A Conceptual Map for Exploring the Landscape of Large Language Models

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 11 October 2025, accepted 31 October 2025, date of publication 5 November 2025, date of current version 12 November 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3629334

RESEARCH ARTICLE

A Conceptual Map for Exploring the Landscape of Large Language Models

FRANCESCO PIERRI[1], ANNA BERNASCONI[1], FLAVIO GIOBERGIA[2], (Member, IEEE), CLAUDIO SAVELLI[2], ELENA BARALIS[2], (Member, IEEE), LUCA CAGLIERO[2], (Member, IEEE), AND STEFANO CERI[1], (Member, IEEE)

[1]Politecnico di Milano, 20133 Milan, Italy
[2]Politecnico di Torino, 10129 Turin, Italy

Corresponding author: Francesco Pierri (francesco.pierri@polimi.it)

This study was carried out within the FAIR-Future Artificial Intelligence Research and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR)-MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3-D.D. 1555 11/10/2022, PE00000013). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

ABSTRACT Selecting and evaluating the most suitable Large Language Model (LLM) for a given task remains a significant challenge, particularly in domain-specific applications such as healthcare and legal research, which require models to provide transparency regarding training data, task specialization, and compliance with ethical standards. Despite the availability of open-source platforms for sharing models and datasets, challenges persist in accessing critical information. Incomplete metadata and inconsistent documentation hinder efficient model discovery, comparison, and adoption. In this paper, we introduce a simple conceptual map for LLMs, designed to assist researchers and practitioners in understanding the complex landscape of generative models. We provide a rationale for our modeling choices and a comprehensive description of the map in terms of four interconnected entities. Our primary objective is to provide all practitioners in the field—not just developers, but managers, testers, and other people who are part of producing and using technology—a clear terminology for expressing how to address the LLM landscape. A secondary objective is a call on industry stakeholders—including collaborative platforms and model providers—to enhance transparency and reproducibility in LLM research and deployment.

INDEX TERMS Conceptual maps, FAIR principles, hugging face, large language models, model metadata, model selection, reproducibility, transparency.

I. INTRODUCTION

Since the launch of ChatGPT in November 2022, Large Language Models (LLMs) have been driving innovation across diverse fields; however, their rapid advancement also brings challenges in evaluating and understanding their full potential and capabilities [1]. Professionals across various fields face difficulties when selecting the most suitable model for their specific tasks. Legal experts require models trained on case law [2], educators need tools for personalized learning [3], and customer support teams depend on multilingual conversational AI [4]. Even experienced researchers in fields like climate science, engineering, and social sciences often struggle with the lack of centralized resources that detail crucial information, such as model training data, domain expertise, and performance benchmarks [5]. The lack of a structured knowledge base properly describing models' training data, domain specialization, and performance metrics forces users to rely on fragmented information, limiting the effective adoption of this emerging technology.

These challenges are further compounded by the growing emphasis on transparency and accountability in AI development, particularly with regulations as the EU AI Act [6], which focus on fairness, safety, and ethics. Insufficient documentation and a lack of transparency of both closed and open models make it difficult for researchers to ensure that their AI systems meet the stringent standards required for responsible and lawful deployment. Moreover, the concept of "openness" in AI models itself is multifaceted, as we discuss later.

The associate editor coordinating the review of this manuscript and approving it for publication was Arianna D'Ulizia.

In this landscape, there is a compelling need for understanding the concepts and terms used in everyday conversations about LLM. Here, we propose a simple conceptual map that organizes and categorizes LLM-related knowledge. This structured approach, based on classical data abstractions [7], supports practitioners in the field in reasoning about model selection, evaluation, and comparison, offering an intuitive and efficient way to navigate the ecosystem of LLMs.

In our vision, this approach should be integrated into existing platforms like Hugging Face (HF, [8]) and Kaggle [9]: these repositories have become vital resources for accessing and sharing models and datasets in an open-source and collaborative manner. HF is the primary hub for collaboration within the machine learning community, enabling researchers and developers to contribute models, datasets, and several applications; however, the platform has notable limitations that prevent efficient information exploration. In particular, HF often lacks crucial information about LLMs, including supported languages, usage license and reference publications. This can be due to both providers' negligence in documentation practices and strategic commercial decisions by major players, who may withhold critical details to maintain competitive advantages or protect proprietary interests.

The remainder of this paper is organized as follows. We first introduce our conceptual map and describe in detail the entities and attributes relevant to understanding LLMs. We also show how our map could be used in practice by stakeholders across different domains, including medicine, finance, law, life sciences, management and software development. Next, we discuss related documentation standards in existing literature and highlight the position of our contribution. Then, we briefly discuss different definitions of openness for LLMs, and analyze the lack of available information on Hugging Face. We conclude with a discussion of various critical aspects in the evolution of LLM, including a call for the community to revise the organization of collaborative platforms, further expanding their capabilities for accessing relevant knowledge.

II. CONCEPTUAL MAP

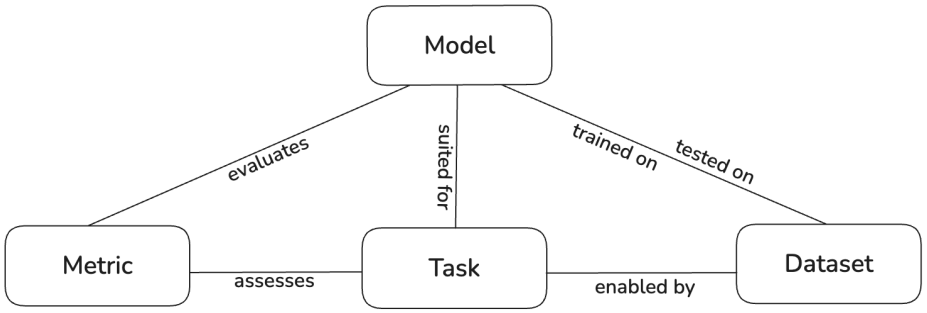
We propose a novel conceptual map to represent the key entities and relationships within the LLM ecosystem. In Figure 1 (top) we identify four main entities: Large Language Model, Dataset, Task, and Metric; six binary relationships interconnect them, explaining how models are tested or trained on datasets, how models are suited for tasks, how tasks are enabled by datasets, and how metrics evaluate models or assess tasks. Fig. 1 (bottom) shows a table listing attributes commonly used to describe the four entities – together with one example value. Attributes belong to three categories: *identifiers*, *descriptive attributes* responding to *FAIR* principles (i.e., contributing to the Findability, Accessibility, Interoperability, and Reuse [10] of related digital objects), and *internal attributes* describing the entity's implementation/deployment. We next describe each entity, with its attributes, and then each relationship.

A. MODEL

This entity describes the main characteristics of Large Language Models. Specifically, a model is uniquely identified by the pair ⟨*Name*, *Version*⟩; note that this distinction is not formalized and is sometimes missing or not exposed in HF (e.g., "Llama-2" in HF is split into "Llama", "2" in our map). Responding to FAIR principles, attributes include the Uniform Resource Identifier (*URI*), to indicate where the model can be accessed or downloaded; the *LicenseToUse* that specifies relevant terms; the *ModelCreator*, that is the original author of a model; the *Developer*, who has fine-tuned the model to serve a specific need – starting from a foundation model. HF allows model providers to freely flag their LLMs as "open-source" without specifying whether this refers to the architecture, training data, or implementation framework. For simplicity, we include a Boolean attribute *OpenSource*, which reflects the tag assigned by developers on the platform. In Section V, we provide an in-depth discussion of the various meanings of LLM openness. Additionally, when available, we include information about *CarbonEmission* in tCO2e (tonnes of carbon dioxide equivalent) as the unit of measurement. To represent implementation characteristics, we detail the used *LibraryFramework*; a list of *Language*s on which the model has been trained; the type of *Architecture* and its *NumberOfParameters*; the *ContextLength*, which indicates the maximum number of tokens that the model can handle in a sentence (e.g., "4k"); the *Tokenization* algorithm that determines how the input prompt (and the output answer) is divided into tokens (e.g., "SentencePiece Byte-Pair Encoding") and the *NumberOfTokens* representing the dictionary employed by the algorithm. The Boolean attributes *Quantization*, *InstructionTuned*, and *HumanAligned* respectively indicate the use of quantization techniques [11], fine tuning on instructions, and optimization w.r.t. human directions.

B. DATASET

This entity describes specific datasets, which can be used for pre-training an LLM (e.g., WebText [12]), for its fine-tuning (e.g., Reddit TL;DR [13]) or its evaluation (e.g., MMLU [14]). The entity is identified by its *Name* and equipped with a *URI* that connects it to its location for retrieval/download. We represent the *LicenseToUse* (detailing terms for the dataset usage) and a *Domain* of reference. Then, we define the *Size* as the number of rows of the dataset, the set of *Language*s of the texts in the dataset, and a Boolean flag *FineTuning* that indicates whether the dataset is suitable for fine-tuning previously trained models. Finally, *Annotation* indicates if the dataset is used for unsupervised (i.e., "Not annotated") or supervised tasks, in which case we capture information about the author of the annotation (e.g., workers, students, domain experts, or unknown crawl) or of



Entity	**Attribute**: Definition →*Example*. Attributes categories color code: identification ( ), FAIR ( ), and implementation ( )
Model	**Name**: Name of the LLM → *Llama*
	Version: Version of the model (identifier together with Name) → *3.2*
	URI: Link or source location to access the model → *https://huggingface.co/meta-llama/Llama-3.2-3B*
	LicenseToUse: Licensing terms for model usage → *llama3.2*
	ModelCreator: Individual/organization who developed the model → *Meta*
	Developer: Individual/team who shared the model → *meta-llama*
	OpenSource: Flag indicating if the model is open-source → *Yes*
	CarbonEmissions: The environmental impact derived from training the model (in tCO2e) → *240*
	LibraryFramework: Library/framework used to implement the model → *PyTorch*
	Language: Language(s) the model is trained to understand → *English, German, French, ...*
	Architecture: Underlying architecture/structure of the model → *Transformer*
	NumberOfParameters: Number of parameters in the model → *3B*
	ContextLength: Maximum input length/context window supported by the model → *128k tokens*
	Tokenization: Tokenization algorithm used to preprocess text for the model → *BPE*
	NumberOfTokens: The number of unique tokens (i.e., dictionary) used by the Tokenization algorithm → *128k*
	Quantization: A Flag that indicates if the model uses quantization techniques → *No*
	InstructionTuned: Indicates if the model has been fine-tuned on instruction-following datasets → *Yes*
	HumanAligned: Indicates if the model has been optimized to align with human preferences → *Yes*
Dataset	**Name**: Identifier of the dataset → *IMDB Reviews*
	URI: Link/source location to access the dataset → *https://ai.stanford.edu/~amaas/data/sentiment/*
	LicenseToUse: Licensing terms for dataset usage → *other*
	Domain: Specific domain/area of interest of the dataset → *Reviews*
	Size: Amount of data (in terms of number of rows) → *50k*
	Language: List of language(s) used in the dataset content → *English, French, Italian, ...*
	FineTuning: Flag indicating if the dataset is suitable for fine-tuning pre-trained models → *Yes*
	Annotation: How the annotation was performed (if applicable) → *Domain experts*
Task	**Name**: Identifier of the downstream task → *Sentiment Analysis*
	Description: Brief text to describe the task → *Determines the sentiment of a given text*
	SubTask: Specification/list of sub-categories of the main task → *Emotion Detection*
Metric	**Name**: Identifier of the metric used to assess performance → *BLEU*
	Description: What the metric measures/evaluates → *Assess goodness of translation*
	ContextFree: Flag indicating if the metric only needs the model's output (vs. also model's input) → *Yes*
	Trained: Flag indicating if the metric has trainable parameters → *No*
	FeatureBased: Flag indicating if the metric employs heuristics (i.e., input features), vs. the data itself (i.e., end-to-end) → *No*
	Granularity: Granularity of the evaluation → *SentenceLevel*

FIGURE 1. Top: Conceptual map describing LLMs, their attributes, and relationships with the surrounding entities.Bottom: Table listing attributes (with definition and example values) grouped by entity according to the above map. We consider the case of Llama 3.2 as a reference when values are available.

the self-annotation process (e.g., through captcha, alternative text).

C. TASK

This entity represents a task addressed by the Large Language Model (e.g., Question Answering, Commonsense Reasoning, Summarization). Each task is identified by its *Name*, clarified by a brief *Description*, and a list of *SubTask*s that progressively refine the main – general – task.

D. METRIC

This entity represents metrics used to evaluate a Language Model on a specific task (e.g., ROUGE [15], BLEU [16], BERTScore [17]). The entity is identified by its *Name*, characterized by a *Description*, and equipped with three Boolean flags, respectively indicating whether (i) they are *ContextFree* (i.e., they only need the model's output and a ground truth text to use as a reference) or context-dependent (i.e., they also need the model's input, such as a table, a document, etc.); (ii) they can be *Trained* (i.e., they have trainable parameters and need human annotations to be trained on) or not—note that context-dependent metrics are usually trained; (iii) they are *FeatureBased* (i.e., use other metrics/heuristics as input features) or can be evaluated in an end-to-end manner (i.e., requiring the input given to the model, the output of the model, and the ground truth). Untrained metrics can operate at the Word-Character level

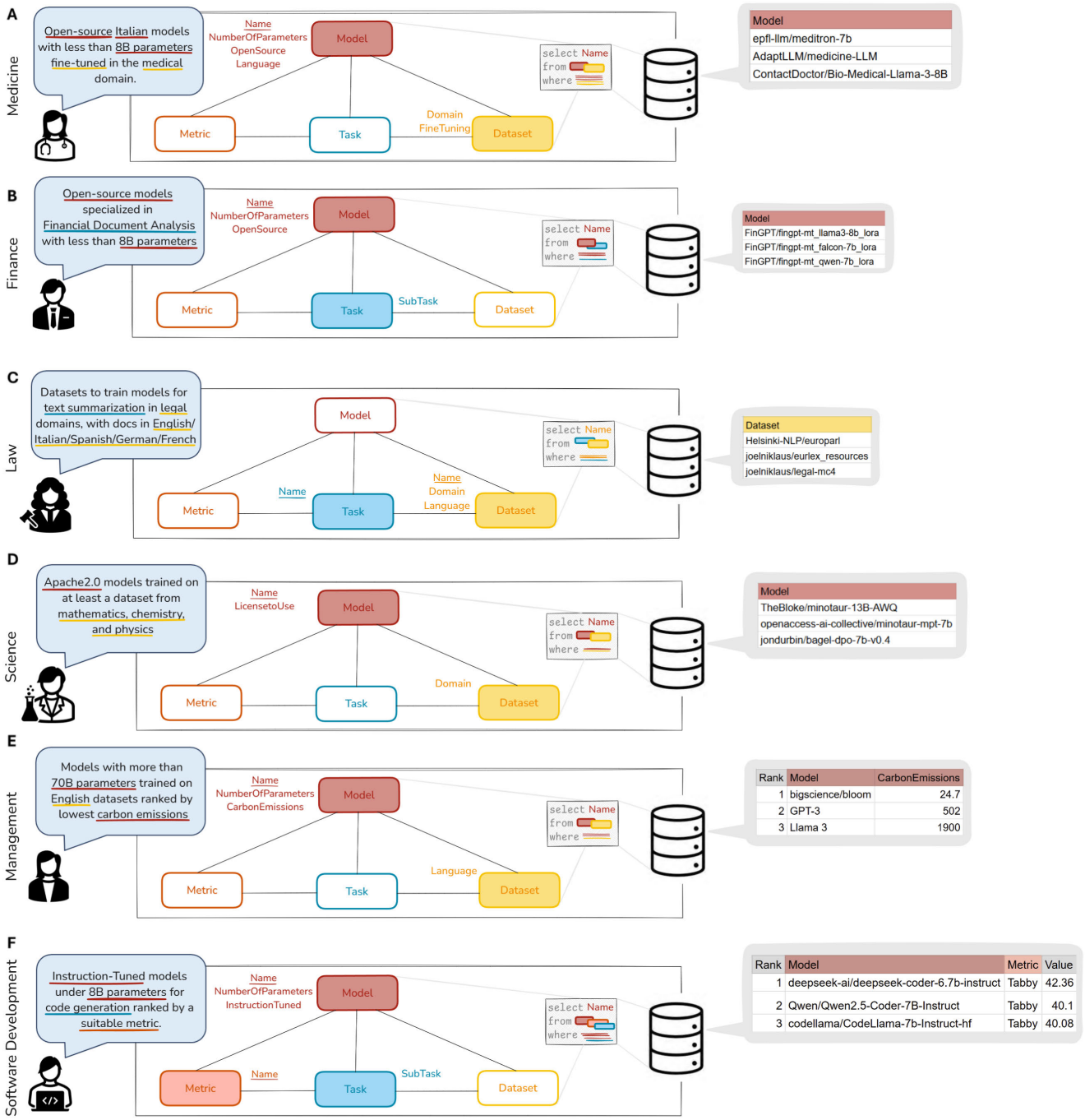


FIGURE 2. Examples of user-based interactions across different domains, mapping a request in natural language to an actual query on our conceptual map. We highlight attributes, entities and relationships involved, using different colors.

or use Embeddings to evaluate the output of the models. Embeddings-based metrics are usually able to capture the semantics of the phrase. We represent this information in the *Granularity* attribute.

E. RELATIONSHIPS

A Model can be `trained` or `tested on` one or more Datasets (e.g., BERT is trained on the Wikipedia Dataset and tested on Natural Questions Dataset [18]). Models are `suited for` resolving one or more Tasks (e.g., GPT-3 → Text Generation). Metrics allow us to `evaluate` Models (e.g., BLEU → GPT-3) and `assess` Tasks (e.g., BLEU → Text Translation, ROUGE → Text Summarization). Finally, Tasks are performed once they are `enabled by` Datasets (e.g., Sentiment Analysis → IMDB Reviews, Question-Answering → SQuAD [19]).

III. NAVIGATING THE CONCEPTUAL MAP

Our conceptual map can be directly translated to a relational database, enabling SQL-like queries to return high-quality information for a variety of stakeholders interested in identifying models and datasets best suited to their tasks. Notably, modern tools now allow relational databases to be queried using natural language [20], making interaction with our map intuitive and accessible.

A focus group was carried out with several users with backgrounds in diverse application domains. Out of this user study, a selection of exemplifying queries was extracted. Figure 2 shows examples of possible user interactions in different domains. Color codes help in visualizing the specific entities and attributes involved. For instance, panel **A** shows a user working in the medical domain who is interested in finding an open-source model that supports the Italian language, has fewer than 8B parameters (hence, of a relatively small size), and is fine-tuned for the medical domain. In this case, the query returns results obtained by navigating through two entities and one relationship. The model identifier is extracted from the Model entity, filtering on the open-sourceness, language and size, while traversing the `trained on` relationship towards the Dataset entity, identifying those used for fine-tuning that are in the medical domain.

Other examples show queries made by stakeholders in diverse domains: finance, law, science, management and software development. Different requests involve different entities and relationships in the map. Note that some user inputs might be expressed in very simple terms (e.g., the legal query on datasets) while others include more technical aspects. To accommodate this range of expertise, we envision an intermediate system that assists users, translating vague or high-level inputs (e.g., ''small size'') into precise query constraints (e.g., ''less than 8B parameters''), thereby making our map accessible to both non-technical and technical users.

A. PROOF OF CONCEPT

To showcase how our map could be used, we built a simple demonstrator accessible at https://geco.deib.polimi.it/LLMap-demo/. This prototype provides readers with a concrete idea of how to navigate a database based on our conceptual map. We show the effect of interacting with such a resource, where exploration of single entities, relationships, and paths is made available and fully implemented; we restrict to a small instance that focuses on the six queries explained above. Developing a full database is out of the scope of this paper, as it would require involvement in the community, tight interaction with actors such as HF and Kaggle, and enforcement of schema constraints by data managers.

B. BUILDING LEADERBOARDS

Leaderboards are user-friendly interfaces for practitioners, gathering all relevant information to support and guide their decision-making processes. With our map, a leaderboard can be built by combining Models, Tasks, Metrics, and Datasets, e.g., for assessing the performance of a given LLM on a given task using a given test dataset, measuring performance according to a given metric. Usually, leaderboards support partial aggregations to compare performance results over multiple datasets and metrics. As of today, there is no official (nor widely adopted) unique platform for leaderboards, which are scattered across the web. For instance, the HF-based open-llm-leaderboard [21] and the Massive Text Embedding Benchmark (MTEB) [22] provide an extensive view of performance metrics for different language models across a variety of datasets and tasks, as question answering and text embedding. Unified benchmarking tools, such as LLMBox [23] and PromptBench [24] and LLM-Uncertainty-Bench [25], offer ad hoc functionalities for LLM assessment, but require end-users to customize dataset and performance metrics.

IV. RELATED DOCUMENTATION STANDARDS AND POSITIONING OF OUR WORK

Our map complements, rather than replaces, widely adopted documentation standards—most notably *Datasheets for Datasets* [26] and *Model Cards for Model Reporting* [27]—and related proposals such as *data statements* for NLP [28]. Whereas datasheets and model cards provide rich, reflective narratives, our contribution is a concise, queryable schema spanning four entities (Model, Dataset, Task, Metric) and their relationships, designed to support discovery, comparison, and governance at scale to diverse stakeholders.

Datasheets for Datasets organize questions across a dataset's lifecycle (motivation, composition, collection, preprocessing/labeling, intended uses, limitations) in order to reduce mismatches between training and deployment contexts—mismatches with societal and financial consequences. They center dataset creators and consumers, while acknowledging other stakeholders (policymakers, and individuals affected by downstream use). Several datasheet elements map directly to attributes in our schema–e.g., *Domain*, *Languages*, *Size*, *LicenseToUse*, and *Annotation* (including who labeled or whether self-annotation was used). Other aspects (rationale, social risks, collection protocols) are intentionally verbose and do not compress into scalar fields; for these, our map serves as a relational backbone with pointers (*URI*) to full narrative artifacts. Because relationships are in the form of `trained on` or `enabled by`, updates to a dataset can propagate to models and tasks, making dependencies explicit for reproducibility and auditing. Although *datasheets* are not designed for automation, many fields in our map could be in principle semi-populated from repository pages and papers; LLM-assisted drafting of datasheets, with human review, could co-evolve with automated population of our schema.

Model Cards similarly standardize model reporting (capabilities, evaluation conditions, intended users, limitations), with strong emphasis on socio-cultural bias and deployment context. Platforms such as Hugging Face already provide

support for "model cards," yet, as our empirical snapshot shows (Fig. 3), critical fields are often missing or inconsistent. Our schema addresses discoverability and comparability by normalizing a minimal set of shared attributes (*LicenseToUse*, *Languages*, *Architecture*, *NumberOfParameters*, *ContextLength*, *InstructionTuned*, *HumanAligned*) and by linking models to concrete Task/Metric/Dataset triplets. This enables platform-level enforcement of required fields and automated queries (e.g., SQL) for stakeholders who must make decisions under constraints such as domain suitability, licensing, hardware, or evaluation provenance.

In short, both *datasheets* and *model cards* provide guidelines to support responsible AI research; our map proposes a *compact, relational layer* that: (i) normalizes the most queried facts; (ii) exposes model–dataset–task–metric dependencies; and (iii) could support automated discovery (leaderboards, filters) while preserving links to the richer documentation where nuanced risks and context can be better described.

To our knowledge, the only systematic investigation of models described within HF has been conducted in the context of *LLM4Code* models, i.e., LLMs for coding tasks [29]. Even the simple selection of such models proved to be a difficult manual process, ultimately leading to the identification of about 400 models as of April 2025. The study involved manual labeling of dependencies, focusing on model reuse, and ranking models by popularity, for instance based on the number of likes. It also experimented with using LLMs to extract metadata, achieving good accuracy only on simple tasks. Although several LLM4Code properties were carefully defined, no data resource systematically reports their values for each assessed model. The discussion further highlights inconsistencies and misalignments in HF metadata, hinting at the need for a more structured process to support the healthy growth of the LLM4Code ecosystem.

V. ON THE DEFINITION OF OPENNESS

Several shades of openness can be defined for LLMs. The most common interpretation for "open model" refers to models that are freely available, which is equivalent to being "open for inference/fine-tuning." Note that, according to Hugging Face's policies, all models on the platform are considered "open" [8]. A model is considered "open for code" when the provider offers free access to the model's implementation (e.g., Falcon and Mistral families). It is "open for derivatives" when the model's output can be reused (e.g., for model distillation). A model is "open dataset for pre-training" when the provider discloses the datasets used for training, and it is "open dataset for instruction-tuning" when users can provide new or different training datasets (e.g., BLOOMZ [30]). Developers are usually encouraged to provide open-source LLM documentation that includes ethical problem statements and use case restrictions [31].

We typically denote a model as "supporting external parametrization" when it is both an open model and

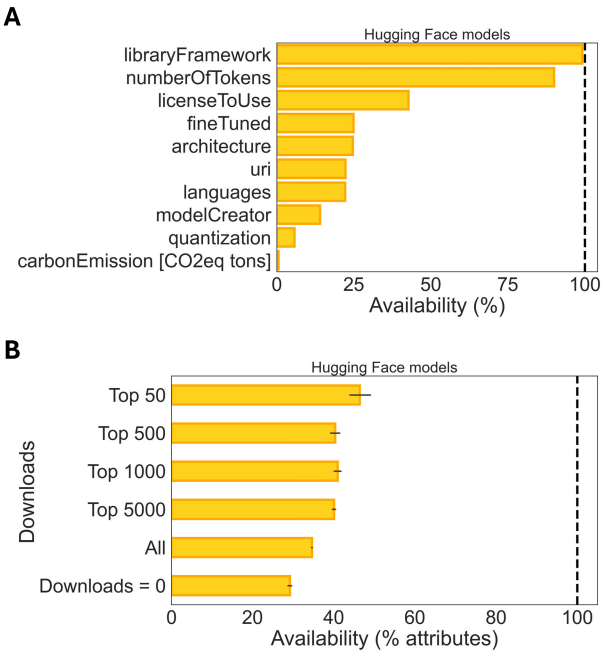


FIGURE 3. **(A) Proportion of available attributes for models on Hugging Face.(B) Average proportion of available attributes for models on Hugging Face depending on their popularity, measured by the total number of downloads. The bars represent 95% C.I.**

open code (e.g., DeepSeek). Several variations of these basic models exist; for example, an open model could be used for commercial applications only under certain legal constraints [32]. Access to pre-training data is regarded as "open" when the training data is open, and as "known" (a weaker form) when the training data includes private data.

Note that entity attributes in our map provide an effective indication of the provider's availability in disclosing the "shade of openness" supported by each model, through flags as *Quantization*, *InstructionTuned*, and *HumanAligned*, or through quantitative indicators like *NumberOfParameters*, *ContextLength*, *Tokenization*, and *NumberOfTokens*. If richer metadata were systematically collected, our map could also enable *quantification* of openness across the ecosystem, in ways similar to the analysis of Hugging Face presented below. For instance, one could measure how many models are open for code but not for datasets, or how many support external parametrization. Anecdotal evidence already shows wide variation: BLOOM and BLOOMZ emphasize dataset transparency, Llama models restrict access and licenses, and commercial APIs such as GPT-4 remain closed. Our framework provides the scaffolding to turn such qualitative comparisons into systematic statistics, which can easily support taxonomical views regarding issues and their mitigation in open-source project ecosystems [33].

VI. ON THE LIMITATIONS OF HUGGING FACE

Founded in 2016, Hugging Face is a widely used online platform that hosts over 530,000 open-source models and

110,000 datasets. While not exclusively focused on LLMs, HF has become a central repository for the AI and NLP communities, providing a suite of tools and resources that enable developers, researchers, and organizations to access, share, and collaborate on machine learning models and datasets [34], [35], [36], [37], [38], [39].

We employed a combination of API calls and web scraping to analyze the availability of information about models hosted on the platform, focusing on attributes that are present in our conceptual map, like the model's name, creator, vocabulary size, usage license and associated publication(s) (see panel **A** of Figure 3. Among the 300,000 + NLP models, on average, information is available less than 30% of the time, with no correlation between popularity and availability (panel **B**). For instance, the top 3 most downloaded models (namely BERT, MPNet, and GPT-2) only provide information about half of their attributes, while across the top 50 downloaded models the most transparent one is Qwen2.5-1.5B-Instruct, which describes 80% of its attributes. This suggests that a lot of critical information remains unavailable, reflecting differing priorities among LLM providers that must balance openness with competitive and commercial considerations.

VII. CONCLUSION

We propose a new conceptual map that provides a well-structured and comprehensive categorization of LLMs, related datasets, tasks, and metrics. In the remainder of this section, we illustrate various perspectives stemming from our approach to data modeling.

A. MODEL USABILITY

Different models are usually optimized for different tasks and use cases, with some performing better in certain scenarios than others. Previous efforts have attempted to quantify model usability in various contexts – for instance, Kaggle provides a score that reflects the completeness of available information and the credibility of the model developer. Our conceptual map builds upon this idea by incorporating additional factors, as the availability of pre-training and fine-tuning data, performance on relevant benchmarks, and hardware constraints (e.g., model size). Given that end users have diverse goals and limitations, we argue that our approach enables more precise, tailored recommendations for selecting the most suitable models for specific use cases.

B. ENVIRONMENTAL COSTS

The carbon footprint of LLMs has been estimated to range from tens to hundreds of tonnes of CO_2-eq [40], [41]. These estimates are typically calculated *a posteriori* by third parties using carbon footprint projection models, as LLMCarbon [41]. By systematically collecting metadata on various LLMs, our proposed conceptual model can be leveraged to generate CO_2-eq estimates automatically across a wide range of models. These estimates provide insights into both the total emissions associated with model training and the projected footprint of inference. This information empowers users to make more informed decisions – either avoiding models with excessive carbon footprints for ethical reasons or adopting strategies to minimize their environmental impact.

C. HUMAN SUPERVISION COSTS

Language models are typically pre-trained using self-supervised tasks, most notably next-token prediction, which allows them to learn without explicit supervision. However, to develop helpful models, LLMs often undergo instruction tuning [42] and alignment with human values, e.g. through Reinforcement Learning from Human Feedback (RLHF) [43]. These fine-tuning steps, unlike pre-training, require human-annotated datasets. As widely recognized in the literature [44], human annotation introduces ethical considerations, particularly regarding the choice of labor sources. We argue that systematically tracking the datasets used for pre-training and fine-tuning LLMs can provide a clearer understanding of the human labor.

D. REPRODUCIBILITY AND TRANSPARENCY

A structured representation of the ecosystem encompassing models, datasets, tasks, and evaluation metrics plays a crucial role in enhancing the reproducibility and transparency of research findings [5]. In scientific publications, it is common for papers to provide only partial – if any – details about the experimental settings used. This lack of clarity hinders the ability of other researchers to replicate results, slowing scientific progress and making it difficult to assess the robustness of proposed methodologies. By introducing a standardized schema, we establish a practical framework that serves both as a checklist for researchers publishing new models and a set of expectations for publishers, whether in the peer-review process or when uploading models to public repositories. Structured metadata can facilitate better model selection, making it easier for organizations to identify models that meet specific criteria regarding performance, ethical considerations, and deployment constraints. Furthermore, by improving model discoverability and usability, our framework fosters collaboration between researchers, developers, and policymakers, ultimately driving responsible AI development. We recognize that companies and organizations may have legitimate reasons to withhold certain information, whether for competitive advantages, regulatory compliance, or intellectual property protection; but this should be their clear choice, and not be hidden by ambiguous or incomplete information.

E. A VISION FOR THE FUTURE

Our conceptual map could be adopted and populated using both semi-automated and fully automated approaches. Large-scale metadata extraction, natural language processing, and LLM-driven automation can systematically retrieve and structure attributes such as model architecture, train-

ing data provenance, licensing terms, and task specializations from heterogeneous sources including documentation, research papers, and repository descriptions. Semi-automated methods, including crowdsourced contributions, could further refine and validate extracted information. As discussed in Section IV, our work is complementary to existing documentation standards, but it emphasizes a structured, queryable schema that enables discovery and comparison at scale. Building on this foundation, we argue that structured metadata collection should become a fundamental practice in the LLM ecosystem. Our map can provide the minimal enforceable core, while narrative artifacts like Datasheets and Model Cards supply the necessary contextual depth. Finally, our work is intended to complement collaborative platforms such as Hugging Face and Kaggle. We call for platform owners to adopt and enforce these design principles, and for providers to use them consistently when describing their models, datasets, metrics, and tasks. Such practices would significantly improve transparency, reproducibility, and usability for both scientific and industrial stakeholders.

ACKNOWLEDGMENT

This article builds upon the results and insights developed during the two-year ChatIMPACT project (https://www.asp-poli.it/chatimpact/), carried out within the Multidisciplinary Project Framework of Alta Scuola Politecnica (https://www.asp-poli.it/). The authors gratefully acknowledge the contributions of students Lorenzo Bertetto, Francesca Bettinelli, Alessio Buda, Marco di Mommio, and Simone di Bari, as well as the industrial tutor Samuele Conti from Advanced Business Solution (A. B. S.) and a faculty members Prof. Andrea Flori and Prof. Piercesare Secchi. Together, they developed several usage scenarios (some of which are illustrated in Figure 2) and conducted multiple focus groups involving a diverse set of stakeholders, including both researchers and practitioners. They also thank Prof. Gianpaolo Cugola and Prof. Elisabetta Raguseo, who supported and evaluated ASP projects during the 20th ASP cycle (2023–2024), for their valuable feedback. These interactions were instrumental in refining the conceptual map presented in this article. This study was carried out within the FAIR-Future Artificial Intelligence Research and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR)-MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3-D.D. 1555 11/10/2022, PE00000013). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

REFERENCES

[1] M. Steyvers, H. Tejeda, A. Kumar, C. Belem, S. Karny, X. Hu, L. W. Mayer, and P. Smyth, "What large language models know and what people think they know," *Nature Mach. Intell.*, vol. 7, no. 2, pp. 1–11, Jan. 2025.

[2] S. Kapoor, P. Henderson, and A. Narayanan, "Promises and pitfalls of artificial intelligence for legal applications," *J. Cross-Disciplinary Res. Comput. Law*, vol. 2, no. 2, pp. 6–7, 2024.

[3] S. Milano, J. A. McGrane, and S. Leonelli, "Large language models challenge the future of higher education," *Nature Mach. Intell.*, vol. 5, no. 4, pp. 333–334, Mar. 2023.

[4] L. Barrault et al., "Joint speech and text machine translation for up to 100 languages," *Nature*, vol. 637, no. 8046, pp. 587–593, 2025.

[5] S. Kapoor, E. M. Cantrell, K. Peng, T. H. Pham, C. A. Bail, O. E. Gundersen, J. M. Hofman, J. Hullman, M. A. Lones, M. M. Malik, P. Nanayakkara, R. A. Poldrack, I. D. Raji, M. Roberts, M. J. Salganik, M. Serra-Garcia, B. M. Stewart, G. Vandewiele, and A. Narayanan, "REFORMS: Consensus-based recommendations for machine-learning-based science," *Sci. Adv.*, vol. 10, no. 18, May 2024, Art. no. eadk3452.

[6] *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 12 July 2024 on Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*, Eur. Union, Jul. 2024.

[7] C. Batini, S. Ceri, and S. Navathe, *Conceptual Database Design: An Entity-Relationship Approach*. Redwood City, CA, USA: Benjamin Cummings, 1992.

[8] (2024). *Hugging Face*. [Online]. Available: https://huggingface.co/

[9] (2024). *Kaggle*. [Online]. Available: https://www.kaggle.com/

[10] M. D. Wilkinson et al., "The FAIR guiding principles for scientific data management and stewardship," *Scientific data*, vol. 3, no. 1, 2016, Art. no. 160018.

[11] D. Patterson, J. Gonzalez, Q. Le, C. Liang, L.-M. Munguia, D. Rothchild, D. So, M. Texier, and J. Dean, "Carbon emissions and large neural network training," 2021, *arXiv:2104.10350*.

[12] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[13] M. Völske, M. Potthast, S. Syed, and B. Stein, "TL;DR: Mining Reddit to learn automatic summarization," in *Proc. Workshop New Frontiers Summarization*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 59–63.

[14] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, "Measuring massive multitask language understanding," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–10. [Online]. Available: https://openreview.net/forum?id=d7KBjmI3GmQ

[15] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: ACL, 2004, pp. 74–81.

[16] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2002, pp. 311–318.

[17] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–14.

[18] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, K. Toutanova, L. Jones, M. Kelcey, M.-W. Chang, A. M. Dai, J. Uszkoreit, Q. Le, and S. Petrov, "Natural questions: A benchmark for question answering research," *Trans. Assoc. Comput. Linguistics*, vol. 7, pp. 453–466, Nov. 2019.

[19] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 2383–2392.

[20] H. Fu, C. Liu, B. Wu, F. Li, J. Tan, and J. Sun, "CatSQL: Towards real world natural language to SQL applications," *Proc. VLDB Endowment*, vol. 16, no. 6, pp. 1534–1547, Feb. 2023.

[21] C. Fourrier, N. Habib, A. Lozovskaya, K. Szafer, and T. Wolf. (2024). *Performances Are Plateauing, Let's Make the Leaderboard Steep Again*. Accessed: Jun. 26, 2024. [Online]. Available: https://huggingface.co/spaces/open-llm-leaderboard/blog

[22] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers, "MTEB: Massive text embedding benchmark," in *Proc. 17th Conf. Eur. Chapter Assoc. Comput. Linguistics*. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 2014–2037.

[23] T. Tang et al., "LLMBox: A comprehensive library for large language models," in *Proc. 62nd Annu. Meeting Assoc. Comput. Linguistics*. Bangkok, Thailand: ACL, Aug. 2024, pp. 388–399. [Online]. Available: https://aclanthology.org/2024.acl-demos.37/

[24] K. Zhu, Q. Zhao, H. Chen, J. Wang, and X. Xie, "PromptBench: A unified library for evaluation of large language models," *J. Mach. Learn. Res.*, vol. 25, no. 1, pp. 1–22, Jan. 2023.

[25] J. Pang, S. Shi, Z. Tu, L. Wang, D. Wong, M. Yang, F. Ye, and E. Yilmaz, "Benchmarking LLMs via uncertainty quantification," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 37, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, Eds., 2024, pp. 15356–15385. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2024/file/1bdcb065d40203a00bd39831153338bb-Paper-Datasets_and_Benchmarks_Track.pdf

[26] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, and K. Crawford, "Datasheets for datasets," *Commun. ACM*, vol. 64, no. 12, pp. 86–92, Dec. 2021.

[27] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, "Model cards for model reporting," in *Proc. Conf. Fairness, Accountability, Transparency*, Jan. 2019, pp. 220–229.

[28] E. M. Bender and B. Friedman, "Data statements for natural language processing: Toward mitigating system bias and enabling better science," *Trans. Assoc. Comput. Linguistics*, vol. 6, pp. 587–604, Dec. 2018.

[29] Z. Yang, J. Shi, P. Devanbu, and D. Lo, "Ecosystem of large language models for code," *ACM Trans. Softw. Eng. Methodol.*, Apr. 2025.

[30] N. Muennighoff, T. Wang, L. Sutawika, A. Roberts, S. Biderman, T. Le Scao, M. S. Bari, S. Shen, Z.-X. Yong, H. Schoelkopf, X. Tang, D. Radev, A. F. Aji, K. Almubarak, S. Albanie, Z. Alyafeai, A. Webson, E. Raff, and C. Raffel, "Crosslingual generalization through multitask finetuning," 2022, *arXiv:2211.01786*.

[31] H. Gao, M. Zahedi, C. Treude, S. Rosenstock, and M. Cheong, "Documenting ethical considerations in open source AI models," in *Proc. 18th ACM/IEEE Int. Symp. Empirical Softw. Eng. Meas.* New York, NY, USA: Association for Computing Machinery, Oct. 2024, pp. 177–188, doi: 10.1145/3674805.3686679.

[32] R. Ghioni, M. Taddeo, and L. Floridi, "Open source intelligence and AI: A systematic review of the GELSI literature," *AI Soc.*, vol. 39, no. 4, pp. 1827–1842, Jan. 2023, doi: 10.1007/s00146-023-01628-x.

[33] Y. Cai, P. Liang, Y. Wang, Z. Li, and M. Shahin, "Demystifying issues, causes and solutions in LLM open-source projects," *J. Syst. Softw.*, vol. 227, Sep. 2025, Art. no. 112452.

[34] J. Castaño, S. Martínez-Fernández, and X. Franch, "Lessons learned from mining the hugging face repository," in *Proc. 1st IEEE/ACM Int. Workshop Methodol. Issues Empirical Stud. Softw. Eng.*, Apr. 2024, pp. 1–6.

[35] J. Castaño, S. Martínez-Fernández, X. Franch, and J. Bogner, "Analyzing the evolution and maintenance of ML models on hugging face," in *Proc. 21st Int. Conf. Mining Softw. Repositories*, Apr. 2024, pp. 607–618.

[36] X. Yang, W. Liang, and J. Zou, "Navigating dataset documentations in AI: A large-scale analysis of dataset cards on huggingface," in *Proc. 12th Int. Conf. Learn. Represent.*, 2024, pp. 1–12.

[37] C. Osborne, J. Ding, and H. R. Kirk, "The AI community building the future? A quantitative analysis of development activity on hugging face hub," *J. Comput. Social Sci.*, vol. 7, no. 2, pp. 1–39, Oct. 2024.

[38] A. Ait, J. L. C. Izquierdo, and J. Cabot, "HFCommunity: A tool to analyze the hugging face hub community," in *Proc. IEEE Int. Conf. Softw. Anal., Evol. Reeng. (SANER)*, Mar. 2023, pp. 728–732.

[39] F. Pepe, V. Nardone, A. Mastropaolo, G. Bavota, G. Canfora, and M. Di Penta, "How do hugging face models document datasets, bias, and licenses? An empirical study," in *Proc. 32nd IEEE/ACM Int. Conf. Program Comprehension*, Apr. 2024, pp. 370–381.

[40] A. S. Luccioni, S. Viguier, and A. Ligozat, "Estimating the carbon footprint of Bloom, a 176B parameter language model," *J. Mach. Learn. Res.*, vol. 24, no. 253, pp. 1–15, 2022.

[41] A. Faiz, S. Kaneda, R. Wang, R. Osi, P. Sharma, F. Chen, and L. Jiang, "LLMCarbon: Modeling the end-to-end carbon footprint of large language models," in *Proc. 12th Int. Conf. Learn. Represent.*, 2023, pp. 1–10.

[42] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, "Finetuned language models are zero-shot learners," in *Proc. Int. Conf. Learn. Represent.*, 2022, pp. 1–12.

[43] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. E. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 27730–27744.

[44] K. Fort, G. Adda, and K. B. Cohen, "Amazon mechanical turk: Gold mine or coal mine?" *Comput. Linguistics*, vol. 37, no. 2, pp. 413–420, Jun. 2011.

FRANCESCO PIERRI received the Ph.D. degree in data analytics and decision sciences from the Politecnico di Milano. He is currently an Assistant Professor with the Data Science Laboratory, Department of Electronics, Information and Bioengineering, Politecnico di Milano. He is also affiliated with Indiana University's Observatory on Social Media and he is a former Visiting Scholar with the Information Sciences Institute, University of Southern California. His research interests include the intersection of computational social science and artificial intelligence, using data-driven methodologies in computer and network science to study large-scale online phenomena. A central theme of his work is exploring the transformative impact of generative AI on digital information ecosystems, particularly its role in amplifying misinformation and disinformation. He was a recipient of the two Runner-up Dissertation Awards for his dissertation (ACM SIGKDD 2023 and the Informatics Europe).

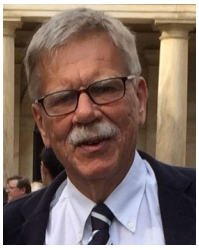
ANNA BERNASCONI received the Ph.D. degree in information technology from the Politecnico di Milano. She was a Visiting Researcher with the Universitat Politécnica de Valéncia, in 2022. She is currently a tenure-track Assistant Professor (RTT) with the Dipartimento di Elettronica, Informazione e Bioingegneria, the Data Science Laboratory and the Bioinformatics Laboratory, Politecnico di Milano. Her research interests include the intersection of bioinformatics, databases, and data science methods, in the context of life sciences and other applied sciences domains, focusing on building conceptual models, open-source tools, and services. For her Ph.D. thesis on genomic data integration, she won the D. N. Chorafas Foundation Prize, the Best Ph.D. Thesis on Big Data and Data Science (CINI Laboratory on Big Data), and the CAiSE Ph.D. Award.

FLAVIO GIOBERGIA (Member, IEEE) received the master's degree in computer engineering from the Politecnico di Milano and the Ph.D. degree in computer and control engineering from the Politecnico di Torino. He is currently a fixed-term Assistant Professor with the Politecnico di Torino and a Lecturer with the Universitá del Piemonte Orientale and the École supérieure de Commerce de Paris (ESCP). His current research interests include the removal of sensitive information from machine learning models via machine unlearning, large language models, and applications of ML techniques in the field of anomaly detection, with a focus on financial crime detection.

CLAUDIO SAVELLI received the bachelor's degree in computer engineering and a master's degree in data science and engineering from the Politecnico di Torino, in 2022 and 2024, respectively, where he is currently pursuing the Ph.D. degree in computer and control engineering. During his master's degree, he enriched his academic experience through the Alta Scuola Politecnica (ASP) excellence program. He was the Team Leader of the MAchine Learning @ PoliTO (MALTO), the official machine learning student team of the Politecnico di Torino. He is one of the organizers of WIPE-OUT, a workshop on privacy-preserving unlearning methods with ECML PKDD, in 2025. His primary research interest includes machine unlearning, the topic of his Ph.D. and the central theme of his master's thesis.

ELENA BARALIS (Member, IEEE) received the master's degree in electrical engineering and the Ph.D. degree in computer engineering from the Politecnico di Torino. She is currently the Deputy Rector of the Politecnico di Torino, where she has been a Full Professor, since January 2005. She has published over 200 papers in international journals and conference proceedings. Her current research interests include data mining, specifically on mining algorithms for big data and stream data analysis.

LUCA CAGLIERO (Member, IEEE) is currently an Associate Professor with the Department of Control and Computer Engineering (DAUIN), Politecnico di Torino, and the Coordinator of the SmartData@PoliTo Interdepartmental Center. He has published more than 200 research papers in conference proceedings, international journals, and book chapters. His main research interests include natural language understanding and generation and multimodal learning. He has been the General Chair of IEEE AICT, in 2024, the Area Chair of IEEE ICDM, in 2024, and a TPC member of several top-tier conferences (NeurIPS, ACL, EMNLP, ACM MM, and ACM SIGMOD). He is an Associate Editor of four international journals edited by IEEE, Springer, and Elsevier.



STEFANO CERI (Member, IEEE) is currently a Professor with the Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB), Politecnico di Milano. He received two AdG ERC grants on Search Computing (SeCo), from 2008 to 2013, and data-driven Genomic Computing (GeCo), from 2016 to 2021. With an H-index 87 on Google Scholar, he has authored more than 500 publications and ten international books; among them *Conceptual Database Design: an Entity Relationship Approach* (with over 2000 citations). His research interests include extending database technology and applying them as a data scientist, with emphasis on human and viral genomics, as well as on finance and legislation. He is an ACM Fellow and an ER Fellow. He received the Edward T. Codd Innovation Award from ACM-SIGMOD.