European Lighthouse of AI for Sustainability

Deliverable number 2.1

Date: 2.1

First release of AI tools for a sustainable society

Project Title	ELIAS - European Lighthouse of AI for Sustainability
Contract No.	101120237
Start of Project	1 September 2023
Duration	48 months

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101120237.

Deliverable title	First release of AI tools for a sustainable society
Deliverable number	D2.1
Deliverable version	Final
Previous version(s)	N/A
Contractual date of delivery	August 31, 2024
Actual date of delivery	August 29, 2024
Deliverable filename	ELIAS_D2.1.pdf
Nature of deliverable	Report
Dissemination level	Public
Number of pages	67
Work Package	WP2
Task(s)	T2.1–T2.4
Parner responsible	UNIMI
Author(s)	Nicolò Cesa-Bianchi (UNIMI)
Editor	zzz (zzz)
Project Officer	Evangelia Markidou

Abstract	This deliverable provides the research results at M12 of the activities in Task 2.1 (Identifying use cases in different communities), Task 2.2 (Development of learning algorithms to protect and secure democracy), Task 2.3 (AI for inclusive and sustainable prosperity), and Task 2.4 (Design of algorithms for supporting efficient and coordinated use of resources). We present in detail the motivations, the developed methods, and the obtained results, including references to the publications and the software developed by the partners.
Keywords	Machine Learning, Sustainability, Computer Vision, Fairness, Large Language Models, Democracy and Prosperity, Fact-checking, Multi-agent Learning, Federated Learning

Copyright

Contributors

NAME	ORGANIZATION
Serge Belongie	UCPH
Raffaela Bernardi	UNITN
Elena Burceanu	BD
Matteo Castiglioni	POLIMI
Roberto Colomboni	POLIMI
Nicolò Cesa-Bianchi	UNIMI
David Alonso Del Barrio	Idiap
Daniel Gatica-Perez	Idiap
Laura Ferrarotti	FBK
César Hidalgo	ANITI
Nicola Gatti	POLIMI
Garima Gaur	Inria
Alberto Marchesi	POLIMI
Joao Pitacosta	JSI
Enver Sangineto	UMORE
Nicu Sebe	UNITN
Roberto Zamparelli	UNITN

Peer Reviews

NAME	ORGANIZATION
Bruno Lepri	FBK
Charlotte Laclau	IPP

The information and views set out in this report are those of the author(s) and do not necessarily reflect the official opinion of the European Union. Neither the European Union institutions and bodies nor any person acting on their behalf.

Table of Abbreviations and Acronyms

Abbreviation	Meaning
AAG	Adaptive Anchor Generator
AI	Artificial Intelligence
ART	Auxiliary Reconstruction Task
CAAPN	Consistency-Aware Anchor Pyramid Network
CDP	Carbon Disclosure Project
CMDP	Constrained Markov Decision Process
CNN	Convolutional Neural Network
CRL	Cascade Region Loss
CSR	Corporate Social Responsibility
ESG	Environmental Social and Governance
GCD	Generalized Category Discovery
LAM	Localizer with Augmented Matching
LLM	Large Language Model
MAE	Mean Absolute Error
MARL	Multi-Agent RL
MDP	Markov Decision Process
MIM	Masked Image Modeling
MSE	Mean Squared Error
NLP	Natural Language Processing
PCT	Political Compass Test
RL	Reinforcement Learning
RoI	Region of Interest
SDG	Sustainable Development Goals
TF-IDF	Term Frequency Inverse Document Frequency

Contents

List of Tables

List of Figures

1 Executive Summary

This deliverable provides the research results at M12 of the activities in Task 2.1 (Identifying use cases in different communities), Task 2.2 (Development of learning algorithms to protect and secure democracy), Task 2.3 (AI for inclusive and sustainable prosperity), and Task 2.4 (Design of algorithms for supporting efficient and coordinated use of resources). We present in detail the motivations, the developed methods, and the obtained results, including references to the publications and the software developed by the partners.

T2.1 (Identifying use cases in different communities) Significant efforts have been dedicated to identifying a major, concrete use case of paramount importance in industrial applications that necessitates foundational research. The chosen use case centers on cyber security in microservices infrastructures. In today's digital landscape, cyber security represents one of the most critical threats to our society. At the same time, microservices architectures are increasingly pervasive in industrial applications, presenting a multitude of unresolved challenges when confronted with cyber attacks. Our primary objective is the development and deployment of a comprehensive dataset that the community can utilize to enhance anomaly detection systems to face cyber attacks. This dataset aims to incorporate a variety of issues that are currently overlooked or underrepresented in existing resources. By addressing these gaps, we intend to provide a robust tool that supports advanced research and practical applications in safeguarding microservices infrastructures against evolving cyber threats.

T2.2 (Development of learning algorithms to protect and secure democracy). Several threads of research have been pursued in this task. In order to force fairness constraints while learning decision-making policies, the work [4] considers the framework of constrained episodic MDPs where the goal is to minimize regret and, simultaneously, minimize the number of constraint violations. A second contribution concerns crowd localization, where the goal is to predict the spatial position of humans in a crowd scenario. This is an essential components in systems for the analysis of crowd intention and behavior. The work [5] introduces a supervision target reassignment strategy for training to reduce ranking inconsistency between training and test, and propose an anchor pyramid scheme to adaptively determine the anchor density, where each anchor predicts a target coordinate offset and a probability of being a target. With the related goal of people tracking and human behavior understanding, the work [6] proposes a new per-object distance estimator, to estimate the distance of a target object from the camera when projected onto the image plane. In the context of understanding polarization and political divisiveness, [7] develops an estimate of the political divisiveness associated with a policy proposal, while [8] explores the creation of agents of augmented democracy using LLMs. Finally, the works [9] and [10] consider latent values and opinions expressed by LLMs when responding to queries, which can impact the users interacting with them. In trying to surface these values and opinions, LLMs are typically prompted to answer survey questions, but these works show that these answers are not just brittle, but can also be steered in terms of political and other biases.

T2.3 (AI for inclusive and sustainable prosperity). A first contribution in this task analyzes European cities' data from open data platforms to understand and assess their sustainability practices and help urban planners prioritize climate resilience and adaptation strategies in response to severe environmental problems. A second contribution introduces a tool that allows to interactively analyze how cities responded to the CDP-ICLEI dataset in order to find patterns, similarities or differences between cities, as well as to report on how a city is performing. This dataset contains cities' responses to a questionnaire, answering various issues related to sustainability, such as what targets are in place, what measures have been taken, what are the action and mitigation plans or what climate risks the various cities are facing. Additional contributions include: a method to estimate bilateral exports and imports for dozens of sectors starting from the corporate revenue data of large digital firms [11]; a machine learning method to estimate the GDP per capita of dozens of countries and hundreds of regions in Europe and North America for the past 700 years [12]; a method to estimate the contribution of famous immigrants, emigrants, and locals to the knowledge specializations of European regions based on data on more than 22,000 historical individuals born between the years 1000 and 2000 [2].

T2.4 (Design of algorithms for supporting efficient and coordinated use of resources). A first set of contributions in this task concerns learning in digital markets: in [13] we studied the problem of bidding in first-price auctions when the value of the auctioned item is discovered only if the auction is won; in [14] we explored how to achieve fairness in repeated bilateral trade, by rewarding the platform with the minimum of the net increase in seller and buyer utilities; in [15] we investigated a variant of the repeated bilateral trade problem where a broker interacts with a sequence of traders who do not have definite seller and buyer roles; instead, they decide to buy or sell their assets based on whether they perceive the brokerage price as too low or too high. Finally, a multi-agent learning setting with partial feedback is explored in [16] to characterize the optimal trade-off between feedback, communication, and regret. In [17], a personalized federated method for generalized category discovery (the task of categorizing unlabeled samples from unknown classes by leveraging labeled data of known classes) is proposed. This method aims not only to improve the personalized abilities of local models, but also to encourage the global model to learn more generic representations.

2 Introduction

Artificial Intelligence is playing a role of increasing importance in the creation and maintenance of a sustainable society. The double-edged nature of AI technologies—which can be used to both defend and undermine the democratic society—creates formidable challenges in which defenders and attackers engage in a confrontation based on a whole array of technologies, including game theory, machine learning and generative AI, signal and sensor processing, computer vision, mechanism design, information design, and others. This complex interplay can be investigated through the analysis of well-chosen use cases endowed with high-quality datasets enabling the study of key issues, including real-world applicability and performance benchmarking. As defenders and attackers are overwhelmingly relying on autonomous agents with learning abilities, an important ingredient in this context is the ability of understanding how a system of interacting learning agents behave, both in a cooperative and in a competitive setting. The design of robust learning agents, which can act acquiring information both from the environment and from the other agents, plays a vital role in this workpackage.

A further leading theme in this workpackage is the combination of the decision-making level with the information (e.g., high-level features) extracted from unstructured data, such as text, images, and video. While the contributions included in this deliverable treat these two levels (information extraction and decision making) separately, some of the use cases—such as cybersecurity of microservices infrastructures—will be designed to explore their integration.

The use of AI for the sustainability of society requires the mapping of opinions to understand polarization and divisiveness. The presence of these biases can then be searched in generative models, for the purpose of understanding their impact on the users and the extent to which the model is robust to manipulation. Alongside the activity on opinion mapping, we apply machine learning and data analysis techniques to map the territory in terms of regional diversification, considering activities and how they are interconnected across multiple geographic, historical, and social scales. This includes understanding patterns and mechanisms involved in sustainable economic development (trades, evolution of GDP, etc.) through the use of advanced machine learning.

In summary, the contributions included in this deliverable reveal the presence of multiple and interconnected levels that AI systems must combine and intervene on: a physical level of sensors and signals, a space of opinions, an economical and geographical level, and an environment defined by abstract features and decisions which is where AI agents operate. Much of the work in this workpackage will be devoted to the exploration of these interconnections for the design of robust and efficient policies to be tested on some concrete use cases.

3 T2.1 Identifying use cases in different communities

3.1 Overview

The objective of T2.1 is to identify significant use cases where artificial intelligence and machine learning can play a pivotal role in promoting societal sustainability. Specifically, we aim to study concrete use cases that are deployable in real-world applications and present technical challenges requiring advancements in foundational research. Through the activities conducted in ELIAS thus far, we have identified a major use case focusing on cyber security within the context of novel software paradigms known as microservices infrastructures [18]. This identification was made in collaboration with a partner company involved in ELIAS, which specializes in cyber security. The primary goal is to develop and deploy a comprehensive dataset that the cyber security community can utilize to enhance the performance of anomaly detection systems when facing cyber attacks. This dataset will be designed to address several critical areas, including real-world applicability, collaborative development, and performance benchmarking. The application of this use case is closely related to Task 2.2 on societal sustainability. Given the pervasive and integral nature of microservices infrastructures, enhancing their robustness against cyber attacks is crucial for securing democracy.

3.2 AI for monitoring the virtual infrastructure

Contributing partners: BD

3.2.1 Introduction

Cyber-attacks represent an ever-escalating danger in today's digital landscape, posing significant threats to individuals, businesses, and entire nations. The relentless pace of technological advancement, while offering numerous benefits, has also paved the way for increasingly sophisticated and frequent cyber-threats. Alarmingly, **approximately** 300,000 **new malware variants are created every day**, showcasing the ingenuity and persistence of cyber-criminals. This staggering figure highlights the constant evolution of malicious software designed to exploit vulnerabilities in systems, steal sensitive data, and disrupt operations.

The frequency of cyber-attacks is another cause for concern. On average, **a business falls victim to a cyber-attack every 39 seconds**. This near-constant barrage underscores the vulnerability of the business sector, regardless of size or industry. Small businesses, often lacking robust cyber-security measures, are particularly at risk, yet even large corporations with extensive security infrastructures are not immune. The sheer volume of attacks highlights the importance of vigilant and proactive cyber-security practices.

The pervasive reach of cyber-attacks is evident in the fact that **71% of organizations have been victims of at least one cyber-attack**. This statistic illustrates that cyber-threats are not isolated incidents but rather a widespread issue affecting a majority of organizations globally. The implications of such high rates are profound, affecting operational continuity, financial stability, and organizational reputation. The economic impact of cyber-attacks can be devastating, with costs associated with data breaches, system repairs, and lost business opportunities.

One of the most troubling aspects of modern cyber-attacks is the time it takes to identify them. On average, **it takes approximately 49 days to detect a cyber-attack**. This significant delay provides cyber-criminals with a substantial window of opportunity to inflict damage, exfiltrate data, and cover their tracks. The prolonged presence of undetected threats increases the potential for extensive harm, making timely detection and response critical components of effective cyber-security strategies.

3.2.1.1 Microservices infrastructure This use case focuses on cyber-attacks in the special scenario of microservices infrastructures which is a paradigm that is now central in the majority of industrial applications. In particular, the transition to microservices infrastructure in the last decade represents a significant shift for organizations, businesses, and various sectors, fundamentally transforming how they design, deploy, and manage their software applications. Unlike the traditional monolithic architecture, where applications are built as a single, interconnected unit, microservices break down applications into smaller, independent services that can be developed, deployed, and scaled individually. This modular approach offers greater flexibility, enabling organizations to adopt a more agile development process. Each microservice focuses on a specific business function, allowing teams to work concurrently on different services without causing disruptions. This independence not only accelerates development cycles, but also simplifies maintenance, as issues affecting a specific service can be addressed without impacting the entire application.

3.2.1.2 Anomalies for microservices The microservices paradigm raises new challenges to face cyber-attacks due to its distributed nature. Indeed, monitoring the interactions inside an infrastructure is crucial in preventing and early detecting cyber-attacks because it provides continuous visibility into the behavior and performance of each service within the application ecosystem. Unlike monolithic architectures, where a single application instance might be easier to monitor as a whole, microservices involve numerous independent services that interact with each other. This complexity can obscure potential security threats, making it essential to have robust monitoring mechanisms in place. Anomalies such as unexpected spikes in traffic, unusual patterns of resource consumption, or deviations from normal operational metrics can be early indicators of cyber-attacks, such as distributed denial-of-service (DDoS) attacks, unauthorized access attempts, or data exfiltration activities.

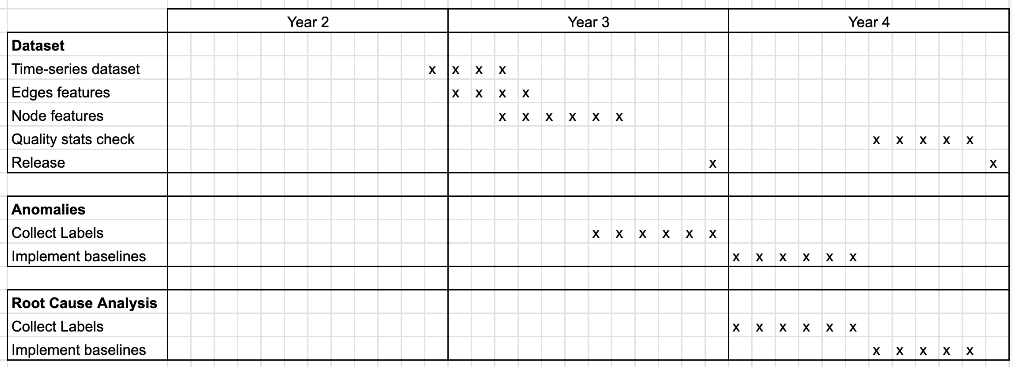
3.2.2 Expected deployment

Our goal is to produce a dataset, collected by Bitdefender, that—hopefully—will allow the community to enhance anomaly detection by incorporating **large, multivariate time-series data**, which captures complex interactions and temporal patterns across multiple variables. Additionally, the dataset should support generalization analysis by reflecting **data distribution shifts over time**, which is crucial for adapting detection models to evolving patterns and trends, thus maintaining their effectiveness. Following the data nature, this corpus will also enable **graph-based approaches for anomaly detection** by mapping the link/requests (edges) between different microservices (nodes), enabling the identification of anomalies that arise from unexpected changes in these connections. This comprehensive dataset will provide a robust foundation for developing adaptive anomaly detection models. The plan for the usecase deployment is shown in Fig. 1.

3.2.3 Open challenges

In this project, we aim to address the following open challenges studied in ELIAS:

- Collect and publish a real, curated dataset for anomaly detection, with high potential impact in the field.
- Cover several months to years of training data (large dataset) to enable valuable research in the field.
- Allow multivariate time-series data analysis.
- Release an open dataset, publicly available.



	Year 2												Year 3												Year 4											
Dataset																																				
Time-series dataset												x	x	x	x																					
Edges features													x	x	x	x																				
Node features															x	x	x	x	x	x																
Quality stats check																															x	x	x	x	x	
Release																								x												x
Anomalies																																				
Collect Labels																			x	x	x	x	x	x												
Implement baselines																									x	x	x	x	x	x						
Root Cause Analysis																																				
Collect Labels																									x	x	x	x	x	x						
Implement baselines																															x	x	x	x	x	

Figure 1. 1. Dataset: we will first collect the time-series data, followed by adding more features for edges and nodes. After validating the dataset meets the wanted quality and enables the research of the wanted problems, it will be publicly released. 2. Anomalies: we first annotate the anomalies for testset, followed by implementing strong baselines. 3. Root Cause Analysis: we will collect labels for testset, using highly skilled humans, followed by baselines implementations.

- Enable data-distribution shifts analysis.
- Enable graph-based approaches for anomaly detection.
- Ensure the anonymity of collected samples, incorporating feedback from the legal department and the ethical advisor.
- Identify multiple features for each node and edge, and determine the total timespan for the corpus.
- Conduct an in-depth analysis and statistical evaluation of the new dataset to ensure it meets our requirements in terms of graph structure, multivariate time-series, and distribution shifts.
- Collect human-labeled data for anomalies.
- Collect highly skilled human labels for Root Cause Analysis [19].
- Develop baselines and new algorithms for detecting anomalies.

4 T2.2 Development of learning algorithms to protect and secure democracy

4.1 Overview

The goal of T2.2 is to utilize and advance machine learning tools and methods to protect and secure democracy. This objective can be pursued through various approaches, ranging from theoretical to experimental contributions. The first contribution addresses ensuring fairness constraints in online learning within sequential decision-making processes [4]. This work provides algorithms with robust theoretical guarantees. The second and third contributions involve the adoption of sensors and real-time signal processing algorithms to monitor people behavior and intentions. Specifically, the second contribution focuses on crowd localization, aiming to pinpoint individuals in crowds using point annotations [5], while the third contribution investigates tracking and understanding human behavior through video and image analysis [6]. The fourth and fifth contributions pertain to analyzing people's opinions and preferences. In particular, the fourth contribution is dedicated to data analysis for understanding political divisiveness [7], while the fifth contribution seeks to uncover latent values and opinions in large language models (LLMs) to identify and mitigate biases [9]. The sixth and final contribution centers on fact-checking, proposing and evaluating a practical system in a real-world application.

4.2 Design of learning algorithms with fairness guarantees

Contributing partners: POLIMI

4.2.1 Introduction and methodology

The framework of *Markov decision processes* (MDPs) [20] has been extensively employed to model sequential decision-making problems. These include, but are not limited to, autonomous driving and users navigation in web platforms. In *reinforcement learning* (RL) [21], the goal is to learn an optimal policy for an agent interacting with an environment modeled as an MDP. A different line of work [22, 23] is concerned with problems in which an agent interacts with an unknown MDP with the goal of guaranteeing that the overall reward achieved during the learning process is as large as possible. This approach is more akin to *online learning* [24], and it is far less investigated than classical RL approaches. Nevertheless, most of the applications arising from web platforms are inherently online, as data become progressively available during users navigation.

In real-world applications, there are usually additional constraints and specifications that an agent has to obey during the learning process, and these cannot be captured by the classical definition of MDP. For instance, autonomous vehicles must avoid crashing while navigating [25, 26], while web platforms must ensure some *fairness* requirements in users navigation. These problems are of great relevance to ELIAS, since they are fundamental in order to protect democracy in our modern society revolving around AI and the web. For instance, in the context of pricing on the web (see, *e.g.*, airlines ticketing and customized offers in e-commerce websites), it is of paramount importance to ensure that pricing is decided in a fair way. Other application examples include bidding agents in ad auctions that are constrained to a given budget [27, 28], while recommender systems should *not* present offending items to users [29]. In order to model such features of real-world problems, [30] introduced *constrained* MDPs (CMDPs) by extending classical MDPs with cost constraints that the agent has to satisfy.

We study *online learning* in episodic CMDPs in which the agent is subject to *long-term* constraints. In such a setting, the goal of the agent is twofold. On the one hand, the agent wants to minimize their (cumulative) *regret*, which is how much reward they lose compared to what they would have obtained by always playing a best-in-hindsight, constraint-satisfying policy. On the other hand, while the agent is allowed to violate the constraints in a given episode, they want that the (cumulative) *constraint violation* stays under control, by growing sublinearly in the number of episodes. Long-term constraints naturally model many features of real-world problems, such as, *e.g.*, budget depletion in automated bidding [31, 32].

All the existing works studying online learning problems in CMDPs with long-term constraints address settings in which the constraints are selected stochastically according to an unknown (stationary) probability distribution. While these works address both the case where the rewards are stochastic (see, *e.g.*, [33, 34]) and the one in which they are adversarial (see, *e.g.*, [35, 36]), to the best of our knowledge there is no work addressing settings with adversarially-selected constraints. Some works (see, *e.g.*, [37, 38]) consider the case in which rewards and constraints are non-stationary, assuming that their variation is bounded. However, these results are *not* applicable to general settings with adversarial constraints.

4.2.2 Main theoretical results

We pioneer the study of CMDPs in which the constraints are selected adversarially. In doing so, we introduce an algorithm that employs a novel primal-dual approach in CMDPs, allowing it to attain *best-of-both-worlds* guarantees, in the flavor of [39]. In particular, our algorithm provides optimal (in the number of episodes T) regret and constraint violation bounds when rewards and constraints are selected either *stochastically* or *adversarially*, without requiring any knowledge of the underling process. While best-of-both-worlds algorithms have been recently introduced in online learning settings subject to constraints (see, *e.g.*, [40, 39]), to the best of our knowledge our algorithm is the first of its kind in CMDPs.[1]

When the constraints are selected stochastically, we show that our algorithm provides $\tilde{\mathcal{O}}(\sqrt{T})$ cumulative regret and constraint violation when a suitably-defined Slater-like condition concerning the satisfiability of constraints is satisfied. Moreover, whenever such a condition does *not* hold, our algorithm still ensures $\tilde{\mathcal{O}}(T^{3/4})$ regret and constraint violation. Instead, whenever the constraints are chosen adversarially, our analysis revolves around a parameter ρ which is related to our Slater-like condition, and in particular to the "margin" by which it is possible to strictly satisfy the constraints. Indeed, under adversarial constraints, [42] show that it is impossible to simultaneously achieve sublinear regret and sublinear cumulative constraint violation. We prove that our algorithm achieves no-α-regret with $\alpha = \rho/(1+\rho)$, while guaranteeing that the cumulative constraint violation is sublinear in the number of episodes. *This matches the regret guarantees derived for other best-of-both-worlds algorithms in (non-sequential) online learning settings* [43, 39].

Differently from previous works on online learning with adversarial constraints, in this work we *relax the strong assumption* that the algorithm has to know the value of the parameter ρ related to Slater's condition. This assumption is ubiquitous in the adversarially-constrained online optimization literature (see, *e.g.*, [44]), but it is *extremely* unreasonable in practice. Indeed, in real-world scenarios, the learner has usually no clue about the "margin" by which a strictly feasible solution satisfies the constraints. Relaxing such an assumption is a non-trivial task from a technical perspective. This is done by proving that our primal-dual algorithm guarantees that dual variables are automatically bounded, by showing that both the primal and the dual regret minimizers attain

[1] Notice that, in the literature on online learning in MDPs, the term *best-of-both-worlds* is sometimes referred to algorithms that achieve optimal instance-dependent regret bounds when rewards are selected *stochastically* and $\tilde{\mathcal{O}}(\sqrt{T})$ regret when rewards are chosen *adversarially* [41]. In this work, we borrow terminology from the literature on online learning with constraints, where the term usually refers to algorithms that achieve optimal regret and constraint violation bounds when the constraints are selected either *stochastically* or *adversarially* [39].

a strong no-regret property, called *no-interval regret*. This is crucial since the classical (weaker) no-regret property is *not* enough to ensure that dual variables are automatically bounded.

A summary of our contributions compared to those of prior works is reported in Table 2.

Table 2. Comparison of our work and the state-of-the-art. We group together previous works that provide similar guarantees. For each group, we only cite the most recent paper. The third column concerns the possibility of learning without the knowledge of the parameter ρ, while the fourth one specifies if the algorithm is capable of learning when the parameter ρ is arbitrarily small. † These works do not *apply to general adversarial settings, but only to settings with bounded non-stationarity.*

	adversarial rewards	adversarial constraints	unknown ρ	without Slater	MDPs
[34]	✗	✗	✓	✓	✓
[36]	✓	✗	✓	✗	✓
[44]	✓	✓	✗	✓	✗
[38]	✗†	✗†	✓	✗	✓
Our Work	✓	✓	✓	✓	✓

4.2.3 Relevant publications

- Stradi, F. E., Germano, J., Genalti, G., Castiglioni, M., Marchesi, A., and Gatti, N. "Online Learning in CMDPs: Handling Stochastic and Adversarial Constraints". In Forty-first International Conference on Machine Learning, 2024 [4].

4.3 Consistency-aware anchor pyramid network for crowd localization

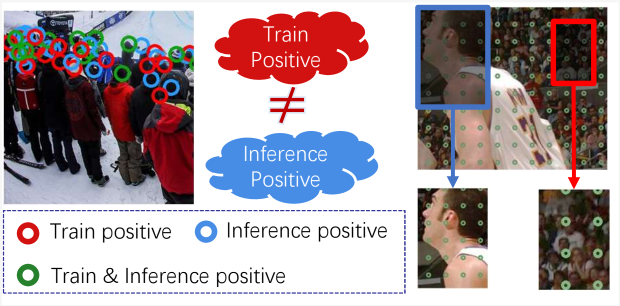
Contributing partners: UNITN

4.3.1 Introduction and methodology

The goal of crowd localization is to localize individuals in crowds using point annotations. This problem has received much attention due to a wide range of applications, such as traffic flow analysis [45], medical cell assay [46], and crowd anomaly detection [47]. In the context of ELIAS, crowd counting and localization could be useful for the analysis of crowd intention and behavior. Despite significant advances that have been made, crowd localization remains challenging partly due to the large variations in density across diverse crowd scenarios.

Existing methods for crowd localization can be broadly categorized into three groups based on their regression targets: detection-based methods, which regress bounding boxes of heads [48, 49, 50, 51, 52]; point regression, which directly regress point annotations [53, 54]; and heuristic methods, which regress heads in a density map [55, 56] or a segmentation map [57, 58, 59, 60].

Detection-based methods formulate crowd localization as a typical object detection task and use the center coordinates of the predicted bounding boxes as head locations. The limited number of bounding box annotations [48, 49, 50] heavily constrains recent advances in detection-based methods. Depth information is used in [52, 51] to estimate head size without bounding box annotations. Heuristic approaches employ various auxiliary maps, such as density maps, segmentation maps, and confidence maps, to capture crowd distribution. These methods require non-differentiable post-processing steps (*e.g.*, finding maxima [61, 62, 55] or finding connected components [57, 58]) to compute head coordinates, making them incapable of being end-to-end trained. On the contrary,



(a) *(b)*

Figure 2. (a) Illustration of the ranking inconsistency of predictions between the training and testing phases, which may lead to sub-optimal inference performance. (b) Excessive or insufficient numbers of evenly distributed anchors across sparse and dense regions in an image cause performance reduction.

point regression methods [53, 54], which also follow the detection paradigm, can directly predict the coordinates of targets. Our work belongs to this category.

Despite significant progress in crowd localization, the performance of prevailing point regression methods is limited in two aspects. One limitation is the ranking inconsistency of predictions between the training and inference phases. During inference, the selection of predictions is solely based on classification scores. However, during training, the top-M (M is the number of targets in the image) predictions are selected based on both spatial distance to targets and classification scores. This inconsistency leads the model to be sub-optimized with respect to its testing. We show one example in Figure 2(a), where part of the predictions used for loss computation (denoted as "train positive") are not selected as final results (marked as "inference positive") for inference and thus distract the training process. The other limitation comes with utilizing a fixed number of evenly distributed anchors. An image may contain diverse crowd densities across regions, as shown in Figure 2(b). Using a fixed number of evenly distributed anchors across an image could lead to excessive predictions in regions with sparse targets and inadequate predictions in regions with dense targets, thereby limiting overall performance. To address these problems, we propose Consistency-Aware Anchor Pyramid Network (CAAPN) for crowd localization, which consists of two main components: an Adaptive Anchor Generator (AAG) and a Localizer with Augmented Matching (LAM) (see Figure 3). The AAG module is designed to generate anchors according to the estimated density in each local region and spatial distribution prior. Therefore, AAG contains a counting branch, which predicts the number of heads in a region. Existing counting loss (*i.e.*, Mean-Squared Error) is susceptible to inevitable shifts in manual annotations, making the predicted density map less precise to guide anchor distribution. To alleviate this issue, we propose a Cascade Region Loss (CRL) to generate a more precise density map.

The distribution prior is gathered from training data in a region-wise manner. The adaptively generated anchors are then fed to the localizer in LAM to make location predictions. As such, the AAG module enables dynamic anchor generation and makes the number and distribution of anchors closer to the target. The LAM module, unlike previous methods, selects two groups of top-M predictions according to independent criteria: one group is chosen according to both distance error and classification score similar to existing methods [53, 54]; and the other group is chosen based solely on classification score to keep consistent with the test phase. To effectively utilize it, we assign this group to a specific ground truth set selected according to inverse probability ranking. Our ablation studies show that this simple design largely alleviates the ranking inconsistency

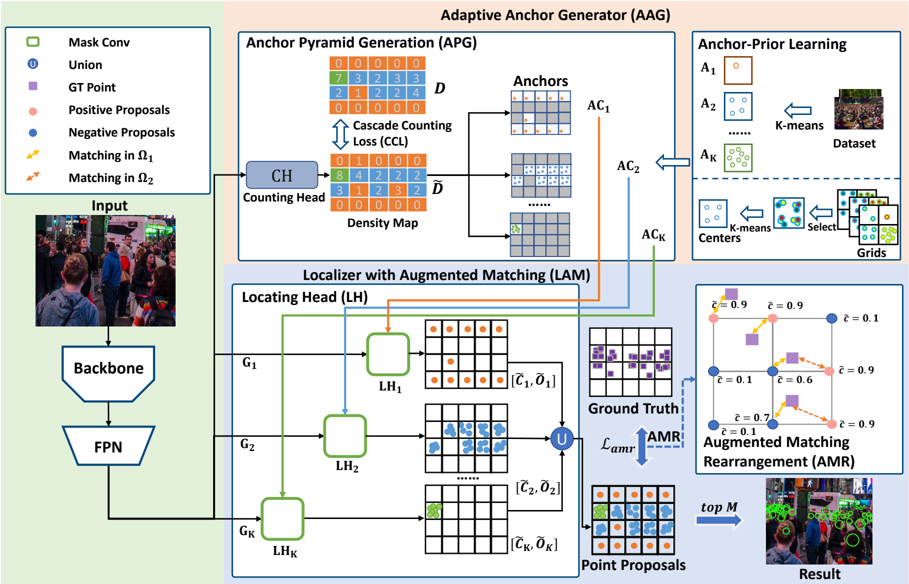


Figure 3. Main architecture of the proposed method. The input image is evenly divided into grids. Then, the Adaptive Anchor Generator generates high-quality anchors according to the estimated head number and spatial distribution prior. Next, these anchors are fed to the Localizer with Augmented Matching module to predict head coordinates. This module is enhanced by introducing a re-matching process of an extra set of target candidates, which alleviates the ranking discrepancy between the training and testing phases.

problem and significantly boosts performance.

4.3.2 Experiments

4.3.2.1 Experimental setup

Datasets. We use ShanghaiTech A and B, UCF-QNRF, JHU-CROWD++, and NWPU-Crowd datasets to evaluate our method. The ShanghaiTech A dataset contains web images with high crowd densities, while the ShanghaiTech B dataset includes street images with relatively sparse crowds. The UCF-QNRF dataset presents a more challenging scenario with high-resolution images and a wide range of human counts, ranging from 49 to 12,865 across 1,525 images. The JHU-CROWD++ dataset covers diverse scenarios and environmental conditions, consisting of 4,250 images with crowd counts ranging from 0 to 7,286. Finally, the NWPU-Crowd dataset provides 5,109 images with a wide range of human counts (including 351 images without humans).

Evaluation metrics. For the counting performance, we adopt the widely used Mean Absolute Error (MAE) and Mean Squared Error (MSE) as metrics. For the localization performance, we use Precision, Recall, and F1-measure (P, R, F1 for short) for evaluation. Following the setting in FIDT [62], different datasets use different criteria for judging a prediction as true positive. Specifically, ShanghaiTech A and B and JHU-CROWD++ datasets adopt two distance thresholds: 4 pixels and 8 pixels. The UCF-QNRF dataset takes a series of thresholds from 1 to 100 with a step size of 1. It computes the average recall, precision, and F1 as the final performance metric. The NWPU-Crowd dataset utilizes thresholds related to the size of targets. For strict localization

Table 3. Localization performance on the NWPU-Crowd dataset. The main metric is F1 under σ_l. The best and second best results are highlighted in red and blue, respectively.

Methods	Features	σ_l	σ_s
		F1 / P / R	F1 / P / R
TinyFaces[63]	ResNet-101	56.7/52.9/61.1	52.6/49.1/56.6
TopoCount[57]	VGG-16	69.1/69.5/68.7	60.1/60.5/59.8
RAZLoc[64]	VGG-16	59.8/66.6/54.3	51.7/57.6/47.0
AutoScale[65]	VGG-16	62.0/67.3/57.4	54.4/59.1/50.4
P2PNet[53]	VGG-16	71.2/72.9/69.5	-/-/-
IIM[58]	HRNet-W48	76.0/82.9/70.2	71.3/77.7/65.8
FIDT[62]	HRNet-W48	75.5/79.7/71.7	70.5/74.4/66.9
DCST[66]	DCST	77.5/82.2/73.4	72.5/76.9/68.6
GMS[67]	HRNet-W48	78.1/79.8/76.5	-/-/-
CAAPN (Ours)	VGG-16	76.5/79.6/73.7	70.4/73.2/67.9
CAAPN (Ours)	ConvNeXt-S	77.8/81.3/74.5	71.5/74.7/68.5
CAAPN (Ours)	HRNet-W48	78.6/80.4/76.8	72.7/74.3/71.1

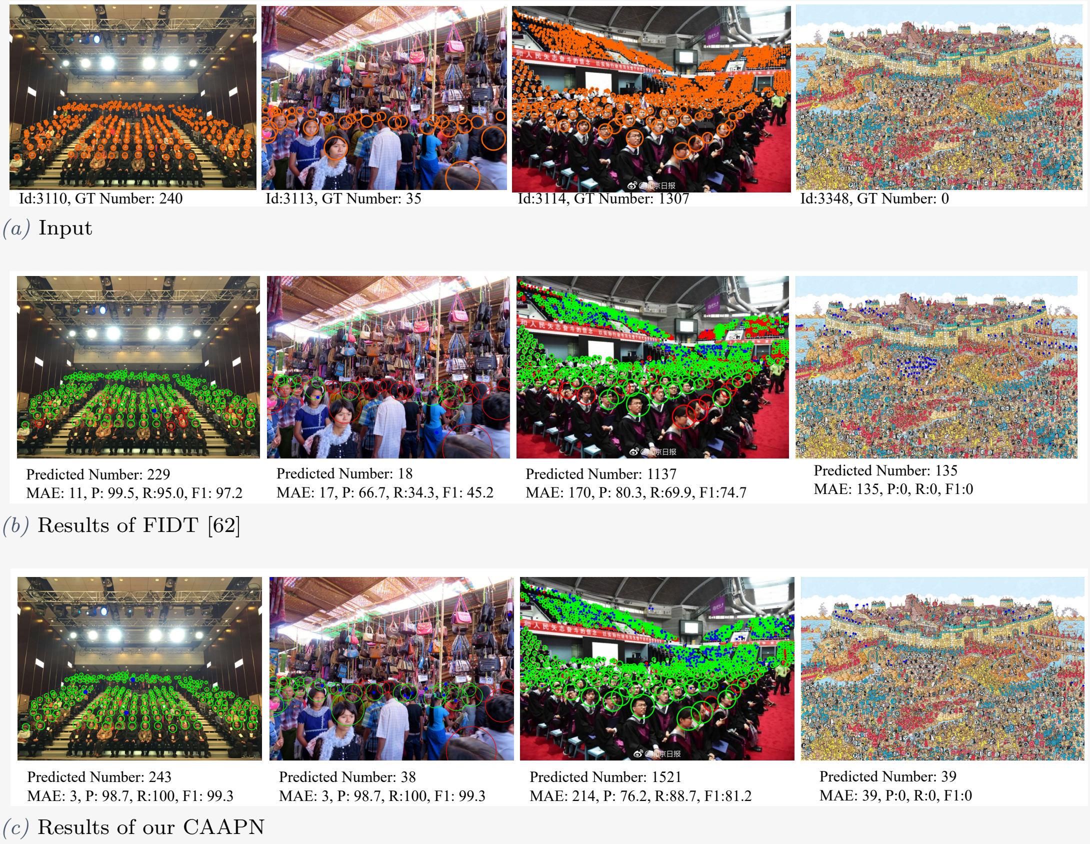
setting, the threshold σ_s^i for ground truth point i is set by $\sigma_s^i = 0.5 \times \min(h_i, w_i)$. For a relatively loose localization setting, the threshold is set to $\sigma_l^i = 0.5 \times \sqrt{h_i^2 + w_i^2}$.

Comparisons to the state-of-the-art methods We note that existing methods utilize different image features. For fair comparisons, we evaluate our method using three different features obtained via VGG-16, HRNet-W48, and ConvNeXt-S, respectively. For the sake of space we present here only the results obtained on the NWPU-Crowd dataset. All the other results can be found in [5].

As shown in Table 3, our method achieves the highest F1 and recall scores under both σ_l and σ_s settings on the test split. Our CAAPN with HRNet-W48 pushes the boundary of F1/R to 78.6/76.8 under the setting σ_l, and to 72.7/71.1 under the setting σ_s.

In Figure 4, we visualize the results on different target densities. For the medium crowded image Id 3110 (level 2 in NWPU-Crowd density label), our CAAPN finds almost all the targets with only 3 incorrect predictions. In this image, most people missed by FIDT are in the front rows and of relatively sparse density. In contrast, our CAAPN can find all these points thanks to the AAG module. For the sparse crowded image Id 3113, which is of various scales (NWPU-Crowd density label 1) in a complex market scene, our method outperforms FIDT by a significant margin on both precision and recall. We attribute this to the region-wise anchor generation and point proposal rearrangement strategy. The image in the third column is not only congested (NWPU-Crowd density label 3) but also low resolution. The density of crowds exceeds the upper bound that FIDT can handle. With our AAG, CAAPN can generate denser anchors in congested regions and thus handle this challenging scenario well. Finally, for the rightmost image, where there are no visible persons, our method still performs well.

Counting performance. Although this work focuses on crowd localization, we also provide the counting performance for comprehensive evaluation. The results are presented in Table 4. Our CAAPN achieves the best performance on four out of five benchmarks in terms of the main metric



(a) Input

(b) Results of FIDT [62]

(c) Results of our CAAPN

Figure 4. Visualization of results obtained by FIDT and our CAAPN on NWPU-Crowd validation set. The predicted True Positives (TP), False Negatives (FN), and False Positives (FP) are denoted as green, blue, and red, respectively.

MAE and ranks second on the dataset STA, slightly behind P2PNet.

4.3.3 Conclusion

The main contributions of this research are as follows:

- We propose an Adaptive Anchor Generator (AAG) to adaptively generate anchors in each region of an image, which can alleviate the anchor deficiency or excess problem.
- We propose a Localizer with Augmented Matching (LAM) for point regression crowd localization, easing ranking inconsistency between training and testing.
- We propose a cascade regression loss (CRL) to relieve the localization shift error.
- Extensive experiments on five benchmarks, ShanghaiTech A&B, UCF-QNRF, JHU-CROWD++, and NWPU-Crowd, demonstrate the effectiveness of our method compared against several state-of-the-art approaches.

Table 4. Comparison of counting performance against state-of-the-art methods. The main metric is MAE. The best and second best results are highlighted in red and blue, respectively.

Methods	Coordinates	Features	NWPU-Crowd	
			MAE	MSE
CSRNet [68]	no	VGG-16	121.3	387.8
BL [55]	no	VGG-19	105.4	454.2
NoisyCC	no	VGG-19	102.6	398.4
DM-Count [56]	no	VGG-19	88.4	388.6
AutoScale [65]	yes	VGG-16	94.2	388.2
P2PNet [53]	yes	VGG-16	72.6	331.6
FIDT [62]	yes	HRNet-W48	86.0	312.5
CAAPN (Ours)	yes	VGG-16	71.5	289.7
CAAPN (Ours)	yes	HRNet-W48	79.7	341.2
CAAPN (Ours)	yes	ConvNeXt-S	76.2	332.0

4.3.4 Relevant publications

- X. Liu, G. Li, Y. Qi, Z. Han, A. van den Hengel, N. Sebe, M-H. Yang, and Q. Huang, Consistency-Aware Anchor Pyramid Network for Crowd Localization, IEEE Transactions on Pattern Analysis and Machine Intelligence, DOI: 10.1109/TPAMI.2024.3392013 [5].
  Zenodo record: https://zenodo.org/record/11864258

4.4 Enhancing local and global features for monocular per-object distance estimation

Contributing partners: UMORE

4.4.1 Introduction and methodology

The Computer Vision community has a long-standing commitment to estimating the *third dimension*, e.g., to estimate the distance of a target object from the camera (or *observer*) when projected onto the image plane, particularly in the context of monocular images. In the context of ELIAS, some of the practical applications of this task include: people tracking and human behavior understanding.

Modern approaches to distance estimation rely on geometric constraints or data-driven strategies. Based on that, per-object distance estimators can be broadly divided into two main categories: **geometric** and **feature-based** methods. The former [69, 70, 71] assumes that objects of the same class (*e.g.*, pedestrians) have consistent sizes. Under such a hypothesis, these methods can exploit projective transformations to approach the task. Namely, it involves regressing the relationship, expected to be roughly linear, between the visual size of an object (such as the height of its bounding box) and its distance. Unfortunately, this assumption does not hold in practice: in real-world scenarios, the dimensions of objects may vary significantly (*e.g.*, from children to adults).

In contrast, **feature-based** approaches [1, 72, 73, 74] incorporate supplementary visual information regarding the target objects and the context of the scene. This is achieved by feeding the entire monocular image to a global encoder (*e.g.*, a Convolutional Neural Network (CNN) [75]) and retaining the activation of the last convolutional block. On top of that, techniques based on Region of Interest (RoI) are used to provide a spatially consistent and fixed-size feature vector for each target object. These approaches can reach a more holistic understanding of the scene, as they leverage local information and the spatial relations between the target and other reference objects of the scene [73]. However, while existing feature-based approaches avoid the shortcomings of the geometric ones, they also come with several architectural drawbacks that are peculiar to CNNs: 1) The large receptive field of CNNs could *wipe out* the fine-grained information tied to the target object; 2) The pooling layers of CNNs downscale the resolution of the feature maps (*e.g.*, by a factor of 32 in ResNets). Unfortunately, it does so to the extent that **smaller** bounding boxes cover only sub-pixel activation areas in latent space; 3) The bottleneck design of CNNs tends to disregard the **spatial relations** between the parts and the whole.

Our work addresses the above limitations by proposing **DistFormer** (Figure 5), a hybrid architecture combining CNNs and Transformer layers. While still feature-based, our proposal can effectively exploit local and global information without giving up the depth of the visual encoding. The **first part** of DistFormer builds upon a **Contextual Encoder** network, that is a CNN equipped with additional layers based upon Feature Pyramid Networks [76] and allow our method to extract high-level representations that retain fine-grained details.

In the **second stage**, we extract per-object representations and pass them to two transformer-based encoders, which focus respectively on *local* cues and *global* relations. In more detail, the former one – the **Local Encoder** – performs self-attention between patches of the same object, disregarding information from other objects. Such a module aims at further enforcing the local visual reasoning and encouraging the extraction of fine-grained details. To do so, it receives an additional self-supervised training signal named **Auxiliary Reconstruction Task** (ART), whose design follows the Masked Image Modeling (MIM) paradigm [77]. Finally, the **Global Encoder** aims at encoding spatial and global relations explicitly, and we achieve this by carrying out self-attention among representations from distinct objects.

In summary, our method is based on the following proposals: *i)* We propose a novel **hybrid architecture** that effectively combines CNNs and Transformer layers. This architecture strikes a balance between local and global information, addressing limitations in existing feature-based methods; *ii)* We introduce an innovative self-supervised component termed **ART** within the Local Encoder. This task enhances object-specific feature learning and encourages each object-specific feature vector to be highly informative, focusing on the object of interest. The ART enforces localized, detailed understanding, boosting the model's performance; *iii)* We employ a **Global Encoder** module that refines local representations by learning mutual relations between objects in the scene.

4.4.2 Experiments

Experimental Setup **Datasets.** We validate the proposed approach by conducting extensive experiments on the real-world datasets KITTI [78] and NuScenes [79], and the synthetic large-scale MOTSynth [80].

Evaluation Metrics. We use the metrics commonly adopted in the per-object distance estimation task, such as the τ-**Accuracy** (δ_τ) [81] (*i.e.*, the maximum allowed relative error), the percentage of objects with relative distance error below a certain threshold ($< 5\%$, $< 10\%$, $< 15\%$) [73] and classical error distances [1]: absolute relative error (**ABS**), square relative error (**SQ**), and root mean squared error in linear and logarithmic space (**RMSE** and $\textbf{RMSE}_{log}$).

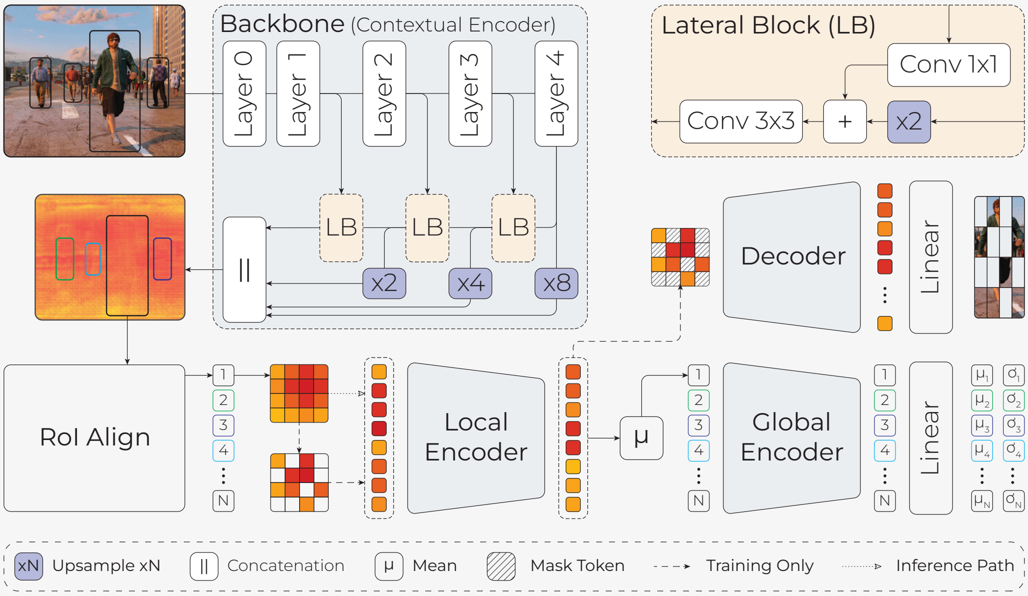


Figure 5. Overview of DistFormer . Initially, we feed the RGB frame **x** *through a classical backbone to obtain an informative feature map. Successively, a* RoIAlign *operation extracts a feature vector for each object in the image, which is then split into tokens and processed by the Local Encoder, extracting intra-object characteristics. During training, we mask k% of the tokens and use a decoder to reconstruct the missing ones. Afterward, the Global Encoder extracts strong scene-level object relations. Finally, an MLP predicts a Gaussian, modeling the distance and its uncertainty.*

Comparisons to the State-of-the-art Methods Tables 5 and 6 present the results of our approach and previous work. Results on KITTI are extracted from their respective papers, while for NuScenes and MOTSynth, we implemented and conducted experiments from scratch. While non-deep geometric methods perform poorly, deep ones perform much better, proving a correlation between the object size and distance from the camera. In addition, visual feature methods improve upon geometric ones, especially on KITTI, which features multiple target classes, showing that more than geometric features are needed for an accurate distance prediction. Our approach achieves state-of-the-art results on the KITTI dataset across all classes except for cars (see Table 5). It is noteworthy that methods surpassing our approach are tailored specifically for the car class or designed for multi-frame scenarios (e.g., Jing *et al.* [72]). In contrast, our approach generalizes over all classes without additional objectives.

The NuScenes dataset presents much more data and unique challenges with its diverse scenes, dynamic scenarios, complex traffic situations, and a maximum distance of over 150 meters. Despite these challenges, our proposed approach demonstrates robust performance, achieving state-of-the-art results across all metrics as depicted in Table 6.

4.4.3 Conclusion

We propose DistFormer, a novel and reliable approach for per-object distance estimation. It includes a local reasoning module performing self-attention between patches of the same object, which captures an object's local and peculiar visual attributes (*e.g.*, shape and texture). Moreover, DistFormer comprises a global module, exploiting self-attention between objects to deliver scene-aware predictions. Overall, we have shown that an additional self-supervised signal greatly benefits the

*Table 5. Experimental comparison on KITTI, following the setting in [1]. (*) our implementation.*

	$\delta_{<1.25}\uparrow$	ABS↓	SQ↓	RMSE↓	$\text{RMSE}_{\text{log}}\downarrow$
Cars					
SVR [69]	34.50%	149.4%	47.7	18.97	1.49
IPM [71]	70.10%	49.70%	1290	237.6	0.45
DisNet [70] *	70.21%	26.49%	1.64	6.17	0.27
Zhu *et al.* [1]	84.80%	16.10%	0.61	3.58	0.22
CenterNet [82]	95.33%	8.70%	0.43	3.24	0.14
PatchNet [83]	95.52%	8.08%	0.28	2.90	0.13
Jing *et al.* [72]	**97.60%**	**6.89%**	0.23	2.50	**0.12**
DistFormer	94.32%	9.97%	**0.22**	**2.11**	0.13
Pedestrian					
SVR [69]	12.90%	149.9%	34.56	21.68	1.26
IPM [71]	68.80%	34.00%	543.2	192.18	0.35
DisNet [70] *	93.24%	7.69%	0.27	3.05	0.12
Zhu *et al.* [1]	74.70%	18.30%	0.65	3.44	0.22
DistFormer	**98.15%**	**5.67%**	**0.08**	**1.26**	**0.09**
Cyclists					
SVR [69]	22.60%	125.1%	31.61	20.54	1.21
IPM [71]	65.50%	32.20%	9.54	19.15	0.37
DisNet [70] *	84.42%	12.13%	0.96	7.09	0.19
Zhu *et al.* [1]	76.80%	18.80%	0.92	4.89	0.23
DistFormer	**95.62%**	**8.01%**	**0.25**	**3.09**	**0.11**
All					
SVR [69]	37.90%	147.2%	90.14	24.25	1.47
IPM [71]	60.30%	39.00%	274.7	78.87	0.40
DisNet [70] *	69.83%	25.30%	1.81	6.92	1.32
Zhu *et al.* [1]	48.60%	54.10%	5.55	8.74	0.51
+ classifier	62.90%	25.10%	1.84	6.87	0.31
DistFormer	**93.67%**	**10.39%**	**0.32**	**2.95**	**0.15**
- W/out ART	93.43%	10.61%	0.34	3.17	0.15

Table 6. Performance comparison on the NuScenes and MOTSynth datasets.

	$\delta_{<1.25}\uparrow$	ABS↓	SQ↓	RMSE↓	$\text{RMSE}_{\text{log}}\downarrow$
		NuScenes			
SVR [69]	32.49%	57.65%	10.48	19.18	4.017
DisNet [70]	76.60%	18.47%	1.646	8.270	0.228
Zhu et al. [1]	84.54%	14.95%	1.244	7.507	0.245
DistFormer	**95.33%**	**8.13%**	**0.533**	**5.092**	**0.146**
- W/out ART	91.10%	11.16%	0.807	6.363	0.165
		MOTSynth			
SVR [69]	26.08%	54.67%	6.758	12.61	0.588
DisNet [70]	94.15%	8.73%	0.266	2.507	0.123
Zhu et al. [1]	98.71%	4.40%	0.116	2.131	0.065
DistSynth [74]	99.13%	3.71%	0.073	1.567	0.142
Monoloco [84]	99.69%	3.59%	0.064	1.488	0.167
DistFormer	**99.70%**	**2.81%**	**0.037**	**1.081**	**0.043**
- W/out ART	99.31%	3.36%	0.046	1.152	0.053

generalization capabilities of the model and synth-to-real knowledge transfer.

4.4.4 Relevant publications

- Aniello Panariello, Gianluca Mancusi, Fedy Haj Ali, Angelo Porrello, Simone Calderara and Rita Cucchiara, DistFormer: Enhancing Local and Global Features for Monocular Per-Object Distance Estimation, under review on NeurIPS 2024. [6]

4.5 Understanding political divisiveness and augmented democracy

Contributing partners: ANITI, TSE

4.5.1 Introduction and methodology

The literature on political polarization differentiates between affective and ideological forms of polarization [85]. Affective polarization involve strong feelings of dislike between groups and has been associated to the emergence of a political identity. Ideological polarization is more about differences and disagreement about belief and ideas. Yet understanding what particular issues drive polarization is challenging. In this reporting period we completed a paper developing an estimate of the political divisiveness associated with a policy proposal [7] and explored the creation of agents of augmented democracy using LLMs [8].

4.5.2 Experiments

4.5.2.1 Understanding political divisiveness In a recent paper [7], completed during this reported period but started a few years earlier, we used data collected during the 2022 presidential elections in France and Brazil[2] to explore the creation of a measure of divisiveness that we could use to identify issues that polarized citizens. The data collection involved the deployment of two collaborative government program builder websites where we asked citizens to select among proposals collected from the government programs of the official candidates of France and Brazil's 2022 presidential elections. In the paper we were able to show that the proposed metric of divisiveness, an estimate of how much the overall ranking of preferences of a population changes when considering citizens that tended to select or not a proposal, is uncorrelated with the voting rules used traditionally in social choice theory, opening a new area of inquiry for the creation of voting rules focused on divisiveness.

4.5.2.2 Exploring the potential of LLMs for augmented democracy In a more recent paper, we used open-source data from [7] to explore the use of LLMs for creating agents for augmented democracy [8]. We used anonymized individual level policy preference data to fine tune several off-the-shelf LLMs (e.g. LLAMA-2, Chat GPT 3.5 Turbo, Falcon 7B, etc.) and explore whether the preferences of the overall population of participants can be better approximated by a probabilistic sample of the data or a probabilistic sample augmented by LLMs. Here augmentation means using the LLMs fine-tuned with the probabilistic sample to estimate preferences unavailable in the sample. We show that LLM augmented data provides better estimates than probabilistic samples alone showing that LLMs could be used to build primitive augmented democracy systems. This paper received a revise and resubmit and is being considered for publication.

[2] Both data collection protocols were approved by ethics boards (IRBs) in Brazil and France, see paper for details: [7]

4.5.3 Relevant publications

- Navarrete, Carlos, Mariana Macedo, Rachael Colley, Jingling Zhang, Nicole Ferrada, Maria Eduarda Mello, Rodrigo Lira et al. ”Understanding political divisiveness using online participation data from the 2022 French and Brazilian presidential elections.” Nature Human Behaviour 8, no. 1 (2024): 137-148. [7]
- Gudiño-Rosero, Jairo, Umberto Grandi, and César A. Hidalgo. ”Large Language Models (LLMs) as Agents for Augmented Democracy.” arXiv preprint arXiv:2405.03452 (2024). [8]

4.5.4 Relevant software/datasets/other outcomes

Data on anonymized policy preferences: Dataverse

4.6 Revealing fine-grained values and opinions in LLMs

Contributing partners: UCPH

4.6.1 Introduction and methodology

Uncovering latent values and opinions in large language models (LLMs) can help identify biases and mitigate potential harm. Recently, this has been approached by presenting LLMs with survey questions and quantifying their stances towards morally and politically charged statements. However, the stances generated by LLMs can vary greatly depending on how they are prompted, and there are many ways to argue for or against a given position. In this work, we propose to address this by analysing a large and robust dataset of 156k LLM responses to the 62 propositions of the Political Compass Test (PCT) generated by 6 LLMs using 420 prompt variations. We perform coarse-grained analysis of their generated stances and fine-grained analysis of the plain text justifications for those stances. For fine-grained analysis, we propose to identify tropes in the responses: semantically similar phrases that are recurrent and consistent across different prompts, revealing patterns in the text that a given LLM is prone to produce. We find that demographic features added to prompts significantly affect outcomes on the PCT, reflecting bias, as well as disparities between the results of tests when eliciting closed-form vs. open domain responses. Additionally, patterns in the plain text rationales via tropes show that similar justifications are repeatedly generated across models and prompts even with disparate stances.

4.6.2 Experiments

Values and opinions embedded into language models have an impact on the opinions of users interacting with them, and can have a latent persuasion effect [86]. Identifying these values and opinions can thus reveal potential avenues for both improving user experience and mitigating harm. Recent works have proposed to evaluate LLM values and opinions using surveys and questionnaires [87, 88, 89, 90], as well as by engaging LLMs in role-playing and adopting the personas of different characters [91]. However, existing approaches suffer from three notable shortcomings.

First, recent work has shown that the responses of LLMs to survey questions depend highly on the phrasing of the question and the format of the answer [92, 93, 94], calling for a more robust evaluation setup for surfacing values embedded in language models. Second, when provided with different personas based on demographic characteristics, LLMs can reflect the social and political biases of the respective demographics [91], highlighting the need for disentangling the opinions embedded into LLMs and their variation when prompted with demographics. Such efforts also aid in aligning language models for different populations and cultures and prevent jailbreaking of LLMs [95]. Lastly, these evaluations focus primarily on quantifying stances towards the survey questions, ignoring justifications and explanations for those decisions. Revealing biases in such data could further demonstrate latent values and opinions as expressed in plain text.

4.6.3 Conclusion

LLMs express latent values and opinions when responding to queries, which can impact the users interacting with them. In trying to surface these values and opinions, LLMs are typically prompted to answer survey questions, but our work shows that these answers are not just brittle but can be steered in terms of political and other biases. We show how some models are more prone to this than others, raising important questions about how the training data and procedures impact embedded opinions and steerability. Additionally, most work on this problem has largely ignored the plain text justifications and explanations for stances towards these survey questions. Our work is a first step towards revealing the fine-grained values and opinions embedded in this text. To accomplish this, we produce a large scale dataset of $156,240$ responses to the Political Compass Test across 6 language models, which we release to the community for further research on this topic. Overall, we argue that while measuring stances towards survey questions can potentially reveal coarse-grained information about latent values and opinions in different settings, these studies should be complemented with fine-grained analyses of the generated text in order to understand how these values and opinions are plainly expressed in natural language.

4.6.4 Relevant publications

- Dustin Wright, Arnav Arora, Nadav Borenstein, Srishti Yadav, Serge Belongie, Isabelle Augenstein, “Revealing Fine-Grained Values and Opinions in Large Language Models,” arXiv preprint 2406.19238 (2024). [9]

- Peter Ebert Christensen, Srishti Yadav, Serge Belongie, “Prompt, Condition, and Generate: Classification of Unsupported Claims with In-Context Learning,” arXiv preprint 2309.10359 (2024). [10]

4.6.5 Relevant software/datasets/other outcomes

- Dataset of 156k LLM responses to the 62 propositions of the Political Compass Test (PCT) generated by 6 LLMs using 420 prompt variations.

- Dataset of 12 debate topics, comprising more than 120k arguments, claims, and comments from heterogeneous sources, each annotated with a narrative label.

4.7 Designing and deploying fact check retrieval pipelines

Contributing partners: INRIA

4.7.1 Introduction and methods

The nature of false news is so that a previously reviewed claim surfaces in different forms. Further, the volume of claims to be reviewed to fight misinformation is overwhelming for Fact-checkers. This has led to the research problem of *fact-check (FC) retrieval* – given a claim and a database of previous checks, find the checks relevant to the claim. To this end, we have built **FactCheck-Bureau**, an end-to-end solution that enables researchers to easily and interactively design and evaluate FC retrieval pipelines. We also present a corpus we have built, which can be used in further research to test fact-check retrieval tools.

4.7.2 Experiments

4.7.2.1 Dataset and FCR pipeline We built a corpus of 98K fact checks (FCs) in 14 languages published by 83 fact-checking agencies recognized as verified signatories by IFCN (International Fact-Checking Network). Further, we have collected 9.1K tweets mentioned in various FC articles and 8K recent tweets from prominent Members of the European Parliament (MEP). We used Google Fact-Check API[3] for collecting FCs. The data returned follows the ClaimReview schema but with some fields omitted as described in Google's documentation[4]. The returned data also includes the URL to the FC article. We also collected the FC article text using these URLs to enrich our dataset. Some of the reputed FC agencies, like Le Monde, do not publish their FCs via Google FC Explorer, or stopped at some point, therefore, we crawled their web pages to collect FCs. For social media post collection, we used a paid subscription of X to gather the 8K tweets from 402 MEP. We focused on social media posts in English and French FCs articles for the aligned pair collection. We collected tweets mentioned in 4.7K English FCs, 1.2K French FCs, and 3.2K FC in other languages. This resulted in 9.1K aligned pairs of social media posts and relevant FC articles, as many FC have two tweets related to them. For some of our FC retrieval experiments described next, we will consider these pairs to be the ground truth, allowing to search for a FC through its associated tweet. For others, the text of the claim described on the ClaimReview schema of the FC is used as its ground-truth pair.

We preload FactCheckBureau with our proposed FC retrieval pipeline that supports the default query interface for non-technical users. In our FC retrieval setting, the collection of FC *articles* servers as the document corpus and tweets serve as the input query. Our pipeline starts with pre-processing a tweet by *removing links, emojis, escape control and special characters*, *standardizing Unicode presentation with more than one representation*, *normalizing numbers and dates* using `num2words`[5] and `dateparser`[6] libraries, and *tokenizing text* using *MPNetTokenizerFast*[7]. We employed the well-established BM25 [96] as our retriever model and used `all-mpnet-base-v2` and `paraphrase-multilingual-mpnet-base-v2` for re-ranking documents in English and French respectively.

4.7.2.2 Inputs: fact-checks and claims Conceptually, there are two main entities. They are `fact-checks`, and `claims`. In principle, a claim can be a social media post, an image specifying a claim, or a simple text phrase. In the corpus we built for our demo, claims are tweets. Therefore, our `claim` is characterized by its *accountHandle* (the Twitter account having published the tweet), *text*, *date*, *language*, *hashTags*, and *URLtoEmbeddedMediaContent*. The attributes of `fact-check` are derived from the *ClaimReview* schema that many FC agencies adopt in their FC articles (https://schema.org/ClaimReview); ClaimReview was promoted by Google, which used to show, next to search results, related FCs[8]. Specifically, the attributes of an FC are: *title*, *claimant*, *publisher*, *dateOfPublication*, *URLtoArticle*, *claimText*, *language* and *rating*. The relationship between these two key entities is captured by a many-to-many relation `claimAboutFC`(*claim-id, FC-id*): a claim

[3] https://toolbox.google.com/factcheck/apis

[4] https://tinyurl.com/25t28phf

[5] https://pypi.org/project/num2words/

[6] https://pypi.org/project/dateparser/

[7] https://huggingface.co/docs/transformers/en/model_doc/mpnet#transformers.MPNetTokenizerFast

[8] That had been discontinued, among others, because some of the shown FC were not semantically close enough to the respective search results [97]. This highlights the importance of the FC retrieval problem.

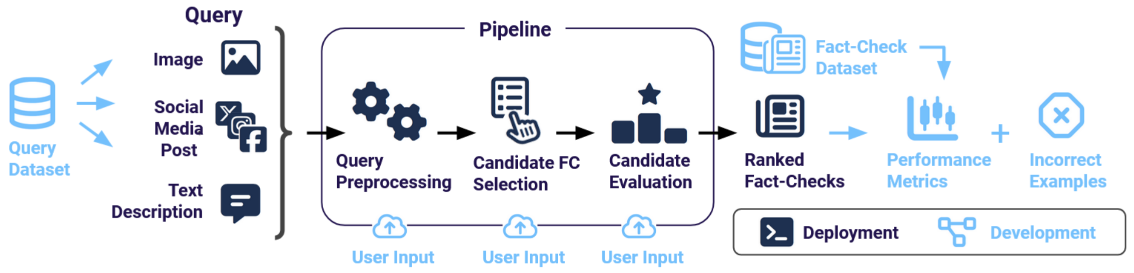
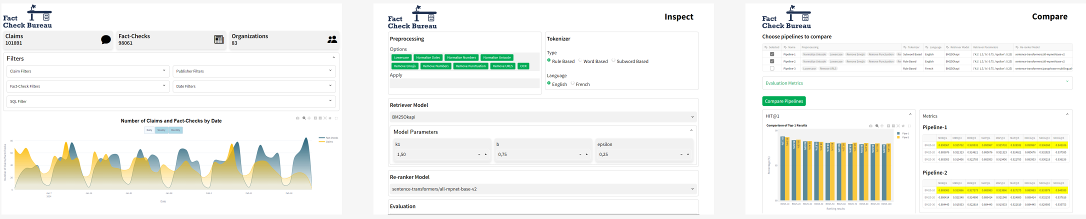


Figure 6. FactCheckBureau architecture in the two use modes (Development and Deployment)



Claims 101891
Fact-Checks 98061
Organizations 83
Filters
Claim Filters
Publisher Filters
Fact-Check Filters
Date Filters
SQL Filter
Number of Claims and Fact-Checks by Date

Preprocessing
Options: Lowercase, Normalize Dates, Normalize Numbers, Normalize Unicode, Remove Emojis, Remove Numbers, Remove Punctuation, Remove URLS, OCR
Apply
Tokenizer
Type: ● Rule Based ○ Word Based ○ Subword Based
Language: ● English ○ French
Retriever Model: BM25Okapi
Model Parameters
k1: 1,50
b: 0,75
epsilon: 0,25
Re-ranker Model: sentence-transformers/all-mpnet-base-v2
Evaluation

Choose pipelines to compare
Evaluation Metrics
Compare Pipelines
HIT@1
Comparison of Top-1 Results
Metrics
Pipeline-1
Pipeline-2

Figure 7. FactCheckBureau in Development mode

and an FC are paired in this way, if (according to a specific automated or manual decision method) the FC is about the claim. We also say the claim and FC are **aligned**.

4.7.2.3 Users' interactions with FactCheckBureau The core task to be solved in FC retrieval pipelines is: given a claim (also called a query) and an FC corpus, find the FCs most relevant to the claim. Specifically, each FC retrieval pipeline contains a *candidate FC retrieval* module and a *candidate ranking* module. If the claim is text, it can be used as such, but other formats may require some *pre-processing*, based on the query type, before being entered into the text-based retrieval pipeline. For instance, if the claim is an image, the text needs to be extracted by OCR, or the image can be captioned; if the claim is a tweet, pre-processing may remove or split hashtags in individual words, normalize numerical data, transcribe emojis to text, etc. To be comprehensive, FactCheckBureau models FC retrieval pipelines as consisting of three stages: pre-processing, retrieval, and ranking.

FactCheckBureau has **two main operation modes**: **deployment** and **development**, shown in Figure 6, where dark navy modules are used in deployment, whereas in development, all the modules (both navy and light blue) may be involved. In development mode, it supports designing, inspecting, and comparing FC retrieval pipelines; in deployment mode, a retrieval pipeline can be deployed and used to query the FC corpus. As explained below, our demonstration will showcase the four use cases (design, inspect, compare, and deploy). We present screenshots of our tool in Figure 7.

Inspect A user builds a retrieval pipeline by choosing or loading: pre-processing modules; a retrieval module; and a ranking module. The user also supplies aligned pairs, and chooses the metric(s) to use to evaluate the quality of the pipeline. Since relevant FC retrieval is a ranked-list search problem, we support the familiar Mean Average Precision (MAP@k), Mean Reciprocal Ranking (MRR@k), Normalized Discounted Cummulative Gain (NDCG@k), and Hits@k (over a set of queries and associated retrieved lists, how many times the expected, i.e., gold standard, FC was within the top-k results, for $k \in \{1, 3, 5, 10\}$ etc.), see also [98]. The user triggers the *evaluation* of the pipeline; for each query, this leads to a list of FCs, ordered by their relevance, the former as computed by the pipeline. FactCheckBureau also presents the values of the chosen metric(s) for different cut-off values k.

For further inspection, a user can choose the *deep-dive* option, where FactCheckBureau enables to inspect test samples where the specified pipeline performed poorly. It also reports the performance of each model in the pipeline in isolation.

- **Input:** Chose models, aligned pairs, metric(s);
- **Output:** Computed metric (plots and tables), the performance of candidate selector only, identify top 5 badly performing examples;
- **Options:** Save the pipeline.

Compare While developing a retrieval solution, it is essential to evaluate and compare different models to find the best possible combination of models for candidate retrieval and ranking. The *compare* mode enables a user to compare previously *saved* pipelines and/or newly specified ones. The user obtains a consolidated performance report comparing all the specified pipelines under a set of chosen metrics. It also provides a *deep-dive* option, to compare models' performance on selected test samples.

- **Input:** Choose pipelines from a list of pipelines;
- **Output:** Single plot of overall performance, plot of candidate identifier performance, 5 worst performance instances for each;
- **Options:** Choose a test instance and check the performance of all the pipelines for the chosen sample.

Design This option is for users who do not intend to develop a pipeline but need to use one. The user can supply an FC corpus, or a default one (the one we prepare for the demo, Section 4.7.2.2) can be used. The user specifies the claim language (or we can auto-detect it[9]), and claim type (post, image, or text).

- **Input:** Specify query type (post, image, text), and dataset language;
- **Output:** A recommended pipeline based on (*i*) the most frequently used components for these inputs, or comparable inputs (same language, same query type) if there is no history of running on the same inputs; (*ii*) simple rules to choose the necessary pre-processing models based on the input type.
- **Options:** Save the pipeline and option for deploying.

Deploy can be used as a search interface for finding the relevant FCs for an input query, or, alternatively, for a specific *topic*, specified as a short phrase, e.g., "Covid". The user chooses a previously specified retrieval pipeline already present in the system and configures the number of relevant documents she wants to retrieve. Then, FactCheckBureau returns a list of the FCs relevant to the claim, respectively, and FCs about the given topic.

[9] https://github.com/Mimino666/langdetect

- **Input:** Choose a pipeline, or select ”auto-design”; specify query or topic;
- **Output:** Querying interface – querying through post, image, text, or topic (will use FC tags available).

4.7.3 Conclusion

The next task that we have taken up is to support our tool FactCheckBureau with an efficient and effective retrieval pipeline. This will ensure that our tool can be used by our primary class of users, viz non-technical users, in particular, journalists. An effective FC retrieval pipeline can avoid redundant efforts put in by Fact checkers to re-reviewing claims that have been reviewed before. We also plan to collect FCs and associated tweets as an ongoing process so that our dataset remains relevant and can support the practical use case of the FC retrieval task.

4.8 Emergent behaviors from LLM-agent simulations

Contributing partners: JSI

4.8.1 Introduction and methodology

Our research proposes that intricate emergent behaviors may develop from multi-agent simulations using Large Language Models (LLMs), potentially mirroring complex societal frameworks. The hypothesis was examined through three increasingly sophisticated simulations, assessing the LLM-agents’ comprehension, task performance, and ability to engage in strategic interactions, including deception. Findings indicate a distinct disparity in reasoning capabilities between models like GPT-3.5-Turbo and GPT-4, particularly in less complex simulations. The research illustrates that emergent behaviors can manifest in LLM-agent simulations, from basic games to complex geopolitical scenarios.

The methodology consists of three primary steps: initially, we convert real-world societal structures and interactions into interactive ecosystems for LLMs. Next, we produce multiple iterations of LLM interactions. In the final phase, we derive significant conclusions from the simulations, offering a detailed analysis of the agents’ behaviors.

Previous research indicates that this approach could yield valuable insights, including generative agents that emulate human behavior by incorporating LLMs into interactive settings, and considering the Theory-of-Mind (ToM) reasoning abilities of LLMs, focusing on GPT-4’s human-like inference patterns.

4.8.2 Experiments

The experiment involved comparing trading decisions between agents using GPT-3.5-Turbo and GPT-4 in a sheep trading scenario. Agents using GPT-3.5-Turbo struggled with the ”Buy Low, Sell High” strategy, while GPT-4 agents effectively employed it. The number of trading rounds left did not affect trading decisions, and adjusting the temperature parameter increased decision variety without drastically changing outcomes. Additionally, increasing the temperature parameter in the models diversified the outcomes without significantly altering the overall game results. The inclusion of few-shot learning examples in prompts showed that game outcomes were highly dependent on the specific examples provided, across all language model variations.

The geopolitical model simulation was run with homogeneous agent identities and goals over 10 rounds to establish a baseline. Agents showed a preference for interacting with the USA initially. In variations, the USA and China aimed to increase military strength, Russia focused on money, and Germany on economic strength. Russia and Germany performed slightly better in their respective focuses, while the USA and China failed to dominate militarily. Another variation involved assigning real-world identities to all agents except Germany, who remained focused on economic strength. Over 10 rounds, economic strength decreased while military strength increased and converged, with agents showing reluctance to change their total money significantly. Simulations with both GPT-3.5-Turbo and GPT-4 yielded similar results (see details in figure 8).

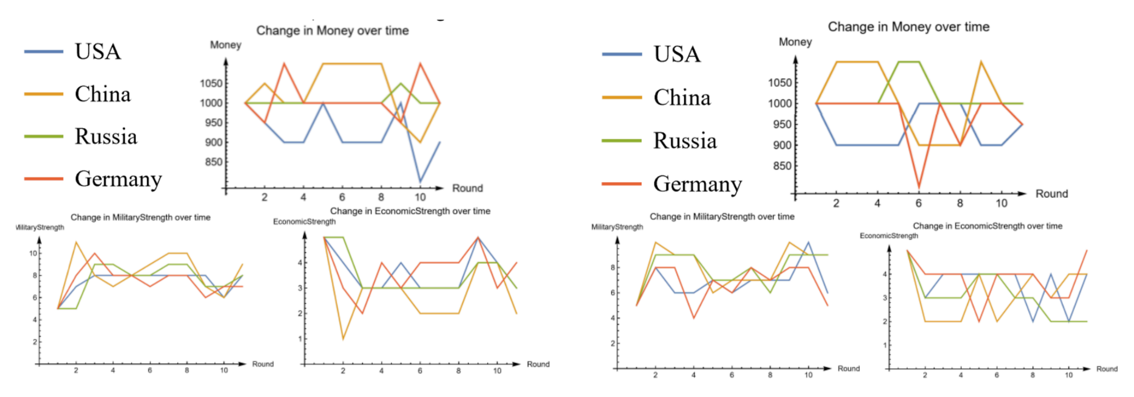


Figure 8. Development of agent attributes over 10 rounds of baseline geopolitics simulation (on the left) and geopolitics simulation (on the right).

4.8.3 Conclusions

Our investigation into multi-agent simulations with LLMs highlights the potential for complex emergent behaviors, mirroring societal structures. Through progressively complex simulations, we assessed the LLMs' understanding, task performance, and strategic interactions, noting that agents exhibited strategic behaviors, decision-making skills, and interaction complexity. Factors such as identities, attributes, actions, goals, past interactions, and few-shot learning examples influenced their performance.

Future research will focus on enhancing agent architecture and simulation complexity. We aim to conduct more controlled and targeted experiments as resources become more accessible. Additionally, larger-scale experiments with more iterations will be undertaken to gain a comprehensive understanding of LLM-agent societies. This research moves us closer to leveraging LLMs for complex simulations and understanding their interactions in sophisticated environments.

4.8.4 Relevant publications

- A. Mladenic Grobelnik, F. Zaman, J. Espigule-Pons, M. Grobelnik. Emergent Behaviors from LLM-Agent Simulations. Presented at the SiKDD Conference, 2023. [99]

5 T2.3 AI for inclusive and sustainable prosperity

5.1 Overview

The goal of T2.3 is to use and contribute machine learning tools and methods to explore patterns and mechanisms involved in sustainable economic development. The task involves the development of novel datasets that can help us understand unexplored aspects of economic development, such as the growing importance of digital trade. The task also includes advancing methodologies in economic complexity, which focuses on the use of machine learning methods to illuminate questions of economic development.

During this reporting period, Task 2.3 has produced important advances including a unique dataset on digital product trade [100], a dataset on validated historical GDP per capita estimates [12], a paper untangling the role played by historical migrants in the development of European knowledge agglomeration [2]. Task 2.3 also included the development of a data analysis framework to understand sustainability practices of European cities from open city data, as well as the development of a dashboard tool.

5.2 Analyzing open data to understand sustainability practices in European cities

Contributing partners: Idiap

5.2.1 Introduction and methodology

Cities are actively considering the repercussions of various human activities on the environment. Analyzing European cities' data from open data platforms is important to understand and assess their sustainability practices, and help urban planners prioritize climate resilience and adaptation strategies in response to severe environmental problems. Our work aims to mine open data to investigate urban sustainability within the scope of Europe to track progress and evaluate the performance of cities.

Data and preprocessing. The original dataset contains 1,085,048 observations answered by 548 cities across 10 CDP (Carbon Disclosure Project) regions, namely the USA, Latin America, Europe, Africa, East Asia, Canada, Oceania, South Asia, Southeast Asia, and the Middle East in 2022. We decided to focus on responses from Europe in order to have a coherent dataset that allowed us to understand, from a European point of view, how different environmental information on the subject was reported. After removing N/A responses and answers with "question not applicable" values, we had 101,274 observations from 138 cities within the scope of Europe. In the case of text responses, there were cases where respondents in some cities answered the questionnaire in the local language of the city. To identify these cases, we used langdetect[10] which enabled us to identify the languages of the texts. Around 10,000 responses in a language other than English were identified, and these cases were translated into English using googletrans[11].

After this first filter, we identified two main types of answers, non-text and text answers. On one hand, the non-text answers were those that required a binary or multiple choice answer, or those defined as a numeric field, where the cities had to answer with a number, such as the amount of solid waste generated. On the other hand, there were answers that we defined as "text answer", which are those answers that the respondent of the questionnaire is expected to write

[10] https://pypi.org/project/langdetect/

[11] https://py-googletrans.readthedocs.io/en/latest/

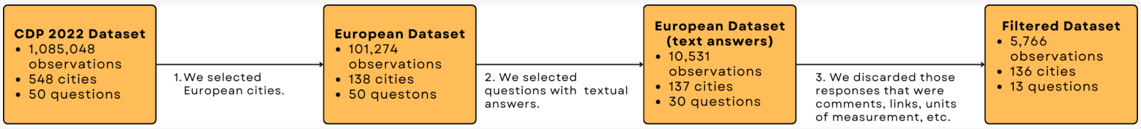


Figure 9. Diagram of the dataset preprocessing and curation steps used in this work.

and explain in more detail the answer to the question. We decided to focus on the text answers to conduct the analysis explained in the following sections. We analyzed textual responses from city officers in different cities. This analysis revealed human subtleties in the data, along with both differences and similarities between city responses. By understanding these factors, we can leverage a rich, consistent dataset that serves as a valuable resource for other cities facing similar climate challenges. Furthermore, within each question of the questionnaire, we filtered all those subquestions that had text as an answer. Later in the analysis, we studied one by one if the text was sufficiently informative to answer the question. We discovered that there were many cases in which a subquestion whose answer was labeled as plain text, was a link to a particular project or initiative; we also discovered cases where the subquestion was simply “Comments”, where a city gave some detail about the answers chosen in the multiple choice subquestions, such that the text did not really answer the question, but rather pointed out some detail; we decided to exclude this type of answers for our analysis. Figure 9 shows a diagram of the steps taken to reach the final version of our dataset. Finally, we curated a dataset where we only took into account for analysis those questions that contained at least one subquestion with a textual answer, and in which each subquestion was of interest for the analysis. Table 7 shows the selected questions, the associated SDGs, and the number of responding cities.

After data preprocessing, we conducted three different types of analysis: similarity analysis, sentiment analysis, and text classification, to compare the similarity of European cities and understand the self-reported sustainability actions in various cities.

City similarity. We wanted to study the similarity (or non-similarity) across cities in the different sections of the dataset. Through this, we could understand the singularities of some cities answering this questionnaire, but also the common points between them. This analysis is of interest because we can understand, from the way they respond, which cities are similar in terms of the measures they are taking or the plans they are developing, but also understand the differences between (or get ideas from) cities with similar characteristics that are responding differently. For analysis, we selected one question from each section, to analyze in detail the response of European cities and to compute their similarity.

Sentiment analysis. The goal of sentiment analysis is to infer the implicit tone in text responses to different questions, in such a way that we can examine which questions are answered with a more negative or positive tone, as well as studying the tone used by the cities in their answers to the questionnaire.

We divided each response into sentences to analyze the sentiment at the sentence level, and thus for each response to a question, we could quantify how many positive, negative and neutral sentences we had. To do the sentiment analysis at the sentence level we used the Financial-RoBERTa model from hugging face [101]: a pre-trained NLP model fine-tuned by encapsulating a large corpus of financial texts. Since the corpus includes Corporate Social Responsibility (CSR) Reports and Environmental, Social, and Governance (ESG) News, this model is able to better analyze sentiment in the context of climate actions and sustainability development ambitions than

Table 7. Question list used for analysis.

Section	SDG	Question	# Responding Cities
Governance	11, 13, 16, 17	0.2: ”Provide information on your jurisdiction’s oversight of climate-related risks and opportunities and how these issues have impacted your jurisdiction’s planning.”	71
	1, 10	0.3: ”Report how your jurisdiction assesses the wider environmental, social and economic opportunities and benefits of climate action.”	69
	17	0.5: ”Report your jurisdiction’s most significant examples of collaboration with government, business, and/or civil society on climate-related issues.”	61
Assessment	17	1.1a: ”Provide details on your climate risk and vulnerability assessment.”	90
	1, 11, 13	1.2: ”Provide details on the most significant climate hazards faced by your jurisdiction.”	130
	11, 13	1.3: ”Identify and describe the most significant factors impacting on your jurisdiction’s ability to adapt to climate change and indicate how those factors either support or challenge this ability.”	122
Targets	11, 13	4.1a; ”Report your jurisdiction’s main adaptation goals.”	111
	-	6.1: ”Provide details of your jurisdiction’s energy-related targets active in the reporting year. In addition, you can report other climate-related targets active in the reporting year.”	101
Planning	7, 11, 13, 17	7.1a ”Report details on the climate action plan or strategy that addresses climate mitigation and/or climate adaptation (resilience) in your jurisdiction.”	92
	12, 13	7.3: ”Does your jurisdiction have a strategy for addressing emissions from consumption of the most relevant goods and services?”	30
	-	7.4: ”Describe any planned climate-related projects within your jurisdiction for which you hope to attract financing.”	98
Actions	11, 13	8.1: ”Describe the outcomes of the most significant adaptation actions your jurisdiction is currently undertaking. Note that this can include those in the planning and/or implementation phase.”	14
	11, 13	9.1: ”Describe the outcomes of the most significant mitigation actions your jurisdiction is currently undertaking. Note that this can include those in the planning and/or implementation phases.”	120

the basic RoBERTa model.

Classification of answers by SDGs. Most of the questions in the dataset have one or more SDGs associated with them. Therefore, we assume that the answers to those questions have also a relationship with the SDGs. Based on this assumption, we defined a classification task where, given a questionnaire answer, we wanted to understand how difficult it is to associate the SDGs to that text.

There are 7 SDGs in our filtered dataset, framing the task to be a multi-class classification. We took text response answers as inputs and predicted SDGs as outputs. It should be noticed that a single text response could correspond to multiple SDGs, which defines the task as multi-label classification task. Before passing the text to the classifier, we need to convert it into vector data that the classifier can handle. Specifically, we removed unnecessary symbols and characters, such as punctuation marks, special characters and stop words. Then we converted the remaining text into numeric statistics by applying TF-IDF [102, 103], which reflects how important a word is to a document in a collection or corpus. Notice that the simplicity of these features is driven by our interest in interpretability of the results. As for predicted SDG labels, each SDG label was transformed into a one-hot encoded vector. The position corresponding to the SDG label of the text is marked as 1, and all other positions are 0. As stated earlier, there could be more than a single 1 in the label vector because a single text answer could correspond to more than one SDG. To implement the classification task, we used Random Forest [104].

With this task, we wanted to study the performance of the classifier and understand whether it treats all cities equally. We also wanted to study what factors affect the classification performance, such as the length of the answers, the number of answers to a given question, or the number of keywords associated with a given SDG. In this way, a classification approach could help city officers understand what aspects to take into account when answering future questions.

5.2.2 Experiments

Similarity analysis. For each of the selected questions we studied how similar the responding cities are. We illustrate using the example of question below:

> **(0.2) ”Provide information on your jurisdiction’s oversight of climate-related risks and opportunities and how these issues have impacted your jurisdiction’s planning.”**

This question has four sub-questions with textual answers, so we decided to analyze each sub-question separately:

- **”Provide further details on your jurisdiction’s oversight of climate-related issues.”** For this subquestion, we obtained a mean cosine similarity value of 0.63. The mean similarity values in the cities range from 0.3 to 0.78. Looking in detail at the highest and lowest similarity cases between cities, we can see that the city of Middelfart (0.43) focuses on mentioning that at national level there is a climate act, while at municipal level the responsibility is not yet being taken: *”At national level we have a climate act, however in reality not much responsibility for climate is allocated to the municipality level yet, but this is changing rapidly.”*, while the city of Kemi emphasizes the opposite in its answer: *”City of Kemi is committed into SDGs since year 2017. Kemi has ISO 14001:2015 environmental certificate since 2019 which includes that we have systematical environmental plans, implementations, follow-ups and continuous improvement of environmental issues.”* Our model estimated the similarity between these two cities (Middelfart-Kemi) as 0.35. From the answers, and based on a subjective evaluation, we can see that RoBERTAa model is generally working properly, creating an embedding of each aggregated answer that represents accurately the content of the answer.

- **”Describe how climate-related issues have impacted your jurisdiction’s development planning.”** In this case, there is an average cosine similarity of 0.58. Brussels stands out with an average cosine similarity of -0.17. In its response, this city states that it is developing a plan of sustainable measures, but emphasizes the difficulty of establishing a link between the problems to be addressed and territorial development: *”Regarding planning, the department concerned works to develop a Communal Sustainable Development Plan (PCDD) which will integrate climatic issues in urban development and territorial planning. This plan is not yet finalized, it is therefore difficult to establish the link between taking into account the issues and territorial development, even if certain climatic issues are already taken into account in planning decisions, even without the publication of this PCDD.”* Perhaps the transparency of the answer makes it stand out from the other cities, where the answers generally emphasize more the positive aspects and how well the taken actions work, as for example in the city of Akureyri: *”Climate-related issues were taken into account during the planning process, as well as during any review or updating of the planning documents. These issues are also discussed during most council meetings, especially during decision making within fields that may impact the climate or be impacted.”*

- **”Describe how climate-related issues have impacted your jurisdiction’s financial planning.”** In this case, there is an average value of cosine similarity of 0.46. The city with the lowest average similarity is Istanbul with -0.19, while the highest is Athens with 0.65. The difference can be seen in the vocabulary used. This is an example of a sentence in Istanbul: *“At this point, budget expenditures due to climate change bring an extra financial burden.”*, while an example of an Athens response is: *”The municipality of Athens has seized the opportunity for funding climate-related issues. National budget that aims at the energy upgrade of buildings or the reinforcement of the role of public spaces in the mitigation and adaptation to global-local warming have been used for respective projects within the municipality (energy upgrade of school buildings, greening and cool materials of public spaces). The municipality has also received a loan from the European Investment Bank for enhancing blue and green infrastructure within the city and for the energy upgrade of municipal buildings.”*

Sentiment analysis. In Fig. 10, we can see the results of the sentiment classification at question level, quantifying the proportion of positive, neutral and negative sentences for each question. In questions [0.3, 0.5] (Governance), [4.1a, 6.1] (Targets), [7.3] (Planning), and [9.1] (Actions), we can observe that more than half of the sentences are classified as positive. In the case of the questions related with the section Targets, it is where we appreciate the biggest content of positive sentences.

In the case of the negative sentences, we see that in most cases it is the minority case, with the exceptions of question 1.2 and question 1.3, both in the Assessment section, where the questions are related to negative concepts such as climate hazards, risks and vulnerability.

Finally, In the case of neutral answers, they account for more than 40% of the sentences in questions [0.2, 0.3, 0.5] (Governance), [1.1a] (Assessment), [7.1, 7.3, 7.4] (Planning), [8.1, 9.1] (Actions).

Text classification. The original CDP dataset comprises 51 questions, of which 44 involve textual responses. Among these, 40 questions are linked to one or more SDGs. For this classification

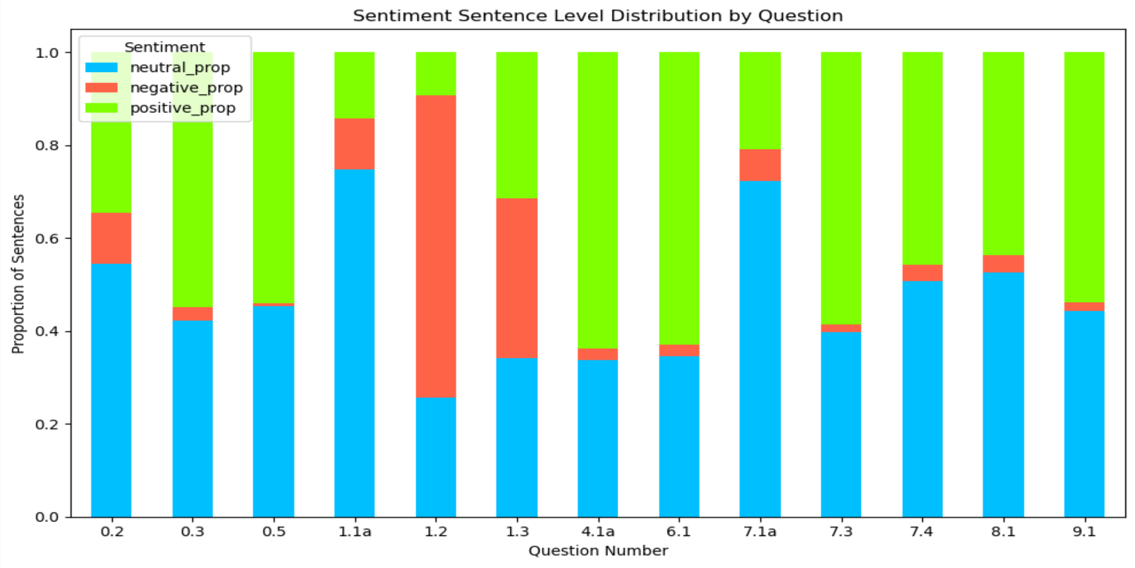


Figure 10. Proportion of positive, neutral, and negative sentences in each question.

task, for our dataset we have selected the responses to these 40 questions, along with their corresponding SDGs. We removed all answers without associated SDGs, and took the remaining text answers and their SDG labels as our dataset. There are 4937 text-SDG pairs in the dataset. Table 8 shows the number of SDG labels in the dataset, representing a case of unbalanced data, which may affect classifier performance.

SDG	SDG1	SDG7	SDG10	SDG11	SDG12	SDG13	SDG16	SDG17
Number of SDG	895	144	130	4424	160	4584	270	637

Table 8. Statistics of SDG labels in the dataset.

In order to train a classifier and evaluate its performance, we randomly split the dataset into train subset (80%) and test subset (20%), having 3949 samples for training and 988 samples for testing. Precision, recall, F1-score, and accuracy (all common evaluation measures) were used to evaluate the performance of classifiers.

Random Forest Classifiers with different max depths were first created in the text classification task. After training, the evaluation results are shown in Table 9. We found that high precision can be seen in most SDGs, while recall and F1-score vary in a wide range. To be more specific, high precision, recall, and F1-score are obtained for certain SDG labels, such as SDG 11 and SDG 13, whose number of samples is much larger than for the the other SDG labels. On the contrary, the classifier performs worse especially on SDG 10 and SDG 16 labels, for which the number of samples is much lower in the test dataset. Precision measures the accuracy of positive inferences by calculating the ratio of true positive inferences to the total number of positive inferences made. In contrast, recall measures the effectiveness of the model in identifying positive instances by determining the ratio of true positive inferences to the total number of actual positive inferences. Essentially, precision focuses on the validity of positive inferences, whereas recall evaluates the ability of the model to identify all relevant instances. Considering the outcomes for SDG 1, SDG

7, SDG 12, and SDG 17, the Random Forest Classifier demonstrates high confidence in its positive inferences, yet overlooked a significant portion of actual positive instances, primarily due to sample imbalance.

Max depth	Metrics	SDG 1	SDG 7	SDG 10	SDG 11	SDG 12	SDG 13	SDG 16	SDG 17	Average
default	Precision	0.96	1.00	0	0.91	1.00	0.93	0	0.86	0.75
	Recall	0.32	0.13	0	1.00	0.21	1.00	0	0.04	0.34
	F1-score	0.48	0.23	0	0.95	0.34	0.96	0	0.09	0.38
24	Precision	0.97	0.00	0	0.91	1.00	0.93	0	0	0.48
	Recall	0.18	0.00	0	1.00	0.04	1.00	0	0	0.28
	F1-score	0.31	0.00	0	0.95	0.08	0.96	0	0	0.29
	Support	174	23	22	896	24	920	65	134	2684 (total)

Table 9. Evaluation results of Random Forest Classifier. Precision, recall,and F1-scores are reported.

In order to address the issue of unbalanced samples in the dataset, we used a weighted loss function. The weight of each SDG label was adjusted inversely proportional to the label frequencies in the training subset. It should be noted that, for multi-label output in our case, the weights of each column of SDG label will be multiplied. By applying the weighted loss function, the evaluation results of the Random Forest are presented in Table 10. A slight improvement of recall can be seen in SDG 1, SDG 7, and SDG 17. In most cases, the weighted loss function did not improve the performance of the classifier. A possible reason is that the default *max_depth* set parameter grows the trees out fully, which makes every leaf node become pure. Limiting *max_depth* could reduce the risk of overfitting and improve the generalization ability of the model.

Max depth	Metrics	SDG 1	SDG 7	SDG 10	SDG 11	SDG 12	SDG 13	SDG 16	SDG 17	Average
default	Precision	0.96	1.00	0	0.91	1.00	0.93	0	0.90	0.71
	Recall	0.40	0.13	0	1.00	0.17	1.00	0	0.07	0.33
	F1-score	0.56	0.23	0	0.95	0.29	0.96	0	0.12	0.39
24	Precision	0.80	0.83	0	0.92	0.75	0.93	0	0.85	0.64
	Recall	0.59	0.43	0	1.00	0.25	1.00	0	0.25	0.44
	F1-score	0.68	0.57	0	0.96	0.38	0.96	0	0.38	0.49
	Support	174	23	22	896	24	920	65	134	2684 (total)

Table 10. Evaluation results of Random Forest Classifier with balanced samples.

5.2.3 Conclusion

In conclusion, the main contributions of this work are the following. First, leveraging the CDP city dataset, we conducted exploratory and descriptive analyses to understand European cities' responsiveness towards questions of environmental themes. Second, we developed a set of text embeddings of the questionnaire answers and used cosine similarity to study the specificities of cities and the common points in their answers. Third, we used a RoBERTa model to detect sentiment across cities and questions, analyzing the sentimennt of the sentences in the dataset, and obtaining an overview of the city's individual attitude towards the selected questions, as well as the sentiment composition of each selected question. Finally, we implemented random forests for SDG classification from text answers, and study a first way the to reduce the bias of classifiers by balancing the dataset.

5.2.4 Relevant publications

A paper describing this work is in preparation.

5.3 Interactive tool for the analysis of textual responses about sustainability of cities in Europe and the US

Contributing partners: Idiap

5.3.1 Introduction and methodology

Buidling on the work described in the previous section, we developed a tool that interactively visualizes how cities respond to the CDP-ICLEI questionnaire to observe patterns of similarity between cities, as well as to understand how a city responds to the questionnaire. This can be very useful for ordinary citizen who would like to understand what their city (and the cities in their country or region) are doing. The tool would also be useful for city officers reporting in this questionnaire, to allow them to compare their city with others, and in this way perhaps take inspiration from what other peeer cities are doing, as well as participate more actively in this type of initiatives.

For the creation of the tool, we reused the sentiment and similarity analysis of European cities done in Python, and adapted it to Dash (a Python framework for building web applications). In addition, we extended the European analysis to the United States, so that the current tool allows both European and American analyses.

The tool has three tabs. The first one is a page where basic instructions on how to use the tool are given. This tab allows to select the region of interest for analysis (whether Europe or US), and shows a map that quantifies the number of cities per country/state in each region.

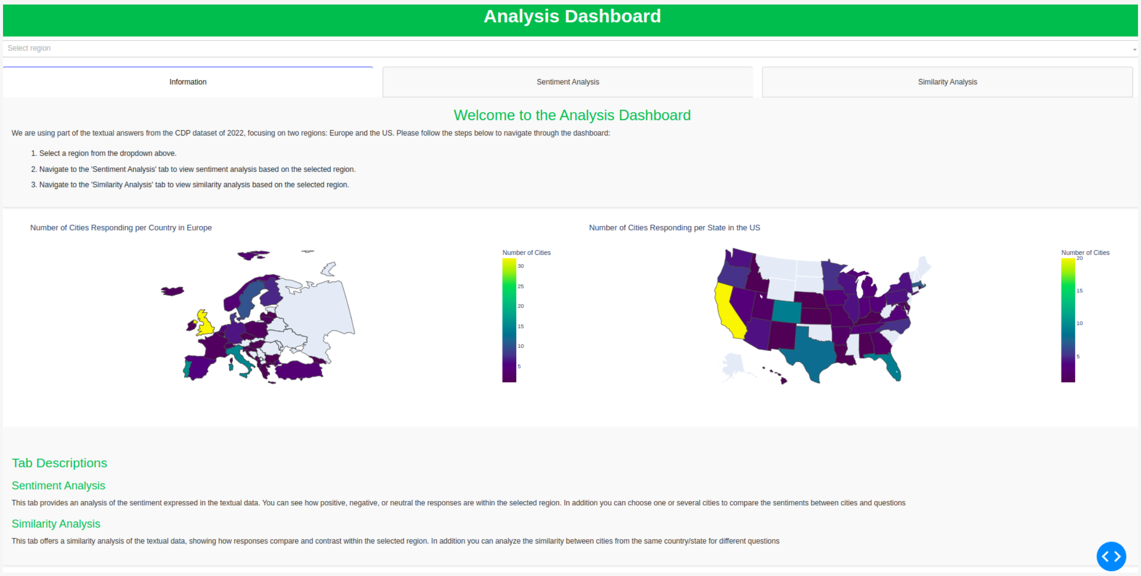
The second tab corresponds to sentiment analysis, where once the European or American regions are chosen, the tools shows a graph with the percentage of positive, negative and neutral statements for each question. Users can also choose one or more cities, and for each city it will show the distribution of positive, negative and neutral sentences for the questions. Finally, a specific button allows, for a given question, to display a couple of examples of positive, negative, and neutral phrases that a city has answered.

Finally, the third tab corresponds to the similarity analysis. After a region (Europe or US) and a question are both chosen, the tools show the pairs of most and least similar cities, as well as the average similarity computation of each city. This is calculated by averaging the similarity of each city with the rest of the cities. The tool also has a city selector, which allows to see the response to a given question, which 3 cities are more similar in their response, which 3 cities are less similar in their response, and within the same country/state, it also shows the similarity with the rest of the cities that have responded to that question.

5.3.2 Experiments

In this section we show some screenshots of the tool. Figure 11 shows the initial tab, with the instructions to follow to use the tool, the two interactive maps (Europe and US) where we show the number of cities that participated in the questionnaire.

Figure 12 presents an example of use of the sentiment analysis tab, where one question and one or more cities were chosen, and some examples of positive, negative, and neutral sentences as well as the distribution of the type of sentences for each question answered by the city, are displayed.



Welcome to the Analysis Dashboard

We are using part of the textual answers from the CDP dataset of 2022, focusing on two regions: Europe and the US. Please follow the steps below to navigate through the dashboard:

1. Select a region from the dropdown above.
2. Navigate to the 'Sentiment Analysis' tab to view sentiment analysis based on the selected region.
3. Navigate to the 'Similarity Analysis' tab to view similarity analysis based on the selected region.

Number of Cities Responding per Country in Europe
Number of Cities
30
25
20
15
10
5
Number of Cities Responding per State in the US
Number of Cities
20
15
10
5

Tab Descriptions

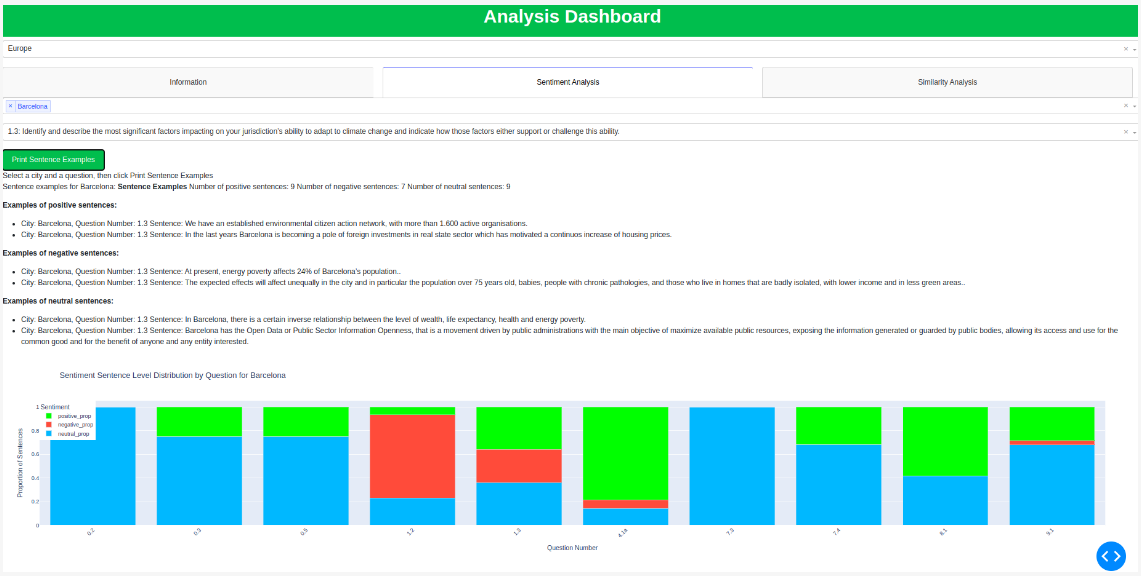
Sentiment Analysis

This tab provides an analysis of the sentiment expressed in the textual data. You can see how positive, negative, or neutral the responses are within the selected region. In addition you can choose one or several cities to compare the sentiments between cities and questions

Similarity Analysis

This tab offers a similarity analysis of the textual data, showing how responses compare and contrast within the selected region. In addition you can analyze the similarity between cities from the same country/state for different questions

Figure 11. Initial tab with the instructions to use the tool.



1.3: Identify and describe the most significant factors impacting on your jurisdiction's ability to adapt to climate change and indicate how those factors either support or challenge this ability.

Select a city and a question, then click Print Sentence Examples
Sentence examples for Barcelona: **Sentence Examples** Number of positive sentences: 9 Number of negative sentences: 7 Number of neutral sentences: 9

Examples of positive sentences:

- City: Barcelona, Question Number: 1.3 Sentence: We have an established environmental citizen action network, with more than 1.600 active organisations.
- City: Barcelona, Question Number: 1.3 Sentence: In the last years Barcelona is becoming a pole of foreign investments in real state sector which has motivated a continuos increase of housing prices.

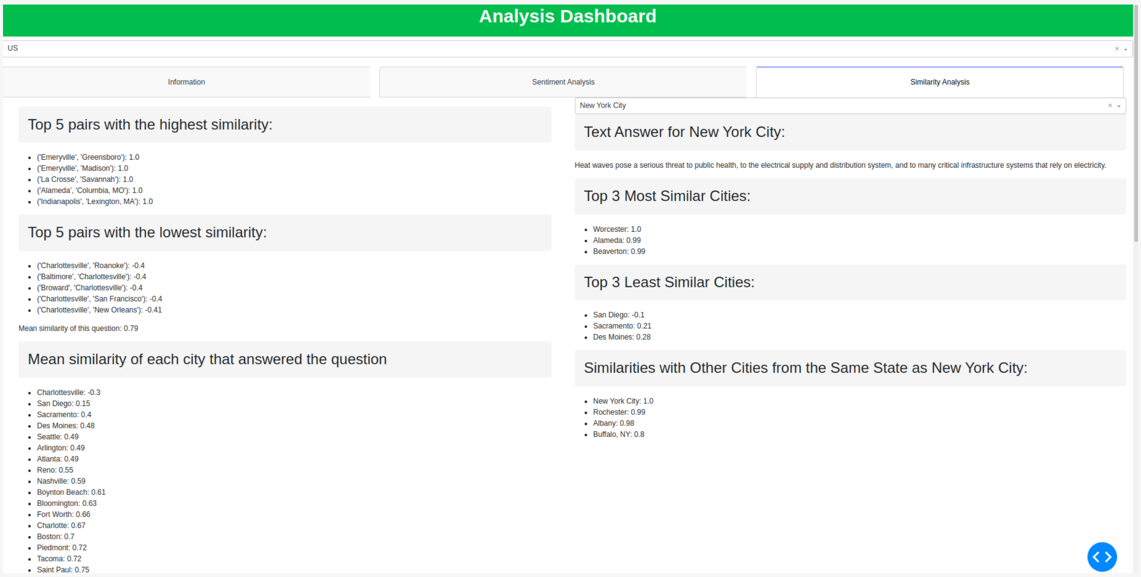
Examples of negative sentences:

- City: Barcelona, Question Number: 1.3 Sentence: At present, energy poverty affects 24% of Barcelona's population..
- City: Barcelona, Question Number: 1.3 Sentence: The expected effects will affect unequally in the city and in particular the population over 75 years old, babies, people with chronic pathologies, and those who live in homes that are badly isolated, with lower income and in less green areas..

Examples of neutral sentences:

- City: Barcelona, Question Number: 1.3 Sentence: In Barcelona, there is a certain inverse relationship between the level of wealth, life expectancy, health and energy poverty.
- City: Barcelona, Question Number: 1.3 Sentence: Barcelona has the Open Data or Public Sector Information Openness, that is a movement driven by public administrations with the main objective of maximize available public resources, exposing the information generated or guarded by public bodies, allowing its access and use for the common good and for the benefit of anyone and any entity interested.

Figure 12. Sentiment analysis tab.



Analysis Dashboard

US

Top 5 pairs with the highest similarity:

- ('Emeryville', 'Greensboro'): 1.0
- ('Emeryville', 'Madison'): 1.0
- ('La Crosse', 'Savannah'): 1.0
- ('Alameda', 'Columbia, MO'): 1.0
- ('Indianapolis', 'Lexington, MA'): 1.0

Top 5 pairs with the lowest similarity:

- ('Charlottesville', 'Roanoke'): -0.4
- ('Baltimore', 'Charlottesville'): -0.4
- ('Broward', 'Charlottesville'): -0.4
- ('Charlottesville', 'San Francisco'): -0.4
- ('Charlottesville', 'New Orleans'): -0.41

Mean similarity of this question: 0.79

Mean similarity of each city that answered the question

- Charlottesville: -0.3
- San Diego: 0.15
- Sacramento: 0.4
- Des Moines: 0.48
- Seattle: 0.49
- Arlington: 0.49
- Atlanta: 0.49
- Reno: 0.55
- Nashville: 0.59
- Boynton Beach: 0.61
- Bloomington: 0.63
- Fort Worth: 0.66
- Charlotte: 0.67
- Boston: 0.7
- Piedmont: 0.72
- Tacoma: 0.72
- Saint Paul: 0.75

New York City

Text Answer for New York City:

Heat waves pose a serious threat to public health, to the electrical supply and distribution system, and to many critical infrastructure systems that rely on electricity.

Top 3 Most Similar Cities:

- Worcester: 1.0
- Alameda: 0.99
- Beaverton: 0.99

Top 3 Least Similar Cities:

- San Diego: -0.1
- Sacramento: 0.21
- Des Moines: 0.28

Similarities with Other Cities from the Same State as New York City:

- New York City: 1.0
- Rochester: 0.99
- Albany: 0.98
- Buffalo, NY: 0.8

Figure 13. Similarity analysis tab.

Finally, Figure 13 presents an example of the similarity analysis, where we have chosen New York City and one of the questions, and the tool can compare its answer with the answer of the rest of the cities.

5.3.3 Conclusion

In conclusion, the contribution of this work is the creation of an interactive tool that allows for aggregated analysis of how cities contributing to the CDP questionnaire responded to this survey. We believe that this tool would be useful for cities to better understand how they responded to this questionnaire in comparison to how other cities did so too. Sustainability is everyone's responsibility, and therefore we believe that cities could support and collaborate with each other; the visualization tool is a small step in that direction.

5.3.4 Relevant software/datasets/other outcomes

We would like to make the tool publicly available in case we obtain authorization from CDP-ICLEI.

5.4 Machine learning methods to understand sustainable economic development

Contributing partners: ANITI, TSE

5.4.1 Introduction and methodology

Machine learning methods have emerged as important tools in the study of economic development thanks to their ability to explain international variations in economic growth, inequality, and emissions[105], as well as due to their ability to generate new sources of data. In this project we contributed to advancing these efforts by generating new datasets on digital trade [100], long term trends in economic development[12], and by studying the historical role played by knowledge diffusion in European knowledge agglomerations [2]

5.4.2 Results

5.4.2.1 Digital trade. We live in a world in which digital trade has become increasingly important. Yet, despite global efforts to harmonize international trade statistics, our understanding of digital trade and its implications remains limited. In a recent paper, we introduced a method to estimate bilateral exports and imports for dozens of sectors starting from the corporate revenue data of large digital firms. This method allows us to provide estimates for digitally ordered and delivered trade involving digital goods (e.g. video games), productized services (e.g. digital advertising), and digital intermediation fees (e.g. hotel rental), which together we call digital products. We used these estimates to explore five key aspects of digital trade, finding that compared to trade in physical goods, digital product exports are more spatially concentrated, have been growing faster, and can offset trade balance estimates, like the United States trade deficit on physical goods. We were also able to show that countries that have decoupled economic growth from greenhouse gas emissions tend to have larger digital exports and that digital exports contribute positively to the complexity of economies. This method, dataset, and findings provide a new lens to understand the impact of international trade in digital products.

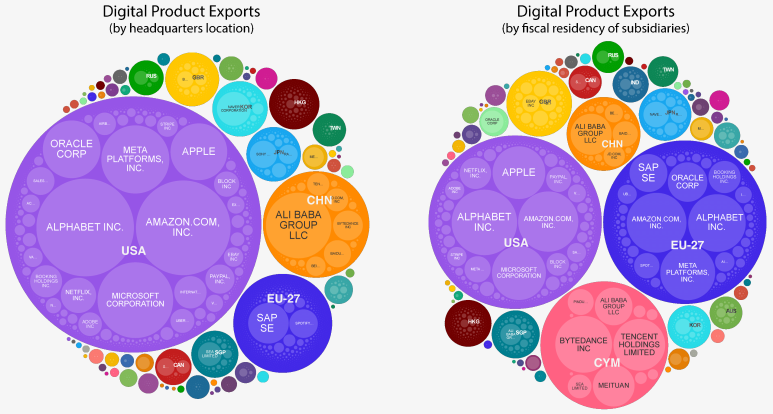


Figure 14. Digital trade estimates based on corporate revenue data using two assignment criteria: subsidiaries and headquarters.

5.4.2.2 Long Term Trends in Economic Development. Understanding long term trends in economic development is a key issue for sustainable development. In a recent study, we asked whether we could use data on the biographies of historical figures to estimate the GDP per capita of countries and regions. In that study we introduced a machine learning method to estimate the GDP per capita of dozens of countries and hundreds of regions in Europe and North America for the past 700 years starting from data on the places of birth, death, and occupations of hundreds of thousands of historical figures. We built an elastic net regression model to perform feature selection and generate out-of-sample estimates that we show explain 90% of the variance in known historical income levels. We used this model to generate GDP per capita estimates for countries, regions, and time periods for which this data is not available and externally validated our estimates by comparing them with four proxies of economic output: urbanization rates in the past 500 years, body height in the 18th century, wellbeing in 1850, and church building activity in the 14th and 15th century. Additionally, we showed our estimates reproduce the well-known reversal of fortune between southwestern and northwestern Europe between 1300 and 1800 and find this is largely driven by countries and regions engaged in Atlantic trade. These findings help validate the use of fine-grained biographical data as a method to produce historical GDP per capita estimates. We publish our estimates with confidence intervals together with all collected source data in a comprehensive dataset.

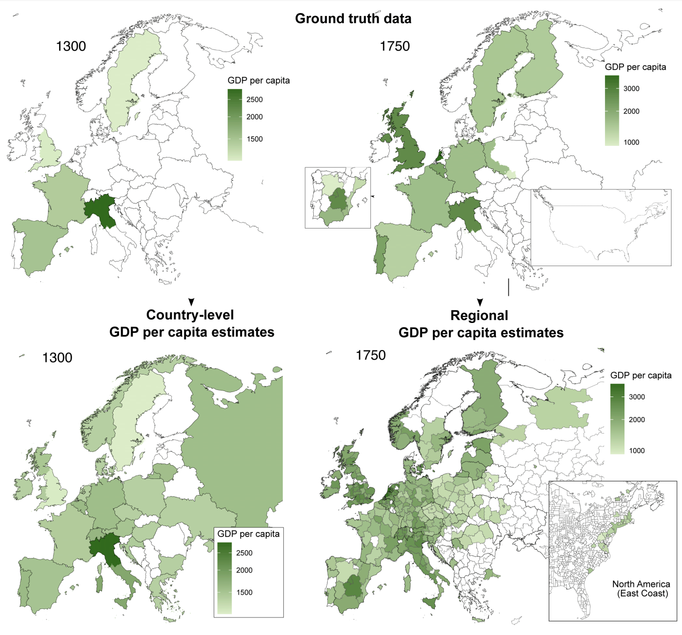


Figure 15. Historical GDP per capita estimates derived from fine grained biographical data. The image compares the ground truth data (top two panels) with the estimates (bottom two panels).

5.4.2.3 The Historical Role Played by Migrants in European Knowledge Agglomerations. Another question that can be explored using economic complexity methods is that of the role of migrants in the historical development of European knowledge agglomerations. In other words, did migrants help make Paris a mecca for the arts and Vienna a beacon of classical music?

Or was their rise a pure consequence of local actors? In a paper published in 2023 we use data on more than 22,000 historical individuals born between the years 1000 and 2000 to estimate the contribution of famous immigrants, emigrants and locals to the knowledge specialisations of European regions. We use measures of relatedness, which are built on collaborative filtering methods, to show that the probability that a region develops or keeps specialisation in an activity grows with both the presence of immigrants with knowledge about that activity and immigrants with knowledge in related activities. In contrast, we do not find robust evidence that the presence of locals with related knowledge explains entries and/or exits. We address some endogeneity concerns using fixed-effects models considering any location–period–activity-specific factors (e.g., the presence of a new university attracting scientists).

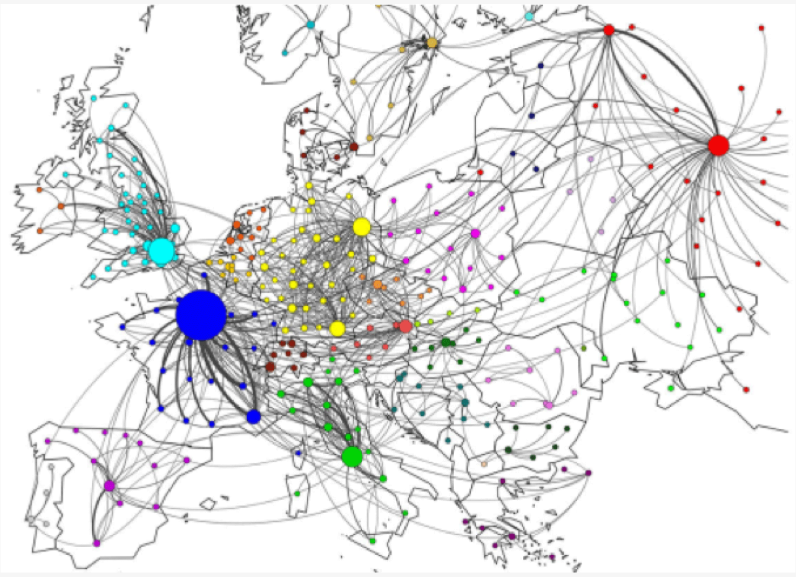


Figure 16. Estimated historical migration network used by [2] to construct historical knowledge flows in Europe and study the role played by migrants in the development of European knowledge agglomerations.

5.4.3 Relevant publications

- Stojkoski, Viktor, Philipp Koch, Eva Coll, and César A. Hidalgo. ”Estimating digital product trade through corporate revenue data.” Nature Communications 15, no. 1 (2024): 5262.
- Koch, Philipp, Viktor Stojkoski, and César A. Hidalgo. ”The role of immigrants, emigrants and locals in the historical formation of European knowledge agglomerations.” Regional Studies (2023): 1-15.
- Koch, Philipp, Viktor Stojkoski, and César A. Hidalgo. ”Quadrupling Historical GDP per Capita Data.” (Under Review)
- Liang, Xiaofan, César A. Hidalgo, Pierre-Alexandre Balland, Siqi Zheng, and Jianghao Wang. ”Intercity connectivity and urban innovation.” Computers, Environment and Urban Systems 109 (2024): 102092.

- Barza, Radu, Edward L. Glaeser, César A. Hidalgo, and Martina Viarengo. Cities as Engines of Opportunities: Evidence from Brazil. No. w32426. National Bureau of Economic Research, 2024.

5.4.4 Relevant software/datasets/other outcomes

Data resources created and delivered during this reporting period include:

- Digital trade: Figshare link
- Historical GDP per capita estimates: Figshare link

6 T2.4 Design of algorithms for supporting efficient and coordinated use of resources

6.1 Learning in digital markets and multi-agent learning

Contributing partners: UNIMI

6.1.1 Introduction and methodology

Digital markets are virtual platforms where agents interact to exchange goods, services, and financial assets. These interactions can take various forms depending on the characteristics of the corresponding mechanisms regulating how the agents interact with each other on the specific platform. For example, brokers striving to facilitate transactions between traders in over-the-counter markets or on ride-sharing platforms, sellers aiming to price their goods efficiently on e-commerce platforms, and buyers participating in auctions to win ad impressions in the ads markets are worth mentioning examples.

The internet enables these markets to operate on a massive scale, generating an enormous stream of data that calls for automated methods. Beyond the challenge of processing vast amounts of information on the fly, a crucial challenge is that agents operate in an only partially known environment where their actions influence the amount of information collected. In turn, this affects their understanding of the environment and, consequently, the identification of the best actions to take. By offering a framework for sequential decision-making in a partially known environment, online learning provides a theoretical perspective through which the stream of data arising in digital markets can be interpreted and managed. In online learning, an agent sequentially chooses an action from a set of possible actions based on the information collected from previous interactions with an otherwise unknown environment. After each interaction, the agent receives a reward and action-dependent feedback that they can use to inform future decisions. The goal of the agent is to maximize their cumulative reward over a certain period, a process formalized through the concept of regret minimization, which measures the difference between the expected cumulative reward of a benchmark strategy and the actual expected cumulative reward the agent earns through their actions during the learning process. Crucially, we can interpret the entanglement in digital markets between action selection and information collection as one of the central themes of online learning, known as the exploration-exploitation dilemma: the need to balance exploration (where agents try new actions to learn more) and exploitation (where they use existing knowledge to maximize reward).

Multi-agent learning extends this online learning framework to scenarios where multiple agents interact within a shared environment. For example, multiple agents may compete for the same resources or rather cooperate to achieve common goals by enhancing the collective performance through shared learning and feedback.

6.1.2 Main theoretical results

Online learning in first-price auctions. We studied the problem of an agent participating in a sequence of first-price auctions to win objects whose value is discovered only when the auction is won, with a particular emphasis on the role that the level of transparency of the auctioneer (i.e., the amount of information disclosed after each auction) plays in how fast the agent can learn to play optimally. This problem has gained significance following the shift to first-price auctions in the ads market by Google AdSense at the end of 2021, following similar switches by Google AdManager and AdMob. By providing a complete characterization of the regret regimes under a variety of assumptions about the properties of the environment (stationarity and/or smoothness) and the level of transparency (all/the highest/no bids are revealed by the auctioneer after each auction), we developed a further understanding of how to devise optimal strategies that software for automatic bidding (autobidders) can use to maximize their returns in these auctions.

Pursuing fairness in repeated bilateral trade. We explored how to achieve fairness in the online learning problem of repeated bilateral trade. In this context, an intermediary platform sets prices for a sequence of seller and buyer pairs, aiming to facilitate transactions. A transaction occurs when the proposed price is higher than the seller's private valuation and lower than the buyer's private valuation.

The existing literature on online learning for repeated bilateral trade [106, 107] has already investigated optimal regret rates under various assumptions, such as stationarity, Lipschitzness of the valuation cumulative distributions, and independence between sellers' and buyers' valuations. These studies primarily focus on maximizing efficiency, modeled as the cumulative net increase in market value, known as the gain from trade. However, this performance metric overlooks fairness considerations: it treats all trades equally as long as they result in a net increase in market value, which is simply the sum of the net increase in utilities for both the seller and the buyer.

In contrast, our research emphasizes fairness in the division of utilities between sellers and buyers. We reward the platform based on a different performance metric: the minimum of the net increase in seller and buyer utilities. By focusing on this metric, the goal is to ensure a more equitable distribution of benefits from trades.

We developed algorithms that achieve optimal regret rates, considering the amount of information the platform obtains after each interaction, under the aforementioned variety of assumptions studied in the literature. Furthermore, our study covers two different feedback mechanisms: full and two-bit. With full feedback, which corresponds to direct-revelation mechanisms, sellers and buyers reveal their private valuations after each interaction. With two-bit feedback, corresponding to posted-price mechanisms, the platform only learns whether the sellers and buyers accepted or rejected the proposed price.

This work is particularly relevant for intermediary platforms, such as ride-sharing services, where online learning methods are required due to the vast amount of information that has to be processed sequentially and where unfair outcomes between sellers and buyers can cause one group to leave the platform, disrupting the service.

Online Learning for brokerage. We investigated a variant of the repeated bilateral trade problem where a broker interacts with a sequence of traders who do not have definite seller and buyer roles. Instead, they decide to buy or sell their assets based on whether they perceive the brokerage price as too low or too high. The ideal outcome is for the asset to pass from the trader who values it less (acting as the seller) to the trader who values it more (acting as the buyer) by paying the brokerage price whenever it lies between their private valuations. We studied this problem under the assumption that the broker's goal is to maximize efficiency, which we modeled using two different benchmarks.

Following the existing literature on repeated bilateral trade, the first approach focuses on maximizing the cumulative gain from trade. The second approach aims to maximize the cumulative number of successful interactions. In stationary environments where the cumulative distribution of traders' asset valuations is Lipschitz, we demonstrated that an agent aiming to maximize the cumulative gain from trade must discover and post the traders' expected valuation for the asset. Conversely, to maximize the cumulative number of successful interactions, the agent should post the median valuation.

We developed algorithms that achieve optimal regret rates based on the amount of information (full feedback or two-bit feedback) the broker discovers after each interaction. Interestingly, allowing traders to have fluid seller and buyer roles leads to a significant improvement in regret rates compared to the standard repeated bilateral trade problem.

Additionally, we explored the contextual case where some information about the traded objects is revealed before each interaction. We characterized the regret regimes when the reward function is the gain from trade, under the assumption of an unknown noisy linear relationship between the context and the traders' valuations, in both the full and two-bit feedback cases.

Our research is particularly relevant for brokerage in over-the-counter markets, decentralized alternatives to traditional financial markets, where brokers play a crucial role in bridging the gap between traders (who may not have direct access to each other) and in performing price discovery.

Cooperative online learning with feedback graphs. We studied the problem of cooperative online learning with feedback graphs, where a network of agents collaboratively solve tasks by sharing feedback through a communication network. This study aims to address the challenges faced in distributed systems, such as geographically distributed learning environments where nodes handle high volumes of prediction requests, as it happens in advertising and finance. Crucially, these systems require methods to optimize performance without global synchronization of locally updated models. By analyzing the interplay between feedback graphs and communication, we provided a characterization of regret in terms of the independence number of the strong product between the feedback graph and the communication network. Our theoretical and empirical findings show that our proposed algorithm EXP3-α2, significantly improves learning rates with respect to isolated learning.

6.1.3 Relevant publications

- N. Cesa-Bianchi, T. Cesari, R. Colomboni, F. Fusco, and S. Leonardi. The role of transparency in repeated first-price auctions with unknown valuations. 56th Annual ACM Symposium on Theory of Computing (STOC), 2024. [13]
- N. Bolić, T. Cesari, R. Colomboni (2023). An online learning theory of brokerage. arXiv preprint arXiv:2310.12107. [15]
- F. Bachoc, N. Cesa-Bianchi, T. Cesari, R. Colomboni (2024). Fair Online Bilateral Trade. arXiv preprint arXiv:2405.13919. [14]
- N. Cesa-Bianchi, T. Cesari, R. Della Vecchia (2024). Cooperative online learning with feedback graphs. TMLR. [16]

6.2 Federated generalized category discovery

Contributing partners: UNITN

6.2.1 Introduction and methodology

Generalized category discovery (GCD) seeks to categorize unlabeled samples from known and unknown classes by leveraging labeled data of known classes. While existing GCD methods [108, 109, 110, 111, 112] have achieved promising performance, they always require centralized training, where all the training data are required to be collected and stored in the center server in advance. However, this condition is not suitable for many security-critical application scenarios, such as

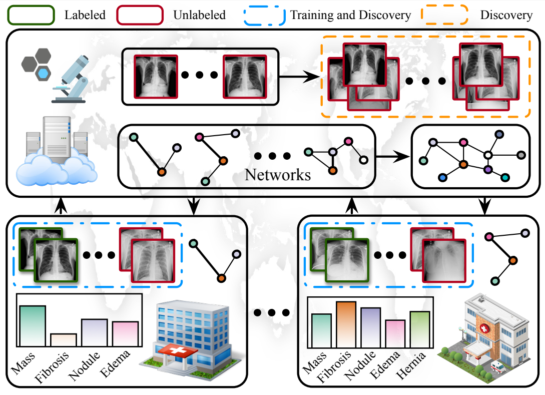


Figure 17. Conceptual diagram of the proposed pFed-GCD with the case of global disease discovery. In this case, the data are distributively collected from different hospitals all over the world, which are partially annotated. Each hospital/client stores both labeled and unlabeled data that may share some common categories with the other clients. Moreover, the raw data in local clients cannot be shared with the central server or other clients, due to data privacy. The goal of pFed-GCD is to jointly improve personalized GCD models in clients and train a robust generic GCD model on the server, via client collaboration under the privacy constraint.

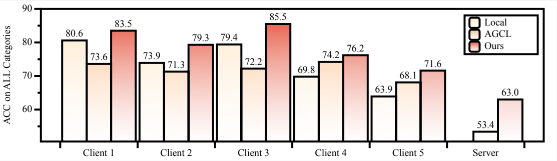


Figure 18. Exploration Experiments. ”Local” indicates that models are trained individually for each client without federated communication. We empirically found that existing Fed-GCD methods (e.g., AGCL [3]) sacrifice the client GCD performance for training a global model.

healthcare, finance, and transportation. For example, a well-established disease diagnosis system (Figure 17) is expected to precisely diagnose known diseases and discover unknown diseases as early as possible, through collaboration and information sharing among local hospitals located in different locations. Nevertheless, the data collected and annotated by different hospitals cannot be shared with others due to different local laws and regulations of privacy protection. Therefore, a trustworthy GCD system is required to be equipped with the capability of decentralized training.

To study the decentralized GCD systems under privacy constraints, Pu et al. [3] propose a Federated GCD (Fed-GCD) task, where the GCD data are individually collected and partially annotated by local clients but cannot be shared with other clients. The objective of Fed-GCD is to train a generic GCD model via collaboration across local clients without sharing local samples. However, as illustrated in Figure 18, we experimentally found that such an objective harms the performances of local models that are applied to individual personalized unlabeled data. This further leads to the degradation of the generic model due to the knowledge collapse.

To solve this limitation, in this research, we focus on a more practical Fed-GCD setting, namely personalized Fed-GCD (pFed-GCD), which aims to not only improve the personalized GCD abilities of local models but also to encourage the global model to learn more generic representations. To this end, we propose a new Personalized Local-graph Contrastive Learning (PLCL) framework with a tailor-made masked KNN-former, to jointly improve each local model’s personalized GCD ability and the global model's generalization ability. Our framework includes personalized Local-Graph Learning (pLGL) and personalized Hybrid Contrastive Learning (pHCL). Specifically, first, we leverage pLGL with a masked KNN-former to learn personalized contrastive relationships between instances for different clients, following a progressive paradigm. Meanwhile, by personalized aggregating and masking KNN-formers, local models can benefit from complementary knowledge from other clients while mitigating the knowledge conflict and collapse. Second, pHCL leverages the personalized contrastive relationships learned by pLGL to enhance local representations on both instance and cluster levels. Furthermore, with adaptive parameter masking, pHCL disentangles personalized and generic knowledge, which can keep personalized performance as well as can learn more comprehensive generic representations.

6.2.2 Experiments

Experimental setup **Dataset.** Apart from the three generic datasets (*i.e.*, CIFAR-10 [113], CIFAR-100 [113] and ImageNet-100 [108]) and the three fine-grained datasets (*i.e.*, CUB-200 [114], Stanford Cars [115], and Oxford-IIIT Pet [116])considering [112] used in [3], we reorganize two more practical long-tailed datasets [117, 118] to verify the effectiveness of pFed-GCD models, where Herbarium 19 [118] is a natural image dataset including 683 types of herbs, and NIH-CXR-LT [117] is an X-ray image dataset, containing 20 types of medical diagnoses. For each dataset, first, we leverage the β-Dirichlet distribution [119] to split the original training set into N^C subsets, separately stored in clients as the local datasets. Then, for each client, we sample a subset of half the classes as "Old" categories in its local dataset, and 50% of instances of each labeled class are drawn to form the local labeled set. The remaining data are regarded as the local unlabeled set. We set N^C=5 in all experiments.

Evaluation protocols Based on the target of our pFed-GCD task, we evaluate performances of methods by measuring both the personalized GCD ability of the local models and the generalization ability of the global model. For the former, we directly use the GCD classifier to predict class labels and measure classification accuracy. For the latter, following [120], we assume that the ground-truth class number is known. Thus, we use k-mean [121] with the ground-truth class number to cluster unlabeled test samples on the server. Then, we use Hungarian algorithm [122] to obtain the optimal assignment between ground-truth labels and predicted class labels, then calculate the clustering accuracy. We measure the accuracies of "All", "Old" and "New" categories for both evaluations. Each experiment is repeated three times and averaged results are reported.

Performance evaluation In this section, we conduct extensive experiments to demonstrate the effectiveness of the proposed PLCL. The compared methods include three groups. First, we use SimGCD [120] to individually train local GCD models without parameter mask and client communication, namely "Local-Sim". Second, we reproduce the recent federated GCD methods (Fed-GCD [3] and AGCL [3]) with non-parametric inference. Third, we compare PLCL with the baseline method based on SimGCD [120], namely "Fed-Sim", and adapt the state-of-the-art centralized GCD methods (DCCL [112] and PromptCAL [123]) into our pFed-GCD framework with parametric classifiers [120], resulting in "Fed-Sim", "Fed-Sim + DCCL" and "Fed-Sim + CAL", respectively. Specifically, for "Fed-Sim + CAL", we inject the contrastive affinity learning objective in [123] into SimGCD [120] for local representation learning without changing the backbone network to prompt version. Similarly, as for "Fed-Sim + DCCL", we plug its cluster-level contrastive learning of DCCL [112] in SimGCD [120] for improving the backbone network.

Table 11. Results on three generic datasets with both personalized evaluation and generalization evaluation.

Methods	Personalized Performance Evaluation (Average of 5 Clients)									Generalization Performance Evaluation								
	CIFAR-10			CIFAR-100			ImageNet-100			CIFAR-10			CIFAR-100			ImageNet-100		
	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New
Local-Sim [120]	95.0$\pm_{0.3}$	95.1$\pm_{0.4}$	93.3$\pm_{0.2}$	73.5$\pm_{0.4}$	75.8$\pm_{0.2}$	65.9$\pm_{0.1}$	78.5$\pm_{0.3}$	90.3$\pm_{0.2}$	71.4$\pm_{0.4}$	-	-	-	-	-	-	-	-	-
Fed-GCD [3]	92.1$\pm_{0.3}$	93.9$\pm_{0.5}$	86.2$\pm_{0.3}$	71.6$\pm_{0.4}$	73.1$\pm_{0.2}$	64.8$\pm_{0.2}$	74.1$\pm_{0.1}$	88.1$\pm_{0.4}$	66.3$\pm_{0.1}$	93.7$\pm_{0.4}$	95.1$\pm_{0.1}$	91.5$\pm_{0.2}$	47.3$\pm_{0.5}$	49.2$\pm_{0.3}$	45.9$\pm_{0.5}$	76.4$\pm_{0.2}$	84.8$\pm_{0.3}$	72.1$\pm_{0.4}$
AGCL [3]	92.7$\pm_{0.2}$	93.5$\pm_{0.5}$	90.2$\pm_{0.4}$	73.2$\pm_{0.1}$	75.1$\pm_{0.4}$	65.3$\pm_{0.4}$	75.2$\pm_{0.3}$	89.4$\pm_{0.1}$	69.7$\pm_{0.1}$	92.5$\pm_{0.4}$	94.4$\pm_{0.4}$	89.2$\pm_{0.5}$	54.1$\pm_{0.4}$	54.3$\pm_{0.3}$	52.5$\pm_{0.4}$	83.1$\pm_{0.5}$	88.1$\pm_{0.5}$	77.0$\pm_{0.2}$
Fed-Sim	96.1$\pm_{0.3}$	95.4$\pm_{0.2}$	96.3$\pm_{0.1}$	74.4$\pm_{0.3}$	76.7$\pm_{0.3}$	67.2$\pm_{0.3}$	79.4$\pm_{0.2}$	91.6$\pm_{0.1}$	73.8$\pm_{0.3}$	95.8$\pm_{0.5}$	96.1$\pm_{0.4}$	93.6$\pm_{0.4}$	55.5$\pm_{0.1}$	58.7$\pm_{0.2}$	50.1$\pm_{0.3}$	81.2$\pm_{0.1}$	85.9$\pm_{0.4}$	75.8$\pm_{0.4}$
+ DCCL [112]	96.4$\pm_{0.5}$	95.5$\pm_{0.1}$	97.0$\pm_{0.2}$	74.1$\pm_{0.1}$	76.3$\pm_{0.2}$	67.9$\pm_{0.3}$	81.1$\pm_{0.4}$	91.7$\pm_{0.5}$	75.2$\pm_{0.4}$	95.6$\pm_{0.3}$	95.7$\pm_{0.2}$	94.1$\pm_{0.2}$	58.4$\pm_{0.3}$	62.3$\pm_{0.4}$	53.8$\pm_{0.3}$	84.7$\pm_{0.3}$	89.2$\pm_{0.3}$	79.5$\pm_{0.1}$
+ CAL [123]	96.5$\pm_{0.4}$	95.8$\pm_{0.5}$	97.1$\pm_{0.4}$	75.8$\pm_{0.3}$	77.9$\pm_{0.4}$	70.7$\pm_{0.2}$	82.3$\pm_{0.5}$	92.1$\pm_{0.2}$	77.9$\pm_{0.4}$	95.8$\pm_{0.3}$	95.9$\pm_{0.3}$	94.3$\pm_{0.3}$	59.7$\pm_{0.1}$	63.2$\pm_{0.1}$	55.5$\pm_{0.1}$	83.9$\pm_{0.4}$	88.5$\pm_{0.1}$	78.9$\pm_{0.5}$
PLCL (Ours)	**97.1**$\pm_{0.3}$	**96.3**$\pm_{0.2}$	**97.2**$\pm_{0.4}$	**80.2**$\pm_{0.1}$	**81.5**$\pm_{0.5}$	**76.6**$\pm_{0.5}$	**83.7**$\pm_{0.2}$	**93.5**$\pm_{0.3}$	**80.3**$\pm_{0.2}$	**96.7**$\pm_{0.3}$	**96.2**$\pm_{0.2}$	**97.5**$\pm_{0.5}$	**62.1**$\pm_{0.4}$	**67.7**$\pm_{0.4}$	**57.9**$\pm_{0.2}$	**86.6**$\pm_{0.3}$	**89.7**$\pm_{0.3}$	**81.4**$\pm_{0.3}$

Table 12. Results on three fine-grained datasets with both personalized evaluation and generalization evaluation.

Methods	Personalized Performance Evaluation (Average of 5 Clients)									Generalization Performance Evaluation								
	CUB-200			Stanford-Cars			Oxford-Pet			CUB-200			Stanford-Cars			Oxford-Pet		
	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New
Local-Sim [120]	73.5$\pm_{1.2}$	77.9$\pm_{1.9}$	70.3$\pm_{2.4}$	43.8$\pm_{0.6}$	56.2$\pm_{1.3}$	41.7$\pm_{2.1}$	83.5$\pm_{0.7}$	87.1$\pm_{1.6}$	82.4$\pm_{0.8}$	-	-	-	-	-	-	-	-	-
Fed-GCD [3]	66.9$\pm_{1.8}$	70.1$\pm_{2.6}$	64.1$\pm_{1.6}$	41.1$\pm_{1.1}$	54.1$\pm_{0.5}$	36.5$\pm_{1.1}$	77.1$\pm_{0.9}$	80.1$\pm_{1.3}$	74.5$\pm_{1.2}$	42.7$\pm_{2.7}$	52.5$\pm_{2.2}$	38.9$\pm_{3.9}$	31.1$\pm_{1.4}$	45.1$\pm_{1.0}$	26.7$\pm_{1.3}$	71.9$\pm_{1.3}$	76.3$\pm_{1.9}$	71.2$\pm_{0.9}$
AGCL [3]	71.9$\pm_{2.3}$	71.2$\pm_{1.5}$	68.7$\pm_{3.1}$	42.4$\pm_{2.0}$	52.9$\pm_{2.5}$	40.8$\pm_{2.9}$	85.7$\pm_{1.3}$	86.0$\pm_{2.3}$	83.7$\pm_{1.1}$	54.0$\pm_{2.6}$	52.7$\pm_{1.5}$	55.4$\pm_{1.6}$	36.4$\pm_{3.9}$	44.9$\pm_{3.4}$	32.8$\pm_{3.6}$	81.2$\pm_{1.7}$	81.7$\pm_{3.1}$	80.5$\pm_{2.4}$
Fed-Sim	75.1$\pm_{1.3}$	76.7$\pm_{0.7}$	73.8$\pm_{1.0}$	45.3$\pm_{1.6}$	56.1$\pm_{2.2}$	43.2$\pm_{1.9}$	84.2$\pm_{1.3}$	87.9$\pm_{1.7}$	82.3$\pm_{2.2}$	50.5$\pm_{3.1}$	53.1$\pm_{1.4}$	49.1$\pm_{0.7}$	32.6$\pm_{0.6}$	45.5$\pm_{2.4}$	29.2$\pm_{0.9}$	76.6$\pm_{1.8}$	77.9 $\pm_{1.1}$	74.7$\pm_{2.1}$
+ DCCL [112]	77.7$\pm_{1.6}$	78.3$\pm_{0.9}$	75.6$\pm_{1.7}$	44.8$\pm_{1.1}$	53.1$\pm_{1.4}$	43.1$\pm_{1.5}$	85.1$\pm_{2.1}$	85.9$\pm_{1.7}$	84.9$\pm_{1.8}$	55.2$\pm_{1.7}$	54.5$\pm_{1.3}$	56.3$\pm_{1.6}$	34.7$\pm_{2.5}$	42.6$\pm_{3.1}$	31.7$\pm_{2.5}$	85.4$\pm_{1.4}$	85.1$\pm_{0.7}$	85.9$\pm_{1.8}$
+ CAL [123]	76.4$\pm_{1.1}$	78.6$\pm_{0.9}$	74.2$\pm_{1.9}$	46.5$\pm_{1.0}$	57.3$\pm_{1.5}$	45.0$\pm_{0.9}$	83.2$\pm_{0.7}$	83.1$\pm_{1.3}$	80.3$\pm_{1.3}$	52.9$\pm_{1.5}$	53.7$\pm_{1.1}$	52.1$\pm_{0.7}$	39.2$\pm_{1.2}$	48.1$\pm_{1.7}$	36.5$\pm_{1.5}$	83.9$\pm_{0.7}$	84.0$\pm_{1.4}$	83.8$\pm_{0.9}$
PLCL (Ours)	**79.2**$\pm_{1.2}$	**82.7**$\pm_{1.5}$	**77.6**$\pm_{1.4}$	**51.1**$\pm_{2.1}$	**63.9**$\pm_{2.3}$	**48.7**$\pm_{1.8}$	**88.3**$\pm_{1.7}$	**88.1**$\pm_{1.1}$	**88.5**$\pm_{0.9}$	**63.0**$\pm_{1.5}$	**60.4**$\pm_{1.2}$	**63.1**$\pm_{1.4}$	**42.8**$\pm_{2.4}$	**59.2**$\pm_{1.9}$	**41.3**$\pm_{2.4}$	**87.5**$\pm_{1.1}$	**87.6**$\pm_{1.2}$	**87.1**$\pm_{1.4}$

Summary The experimental results in Table 11, 12, and 13 demonstrate that 1) our PLCL achieves superior performance compared to three group competitors on all datasets. Especially for Herbarium 19 [118], PLCL outperforms the baseline method by 15.3% on generalization evaluation. 2) Although DCCL [112] and CAL [123] significantly improve centralized GCD models, they are still struggling to generate effective supervision for personalized contrastive representation learning, due to the uncertain data distribution. Our PLCL learns the contrastive relationship from data instead of a fixed algorithm, which is more flexible and suitable for open-world scenarios.

6.2.3 Conclusion

The main contributions of this research are as follows:

- We identify the issue of knowledge collapse, a problem significantly neglected in the current Fed-GCD setting, and propose the necessity of investigating personalized Fed-GCD algorithms.
- We introduce a novel Personalized Local-graph Contrastive Learning (PLCL) framework equipped with a masked KNN-former, which can effectively enhance both generic and personalized GCD performance.
- We conduct experiments on various datasets, including the generic, fine-grained, long-tailed natural, and medical image sets, demonstrating that our PLCL achieves state-of-the-art performance across all settings.

6.2.4 Relevant publications

- N. Pu, W. Li, X. Ji, Y. Qin, N. Sebe, and Z. Zhong, Federated Generalized Category Discovery, CVPR 2024 [17].
  Zenodo record: https://zenodo.org/record/11865219

6.3 Cooperative rewards design for efficient resource allocation

Contributing partners: FBK

Table 13. Results on two long-tail datasets with both personalized evaluation and generalization evaluation.

Methods	Personalized Performance Evaluation (Average of 5 Clients)						Generalization Performance Evaluation					
	Herbarium 19			NIH-CXR-LT			Herbarium 19			NIH-CXR-LT		
	All	Old	New	All	Old	New	All	Old	New	All	Old	New
Local-Sim [120]	$38.8\pm_{0.7}$	$51.3\pm_{0.3}$	$28.9\pm_{1.6}$	$29.4\pm_{0.8}$	$30.1\pm_{0.5}$	$15.3\pm_{1.8}$	-	-	-	-	-	-
Fed-GCD [3]	$32.7\pm_{1.5}$	$33.3\pm_{0.3}$	$25.5\pm_{0.9}$	$21.6\pm_{0.5}$	$23.1\pm_{1.9}$	$9.5\pm_{0.6}$	$22.9\pm_{1.2}$	$16.8\pm_{1.5}$	$38.2\pm_{1.1}$	$16.4\pm_{0.6}$	$18.0\pm_{0.8}$	$4.2\pm_{0.6}$
AGCL [3]	$37.0\pm_{1.5}$	$35.8\pm_{0.4}$	$40.4\pm_{1.6}$	$24.2\pm_{1.8}$	$24.4\pm_{0.2}$	$10.3\pm_{1.3}$	$33.4\pm_{0.7}$	$24.1\pm_{0.4}$	$45.0\pm_{1.1}$	$19.6\pm_{1.4}$	$18.9\pm_{0.9}$	$5.4\pm_{1.3}$
Fed-Sim	$35.6\pm_{0.5}$	$35.5\pm_{0.9}$	$36.0\pm_{1.8}$	$27.3\pm_{0.4}$	$28.3\pm_{0.5}$	$13.4\pm_{0.2}$	$28.3\pm_{0.6}$	$28.1\pm_{1.1}$	$29.5\pm_{1.8}$	$22.6\pm_{0.9}$	$23.4\pm_{0.4}$	$8.2\pm_{1.3}$
+ DCCL [112]	$37.1\pm_{0.9}$	$38.4\pm_{0.6}$	$37.0\pm_{1.4}$	$24.9\pm_{1.4}$	$25.3\pm_{0.5}$	$11.6\pm_{1.6}$	$39.7\pm_{1.6}$	$39.1\pm_{0.9}$	$41.3\pm_{0.3}$	$19.9\pm_{1.3}$	$20.4\pm_{0.6}$	$6.8\pm_{0.3}$
+ CAL [123]	$33.2\pm_{1.7}$	$32.8\pm_{1.2}$	$33.0\pm_{1.4}$	$29.1\pm_{0.2}$	$29.8\pm_{0.3}$	$14.7\pm_{1.9}$	$35.1\pm_{0.3}$	$33.7\pm_{0.3}$	$37.6\pm_{0.3}$	$24.1\pm_{0.9}$	$24.9\pm_{0.4}$	$9.6\pm_{1.4}$
PLCL (Ours)	$\mathbf{43.5}\pm_{0.7}$	$\mathbf{55.1}\pm_{1.1}$	$\mathbf{40.8}\pm_{0.9}$	$\mathbf{33.7}\pm_{1.7}$	$\mathbf{34.2}\pm_{1.4}$	$\mathbf{18.5}\pm_{0.7}$	$\mathbf{43.6}\pm_{0.4}$	$\mathbf{41.3}\pm_{1.3}$	$\mathbf{47.8}\pm_{1.8}$	$\mathbf{28.8}\pm_{1.1}$	$\mathbf{29.4}\pm_{0.6}$	$\mathbf{13.3}\pm_{0.8}$

6.3.1 Introduction and methodology

The goal of achieving autonomous efficient allocation is a central challenge in the study of multi-agent systems interacting within a shared environment comprising common spaces and resources. In these systems, each agent is inherently selfish, focused on attaining its individual objectives. This dynamic often leads to commons dilemmas, where the pursuit of personal goals by each agent can result in suboptimal outcomes for the group as a whole. By training agents to cooperate, it is possible to align their individual actions with collective welfare, thereby enhancing the efficiency and sustainability of resource allocation in the shared environment.

Multi-Agent Reinforcement Learning (MARL) [124, 125, 126] presents a flexible framework for developing intelligent agents capable of making decisions in complex, dynamic environments. In decentralized scenarios, each agent operates based on its local observations, retaining its state information and individual rewards privately. This independence poses a challenge for fostering cooperative behaviors essential for achieving global objectives. To address this, designing reward structures that encapsulate the impact of an agent's actions on the broader society becomes imperative [127, 128, 129, 130, 131, 131, 132, 133]. Drawing inspiration from human social cooperation, this approach aims to encourage agents to consider the welfare of others, thus enhancing overall system performance and harmony. Our aim is to explore the integration of socially-aware reward mechanisms within MARL, to achieve more effective and cohesive multi-agent interactions, and hence improve collective environment exploitation, avoiding scenarios of resource depletion.

6.3.2 Scenarios and work in progress

Our research focuses on exploiting communication among agents to further improve coordination and cooperation for resource collection. By incorporating feedback from peers into the reward mechanisms, agents can receive more nuanced and socially-informed evaluations of their actions. This approach leverages the collective intelligence of the agent network, promoting behaviors that are beneficial not only to individual agents but also to the group as a whole. The integration of rewards based on peer feedback aims to create a more cohesive multi-agent system, where the success of each agent is interlinked with the well-being and performance of its peers. The underlying idea is to enhance the MARL framework by fostering a deeper level of social awareness and cooperation among agents.

Our reward design based on exchange of feedback among peers is currently being tested on the Harvest benchmark [131]. The objective of the Harvest game is to collect apples scattered across a common map, each of which provides a reward of 1. The regrowth rate of apples varies across the map and depends on the spatial distribution of uncollected apples: the more nearby apples, the higher the local regrowth rate. Additionally, agents are allowed to generate a firing beam paying a cost of -1 that, when hitting other individuals, causes them damage, in the form of a penalty of -50, and the exclusion from the harvesting game for a window of time. In this game the short-term interest of each individual agent is to harvest apples as quickly as possible. However, the long-term interest of the group is better served if individuals refrain from rapid harvesting, particularly when many agents are concentrated in the same area. Such situations are precarious because the more agents there are harvesting, the higher the risk of permanently depleting local resources. Cooperating agents must forgo immediate personal benefits for the collective good of the group, and hence avoiding or delaying apple depletion in the environment. The underlying MARL algorithm currently implemented in this scenario is Deep Q-Network (DQN) [134], a model-free Reinforcement Learning (RL) algorithm that combines Q-learning with deep neural networks to handle high-dimensional state spaces, but our method generalizes to all the RL algorithms provided of a critic module. Each agent utilizes a DQN to estimate the optimal action-value function, guiding its decisions to maximize long-term rewards. We are testing our method considering both a full communication network to exchange feedback among agents in the system, and a more challenging scenario in which agents can communicate only when they are in each other's field of vision in the shared environment. For testing purposes, we consider measures of efficiency (average episodic reward achieved), sustainability (average and last reward acquisition time), peace (average episodic life duration), and equality (statistics over the difference in episodic rewards achieved by different agents).

7 Conclusions

The activity carried out so far in this workpackage has been quite intense, with a fairly large number of publications in high-profile journals and conferences. In particular, there are seven journal papers, seven conference papers, and seven submitted manuscripts. These achievements are well distributed among the contributing partners and their number and overall quality reflects the active involvement of the partners towards fulfilling the goals of the workpackage. The high number of submissions also attests the amount of ongoing work, some of which will be included in the next deliverable. While this document covers only the initial results, the current trajectory convincingly shows that we are on track with respect to the original planning (also considering that Tasks 2.2–2.4 only officially started at M8). The deliverable D2.2, which is due at M30, will include the outcomes of the ongoing work related to the tasks covered in this workpackage.

An aspect that this deliverable is clearly missing is the reporting of cross collaborations between partners. This is understandable given that this deliverable comes at M12 and many tasks in this workpackage started only at M8. There are, however, a few natural opportunities for starting collaborations. The dataset curated by BD in T2.1 will be used in T2.2 in connection with the topic of societal sustainability, so we may expect several interactions between BD and the partners contributing to T2.2. Also, some of the topics developed by POLIMI in T2.2 (e.g., design of learning algorithms with fairness guarantees) have close connections with the work by UNIMI in T2.4, especially for the part related to learning in digital markets and multi-agent learning. Indeed, the paper [13] is already a collaboration between POLIMI and UNIMI. Moreover, the work on computer vision carried out by UNITN and UNIMORE in T2.2 has the potential of creating collaborations between these two partners. Similarly, the work done by UCPH and JSI in T2.2 on LLMs also lends itself to a cooperation between the partners. Finally, a very interesting direction for collaboration is connecting the thread of research pursued by FBK in T2.4 on cooperative MARL with the work on cooperative online learning carried out by UNIMI in the same task.

References

[1] J. Zhu and Y. Fang, "Learning object-specific distance from a monocular image," in *ICCV*, 2019.

[2] P. Koch, V. Stojkoski, and C. A. Hidalgo, "The role of immigrants, emigrants and locals in the historical formation of european knowledge agglomerations," *Regional Studies*, pp. 1–15, 2023.

[3] N. Pu, Z. Zhong, X. Ji, and N. Sebe, "Federated generalized category discovery," *arXiv preprint arXiv:2305.14107*, 2023.

[4] F. E. Stradi, J. Germano, G. Genalti, M. Castiglioni, A. Marchesi, and N. Gatti, "Online learning in cmdps: Handling stochastic and adversarial constraints," in *Forty-first International Conference on Machine Learning*, 2024.

[5] X. Liu, G. Li, Y. Qi, Z. Han, A. van den Hengel, N. Sebe, M.-H. Yang, and Q. Huang, "Consistency-aware anchor pyramid network for crowd localization," *IEEE Transactions on Pattern Analysis and Machine Learning Intelligence*, 2024.

[6] A. Panariello, G. Mancusi, F. H. Ali, A. Porrello, S. Calderara, and R. Cucchiara, "Distformer: Enhancing local and global features for monocular per-object distance estimation," *arXiv preprint arXiv:2401.03191*, 2024.

[7] C. Navarrete, M. Macedo, R. Colley, J. Zhang, N. Ferrada, M. E. Mello, R. Lira, C. Bastos-Filho, U. Grandi, J. Lang, *et al.*, "Understanding political divisiveness using online participation data from the 2022 french and brazilian presidential elections," *Nature Human Behaviour*, vol. 8, no. 1, pp. 137–148, 2024.

[8] J. Gudiño-Rosero, U. Grandi, and C. A. Hidalgo, "Large language models (llms) as agents for augmented democracy," *arXiv preprint arXiv:2405.03452*, 2024.

[9] D. Wright, A. Arora, N. Borenstein, S. Yadav, S. Belongie, and I. Augenstein, "Revealing fine-grained values and opinions in large language models," *arXiv preprint arXiv:2406.19238*, 2024.

[10] P. E. Christensen, S. Yadav, and S. Belongie, "Prompt, condition, and generate: Classification of unsupported claims with in-context learning," *arXiv preprint arXiv:2309.10359*, 2023.

[11] V. Stojkoski, P. Koch, E. Coll, and C. A. Hidalgo, "The growth, geography, and implications of trade in digital products," *Nature Communications (In press)*, 2024.

[12] P. Koch, V. Stojkoski, and C. A. Hidalgo, "Quadrupling the availability of historical gdp per capita estimates through machine learning," *(Under Review)*, 2024.

[13] N. Cesa-Bianchi, T. Cesari, R. Colomboni, F. Fusco, and S. Leonardi, "The role of transparency in repeated first-price auctions with unknown valuations," in *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pp. 225–236, 2024.

[14] F. Bachoc, N. Cesa-Bianchi, T. Cesari, and R. Colomboni, "Fair online bilateral trade," *arXiv preprint arXiv:2405.13919*, 2024.

[15] N. Bolić, T. Cesari, and R. Colomboni, “An online learning theory of brokerage,” AAMAS ’24, (Richland, SC), p. 216–224, International Foundation for Autonomous Agents and Multiagent Systems, 2024.

[16] N. Cesa-Bianchi, T. Cesari, and R. D. Vecchia, “Cooperative online learning with feedback graphs,” *Transactions on Machine Learning Research*, 2024.

[17] N. Pu, W. Li, X. Ji, Y. Qin, N. Sebe, and Z. Zhong, “Federated generalized category discovery,” in *CVPR*, 2024.

[18] S. Jacob, Y. Qiao, Y. Ye, and B. Lee, “Anomalous distributed traffic: Detecting cyber security attacks amongst microservices using graph convolutional networks,” *Computers & Security*, vol. 118, p. 102728, 2022.

[19] B. Andersen and T. Fagerhaug, *Root cause analysis.* Quality Press, 2006.

[20] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming.* John Wiley & Sons, 2014.

[21] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction.* MIT press, 2018.

[22] E. Even-Dar, S. M. Kakade, and Y. Mansour, “Online markov decision processes,” *Mathematics of Operations Research*, vol. 34, no. 3, pp. 726–736, 2009.

[23] G. Neu, A. Antos, A. György, and C. Szepesvári, “Online markov decision processes under bandit feedback,” *Advances in Neural Information Processing Systems*, vol. 23, 2010.

[24] F. Orabona, “A modern introduction to online learning,” *CoRR*, vol. abs/1912.13213, 2019.

[25] L. Wen, J. Duan, S. E. Li, S. Xu, and H. Peng, “Safe reinforcement learning for autonomous vehicles through parallel constrained policy optimization,” in *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1–7, IEEE, 2020.

[26] D. Isele, A. Nakhaei, and K. Fujimura, “Safe reinforcement learning on autonomous vehicles,” in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1–6, IEEE, 2018.

[27] D. Wu, X. Chen, X. Yang, H. Wang, Q. Tan, X. Zhang, J. Xu, and K. Gai, “Budget constrained bidding by model-free reinforcement learning in display advertising,” in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pp. 1443–1451, 2018.

[28] Y. He, X. Chen, D. Wu, J. Pan, Q. Tan, C. Yu, J. Xu, and X. Zhu, “A unified solution to constrained bidding in online display advertising,” in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 2993–3001, 2021.

[29] A. Singh, Y. Halpern, N. Thain, K. Christakopoulou, E. Chi, J. Chen, and A. Beutel, “Building healthy recommendation sequences for everyone: A safe reinforcement learning approach,” in *Proceedings of the FAccTRec Workshop, Online*, pp. 26–27, 2020.

[30] E. Altman, *Constrained Markov Decision Processes.* Chapman and Hall, 1999.

[31] S. R. Balseiro and Y. Gur, “Learning in repeated auctions with budgets: Regret minimization and equilibrium,” *Management Science*, vol. 65, no. 9, pp. 3952–3968, 2019.

[32] R. Gummadi, P. Key, and A. Proutiere, "Repeated auctions under budget constraints: Optimal bidding strategies and equilibria," in *the Eighth Ad Auction Workshop*, vol. 4, Citeseer, 2012.

[33] L. Zheng and L. Ratliff, "Constrained upper confidence reinforcement learning," in *Proceedings of the 2nd Conference on Learning for Dynamics and Control* (A. M. Bayen, A. Jadbabaie, G. Pappas, P. A. Parrilo, B. Recht, C. Tomlin, and M. Zeilinger, eds.), vol. 120 of *Proceedings of Machine Learning Research*, pp. 620–629, PMLR, 10–11 Jun 2020.

[34] Y. Efroni, S. Mannor, and M. Pirotta, "Exploration-exploitation in constrained mdps," 2020.

[35] X. Wei, H. Yu, and M. J. Neely, "Online learning in weakly coupled markov decision processes: A convergence time study," *Proc. ACM Meas. Anal. Comput. Syst.*, vol. 2, apr 2018.

[36] S. Qiu, X. Wei, Z. Yang, J. Ye, and Z. Wang, "Upper confidence primal-dual reinforcement learning for cmdp with adversarial loss," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 15277–15287, Curran Associates, Inc., 2020.

[37] Y. Ding and J. Lavaei, "Provably efficient primal-dual reinforcement learning for cmdps with non-stationary objectives and constraints," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, pp. 7396–7404, 2023.

[38] H. Wei, A. Ghosh, N. Shroff, L. Ying, and X. Zhou, "Provably efficient model-free algorithms for non-stationary cmdps," in *International Conference on Artificial Intelligence and Statistics*, pp. 6527–6570, PMLR, 2023.

[39] S. R. Balseiro, H. Lu, and V. Mirrokni, "The best of many worlds: Dual mirror descent for online allocation problems," *Operations Research*, vol. 71, no. 1, pp. 101–119, 2023.

[40] N. Liakopoulos, A. Destounis, G. Paschos, T. Spyropoulos, and P. Mertikopoulos, "Cautious regret minimization: Online optimization with long-term budget constraints," in *International Conference on Machine Learning*, pp. 3944–3952, PMLR, 2019.

[41] T. Jin, L. Huang, and H. Luo, "The best of both worlds: stochastic and adversarial episodic mdps with unknown transition," *Advances in Neural Information Processing Systems*, vol. 34, pp. 20491–20502, 2021.

[42] S. Mannor, J. N. Tsitsiklis, and J. Y. Yu, "Online learning with sample path constraints," *Journal of Machine Learning Research*, vol. 10, no. 20, pp. 569–590, 2009.

[43] M. Castiglioni, A. Celli, and C. Kroer, "Online learning with knapsacks: the best of both worlds," in *Proceedings of the 39th International Conference on Machine Learning* (K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, eds.), vol. 162 of *Proceedings of Machine Learning Research*, pp. 2767–2783, PMLR, 17–23 Jul 2022.

[44] M. Castiglioni, A. Celli, A. Marchesi, G. Romano, and N. Gatti, "A unifying framework for online optimization with long-term constraints," *Advances in Neural Information Processing Systems*, vol. 35, pp. 33589–33602, 2022.

[45] Z. Zhao, H. Li, R. Zhao, and X. Wang, "Crossing-line crowd counting with two-phase deep neural networks," in *ECCV*, pp. 712–726, 2016.

[46] Z. Huang, Y. Ding, G. Song, L. Wang, R. Geng, H. He, S. Du, X. Liu, Y. Tian, Y. Liang, *et al.*, “Bcdata: A large-scale dataset and benchmark for cell detection and counting,” in *MICCAI*, pp. 289–298, 2020.

[47] A. Aldayri and W. Albattah, “Taxonomy of anomaly detection techniques in crowd scenes,” *Sensors*, vol. 22, no. 16, p. 6080, 2022.

[48] J. Wang, L. Song, Z. Li, H. Sun, J. Sun, and N. Zheng, “End-to-end object detection with fully convolutional network,” in *CVPR*, pp. 15849–15858, 2021.

[49] X. Chu, A. Zheng, X. Zhang, and J. Sun, “Detection in crowded scenes: One proposal, multiple predictions,” in *CVPR*, pp. 12214–12223, 2020.

[50] J. Wan, Z. Liu, and A. B. Chan, “A generalized loss function for crowd counting and localization,” in *CVPR*, pp. 1974–1983, 2021.

[51] D. Lian, J. Li, J. Zheng, W. Luo, and S. Gao, “Density map regression guided detection network for rgb-d crowd counting and localization,” in *CVPR*, pp. 1821–1830, 2019.

[52] D. Lian, X. Chen, J. Li, W. Luo, and S. Gao, “Locating and counting heads in crowds with a depth prior,” *IEEE TPAMI*, 2021.

[53] Q. Song, C. Wang, Z. Jiang, Y. Wang, Y. Tai, C. Wang, J. Li, F. Huang, and Y. Wu, “Rethinking counting and localization in crowds: A purely point-based framework,” in *ICCV*, pp. 3365–3374, 2021.

[54] D. Liang, W. Xu, and X. Bai, “An end-to-end transformer model for crowd localization,” in *ECCV*, pp. 38–54, 2022.

[55] Z. Ma, X. Wei, X. Hong, and Y. Gong, “Bayesian loss for crowd count estimation with point supervision,” in *ICCV*, pp. 6142–6151, 2019.

[56] B. Wang, H. Liu, D. Samaras, and M. H. Nguyen, “Distribution matching for crowd counting,” in *NeurIPS*, pp. 1595–1607, 2020.

[57] S. Abousamra, M. Hoai, D. Samaras, and C. Chen, “Localization in the crowd with topological constraints,” in *AAAI*, 2021.

[58] J. Gao, T. Han, Y. Yuan, and Q. Wang, “Learning independent instance maps for crowd localization,” *arXiv preprint arXiv:2012.04164*, 2020.

[59] C. Arteta, V. Lempitsky, and A. Zisserman, “Counting in the wild,” in *ECCV*, pp. 483–498, 2016.

[60] T. Han, J. Gao, Y. Yuan, X. Li, *et al.*, “Ldc-net: A unified framework for localization, detection and counting in dense crowds,” *arXiv preprint arXiv:2110.04727*, 2021.

[61] K. Kim and H. S. Lee, “Probabilistic anchor assignment with iou prediction for object detection,” in *ECCV*, pp. 355–371, 2020.

[62] D. Liang, W. Xu, Y. Zhu, and Y. Zhou, “Focal inverse distance transform maps for crowd localization,” *IEEE TMM*, 2022.

[63] P. Hu and D. Ramanan, “Finding tiny faces,” in *CVPR*, 2017.

[64] C. Liu, X. Weng, and Y. Mu, "Recurrent attentive zooming for joint crowd counting and precise localization," in *CVPR*, pp. 1217–1226, 2019.

[65] C. Xu, D. Liang, Y. Xu, S. Bai, W. Zhan, X. Bai, and M. Tomizuka, "Autoscale: Learning to Scale for Crowd Counting," *IJCV*, vol. 130, no. 2, pp. 405–434, 2022.

[66] J. Gao, M. Gong, and X. Li, "Congested crowd instance localization with dilated convolutional swin transformer," *Neurocomputing*, vol. 513, p. 94–103, 2022.

[67] J. Wang, J. Gao, Y. Yuan, and Q. Wang, "Crowd localization from gaussian mixture scoped knowledge and scoped teacher," *IEEE TIP*, vol. 32, pp. 1802–1814, 2023.

[68] Y. Li, X. Zhang, and D. Chen, "Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes," in *CVPR*, pp. 1091–1100, 2018.

[69] F. Gökçe, G. Üçoluk, E. Şahin, and S. Kalkan, "Vision-based detection and distance estimation of micro unmanned aerial vehicles," *Sensors*, 2015.

[70] M. A. Haseeb, J. Guan, D. Ristic-Durrant, and A. Gräser, "Disnet: a novel method for distance estimation from monocular camera," *10th Planning, Perception and Navigation for Intelligent Vehicles (PPNIV18), IROS*, 2018.

[71] S. Tuohy, D. O'Cualain, E. Jones, and M. Glavin, "Distance determination for an automobile environment using inverse perspective mapping in opencv," in *IET Irish Signals and Systems Conference*, 2010.

[72] L. Jing, R. Yu, H. Kretzschmar, K. Li, C. R. Qi, H. Zhao, A. Ayvaci, X. Chen, D. Cower, Y. Li, *et al.*, "Depth estimation matters most: improving per-object depth estimation for monocular 3d detection and tracking," in *ICRA*, 2022.

[73] Y. Li, T. Chen, M. Kabkab, R. Yu, L. Jing, Y. You, and H. Zhao, "R4d: Utilizing reference objects for long-range distance estimation," in *ICLR*, 2022.

[74] G. Mancusi, A. Panariello, A. Porrello, M. Fabbri, S. Calderara, and R. Cucchiara, "Trackflow: Multi-object tracking with normalizing flows," in *ICCV*, 2023.

[75] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[76] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *CVPR*, pp. 2117–2125, 2017.

[77] H. Bao, L. Dong, S. Piao, and F. Wei, "BEiT: BERT pre-training of image transformers," *ICLR*, 2021.

[78] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *CVPR*, 2012.

[79] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *CVPR*, 2020.

[80] M. Fabbri, G. Brasó, G. Maugeri, A. Ošep, R. Gasparini, O. Cetintas, S. Calderara, L. Leal-Taixé, and R. Cucchiara, "Motsynth: How can synthetic data help pedestrian detection and tracking?," in *ICCV*, 2021.

[81] L. Ladicky, J. Shi, and M. Pollefeys, “Pulling things out of perspective,” in *CVPR*, 2014.

[82] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, “Centernet: Keypoint triplets for object detection,” in *ICCV*, pp. 6569–6578, 2019.

[83] X. Ma, S. Liu, Z. Xia, H. Zhang, X. Zeng, and W. Ouyang, “Rethinking pseudo-lidar representation,” in *ECCV*, 2020.

[84] L. Bertoni, S. Kreiss, and A. Alahi, “Monoloco: Monocular 3d pedestrian localization and uncertainty estimation,” in *ICCV*, 2019.

[85] S. Iyengar, Y. Lelkes, M. Levendusky, N. Malhotra, and S. J. Westwood, “The origins and consequences of affective polarization in the united states,” *Annual review of political science*, vol. 22, pp. 129–146, 2019.

[86] M. Jakesch, A. Bhat, D. Buschek, L. Zalmanson, and M. Naaman, “Co-writing with opinionated language models affects users’ views,” in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI ’23, (New York, NY, USA), Association for Computing Machinery, 2023.

[87] A. Arora, L.-A. Kaffee, and I. Augenstein, “Probing pre-trained language models for cross-cultural differences in values,” in *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)* (S. Dev, V. Prabhakaran, D. Adelani, D. Hovy, and L. Benotti, eds.), (Dubrovnik, Croatia), pp. 114–130, Association for Computational Linguistics, May 2023.

[88] E. Durmus, K. Nyugen, T. I. Liao, N. Schiefer, A. Askell, A. Bakhtin, C. Chen, Z. Hatfield-Dodds, D. Hernandez, N. Joseph, *et al.*, “Towards measuring the representation of subjective global opinions in language models,” *arXiv preprint arXiv:2306.16388*, 2023.

[89] E. Hwang, B. Majumder, and N. Tandon, “Aligning language models to user opinions,” in *Findings of the Association for Computational Linguistics: EMNLP 2023* (H. Bouamor, J. Pino, and K. Bali, eds.), (Singapore), pp. 5906–5919, Association for Computational Linguistics, Dec. 2023.

[90] G. Pistilli, A. Leidinger, Y. Jernite, A. Kasirzadeh, A. S. Luccioni, and M. Mitchell, “CIVICS: building a dataset for examining culturally-informed values in large language models,” *CoRR*, vol. abs/2405.13974, 2024.

[91] L. P. Argyle, E. C. Busby, N. Fulda, J. R. Gubler, C. Rytting, and D. Wingate, “Out of one, many: Using language models to simulate human samples,” *Political Analysis*, vol. 31, no. 3, pp. 337–351, 2023.

[92] X. Wang, B. Ma, C. Hu, L. Weber-Genzel, P. Röttger, F. Kreuter, D. Hovy, and B. Plank, “”my answer is c”: First-token probabilities do not match text answers in instruction-tuned language models,” 2024.

[93] P. Röttger, V. Hofmann, V. Pyatkin, M. Hinck, H. R. Kirk, H. Schütze, and D. Hovy, “Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models,” 2024.

[94] F. Motoki, V. Pinho Neto, and V. Rodrigues, “More human than human: Measuring chatgpt political bias,” *Public Choice*, vol. 198, no. 1, pp. 3–23, 2024.

[95] A. Wei, N. Haghtalab, and J. Steinhardt, “Jailbroken: How does LLM safety training fail?,” in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023* (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), 2023.

[96] S. Robertson and H. Zaragoza, “The probabilistic relevance framework: Bm25 and beyond,” *Foundations and Trends in Information Retrieval*, vol. 3, no. 4, 2009.

[97] E. Lurie and E. Mustafaraj, “Highly partisan and blatantly wrong: Analyzing news publishers’ critiques of google’s reviewed claims,” in *Truth and Trust Online Conference*, Hacks Hackers, 2020.

[98] B. Wang, “Ranking evaluation metrics for recommender systems,” *Medium*, 2021.

[99] A. Mladenic Grobelnik, F. Zaman, J. Espigule-Pons, and M. Grobelnik, “Emergent behaviors from llm-agent simulations,” in *SiKDD Conference*, 2023.

[100] V. Stojkoski, P. Koch, E. Coll, and C. A. Hidalgo, “Estimating digital product trade through corporate revenue data,” *Nature Communications*, vol. 15, no. 1, p. 5262, 2024.

[101] “Financial roberta dataset,” 2022.

[102] G. Salton and C. Buckley, “Term-weighting approaches in automatic text retrieval,” *Information processing & management*, vol. 24, no. 5, pp. 513–523, 1988.

[103] J. Ramos *et al.*, “Using tf-idf to determine word relevance in document queries,” in *Proceedings of the first instructional conference on machine learning*, vol. 242, pp. 29–48, Citeseer, 2003.

[104] L. Breiman, “Random forests,” *Machine learning*, vol. 45, pp. 5–32, 2001.

[105] C. A. Hidalgo, “Economic complexity theory and applications,” *Nature Reviews Physics*, vol. 3, no. 2, pp. 92–113, 2021.

[106] N. Cesa-Bianchi, T. Cesari, R. Colomboni, F. Fusco, and S. Leonardi, “Bilateral trade: A regret minimization perspective,” *Mathematics of Operations Research*, vol. 49, no. 1, pp. 171–203, 2024.

[107] N. Cesa-Bianchi, T. R. Cesari, R. Colomboni, F. Fusco, and S. Leonardi, “Repeated bilateral trade against a smoothed adversary,” in *The Thirty Sixth Annual Conference on Learning Theory*, pp. 1095–1130, PMLR, 2023.

[108] S. Vaze, K. Han, A. Vedaldi, and A. Zisserman, “Generalized category discovery,” in *CVPR*, 2022.

[109] Y. Fei, Z. Zhao, S. Yang, and B. Zhao, “Xcon: Learning with experts for fine-grained category discovery,” in *BMVC*, p. 96, BMVA Press, 2022.

[110] M. Yang, Y. Zhu, J. Yu, A. Wu, and C. Deng, “Divide and conquer: Compositional experts for generalized novel class discovery,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14268–14277, 2022.

[111] Y. Sun and Y. Li, “Opencon: Open-world contrastive learning,” in *Transactions on Machine Learning Research*, 2022.

[112] N. Pu, Z. Zhong, and N. Sebe, "Dynamic conceptional contrastive learning for generalized category discovery," in *CVPR*, 2023.

[113] A. Krizhevsky, G. Hinton, *et al.*, "Learning multiple layers of features from tiny images," 2009.

[114] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011.

[115] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3d object representations for fine-grained categorization," in *ICCV Workshop*, 2013.

[116] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. Jawahar, "Cats and dogs," in *CVPR*, 2012.

[117] G. Holste, S. Wang, Z. Jiang, T. C. Shen, G. Shih, R. M. Summers, Y. Peng, and Z. Wang, "Long-tailed classification of thorax diseases on chest x-ray: A new benchmark study," in *MICCAI Workshop on Data Augmentation, Labelling, and Imperfections*, 2022.

[118] K. C. Tan, Y. Liu, B. Ambrose, M. Tulig, and S. Belongie, "The herbarium challenge 2019 dataset," *arXiv preprint arXiv:1906.05372*, 2019.

[119] T.-M. H. Hsu, H. Qi, and M. Brown, "Federated visual classification with real-world data distribution," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pp. 76–92, Springer, 2020.

[120] X. Wen, B. Zhao, and X. Qi, "Parametric classification for generalized category discovery: A baseline study," in *ICCV*, 2023.

[121] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1967.

[122] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[123] S. Zhang, S. Khan, Z. Shen, M. Naseer, G. Chen, and F. S. Khan, "Promptcal: Contrastive affinity learning via auxiliary prompts for generalized novel category discovery," in *CVPR*, 2023.

[124] L. Busoniu, R. Babuska, and B. De Schutter, "A comprehensive survey of multiagent reinforcement learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 38, no. 2, pp. 156–172, 2008.

[125] Y. Yang and J. Wang, "An overview of multi-agent reinforcement learning from game theoretical perspective," *arXiv preprint arXiv:2011.00583*, 2020.

[126] K. Zhang, Z. Yang, and T. Başar, "Multi-agent reinforcement learning: A selective overview of theories and algorithms," *Handbook of reinforcement learning and control*, pp. 321–384, 2021.

[127] N. Jaques, A. Lazaridou, E. Hughes, C. Gulcehre, P. Ortega, D. Strouse, J. Z. Leibo, and N. De Freitas, "Social influence as intrinsic motivation for multi-agent deep reinforcement learning," in *International conference on machine learning*, pp. 3040–3049, PMLR, 2019.

[128] T. Eccles, E. Hughes, J. Kramár, S. Wheelwright, and J. Z. Leibo, "Learning reciprocity in complex sequential social dilemmas," *arXiv preprint arXiv:1903.08082*, 2019.

[129] J. N. Foerster, R. Y. Chen, M. Al-Shedivat, S. Whiteson, P. Abbeel, and I. Mordatch, "Learning with opponent-learning awareness," *arXiv preprint arXiv:1709.04326*, 2017.

[130] T. Willi, A. H. Letcher, J. Treutlein, and J. Foerster, "Cola: consistent learning with opponent-learning awareness," in *International Conference on Machine Learning*, pp. 23804–23831, PMLR, 2022.

[131] E. Hughes, J. Z. Leibo, M. Phillips, K. Tuyls, E. Dueñez-Guzman, A. García Castañeda, I. Dunning, T. Zhu, K. McKee, R. Koster, *et al.*, "Inequity aversion improves cooperation in intertemporal social dilemmas," *Advances in neural information processing systems*, vol. 31, 2018.

[132] T. Eccles, Y. Bachrach, G. Lever, A. Lazaridou, and T. Graepel, "Biases for emergent communication in multi-agent reinforcement learning," *Advances in neural information processing systems*, vol. 32, 2019.

[133] A. Yaman, J. Z. Leibo, G. Iacca, and S. W. Lee, "The emergence of division of labor through decentralized social sanctioning," *arXiv preprint arXiv:2208.05568*, 2022.

[134] I. Osband, C. Blundell, A. Pritzel, and B. Van Roy, "Deep exploration via bootstrapped dqn," *Advances in neural information processing systems*, vol. 29, 2016.