - Senior Program Chairs of: IJCAI (International Conference in Artificial Intelligence), AAAI (Conference of the Association for the Advancement of Artificial Intelligence), ECAI (European Conference in Artificial Intelligence)
- Elena Cabrio has also been Senior Area Chair for COLING 2025, track "Sentiment Analysis, Opinion and Argument Mining".
- Victor David was Senior Area Chair for KR-2025.
- Federica Granese was part of the program committee for *Prestigious Applications of Intelligent Systems (PAIS 2025)* in association with the 28th European Conference of Artificial Intelligence (ECAI-2025). Bologna, Italy, October 25-30, 2025.
- Serena Villata has also been Program Committee member for KR-2025 and SAC-2025.

## Reviewer

- Victor David was reviewer for AAAI-2025 and ECAI-2025.
- Federica Granese was reviewer for:
  - *ACM TheWebConf 2026 (Industry Track)*. Dubai, United Arab Emirates, April 13-17, 2026.
  - *International Conference on Artificial Intelligence and Statistics (AISTATS 2026)*. Tangier, Morocco, May 2nd – May 5th, 2026.
  - *International Conference on Learning Representations (ICLR 2026)*. Rio de Janeiro, Brazil, April 23-27, 2026.
  - *ACM Knowledge Discovery and Data Mining, Applied Data Science (ADS) Track Papers (KDD 2026, I/II Cycle)*. Jeju, Korea, August 9-13, 2026.

### 9.1.3 Journal

#### Member of the editorial boards

- Elena Cabrio is part of the editorial boards of:
  - Italian Journal of Computational Linguistics (IJCoL ISSN 2499-4553)
  - French journal Traitement Automatique des Langues (TAL).
- Serena Villata is part of the editorial boards of:
  - Artificial Intelligence and Law
  - Argument and Computation
  - Journal of Web Semantics

#### Reviewer - reviewing activities

- Elena Cabrio was:
  - Appointed Reviewer of the Computational Linguistics journal (two-year term, ending in June 2026).
  - Appointed Reviewer of the Transactions of the Association for Computational Linguistics journal (two-year term, ending in June 2026).
  - Reviewer for the AI and Law journal, Argument and Computation journal, Journal on Multimodal User Interfaces
- Victor David was reviewer for:
  - Journal Argumentation & Computation
  - EAAI: Engineering Applications of Artificial Intelligence
- Federica Granese was reviewer for:
  - IEEE Transactions on Dependable and Secure Computing.
  - IEEE Transactions on Information Forensics and Security.
- Anais Ollagnier was reviewer for:
  - Journal of Social Computing
  - Information Processing & Management
  - AI & Law Journal

#### 9.1.4 Invited talks

Federica Granese delivered the following invited talks:

- *When Pragmatic AI Agents Misalign: Safety Risks in Multi-Agent AI Systems – A Focus on Deceptive Behavior*. Inria–IVADO Closing Event, December 2–3, 2025, Montréal, Québec, Canada. Franco–Canadian Dialogue on AI. Round table on AI and Safety, November 28, 2025, Mila – Quebec AI Institute, Montréal, Québec, Canada; English.
- *When Pragmatic AI Agents Misalign: Safety Risks in Multi-Agent AI Systems – A Focus on Deceptive Behavior*. Invited seminar, International Laboratory on Learning Systems (ILLS), November 27, 2025, Montréal, Québec, Canada.

Serena Villata delivered the following invited talks:

- *Generative AI for fighting online abusive content: results and open challenges*, Séminaire IA Génératives: Promesses et Défis, CNRS, March 12th, 2025.
- *Argumentation for medicine: extracting arguments and generating explanations for a better human understanding*, Inria Meetup NLP Santé, April 8th, 2025.
- *Intelligence Artificielle : questions techniques et d'éthique*, Journée d'étude sur « l'impact de l'IA sur la pratique de l'enseignement universitaire » Académie Royale de Belgique, Bruxelles, September 29th, 2025.
- *L'IA contre les fake news et le cyberharcèlement*, Allianz Conference, Paris, October 16th, 2025.
- *Generative AI: technical and ethical challenges in game applications*, First Workshop on Artificial Intelligence for Human-Game Interaction, co-located with the 28th European Conference of Artificial Intelligence (ECAI '25) Bologna, October 25th-30th, 2025.

#### 9.1.5 Leadership within the scientific community

- Elena Cabrio is a chairholder of the 3IA Institute, project title: “Advanced Natural Language Understanding” (Chair renewed in 2025).
- Anais Ollagnier is member of the Academy 1 Board: Networks, Information, Digital Society (term 2025–2026).
- Anais Ollagnier is contributor to the European EMAI4EU project, as part of the EIT Digital Master School, a European master’s-level training program in deep-tech technologies, combining academic excellence, entrepreneurship, and innovation, and involving a network of university partners across Europe.
- Serena Villata is a chairholder of the 3IA Institute, project title: “Artificial Argumentation for Humans”.
- Since June 2025, Serena Villata is nominated member of the new Conseil de l’intelligence artificielle et du numérique **CIANum**. The Conseil is an independent instance reporting to the Minister responsible for artificial intelligence and digital technology. Its mission is to study all issues relating to the development of digital technology and artificial intelligence, as well as their impact on society, the economy, and local areas.
- Till June 2025, Serena Villata has been member of the scientific committee of the Direction Generale des Finances Publiques. The committee had the goal to discuss with the Direction Generale des Finances Publiques about their strategy and open issues about digital transformation and data science.

### 9.1.6 Scientific expertise

Elena Cabrio has been:

- Reviewer for the ERC Advanced Grant 2024 Call.
- Reviewer for Research Foundation Flanders (FWO).
- Reviewer for the scientific project call from COMUE Université de Toulouse.
- Expert tutor for the AI Projects of the Fondation de recherche pour l'aéronautique et l'espace and IRT ANTOINE SAINT EXUPERY.
- Member of the advisory committee of the MultiPOD projet (Multilingual and Multicultural Spaces for Political Deliberation), funded by the EU.
- Member of the AI Advisory Council of the company ENGIE Energy.
- Member of the Scientific Committee of ISTEEX.fr (CNRS), preparing the platform's application as a national research infrastructure.

Anais Ollagnier has been member of the selection committee for:

- the scientific projects AAP 2025 – DIM AI4IDF
- the ATALA thesis prize, 2025

Serena Villata has been member of:

- the recruitment committee for the Inria Paris CRCN-ISFP positions, 2025 (7 applications to review, participation to the jury meetings, oral interviews of the pre-selected candidates, deliberation).
- the HCRES committee for the LTCI Laboratory, 2025 (3 teams to evaluate, participation to the committee meetings, oral interviews of the teams, report writing).

### 9.1.7 Research administration

- Elena Cabrio is Appointed Member of the Academic Board of Université Côte d'Azur.
- Victor David managed the Inria–UCL associated team.
- Anais Ollagnier obtained and managed research and training funding, more precisely, the Call for Expressions of Interest in "Digital Transformation" for the Introduction to Deep Learning course, and a six-month internship grant (AAP MIDI 02) on the DataLens project (a web-based platform that integrates faceted search with advanced information visualization techniques to facilitate the search and exploration of machine learning (ML) datasets).
- Serena Villata has been Deputy Scientific Director of the Interdisciplinary Institute of Artificial Intelligence (3IA Côte d'Azur) till October 2025. The 3IA Côte d'Azur Institute was one of the four Interdisciplinary Institutes of Artificial Intelligence labeled in April 2019 by the Ministry of Higher Education, Research and Innovation, and it is now one of the nine French National Centers of Excellence in AI labeled "IA Cluster" (since May 2024). Since November 2025, Serena Villata took the responsibility of Scientific Director of the Institute.
- Federica Granese has been the organizer and coordinator of MARIANNE Journal Club (Exploring LLMs from the ground up, covering seminal papers, key advancements, and their limitations)<sup>1</sup>.

## 9.2 Teaching - Supervision - Juries - Educational and pedagogical outreach

The team is also highly involved in PhD supervision, academic teaching activities and participation to research juries. This section details the involvement of each team members in such activities.

---

<sup>1</sup>[Link to the MARIANNE journal club webpage](#)

### 9.2.1 Teaching

- Elena Cabrio delivered the following teaching hours: *Natural Language Processing*, Master 1 Computer Science. 30 hours (eq. TD); *Natural Language Processing*, Master 2 MIAGE - IA2: Applied Artificial Intelligence. 8 hours; *Introduction to Computational Linguistics*, Master 1 EUR CREATES, Parcours Linguistique, traitements informatiques du texte et processus cognitifs. 30 hours; *Introduction to Python*, Master 1 EUR CREATES, Parcours Linguistique, traitements informatiques du texte et processus cognitifs. 15 hours; *Natural Language Processing*, DUT IA et santé. 2 hours; *Natural Language Processing*, Lic. 3 Sciences et Technologies: Intelligence Artificielle. 11 hours; *Introduction to AI*. Licence 2. IUT. 22 hours; *Introduction to DataBases*. Licence 1. IUT. 90 hours; *Web interfaces*. Licence 1. IUT. 23.5 hours.
- Serena Villata delivered the following teaching hours: Master II Droit de la Création et du Numérique - Sorbonne University: Approche de l'Elaboration et du Fonctionnement des Logiciels, 15 hours (CM), 30 students; Master 2 MIAGE IA - University Côte d'Azur: I.A. et Langage : Traitement automatique du langage naturel, 28 hours (CM+TP), 30 students; Master 2 Algorithmic law and data governance - University Côte d'Azur: Introduction to AI, 12 hours (CM), 20 students; Master NeuroMod - University Côte d'Azur: Text Analysis, Deep Learning and Statistics, 15 hours (CM), 10 students.
- Anais Ollagnier delivered the following teaching hours: EUR DS4H, Enseignement mutualisé (S2 et S1), M1 Droit des affaires, UE Artificial Intelligence: Introduction to Deep Learning (KMUIAIUI) (CM : 12,5 h ; TD : 16 h) x2; EFELIA, M1 Anthropologie des arts vivants (S1), UE Artificial Intelligence and Societal Transformation (BMUIST1), CM : 12 h; EUR DS4H – Enseignement mutualisé, M1 Informatique – parcours informatique (S2), UE Individual Project S2 (EN) (KMUMP22U), TD : 2 h; Portail Sciences et Technologies – Enseignement mutualisé, PO1 Sciences, ingénierie, technologie, environnement (S2), UE INFO : Système 1 (SPUF201), TP : 60 h; EIT Digital Master School / EMAI4EU, Conception et dispensation de cours en Traitement Automatique des Langues. Cours : Introduction to Natural Language Processing (enseignement en ligne, public international, niveau master).
- Victor David delivered the following teaching hours: University Tutor for Internships and Apprenticeships, Université Côte d'Azur & Polytech Nice Sophia; Fundamentals of AI, Practical Sessions (18 hours), DUT2, IUT Nice Côte d'Azur (x 3 groups).
- Sofiane Elguendouze delivered the following teaching hours: Advanced NLP course - M2 Computer Science (27h): NLP basics, deep neural networks & backpropagation, recurrent neural networks, sequence to sequence models and attention, transformers, LLMs and prompt engineering; Introduction to AI - M1 applied foreign languages (15h): definition, demystification, introduction to machine learning, introduction to rule-based/statistical and neural NLP, introduction to language modeling; Applied AI - M2 applied foreign languages (15h): reminder about AI and machine learning, reminder about NLP, introduction to language models and prompting; Introduction to AI - M2 humanities and social sciences (15h): introduction to AI, introduction to machine learning, introduction to deep learning, introduction to NLP, language modeling and language models, ethical and societal limits and challenges of AI methods; Introduction to AI - M2 archaeology (12h): introduction to deep learning and computer vision, applied computer vision techniques to archaeological learning problems.
- Greta Damo: In 2025, as part of her teaching duty, Greta helped evaluating the group project presentations for the course I.A. et langage at the M2 Intelligence Artificielle Appliquée program (Semester 2 of the 2024/25 academic year), which focused on natural language processing topics and techniques, dedicating 5 hours to this activity. Additionally, she taught both theoretical and practical sessions for the Data Analysis and Visualization course at CentraleDigitalLab, an AI postgraduate program at Centrale Méditerranée, during the first semester of the 2025/26 academic year, totaling 22 hours.
- Ekaterina Sviridova: Ekaterina delivered 12 hours of TD (Travaux dirigés) for Master 2 students in Traitement automatique des textes (EUR CREATES), in the course HMEDAC3 – ECUE Annotation de corpus. The course introduced the role of annotated data in NLP, presented different types of

annotation tasks and tools, and included hands-on annotation sessions. It also covered post-processing and preprocessing of annotated data for language models, followed by a practical project in which students implemented the simple pipeline from raw data annotation to model training to address a specific application task.

- Deborah Dore delivered the following teaching hours: Data Analysis and Visualization (22H), Centrale Méditerranée, Theory and laboratory sessions for students from different backgrounds. The course focused on how data visualization can provide new insights and support data driven analysis; Traitement du Langage Naturel (24H), Université Côte d'Azur, MIAGE. Laboratory courses focused on exploring the field of natural language processing, from text preprocessing to the creation of AI models; Traitement du Langage Naturel (14H), Université Côte d'Azur, L3 Informatique. Laboratory courses focused on exploring the field of natural language processing, from text preprocessing to the creation of AI models.

### 9.2.2 Supervision

- Elena Cabrio and Anais Ollagnier co-supervised the doctoral thesis of Thi Thao Ha, “Toward Culturally and Contextually Grounded Language Models”.
- Elena Cabrio and Serena Villata co-supervised the PhD thesis of Greta Damo (counter-narratives against hate speech), Deborah Dore (political argumentation evolving in time), Ekaterina Sviridova (implicitness in argumentation), Cyprien Michel-Deletie (bridging natural language argumentation with formal reasoning), Benjamin Ocampo (implicit and subtle hate speech detection), Xiaoou Wang (argumentation-based fact checking).
- Victor David supervised the two Master’s level internships (5 and 6 months) of Loup Doinel and Nino Pireaud. Victor also supervised the activity of the research engineer Theo Alkibiades Collias in the context of the AGGREY ANR project.
- Anais Ollagnier supervised the scientific activity of post-doctoral research Yingxue Fu on “Improving argument mining through contextual information synthesis”.
- Anais Ollagnier also supervised the two Master’s level internships of Hajar Bakarou and Mohamed Sine El Messoussi.
- Serena Villata supervised Sofiane Elguedouze as post-doc on the CIGAIA and then the ORBIS project, on argument mining for online deliberation and dispute resolution, Mariana Chavez as engineer on the AI4MEDIA and then CIGAIA project, working on fallacy identification in online social media.

### 9.2.3 Juries

Elena Cabrio took part to the following juries:

- President of the jury for a tenure track position at Pompeu Fabra University, Spain.
- Jury member for a tenure track position at Fondazione Bruno Kessler, Italy.
- Reviewer of the PhD thesis of Arij Riabi, Sorbonne Université.
- Reviewer of the PhD thesis of Amir Reza Jafari, Institut-Mines Télécom, Télécom SudParis.
- Reviewer of the PhD thesis of Andrea Zaninello, Bolzen University, Italy.
- Examiner of the PhD thesis of Virgile Rennard, LIX, École Polytechnique.

Serena Villata took part to the following juries:

- Marta Marchiori Manerba, University of Pisa. Title of the thesis: "Fairness Auditing, Explanation and Debiasing in Linguistic Data and Language Models". Role: reviewer. PhD defense: 2025.

- Mustapha Bounoua, EURECOM. Title of the thesis: "Harnessing Multimodality: Diffusion based Generative Modeling and Information Estimation". Role: reviewer. PhD defense: 2025.
- Youri Peskine, EURECOM. Title of the thesis: "Using Knowledge Graph To Detect And Explain Misinformation Spread On The Web". Role: reviewer. PhD defense: 2025.

### 9.3 Popularization

As the research carried out in MARIANNE is highly interdisciplinary and that the targeted application scenarios tackle societal challenges, the team members have been solicited to participate to different popularization events and productions. This section details the contributions of the team on this matter.

#### 9.3.1 Productions (articles, videos, podcasts, serious games, ...)

Serena Villata has been interviewed by the newspaper Epsilon, in the article "La fin d'Internet" by Pierre-Yves Bocquet, March 26, 2025. [Link to the article](#).

#### 9.3.2 Participation in Live events

- Elena Cabrio took part in the following live events:
  - Panelist in the roundtable “The metamorphosis of democracy” at the World AI Cannes Festival (WAICF), 2025.
  - Festival des Sciences de Nice, 2025.
- Federica Granese took part to the live event "Attractivité de la France dans le domaine de la recherche en IA" (3 minutes interview, [link to the online event](#)).
- Serena Villata participated to the following popularization events:
  - Round table at the Conférence Intelligence Artificielle de l'Association OBJECTIF SCIENCE: "L'humanité à l'heure de l'IA : Entre Maîtrise & Dépendance (social, démographie, santé, éthique, démocratie, liberté, informations et communications...)", Isle sur la Sorgue, April 4th, 2025.
  - Round table at the Printemps des technologies 2025, March 2025, Centre de Congres Ville Saint Raphael, "Faut-il tout accepter de la techno ?".

## 10 Scientific production

### 10.1 Major publications

- [1] J. Ben-Naim, V. David and A. Hunter. ‘An Axiomatic Study of a Modular Evaluation of Enthymeme Decoding in Weighted Structured Argumentation’. In: *Proceedings of the 22nd International Conference on Principles of Knowledge Representation and Reasoning*. 22nd International Conference on Principles of Knowledge Representation and Reasoning. Melbourne, Australia, 2025. URL: <https://hal.science/hal-05422770>.
- [2] S. Bistarelli, V. David, P. Monnin, F. Santini and C. Taticchi. ‘Fast Computing of Dung Semantics in Acyclic Probabilistic Argumentation Frameworks’. In: *The 39th Annual AAAI Conference on Artificial Intelligence (AAAI 2025)*. Thirty-Ninth AAAI Conference on Artificial Intelligence, AAAI 2025. Philadelphia, United States, 25th Feb. 2025. URL: <https://inria.hal.science/hal-04876258>.
- [3] V. David and A. Hunter. ‘A Logic-based Framework for Decoding Enthymemes in Argument Maps Involving Implicitness in Premises and Claims’. In: *IJCAI 2025 - Thirty-Fourth International Joint Conference on Artificial Intelligence*. Montreal, Canada: International Joint Conferences on Artificial Intelligence Organization, 16th Aug. 2025, pp. 4445–4453. DOI: [10.24963/ijcai.2025/495](https://doi.org/10.24963/ijcai.2025/495). URL: <https://inria.hal.science/hal-05453430>.

- [4] S. Elguendouze, L. Anastasiou, E. Hain, E. Cabrio, A. de Liddo and S. Villata. ‘AM4DSP: Argumentation Mining in Structured Decentralized Discussion Platforms for Deliberative Democracy’. In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Suzhou, China: Association for Computational Linguistics, Nov. 2025, pp. 796–805. DOI: [10.18653/v1/2025.emnlp-demos.61](https://doi.org/10.18653/v1/2025.emnlp-demos.61). URL: <https://hal.science/hal-05414836>.
- [5] Y. Fu, M.-J. Nederhof and A. Ollagnier. ‘A Topicality-Driven QUD Model for Discourse Processing’. In: *ACL Anthology*. SIGDIAL 2025 - The 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue. Vol. 2025.sigdial-1.0. Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue. Avignon, France, 25th Aug. 2025, pp. 214–230. URL: <https://hal.science/hal-05288113>.
- [6] P. Goffredo, D. Dore, E. Cabrio and S. Villata. ‘DISPUTool 3.0: Fallacy Detection and Repairing in Argumentative Political Debates’. In: *ACL Anthology*. 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025). Vol. 3. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations). Vienne, Austria: Association for Computational Linguistics, 27th July 2025, pp. 472–480. DOI: [10.18653/v1/2025.acl-demo.45](https://doi.org/10.18653/v1/2025.acl-demo.45). URL: <https://hal.science/hal-05210504>.
- [7] F. Granese, B. Navet, S. Villata and C. Bouveyron. ‘Merging Embedded Topics with Optimal Transport for Online Topic Modeling on Data Streams’. In: *ECML PKDD 2025 - Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Vol. 16019. Lecture Notes in Computer Science. Porto, Portugal: Springer Nature Switzerland, 1st Sept. 2025, pp. 290–307. DOI: [10.1007/978-3-032-06109-6\\_17](https://doi.org/10.1007/978-3-032-06109-6_17). URL: <https://hal.science/hal-05468591>.
- [8] X. Wang, E. Cabrio and S. Villata. ‘SAFE: Structured Argumentation for Fact-checking with Explanations’. In: *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)*. 34th International Joint Conference on Artificial Intelligence (IJCAI-25). Montreal, Canada: International Joint Conferences on Artificial Intelligence, 16th Aug. 2025, pp. 11114–11118. DOI: [10.24963/ijcai.2025/1274](https://doi.org/10.24963/ijcai.2025/1274). URL: <https://hal.science/hal-05370780>.

## 10.2 Publications of the year

### International journals

- [9] G. Damo, N. B. Ocampo, E. Cabrio and S. Villata. ““Detectors Lead, LLMs Follow”: Integrating LLMs and traditional models on implicit hate speech detection to generate faithful and plausible explanations’. In: *Data and Knowledge Engineering* 162 (24th Nov. 2025). DOI: [10.1016/j.datak.2025.102535](https://doi.org/10.1016/j.datak.2025.102535). URL: <https://hal.science/hal-05386904> (cit. on p. 23).
- [10] D. Gorla, L. Jalouzot, F. Granese, C. Palamidessi and P. Piantanida. ‘On Estimating the Strength of Differentially Private Mechanisms in a Black-Box Setting’. In: *IEEE Transactions on Dependable and Secure Computing* 22.5 (Sept. 2025), pp. 5494–5507. DOI: [10.1109/TDSC.2025.3568160](https://doi.org/10.1109/TDSC.2025.3568160). URL: <https://hal.science/hal-05454091> (cit. on p. 27).
- [11] E. Sviridova, E. Cabrio and S. Villata. ‘Mining implicit arguments for reasoning: A survey’. In: *Argument and Computation* (30th June 2025). DOI: [10.1177/19462174251344764](https://doi.org/10.1177/19462174251344764). URL: <https://hal.science/hal-05435118> (cit. on p. 21).
- [12] X. Wang, E. Cabrio and S. Villata. ‘When automated fact-checking meets argumentation: unveiling fake news through argumentative evidence’. In: *Argument and Computation* 16 (18th Apr. 2025). DOI: [10.1177/19462174251330980](https://doi.org/10.1177/19462174251330980). URL: <https://hal.science/hal-05017906> (cit. on p. 26).

### International peer-reviewed conferences

- [13] J. Ben-Naim, V. David and A. Hunter. ‘An Axiomatic Study of a Modular Evaluation of Enthymeme Decoding in Weighted Structured Argumentation’. In: *Proceedings of the 22nd International Conference on Principles of Knowledge Representation and Reasoning*. 22nd International Conference on Principles of Knowledge Representation and Reasoning. Melbourne, Australia, 2025. URL: <https://hal.science/hal-05422770> (cit. on p. 20).

- [14] S. Bistarelli, V. David, P. Monnin, F. Santini and C. Taticchi. ‘Fast Computing of Dung Semantics in Acyclic Probabilistic Argumentation Frameworks’. In: The 39th Annual AAAI Conference on Artificial Intelligence (AAAI 2025). Thirty-Ninth AAAI Conference on Artificial Intelligence, AAAI 2025. Philadelphia, United States, 25th Feb. 2025. URL: <https://inria.hal.science/hal-04876258> (cit. on p. 19).
- [15] M. Chaves, E. Cabrio and S. Villata. ‘FALCON: A multi-label graph-based dataset for fallacy classification in the COVID-19 infodemic’. In: SAC 2025 - ACM/SIGAPP Symposium on Applied Computing. Catania, Italy, 2025. DOI: [10.1145/3672608.3707913](https://doi.org/10.1145/3672608.3707913). URL: <https://hal.science/hal-04834405> (cit. on p. 16).
- [16] G. Damo, E. Cabrio and S. Villata. ‘Effectiveness of Counter-Speech against Abusive Content: A Multidimensional Annotation and Classification Study’. In: *IEEE Xplore*. WI-IAT 2025 - 24th IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology. London, United Kingdom, 15th Nov. 2025. URL: <https://hal.science/hal-05353882> (cit. on p. 23).
- [17] V. David, J. Delobelle and J.-G. Mailly. ‘Similarity Measures for First-Order Logical Arguments’. In: *CEUR workshop*. NMR 2025 - 23rd International Workshop on Nonmonotonic Reasoning. Vol. CEUR-4071. Melbourne, Australia, 11th Nov. 2025. URL: <https://inria.hal.science/hal-05453458> (cit. on p. 21).
- [18] V. David and A. Hunter. ‘A Logic-based Framework for Decoding Enthymemes in Argument Maps Involving Implicitness in Premises and Claims’. In: IJCAI 2025 - Thirty-Fourth International Joint Conference on Artificial Intelligence. Montreal, Canada: International Joint Conferences on Artificial Intelligence Organization, 16th Aug. 2025, pp. 4445–4453. DOI: [10.24963/ijcai.2025/495](https://doi.org/10.24963/ijcai.2025/495). URL: <https://inria.hal.science/hal-05453430> (cit. on p. 20).
- [19] D. Dore, S. Faralli and S. Villata. ‘Leveraging Graph Structural Knowledge to Improve Argument Relation Prediction in Political Debates’. In: 12th Argument Mining Workshop. Proceedings of the 12th Argument Mining Workshop. Vienne, Austria: Association for Computational Linguistics, July 2025, pp. 74–86. DOI: [10.18653/v1/2025.argmining-1.7](https://doi.org/10.18653/v1/2025.argmining-1.7). URL: <https://hal.science/hal-05210499> (cit. on p. 18).
- [20] S. Elguendouze, L. Anastasiou, E. Hain, E. Cabrio, A. de Liddo and S. Villata. ‘AM4DSP: Argumentation Mining in Structured Decentralized Discussion Platforms for Deliberative Democracy’. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Suzhou, China: Association for Computational Linguistics, Nov. 2025, pp. 796–805. DOI: [10.18653/v1/2025.emnlp-demos.61](https://doi.org/10.18653/v1/2025.emnlp-demos.61). URL: <https://hal.science/hal-05414836> (cit. on p. 16).
- [21] Y. Fu, M.-J. Nederhof and A. Ollagnier. ‘A Topicality-Driven QUD Model for Discourse Processing’. In: *ACL Anthology*. SIGDIAL 2025 - The 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue. Vol. 2025.sigdial-1.0. Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue. Avignon, France, 25th Aug. 2025, pp. 214–230. URL: <https://hal.science/hal-05288113> (cit. on p. 19).
- [22] P. Goffredo, D. Dore, E. Cabrio and S. Villata. ‘DISPUTool 3.0: Fallacy Detection and Repairing in Argumentative Political Debates’. In: *ACL Anthology*. 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025). Vol. 3. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations). Vienne, Austria: Association for Computational Linguistics, 27th July 2025, pp. 472–480. DOI: [10.18653/v1/2025.acl-demo.45](https://doi.org/10.18653/v1/2025.acl-demo.45). URL: <https://hal.science/hal-05210504> (cit. on p. 23).
- [23] P. Goffredo, D. Dore, E. Cabrio and S. Villata. ‘Repairing Fallacious Argumentation in Political Debates’. In: Argumentation in the Digital Society: Proceedings of the 5th European Conference on Argumentation. Warsaw, Poland, 23rd Sept. 2025. URL: <https://hal.science/hal-05063601> (cit. on p. 24).

- [24] F. Granese, B. Navet, S. Villata and C. Bouveyron. ‘Merging Embedded Topics with Optimal Transport for Online Topic Modeling on Data Streams’. In: ECML PKDD 2025 - Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Vol. 16019. Lecture Notes in Computer Science. Porto, Portugal: Springer Nature Switzerland, 1st Sept. 2025, pp. 290–307. doi: [10.1007/978-3-032-06109-6\\_17](https://doi.org/10.1007/978-3-032-06109-6_17). URL: <https://hal.science/hal-05468591> (cit. on pp. 18, 19).
- [25] B. Molinet, E. Cabrio and S. Villata. ‘Assessing Argument-based Natural Language Explanations in Medical Text’. In: The 40th ACM/SIGAPP Symposium On Applied Computing (SAC 2025). Catania, Sicily, Italy, 31st Mar. 2025. URL: <https://hal.science/hal-05051047>.
- [26] N. B. Ocampo, E. Cabrio and S. Villata. ‘From Hidden to Harmful: Connecting Implicit and Explicit Hate Through Implied Statements’. In: WI-IAT 2025 - 24th IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology. London, United Kingdom, 15th Nov. 2025. URL: <https://hal.science/hal-05453245>.
- [27] X. Wang, E. Cabrio and S. Villata. ‘Leveraging Argumentation Schemes in Justification Generation for Automated Fact-checking’. In: The 24th IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT 2025). Londres, United Kingdom, 15th Nov. 2025. URL: <https://hal.science/hal-05370762> (cit. on p. 26).
- [28] X. Wang, E. Cabrio and S. Villata. ‘SAFE: Structured Argumentation for Fact-checking with Explanations’. In: *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)*. 34th International Joint Conference on Artificial Intelligence (IJCAI-25). Montreal, Canada: International Joint Conferences on Artificial Intelligence, 16th Aug. 2025, pp. 11114–11118. doi: [10.24963/ijcai.2025/1274](https://doi.org/10.24963/ijcai.2025/1274). URL: <https://hal.science/hal-05370780> (cit. on p. 26).

#### Conferences without proceedings

- [29] B. Calvo Figueras, J. Bengoetxea, M. Heredia, E. Sviridova, E. Cabrio, S. Villata and R. Agerri. ‘Overview of the Critical Questions Generation Shared Task’. In: Proceedings of the 12th Argument mining Workshop. Vienna, Austria: Association for Computational Linguistics, July 2025, pp. 243–257. doi: [10.18653/v1/2025.argmining-1.23](https://doi.org/10.18653/v1/2025.argmining-1.23). URL: <https://hal.science/hal-05449766> (cit. on p. 26).
- [30] F. Granese, S. Villata and C. Bouveyron. ‘Stick-Breaking Embedded Topic Model with Continuous Optimal Transport for Online Analysis of Document Streams’. In: 29th International Conference on Artificial Intelligence and Statistics (AISTATS). Tangier,, Morocco, 2026. URL: <https://hal.science/hal-05325837> (cit. on p. 19).

#### Doctoral dissertations and habilitation theses

- [31] N. B. Ocampo. ‘Unmasking implicit and subtle hate speech : NLP approaches for detecting and countering online harm’. Université Côte d’Azur, 5th June 2025. URL: <https://hal.science/tel-05247463>.

#### Reports & preprints

- [32] H. Bakarou, M. Sinane El Messoussi and A. Ollagnier. *Addressing Antisocial Behavior in Multi-Party Dialogs Through Multimodal Representation Learning*. 16th Oct. 2025. doi: [10.1145/nnnnnnn.nnnnn](https://doi.org/10.1145/nnnnnnn.nnnnn). URL: <https://hal.science/hal-05312340> (cit. on p. 22).
- [33] G. Damo, E. Cabrio and S. Villata. *Beating Harmful Stereotypes Through Facts: RAG-based Counter-speech Generation*. 14th Oct. 2025. URL: <https://hal.science/hal-05452643> (cit. on p. 25).
- [34] D. Dore, E. Cabrio and S. Villata. *RooseBERT: A New Deal For Political Language Modelling*. 7th Oct. 2025. URL: <https://inria.hal.science/hal-05450974> (cit. on p. 17).

- [35] Y. Fu and A. Ollagnier. *Contextualizing Toxicity: An Annotation Framework for Unveiling Pragmatics in Conversations of Online Discussion Forums*. 16th Dec. 2025. URL: <https://hal.science/hal-05416782> (cit. on p. 25).
- [36] A. Ollagnier. *Before the Outrage: Challenges and Advances in Predicting Online Antisocial Behavior*. 25th July 2025. URL: <https://hal.science/hal-05175061> (cit. on p. 22).
- [37] A. Ollagnier and A. Menin. *DataLens: Enhancing Dataset Discovery via Network Topologies*. 30th July 2025. URL: <https://hal.science/hal-05189348>.