IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 26 December 2025, accepted 26 January 2026, date of publication 6 February 2026, date of current version 11 February 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3661027

SURVEY

Agent-Based Hybrid AI Models and Technologies: A Systematic Literature Review

ELIA PACIONI[1,2], (Graduate Student Member, IEEE), ANDREI C. COMAN[1], DAVIDE CALVARESI[1], GAETANO MANZO[1], AND MICHAEL IGNAZ SCHUMACHER[1]

[1]University of Applied Sciences and Arts Western Switzerland HES-SO, 3960 Sierre, Switzerland
[2]University of Extremadura, 06800 Mérida, Spain

Corresponding author: Elia Pacioni (elia.pacioni@hevs.ch)

This work was supported by the HES-SO RCSO ISNet Hybrid Ai foR RelIable perSonalized cOachiNg (HARRISON - Project No. 24-06) grant.

ABSTRACT Personalized *hybrid* agent-based systems leverage data-driven and symbolic components to provide tailored, context-aware decision support in multiple domains. Yet, the field lacks a consolidated and evidence-based overview of current approaches, their maturity, and the open challenges they present. Method. This study presents a Systematic Literature Review (2018–2025) combining Kitchenham's protocol with a Goal–Question–Metric (GQM) framework. Searches in peer-reviewed and indexed repositories (e.g., Scopus, Web of Science, and IEEE Xplore) returned 9,733 records, reduced to 46 primary studies after an initial screening. The main research question targets how personalized, agent-based Hybrid AI are currently conceived and implemented. In particular, the study is organized around 10 Structured Research Questions (SRQs) focusing on demographics, abstraction, domains, objectives, users, hybridization, technologies, advantages, limitations, evaluation, and future challenges. Results. Three dominant integration strategies surfaced: (1) concatenated pipelines that serially couple ML outputs to rule engines; (2) shared-representation models that embed symbolic knowledge in neural architectures; and (3) agent-level orchestration where heterogeneous components interact via message passing. While recommendation and adaptive coaching are the main use-cases, 71% of contributions remain at the prototype-level and lack large-scale or longitudinal evaluations. Technical barriers include scalability, semantic interoperability, and explainability; only 15% of studies report user-centered validation. Conclusions. The review reveals a growing but fragmented landscape. The paper proposes a research roadmap calling for: (i) publicly available benchmark datasets for hybrid personalization, (ii) standardized Hybrid Artificial Intelligence protocols for agent-to-human and agent-to-agent explanations, and (iii) design guidelines that combine symbolic guarantees with data-driven adaptability. Addressing these gaps is essential for a trustworthy deployment of Hybrid AI in sensitive settings.

INDEX TERMS Hybrid AI, personalized AI, agent-based AI, rule-based systems, data-driven models, LLMs.

I. INTRODUCTION

Artificial Intelligence (AI) applications have migrated from controlled laboratories to everyday, high-stakes settings such as personalized health coaching [1], finance [2], and autonomous mobility [3]. In these contexts, *two competing requirements emerge*: systems must *adapt* to rich sensor and user data while remaining *transparent, auditable, and controllable*. On one hand, purely data-driven (sub-symbolic) models excel at adaptation but behave as black boxes. On the other hand, mainly rule-based (symbolic) systems (i.e., Multi Agent Systems – MAS) offer distributed intelligence and interpretability but struggle with noisy or high-dimensional data. Hybrid Artificial Intelligence (Hybrid AI) combines symbolic knowledge representations and data-driven learning mechanisms within a single architecture. This notion dates back to early hybrid learning systems of the 1990s (KBANN [4]), which were characterized by the joint use

The associate editor coordinating the review of this manuscript and approving it for publication was Orazio Gambino.

© 2026 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

of hand-crafted domain theories and empirical learning algorithms, enabling mutual interaction between prior knowledge and data-driven adaptation [4], [5]. Yet, the recent availability of powerful Large Language Models (LLMs) and fine-grained regulatory guidance (illustrated, for instance, by the EU AI Act 2024/1689[1] and the NIST AI Risk-Management Framework[2]) has turned Hybrid AI from an academic aspiration into a practical necessity.

A. WHY HYBRID?

The explosion of open-weights LLMs (e.g., Llama [6], DeepSeek [7], Gemma [8]) and increasingly affordable commercial APIs[3] has democratized advanced AI. Yet, entirely data-driven models still exhibit critical shortcomings:

- **Safety.** In high-risk settings the absence of explicit control logic leads to brittle or unpredictable behavior [9].
- **Transparency.** The ''black-box'' nature of LLMs limits explanation, auditing, and bias detection [10].
- **Ethics and privacy.** Training corpora of opaque provenance raise questions of representativeness, intellectual property, and compliance with data-protection laws such as the GDPR.[4]

Symbolic layers could, in principle, mitigate these issues by enforcing machine-interpretable constraints and providing verifiable rationales.

Beyond technical considerations, these limitations are increasingly reflected in emerging regulatory and governance frameworks. Several jurisdictions have begun to codify expectations for trustworthy AI. For example, Regulation (EU) 2024/1689 (''AI Act'') emphasized traceability and risk assessment for high-risk systems, the United States introduced the *NIST AI Risk Management Framework* [11] and Executive Order 14110 (2023), Switzerland published federal *AI Guidelines*, and more than forty countries have adopted the *OECD AI Principles*.[5] Although these instruments differ in scope and legal force, they converge on requirements, such as explainability and accountability, that hybrid architectures are well placed to satisfy. This paper references them here purely as *illustrative examples*, mindful that compliance obligations ultimately depend on the application's geographical and sector-specific context.

B. PURPOSE AND CONTRIBUTION OF THIS REVIEW

This paper conducts a **Systematic Literature Review (SLR)** of personalized, agent-based Hybrid AI systems published between 2018 and 2025.

Using Kitchenham's SLR guidelines [12] in combination with a Goal-Question-Metric (GQM) framework [13], we screened 9,733 bibliographic records and retained 46 primary studies that met strict criteria of methodological rigor and replicability.

1) *Three recurrent architectural patterns* dominate: (i) serial hybrid pipelines, (ii) shared representation spaces, and (iii) multi-agent orchestration.
2) *Validation gap:* 71% of contributions remain prototype-level, with scarce large-scale or longitudinal evaluations.
3) *Outstanding challenges* include standardized transparency protocols, GDPR-compatible privacy safeguards, and continuous neuro-symbolic integration.

Building on these observations, we outline a research agenda that prioritizes (i) public benchmarks dedicated to hybrid personalization, (ii) cross-paradigm explanation standards for both agent–to–agent and agent–to–human communication, and (iii) design guidelines that blend symbolic guarantees with data-driven adaptability.

C. PAPER ORGANIZATION

The rest of the paper is organized as follows: Section II provides an overview of the methodology used to perform this SLR. Section III introduces the planning phase of the review, including the formulation of the protocol, macro areas, and articulation of the RQs. Section IV reviews the results from the implemented methodology, organized according to the macro areas and RQs. Section V discusses the results acquired, including those indicated in the primary studies and those inferred by the authors. Finally, Section VI concludes the paper.

II. SYSTEMATIC LITERATURE REVIEW METHODOLOGY

The methodology is illustrated in Figure 1 and structured into three distinct phases:

1) **Review Planning**: This initial phase focuses on organizing the macro areas of interest that define the milestones of our study. To explore the areas in more depth, we adopted the GQM approach proposed by Galster [13], a methodology that has been applied in various domains, including software measurement initiatives and virtual reality simulators [14], digital twins in manufacturing [15], tourism applications [16], and decentralized systems such as Multi-Agent Systems (MAS) and blockchain technologies [17]. Through the GQM approach, we formulate the main Research Question (RQ), from which Structured Research Questions (SRQs) are derived. This process also involves defining and validating the search protocol, ensuring its alignment with established standards of rigor and reproducibility.
2) **Reviews and analysis**: In this stage, the review is carried out through the systematic execution of predefined tasks, including the identification, selection, and detailed examination of relevant literature and

[1] https://eur-lex.europa.eu/legal-content/EN/LSU/?uri=oj:L_202401689
[2] https://www.nist.gov/itl/ai-risk-management-framework
[3] https://openai.com/api/pricing, https://ai.google.dev/gemini-api/docs/pricing, https://ai-claude.net/pricing, https://mistral.ai/pricing#api-pricing (all accessed December 2, 2025).
[4] https://gdpr-info.eu
[5] https://www.oecd.org/en/topics/ai-principles.html (Accessed December 2, 2025).

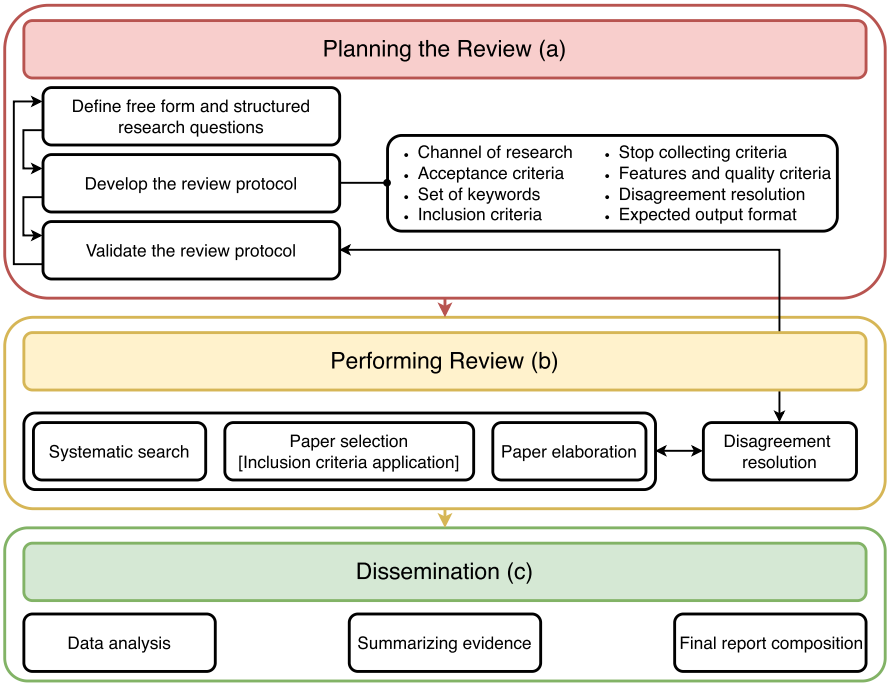


FIGURE 1. SLR methodology adapted from [12], showing the Planning (Sections II, III), Execution (Sections III, IV), and Dissemination stages (Sections IV, V).

resolving any disagreements that may arise during the process.

3) **Discussion and Conclusions**: The final phase involves analyzing, documenting, and reporting the results in a structured manner, along with synthesizing the lessons learned, thereby supporting a comprehensive understanding of the findings.

III. REVIEW PLANNING

This section defines the RQ, the macro areas, and the SRQs. It also details the search strategy, inclusion and exclusion criteria, results produced from each query, the procedures for studies selection, and the mechanisms used to resolve potential disagreements.

A. MACRO AREAS & RESEARCH QUESTIONS

As mentioned in Section I, recent years have seen a growing interest within the research community in applying personalized Hybrid AI across different domains and stakeholders. Within this context, the overarching research question can be set as: ***How is personalized, agent-based Hybrid AI currently conceived and implemented?***

To structure and guide the formulation of the SRQs, we identified four key thematic dimensions that reflect the core concerns and developments in personalized, hybrid, agent-based systems. These dimensions emerged from an initial exploratory literature analysis and serve as analytical lenses for the review process. They are briefly described below:

1) **Hybrid Personalized Agent-Based Methodologies and Frameworks:** Examination of current methodologies and frameworks that integrate data-driven and rule-based models within personalized agent-based systems, including how such hybrid systems leverage Machine Learning (ML), neuro-symbolic methods, and LLMs to enhance effectiveness and reliability.
2) **Use Cases and Domain-Specific Application Scenarios:** Mapping of primary real-world scenarios and contexts in which personalized hybrid agent-based systems have demonstrated significant value, with attention to how these applications vary across different domains.
3) **Frameworks, Evaluation Metrics, and Benchmarking Standards:** Focus on available technological stack and benchmarks to systematically evaluate hybrid agent-based systems, identification of the most effective metrics to assess performance across dimensions such as accuracy, interpretability, and adaptability, as well as comparative analysis of how current systems perform against these benchmarks.
4) **Limitations, Failure Modes, and Open Research Challenges:** Focus on key limitations faced by hybrid agent-based modeling systems – including computational complexity, validation difficulties, data requirements, scalability, and interoperability – and identification of open challenges that future research must address to enhance their effectiveness and reliability.

Building on these macro areas, we structure 10 SRQs to address demographics and application domains, types of users, objectives and system requirements, levels of abstraction, employed technological frameworks and methodologies, as well as semantic capabilities. Finally, we assess their benefits, limitations, proposed mitigations, and future developments. Such SRQs are formalized as follows:

SRQ1 Demographics. *What is the temporal and geographical distribution of research works on personalized agent frameworks?*

SRQ2 Abstraction. *What is the abstraction level of the scientific contributions on personalized agent frameworks? Are they conceptual models (C), prototypes (P), or tested/evaluated systems (T)?*

SRQ3 Domains. *In which domains have Hybrid AI-based personalized agent frameworks been applied?*

SRQ4 Objectives. *What are the objectives set for Hybrid AI agent-based frameworks?*

SRQ5 Users. *Who are the primary users of Hybrid AI agent-based frameworks, and what are their roles?*

SRQ6 Hybridization. How are hybrid models conceptualized and designed to integrate data-driven and rule-based components in Hybrid AI agent-based frameworks?

SRQ7 Advantages. *What advantages do Hybrid AI agent-based frameworks provide for users?*

SRQ8 Limitations. *What are the limitations in implementing Hybrid AI?*

SRQ9 Evaluation. *What evaluation metrics and benchmarking techniques have been used to assess the effectiveness of Hybrid AI agent-based frameworks?*

SRQ10 Future Challenges. *What are the main challenges and open RQs for Hybrid AI agent-based frameworks?*

B. REVIEW PROTOCOL

The review protocol establishes the methodological foundation for this study, ensuring a comprehensive, unbiased, and reproducible process. It includes five key components: (i) a search strategy that specifies information sources and search terms, (ii) defined inclusion and exclusion criteria for determining study eligibility, (iii) a structured procedure for screening and selecting relevant literature, (iv) guidelines for addressing potential disagreements among reviewers to minimize bias, and (v) an analysis of methodological limitations.

1) SEARCH STRATEGY

The search strategy involved querying multiple scientific databases, including IEEE Xplore,[6] ScienceDirect,[7] ACM Digital Library,[8] CiteSeerX,[9] PubMed,[10] Google Scholar[11] to access the most important conferences not included in other databases, and arXiv[12] in order to account for recent and fast-evolving contributions in the field. Preprints were considered only when they provided sufficient technical detail and were not superseded by peer-reviewed versions at the time of the review. Identifying relevant keywords was grounded in the reviewers' expertise and in-depth knowledge of the research landscape concerning Hybrid AI, Personalized AI, and Agent-based AI. Specifically, the keywords employed are: *agent-based, multiagent systems, Hybrid AI, personalized framework, support, assistive, framework, data-driven, rule-based, LLM, agentic, ethical, legal aspects, compliance, privacy, explainability*. These keywords were combined in various ways to maximize coverage and thematic relevance. The search continued until thematic saturation was achieved. This means that subsequent queries consistently yielded either papers already selected or papers that fell outside the established inclusion criteria. No new relevant concepts, frameworks, or application domains have been identified in the literature, indicating that the search has successfully covered the necessary material. The research time frame is limited to 2018 and February 2025, as this period reflects the major shifts introduced by the attention mechanism and LLMs. Table 1 reports all the search queries used in this study, along with the number of results returned by each query and the number of papers selected after preliminary screening. The queries were formed by combining the predefined keyword sets (listed above) using Boolean logic. Collectively, the set of queries covers keywords related to agent-based systems, Hybrid AI, and application domains such as personalized support. While most queries combine terms from these dimensions, some intentionally focus only on agent-based and Hybrid AI concepts to maximize recall and capture candidate architectures that are later filtered through the inclusion and exclusion criteria.

The reviewers conducted a preliminary screening for each query to evaluate the retrieved papers' relevance to the research objectives. This assessment primarily focused on titles and abstracts and followed the eligibility criteria outlined in the following section.

2) INCLUSION AND EXCLUSION CRITERIA

The initial search process yielded a total of 114 selected papers. A subsequent filtering phase was carried out based on the following criteria:

- exclusion of duplicate or incremental publications describing substantially the same work;
- restriction of the publication timeframe to discard outdated or less relevant contributions, considering technological evolution;

[6] http://ieeexplore.ieee.org
[7] http://www.sciencedirect.com/
[8] http://dl.acm.org/
[9] http://citeseerx.ist.psu.edu/index
[10] http://www.ncbi.nlm.nih.gov/pubmed
[11] https://scholar.google.com
[12] https://arxiv.org

TABLE 1. Queries used in the search process, papers retrieved, and papers selected.

Queries	N. of results	Selected
(“agent-based” OR “multiagent systems”) AND “personalized framework”	30	9
(“agent-based” OR “multiagent systems”) AND “framework” AND “Hybrid AI”	375	14
(“agent-based” OR “multiagent systems”) AND “hybrid ai”	382	4
(“agent-based” OR “multiagent systems”) AND “hybrid ai” AND (“assistive” OR “support”)	359	3
(“agent-based” OR “multiagent systems” OR “agentic”) AND “data-driven” AND “rule-based”	2,580	15
(“agent-based” OR “multiagent systems” OR “agentic”) AND “LLM”	4,138	58
(“agent-based” OR “multiagent systems” OR “agentic”) AND “LLM” and “rule-based”	776	5
(“agent-based” OR “multiagent systems” OR “agentic”) AND “personalized” AND (“ethical” OR “legal aspects” OR “compliance” OR “privacy” or “explainability”)	1093	6
Total	9,733	114

- selection of studies that directly address the specific research topic;
- inclusion of papers offering substantial theoretical or practical contributions while excluding purely visionary or speculative works.

3) STUDY SELECTION

Three reviewers independently applied the inclusion and exclusion criteria to identify relevant publications. The work was organized in batches, and two reviewers examined the papers for each batch. After completing the individual screening processes, a consensus-based validation was conducted. As a result, 43 articles were included, 43 were discarded, and 28 had conflicting opinions between the two reviewers. To resolve conflicts, The third reviewer independently examined the articles for which the two reviewers reported conflicting assessments and reviewed their feedback. Where necessary, the three reviewers jointly compared the eligibility criteria with their respective screening assessments to reach a final decision. Specifically, of the 28 articles with conflicting assessments, 10 were included and 18 were excluded.

As a result, a final set of 53 articles was identified and will be referred to as the primary studies throughout the remainder of this work. Following the demographic analysis, 7 additional articles were discarded. Although these studies appeared relevant based on title, abstract, and keywords, the in-depth analysis revealed that they did not adequately address the research objectives of this SLR or lacked sufficient methodological or contextual alignment with the defined RQs (e.g. issues identified only during full-text analysis). Therefore, from the abstraction analysis (SRQ2) onward, the analysis focused on 46 selected studies.

4) BIASES AND LIMITATIONS OF THE SLR METHODOLOGY

The limitations associated with the adopted SLR methodology can be summarized as follows:

- **Accessibility**: Some potentially relevant primary studies may have been inaccessible or unintentionally omitted.

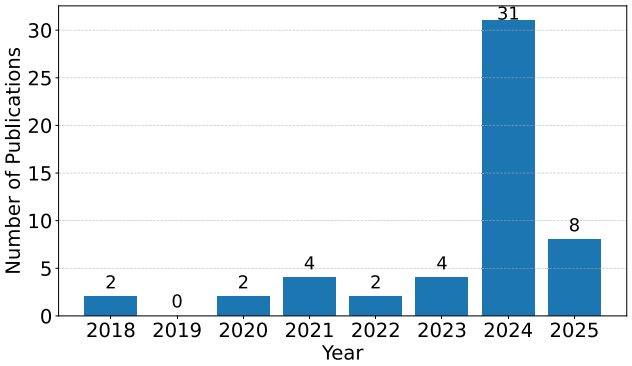


FIGURE 2. Temporal distribution of selected studies.

- **Timeliness**: Recently published studies may not have been captured due to the temporal constraints of the review process.
- **Clarity**: Extracting key information was occasionally limited by a lack of clarity in how authors reported the limitations of their studies. To address this, the reviewers relied on their expertise to interpret the content and supplemented the analysis with additional insights provided in Section V.

Procedures were established to mitigate bias among reviewers and to address disagreements throughout the selection process. The three reviewers cross-validated the inclusion and exclusion criteria to enhance the methodological rigor at each stage. Additionally, regular meetings were held to resolve ambiguities and reach consensus.

IV. REVIEW RESULTS AND ANALYSIS

In the following subsections, we present the SLR outcomes based on the macro areas and the SRQs introduced in Section III-A.

A. SRQ1 - DEMOGRAPHICS

Figures 2, 3, and 4 address SRQ1 by providing a temporal and geographical overview of the selected papers.

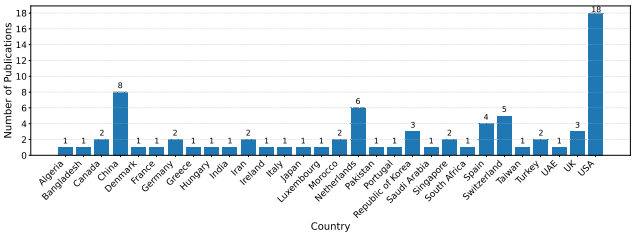


FIGURE 3. Distribution of paper by country.

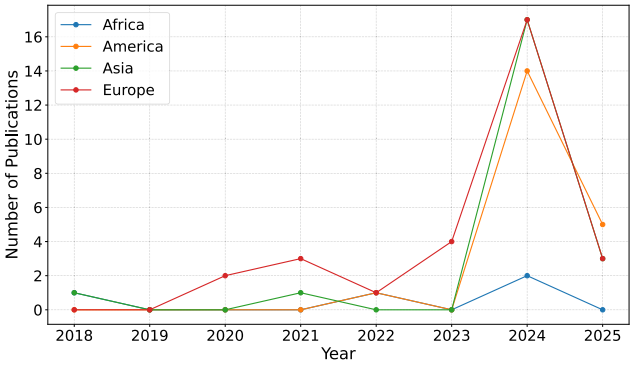


FIGURE 4. Distribution of papers by continent and year.

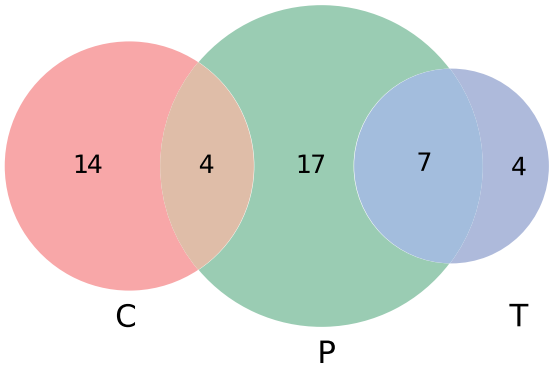


FIGURE 5. Types of studies. C: conceptual, P: prototype, T: tested. Permitted intersections are only $C \cap P$, and $P \cap T$.

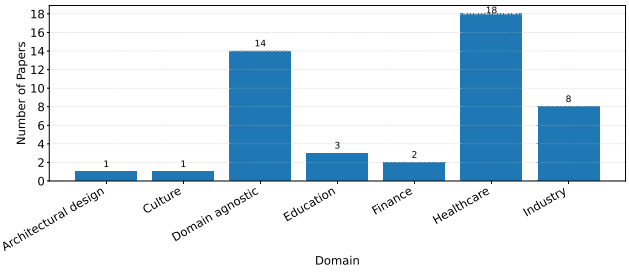


FIGURE 6. Distribution of papers by domains.

Figure 2 shows the temporal distribution of the selected studies. Most publications are concentrated in 2024, whereas in earlier years the number of papers remains relatively low. Since the primary study search was conducted in February 2025, a roughly linear growth can be hypothesized, suggesting a continued increase in publications through 2025 and pointing to a growing interest in hybrid, personalized, and agent-based AI.

Figures 3 and 4 provide geographical insights. All authors' affiliations are considered. However, a country is counted only once for each paper, even if shared by multiple authors.

Figure 3 shows the distribution by country. The United States and China are the most productive countries, followed by the Netherlands, Switzerland, and Spain. The United States produced more than twice as many papers as China.

Figure 4 aggregates data by continent, showing the number of papers published yearly. Europe emerged as the most prolific continent in this field and shared the top position with Asia in 2024. This aggregated view suggests that Europe, as a whole, is producing more studies than the United States.

Overall, temporal and geographical distributions indicate that the adoption of Hybrid AI is rapidly expanding, with expectations for further growth in the coming years.

B. SRQ2 - PAPERS ABSTRACTION

To address SRQ2, Figure 5 shows the distribution of the abstraction levels among the 46 selected studies. Most contributions are prototypes (37%) and conceptual models (30%). Thoroughly tested or evaluated systems are still a minority (9%).

Intersections between categories reveal the progression of research efforts. Specifically, 9% of studies are positioned between conceptual and prototypical ($C \cap P$), indicating theoretical works that have started an initial implementation phase but have not yet reached a fully functional prototype stage. Similarly, 15% fall between prototyping and testing ($P \cap T$), representing systems that provide some empirical evidence (e.g., performance metrics) but lack comprehensive evaluations to be classified as thoroughly tested.

Overall, more than half of the studies (61%) involve some level of implementation (P, P/T, and T), suggesting a shift toward applied research. However, the relatively low number of fully validated systems highlights that the field is still in an exploratory and development-oriented phase.

C. SRQ3 - APPLICATION DOMAINS

The field of Hybrid AI crosses many research and application domains. Figure 6 answers SRQ3 and analyzes the distribution of studies across domains. The distribution suggests a prevalence of studies in the healthcare domain, followed by domain-agnostic research and the industrial domain.

Specifically, the industrial domain includes autonomous driving [3], drone navigation [9], robotics [18], the construction sector [19], and 3D design [20]. The high presence of domain-agnostic research shows that the field of personalized Hybrid AI is still in an early stage and is experiencing a strong expansion that has yet to be fully transferred to application domains. There is an increasing focus on AI that includes ethics [21], [22], technologies integration [23],

TABLE 2. Goals organized in categories.

Category	N. of papers	Studies
Personalized coaching and health support	8	[32]–[39]
Decision support systems	10	[2], [3], [9], [28], [29], [40]–[44]
Recommendation systems	8	[1], [20], [26], [29], [30], [39], [45], [46]
User Modeling, Monitoring and Adaptation	4	[2], [30], [31], [47]
Hybrid Architectures, Frameworks and Theoretical Contributions	15	[18], [19], [21]–[24], [36], [46], [48]–[54]
Ethical and Human-Centered AI	4	[1], [39], [40], [54]
Overcoming Limitations of LLMs in Complex Reasoning	10	[20], [27], [32], [34], [37], [40], [44], [55]–[57]
Human-AI Interaction / Conversational Agents	6	[25], [26], [34], [57]–[59]

comprehension [24], and conversation with increasingly natural language [25], [26] and personalization [27]. The education sector explores subdomains of personalized recommendations [28], [29] and adaptive e-learning [30]. In the finance sector, research focuses on trading strategies [2] and digital banking [31]. Finally, the healthcare domain, which has the strongest concentration, is examined as a specific use case in Section V.

D. SRQ4 - GOALS

SRQ4 focuses on analyzing goals in the field of agent-based Hybrid AI. To answer this research question, we grouped the objectives extracted from the primary studies and categorized them, as shown in Figure 7. Below, we analyze the main categories; it is essential to note that an article can have multiple objectives and thus fall into various categories. Table 2 shows an exhaustive classification of each item studied.

Numerous studies target the main architectural or theoretical contributions, which are the methodological basis for integrating different AI models. Schmid [23] introduces a systematic approach to the design of hybrid modular architectures. van Bekkum et al. [18] explore more than 15 design patterns and define a taxonomy for Hybrid AI. Hu et al. [50] present a survey of the vulnerabilities and challenges of LLM-MAS systems and propose a human-centered approach to ensure security and reliability. Spivack et al. [24] describe the minimum functional requirements for a cognitive architecture and argue why general AI cannot emerge from probabilistic models alone.

Building on these foundations, an emerging stream of research addresses the limitations of LLMs, particularly in complex, multi-step reasoning. Hong et al. [40] introduce a MAS framework for the argumentation, validation, and reflection stages to mitigate bias and forgetting. Rasal [27] proposes an orchestrator for context maintenance and conversational memorization, focusing on personalization. Other studies, such as Aggarwal [55], explore strategies to increase reasoning, memory, and coherence in LLM-based systems.

These architectural and reasoning advancements support the development of decision support systems across healthcare, educational, and logistical processes. The study [3] explores the collaboration between ML models and rule-based systems to optimize decisions in the context of autonomous driving. Kwon and Lee [2] introduce a system to support finance decisions while minimizing risk. Meanwhile, Li et al. [41] define a distributed approach that supports decisions through shared rules, helping to enforce standard protocols.

Closely related to decision support, personalized recommendation systems represent another key area, often integrating content-based techniques, collaborative filtering, and hybrid architectures. For example, Amin et al. [29] present a system for recommending educational courses using Reinforcement Learning (RL) and rule-based systems. Portugal et al. [46] propose an LLM-based MAS framework for intelligent and personalized recommendations.

Many decision, recommendation, and support systems rely on modeling, monitoring, and adapting to user behavior. Bourkoukou and El Bachari [30] propose a model that dynamically customizes training paths based on student profiles, demonstrating how behavioral data can drive adaptation in educational contexts.

Building on these user modeling capabilities, several studies explore personalized coaching and health support applications, where interventions are tailored using clinical data and individual goals. For instance, Calvaresi et al. [33] present a decentralized, AI-based system for managing personal health trajectories. Mitchell et al. [35] explore micro-coaching dialogues in nutrition, comparing hybrid and data-driven approaches. In medical education, Yu et al. [38] introduce a virtual patient capable of generating realistic, dynamic, and personalized scenarios.

Several studies also focus on Human-AI interaction, often through conversational agents, chatbots, or voice assistants. For example, Varitimiadis et al. [57] propose a chatbot for museums that can handle collaborative and informed conversations. Wang et al. [25] present a conversational assistant that aids academic researchers in managing literature.

Finally, many contributions highlight Hybrid AI systems' ethical and human-centered dimensions, focusing on transparency, explainability, trust, and fairness. Buzcu et al. [1] explore interactive explainability to increase recommendations and user experience transparency. Yang et al. [54] propose a new framework in healthcare to improve privacy, interpretability, and transparency through the combined use of LLMs, Knowledge Graphs (KGs), and Causal Graphs.

E. SRQ5 - INTENDED USERS

This section analyzes the categories of end users targeted by agent-based personalized Hybrid AI applications. End users are defined as those who benefit from the advantages of the

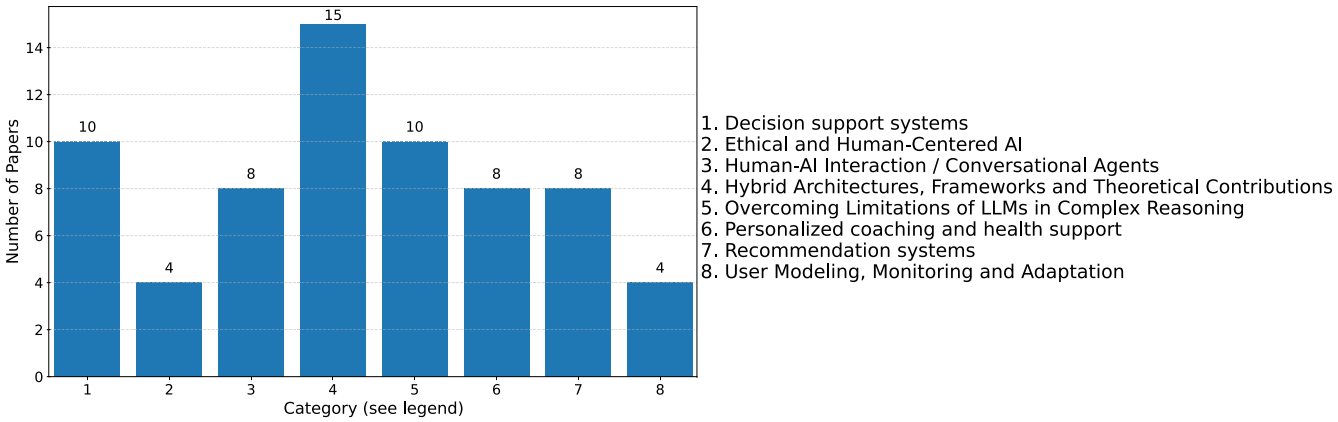


FIGURE 7. Objectives of primary studies organized by category.

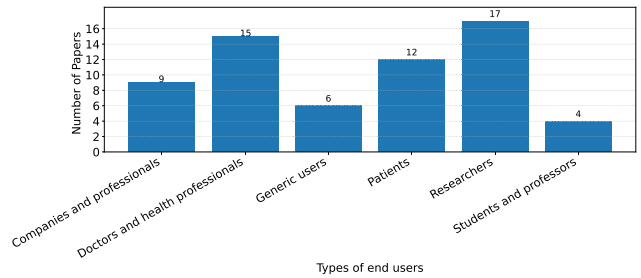


FIGURE 8. End users of selected papers.

TABLE 3. Items classified by end-user type.

Type of users	N. of papers	Studies
Companies and professionals	9	[2], [19], [20], [23], [28], [44], [52], [58], [59]
Doctors and health professionals	15	[33]–[43], [48], [49], [54], [59]
Generic users	6	[1], [22], [27], [31], [46], [57]
Patients	12	[1], [32], [33], [35]–[37], [39], [41], [45], [48], [49], [54]
Researchers	17	[3], [9], [18], [21], [24]–[26], [38], [47], [50]–[56], [59]
Students and professors	4	[28]–[30], [38]

paper. For this reason, researchers have been included, as they are the main users of conceptual papers. Figure 8 shows that most end users are researchers (26.98%), confirming that many selected studies are still at a conceptual or prototype stage. Right after, we find doctors and health professionals (23.81%), reflecting the high incidence of applications in the medical field. They are followed, tied, by patients (19.05%), and companies/professionals with a 14.29%, indicating an interest in clinical and industrial adoption. General users account for 9.52%, while students and faculty stand at 6.35%, indicating some educational and academic involvement. Table 3 shows a classification of items by type of end-user.

F. SRQ6 - HYBRIDIZATION MODELS

This section analyzes the nature of the hybrid models employed in the selected studies. It focuses on the technologies used, how data-driven and rule-based components are combined, and the architectural and interaction strategies through which their integration is achieved. The analysis shows that the hybrid models in the literature include a wide range of architectural, functional, and semantic combinations. The most common forms of hybridization include integrating data-driven learning components with agent-based architectures and symbolic/knowledge-based reasoning modules. The technological framework is dominated by data-driven approaches: thirty-eight papers (82.61%) rely on ML methods, deep neural networks, RL, or LLMs, while twenty-eight (65.22%) use agent-based infrastructures to coordinate specialized sub-components. Symbolic and knowledge-based reasoning appear at 41.30%, typically imposing logical constraints or providing explanations. Emerging technologies such as Responsible AI and Human-AI Interaction remain marginal, appearing in four and three cases, respectively. The use of evolutionary techniques and heuristics results in a few isolated cases.

1) ARCHITECTURAL COMPLEXITY AND COMBINATION PATTERNS

Concerning architectural complexity, the number of integrated techniques per paper can reach 4, with an average of 2.66 and a median of 3; nearly 45% of the studies employ 3 techniques.

The most widely used combination employs MAS with a data-driven approach, creating a purely data-driven system without integrating other knowledge-based or symbolic AI techniques. The study of Rasal [27], for example, presents a multi-LLM approach, i.e., an architecture that combines multiple language models with distinct functional roles coordinated by MAS. Although they present themselves as ‘‘hybrids’’, these approaches do not include explicit symbolic components or formal logical or cognitive structures. They

are, therefore, a form of hybridization within the sub-symbolic paradigm. Although interesting on the application level, such systems do not involve the fusion of distinct cognitive paradigms.

From an architectural perspective, most studies adopt a modular approach, in which symbolic and statistical components are designed as functionally distinct but semantically coordinated modules. In this scheme, hybridization is achieved through cooperation or concatenation between modules that play different roles within the system. Some studies employ a layered architecture, in which symbolic reasoning is used to filter inputs or refine outputs of data-driven models [2]. Others adopt an orchestration module that mediates between an ML system and an rule-based layer [3].

Van Bekkum et al. [18] propose a significant contribution as it addresses the problem of the lack of a shared conceptual and architectural language for neuro-symbolic systems. The work presents 15 modular design patterns, which can be combined to describe complex hybrid architectures. The idea comes from design patterns in software engineering, which are reusable models for solving common problems.

A taxonomy is proposed to support the description of modular systems:

- Instances: processed data, divided into symbols (e.g., labels, logical relationships) and raw data (e.g., images, text).
- Models: representations constructed by algorithms, which can be statistical (e.g., neural networks) or semantic (e.g., ontologies, rules).
- Processes: operations performed on data and models, such as training, inference, transformation, and generation.
- Actors: autonomous entities (humans or software agents) that execute processes.

This taxonomy facilitates the representation of each design pattern through graphical notation. The foundational eight patterns serve as the core components to construct complex hybrid architectures. Table 4 shows the patterns characterization. These basic patterns, when combined, give rise to composite patterns. Below, we take a closer look at three particularly representative advanced patterns: (i) Learning an Intermediate Abstraction; (ii) Learning to Reason; (iii) Meta-Reasoning for Control.

a: LEARNING AN INTERMEDIATE ABSTRACTION

It describes architectures in which the system does not directly learn a mapping from raw data to the final output. However, it introduces an intermediate step: a symbolic or conceptual representation. A typical example is the addition of handwritten digits: instead of training a model that directly sums the images of the numbers, a first module is used to recognize the digits (e.g., '5', '3') and a second module to perform the sum. This approach enhances robustness, modularity, and transferability by separating perception understanding from task logic.

b: LEARNING TO REASON

It focuses on systems in which a neural network directly learns to perform logical reasoning. Starting from a symbolic basis (e.g., an RDF graph or a theory in descriptive logic) and observing the results of a logical reasoner, the statistical system learns to replicate those inferences. Its goal is to create models that are more scalable and robust than classical logical reasoners, especially in the presence of noise or incomplete data. Concrete examples are Relational Tensor Networks [60] and neural models for query answering on knowledge graphs. These systems enable approximate yet rapid inferences, learning directly from logical examples.

c: META-REASONING FOR CONTROL

AI modules are guided via a symbolic control module. For example, a knowledge base can determine which hyperparameters to use for training a neural network or select the order of examples to provide for maximum learning (curriculum learning). This approach enhances adaptability and transparency in complex systems and can also be applied in AutoML [61] contexts.

2) FORMALIZATION AND THEORETICAL FRAMEWORKS

Some theoretical contributions present the formalization of hybrid architectures, e.g., [23] presents a taxonomy for Hybrid AI and a modular architecture that divides modules according to the methods used, the cognitive capabilities emulated, and the level of risk and criticality. The study [18] aims to provide a high-level description of Hybrid AI to unify the terminology used in various domains and present reusable modular design patterns. Meyer-Vitali et al. [21] focus on transparency at the design pattern level to simplify the engineering of complex Hybrid AI models.

Integration Strategies Between Data-driven and Rule-based Components. Hybridization does not only depend on the co-presence of heterogeneous components, but also on the strategies adopted to enable their interaction. Across the analyzed studies, three recurrent integration strategies can be identified.

- **Pipelined or concatenated integration.** Data-driven and symbolic modules are organized sequentially, such that the output of one component becomes the input of the next. This strategy is common in modular hybrid architectures and is often implemented through explicit interfaces or coded transitions. Integration may be unidirectional or bidirectional, allowing iterative refinement between statistical inference and symbolic reasoning.
- **Shared representations and interfaces.** Some architectures rely on a common representational space to mediate interaction between paradigms. In these

TABLE 4. Description and classification of patterns by purpose, conceptual diagram, and typical example.

Purpose	Pattern	Conceptual Schema	Typical Example
Train a statistical model from raw data.	Train	Data ⇒ train ⇒ Statistical model	CNN trained on images.
Learn a symbolic model from symbolic examples.	Train	Symbols ⇒ train ⇒ Semantic model	Inductive Logic Programming.
Manual construction of a semantic model by experts.	Manual (model:sem)	Expert ⇒ Ontology/Rule base	Competency ontology.
Conversion between representations.	Transform	Instance (data / symbol) ⇒ transform ⇒ New instance	Word2Vec for text; KG embedding.
Deductive use of a statistical model on new data.	Infer (data+model:stat)	Data + Statistical model ⇒ symbol	Image classification.
Symbolic deduction using a semantic model.	Infer (symbol+model:sem)	Symbols + Semantic model ⇒ New symbols	Reasoner over ontology.
Induction of symbols from data with model support.	Induce/Infer (data ⇒ symbol)	Data + Model ⇒ Symbolic rules/labels	Generating rules for KG completion.
Semantic model transformation.	Transform (model:sem)	Symbols + Semantic model ⇒ transform ⇒ data	Serializing an ontology in RDF.

systems, symbolic structures (e.g., knowledge graphs) and learned representations (e.g., embeddings) coexist and are jointly exploited by downstream components, such as retrieval-augmented generators or graph-based reasoning modules.

- **Agent-level orchestration.** In MAS, each agent may adopt a distinct paradigm (e.g., a symbolic agent for reasoning and an ML agent for prediction). An orchestrator or communication protocol, therefore, handles coordination. In [56], a meta-planner assigns agent roles according to their underlying paradigm and task constraints. At the same time, integration is achieved through a dependency-aware workflow graph monitored by edge and validation agents.

Despite this variety, only a few studies propose a rigorous formalization of the integration process. There is often a lack of explicit modeling of semantic compatibility, conflict resolution, or inferential consistency between symbolic and statistical-element modules, representing open future challenges.

3) USE CASES – APPLIED HYBRID ARCHITECTURES.

The architecture presented by Jeong [58] proposes an advanced Retrieval-Augmented Generation (RAG) [62] system enhanced through an agent-based and graph-based framework. This solution integrates graph technologies, agent orchestration (LangChain [63]), and an LLM (OpenAI GPT-4-turbo [64]) to improve accuracy and reliability in response generation. The architecture leverages a multi-step approach consisting of initial information retrieval, relevance assessment using intelligent agents, possible automatic query reformulation, and dynamic integration of results from Web searches. Information is managed through vector databases (ChromaDB [65]), while advanced embedding techniques (OpenAI Embeddings [66]) ensure effective knowledge management. Workflow is represented and managed through graph structures, which allow the definition of complex flows with clear and precise transitions.

The study of San-Segundo et al. [9] proposes a Hybrid AI architecture for autonomous drone navigation. This solution combines Deep Learning (DL) models, trained through RL algorithms, with an expert rule-based system to handle specific critical situations. The architecture includes a navigation module that dynamically alternates between automatically learned strategies (via Proximal Policy Optimization (PPO) [67]) and predefined rules to avoid obstacles and ensure reliability. Also integrated is an explainability module that uses techniques such as LIME [68] and SHAP [69] to clarify drone decisions according to the local observation space. Finally, human interaction capabilities allow a supervisor to manually intervene on the drone during navigation, providing additional control and operational flexibility. This study highlights that the combination of AI techniques is exclusive, with the rule system ready to intervene in critical situations.

Rahbar et al. [52] propose a hybrid architecture for automating two-dimensional architectural layouts, combining agent-based models and DL techniques using Generative Adversarial Networks (GANs) [70]. The architecture addresses topological and geometric constraints by integrating explicit rules and implicit learning from data. In the first phase, it generates a bubble diagram that respects the topological constraints explicitly defined by the designer;

this result is transformed into a heat map and is passed to GAN-based model. In the second stage, the model translates the heat map into a detailed spatial layout, respecting implicitly learned geometric constraints such as dimensions and proportions. This approach ensures high topological control and geometric quality, although its effectiveness depends on the dataset's quality and consistency between the two stages of the process.

Another interesting approach is proposed with AGENTi-Graph software [59], a platform for interaction between LLMs and KGs based on a hybrid MAS architecture. The system addresses the problems of information incompleteness, hallucination, and limited reasoning ability typical of LLMs by integrating Natural Language Processing techniques, semantic knowledge management, and intelligent agent orchestration. The architecture is based on a system of specialized agents, each dedicated to specific tasks: interpretation of user intent through Few-Shot Learning and Chain-of-Thought reasoning, extraction of key concepts through Named Entity Recognition (NER), and Relation Extraction (RE) based on BERT [71] embeddings, task scheduling, dynamic interaction with KGs through the ReAct framework, application of logical inference, and real-time updating of the knowledge base on Neo4j [72].

Zhang et al. [44] propose Planning with Multi-Constraints (PMC), a hybrid architecture for MAS based on LLMs that addresses the problem of complex planning with multiple constraints. PMC introduces a structured task planning approach in which a manager agent decomposes the main task into a hierarchy of related sub-tasks by constructing a sub-task graph. Executor agents subsequently translate each sub-task into a sequence of concrete actions using established planning techniques such as ReAct [73]. The hybridization of the architecture is achieved by combining rule-based components (explicit structuring of tasks, definition of local and global constraints) with data-driven components (zero-shot reasoning of LLMs to generate and execute dynamic plans). In addition to managers and executors, the system includes a supervisor agent that refines sub-tasks based on the partial results obtained and a deliverer agent that aggregates the final results, respecting the global constraints of the problem.

Another approach that combines data-driven and RL is the one of Chang [56], where a system called MACI (Multi-Agent Collaborative Intelligence) is presented, which proposes an architecture to overcome the limitations of LLMs in complex planning and managing time constraints. The approach is based on a hybrid model integrating rule-based components, specialized agents, and adaptive reasoning modules. MACI's architecture consists of three main components: (i) Meta-Planner that automatically generates a planned workflow by building a task dependency graph based on the objectives, roles, and constraints extracted from the task. (ii) Repository of Agents: contains specialized agents distinguished between common agents (for constraint checking, common sense integration, validation of schedules) and task-specific agents (for specific application domains such as logistics or medicine). Each agent is designed to operate on narrow context windows, minimizing the risk of attention bias. (iii) Run-Time Monitor: oversees the execution of the plan, adapting it in real-time to unexpected events through dynamic changes to roles and dependencies. Hybridization emerges from combining (i) rule-based techniques with explicit constraints between tasks and plan management via a graph. (ii) data-driven techniques: use of LLMs for reasoning about specific tasks and inference not explicitly encoded.

G. SRQ7 - ADVANTAGES

To address SRQ7, we categorized the advantages that primary studies bring to users. Each study may highlight more than one benefit and, consequently, be included in more than one category. Figure 9 illustrates the identified categories and their intersections, showing that 10 studies have benefits that can be attributed to only one category (shown in red in the graph). In the other cases, the advantages range from 2 to 4 categories, with an average of 2 categories per study.

Most studies focus on the ''Personalization'' and ''Efficiency and performance'' categories, with 27 and 28 associated studies, respectively, and an intersection of 10 studies falling into both. This finding highlights that personalization is an area of particular research interest, with an increasing focus on developing solutions tailored to users' specific needs. At the same time, a strong push toward continuous improvement in the quality and efficiency of proposed systems emerges.

The categories ''Quality of service or user experience'', ''Security, privacy, control, risk reduction'', and ''Transparency, explainability, interpretability'' followed, with 13, 11, and 9 studies, respectively. These results further confirm the central role of the user in system design and show how Hybrid AI is oriented toward providing increasingly secure, reliable, transparent, and understandable solutions.

Finally, only six studies represent the ''Flexibility and modularity'' category. Although less frequent, this category is strategically essential: flexibility and modularity are crucial to ensuring systems' adaptability to diverse and evolving application contexts.

H. SRQ8 - LIMITATIONS AND PROPOSED SOLUTIONS

SRQ8 focuses attention on limitations in Hybrid AI implementation. Many studies do not report the limitations; only 27 out of 46 papers present them, and a subset present possible solutions. Table 5 shows the analysis for each primary study.

I. SRQ9 - EVALUATION

In evaluating each primary study, it is critical to identify the metrics and benchmarks used. The analysis is organized initially by level of abstraction and then cross-sectionally, as this level directly affects the evaluation methods adopted.

The studies classified as conceptual [18], [21], [22], [23], [24], [33], [45], [46], [48], [49], [50], [51], [53], [54], are

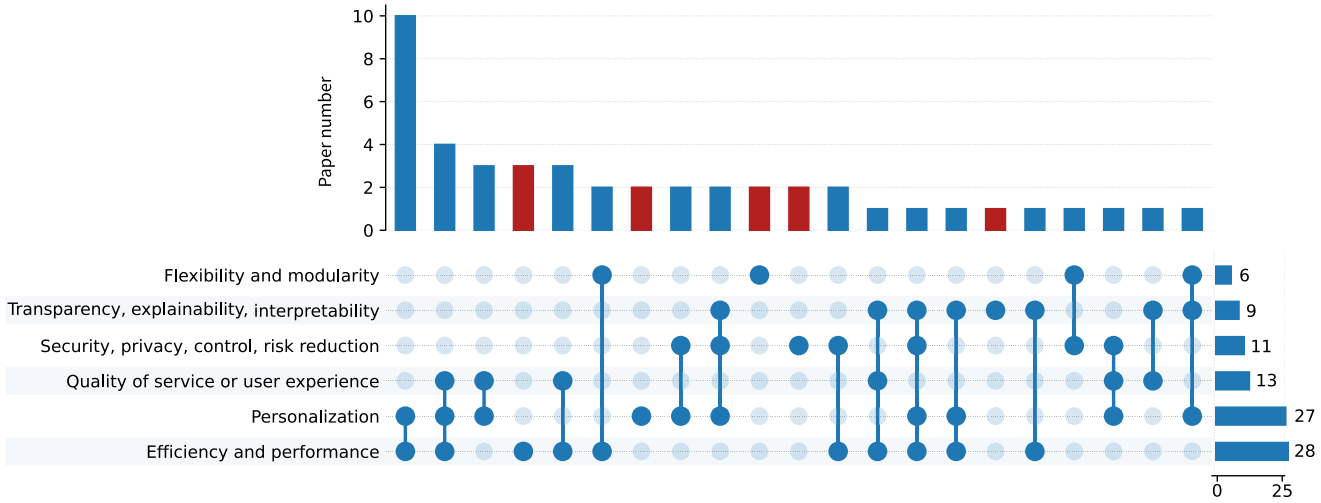


FIGURE 9. Advantages for users organized into categories. The graph shows the intersections between categories, and the studies in red belong to only one category.

consistent with their theoretical nature, and do not include metrics or benchmarks. However, two of them propose potential metrics for use in future implementations.

Studies placed at the intersection of conceptual and prototypical show early-stage applications, and they start using metrics. Rahbar et al. [52], for example, present a subjective evaluation conducted by experts without automated metrics. Varitimiadis et al. [57] adopt the AHP method, a multi-criteria evaluation technique, which is not an objective metric in the strict sense. In Antunes et al. [47], system productivity is measured by the number of artifacts produced per minute, accompanied by subjective evaluations. The study of Easin et al [31] does not report metrics of any kind.

Prototypes comprise most primary studies (see Section IV-B). Except for the studies of Nzomo [36] and Yao et al. [28], all have evaluation metrics. About 35% also employ comparative benchmarks. For example, [34] compares its system with GPT-4, Claude, and a traditional RAG; [27] compares it with Mistral-Instruct-7B, Llama-3.1-Chat-70B, and GPT-4o. Study [42] introduces a custom benchmark called MedChain, while [37] uses 50 public prompts collected during lockdown by COVID-19. Study [56] is based on the Traveling Salesman Problem (TSP) benchmark, while [20] adopts the ASTM-D638-14 Type IV benchmark. The studies [19], [20], [27], [34], [35], [39], [42], [43], [59] adopt standard metrics such as loss, accuracy, F1-score, precision, recall, and Intersection over Union (IoU). Meanwhile, [1], [34] integrate subjective metrics such as usability or user satisfaction questionnaires. San-Segundo et al. [9] propose customized qualitative metrics for the specific domain, including processing time measurements. Finally, works of [30], [37], and [58] do not use standard metrics, relying solely on subjective evaluations or questionnaires.

At the intersection of prototypes and tested applications, we find seven primary studies such as those of [2], [3], [25], [29], [32], [40], and [41], which are characterized by specific and objective metrics and, in five cases, also limited benchmarks. For example, Buzcu et al. [32] conducted experiments with patients; however, as the authors stated, there is a lack of long-term studies to confirm the validity and robustness of the proposed approach.

The four papers classified as tested [26], [38], [44], [55] present results tested on public benchmarks, three of which use at least two benchmarks. The limited number of studies that present well-defined, objective, and benchmark-validated metrics suggests that the field of Hybrid AI is still at an exploratory stage. In this context, the scientific community must begin to converge on shared evaluation practices to support comparisons between approaches, promote replicability, and accelerate research progress.

J. SRQ10 - FUTURE CHALLENGES

The analysis of the reviewed papers revealed a multifaceted set of future challenges spread across several critical axes: architectural integration, scalability, interpretability, validation, human-computer interaction, and ethical and social sustainability of systems. In response to SRQ10, we examine the classification of each contribution.

Architectural integration and communication between modules: Several studies [2], [3], [40] highlight the urgency of refining the mechanisms of interaction between symbolic modules (e.g., BDI, formal logics) and sub-symbolic modules (e.g., ML, RL, LLMs), overcoming the current predominance of unidirectionally controlled structures. The goal is to build bidirectional architectures capable of dynamically adapting decision-making strategies according to context, supporting effective synergy between explicit reasoning and implicit learning. This involves the development of smarter orchestrators, mechanisms for knowledge transfer between modules, and the standardization of modular design practices [18], [21], [23].

TABLE 5. Limitations and proposed solutions.

Study	Limitations	Solutions
[3]	the technique presented is not generalizable, it is ad-hoc for the field of application.	No direct solutions proposed
[29]	Few open source environments, weak reward system.	Improve the DRL system.
[47]	(i) inaccuracies without human supervision.(ii) Difficulty in complex reasoning tasks. (iii) Inherent limitations of LLMs that impact the method: repetition, hallucinations, and inconsistency.	Possible solutions not provided.
[49]	the paper is conceptual and presents the field's challenges, but not real limitations.	
[30]	(i) Generic limitations about Recommended Systems (ii) Absent consideration of pedagogical and motivational factors	Include more sophisticated recommendation methods to take into account students' motivation and level.
[32]	Focus only on the short-term impact of the developed system.	It would be interesting to examine a long-term usage of such chatbot interaction where the agent proactively engages in the dialogue instead.
[1]	(i) it is impossible to distinguish which effect, added explanation or added feedback, is responsible for the results. (ii) The recommendation algorithm is limited.	(i) study the effect of the precise moment when the explanation is generated.
[33]	(i) Data Heterogeneity, (ii) Large-scale integration and aggregation, (iii) Unaligned terminology, (iv) Reduced computational capabilities for wearable devices	(i) Representing trajectories as KGs based on standard ontologies, (ii) Use aggregator-type τAgents that negotiate the collection of trajectories from multiple patient-agents and prepare them for ML analysis, (iii) Apply ontology matching and mapping techniques (iv) real-time support and lightweight scheduling strategies
[34]	(i) Problems in automatic evaluation. (ii) Slow responses in complex cases. (iii) Lack of validation in real scenarios.	(i) Integration of human feedback. (ii) Implementation of caching and orchestration mechanisms.
[20]	(i) Budget and computational power limitations; (ii) limited data; (iii) need to update the KG continuously.	The possible proposed solution involves moving to a specialized multi-agent architecture with targeted fine-tuning of models. In addition, the authors want to add multimodal data and build an adaptive KG with semi-automatic updates.
[40]	(i) Lack of more expressive logic. (ii) Absence of personalization.	(i) Use of formal and probabilistic logics. (ii) Interactivity. Use of user feedback.
[58]	(i)The LangGraph-based system is optimized for specific domains, which may result in performance degradation when applied to other fields. (ii) The system's complexity may require additional resources for implementation and maintenance. (iii) Further validation processes are necessary to ensure the accuracy and reliability of real-time data, which could impact overall system performance.	Experimentation in cross-domain environments to improve the generalization of the algorithm.
[2]	domain problem: The system assumes that the agent's transactions do not affect prices; this makes it suitable only for large-cap stocks, not small-caps.	Possible solutions not specified.
[41]	(i) Scalability when using more than 80 agents. (ii) External intervention is required if problems deriving rules from data exist. (iii) Low generalization.	(ii) exploring the general mathematical properties of the model's internal probabilities.
[42]	(i) Data from one context may not reflect the actual scenario. (ii) Limitations in patient responses as they are generated with Gemma 2.	(i) Use of data from different sources. (ii) studying an advanced simulation system to generate responses with different communication styles and behaviors.
[35]	(i) Domain limitation, (ii) this study only focused on perceptions of the coaching chatbot, and not on their impact on individuals' behaviors and health.	
[36]	No analysis of ethical implications is present.	They put this task on future work.
[51]	Conceptual paper. (i) Limitation of FL aggregation techniques on non-IID data. (ii) costs of communication in the FL.	(i) Use GP to define a new aggregation function. (ii) Use MAS to manage the communication. Define a distance metric to communicate only when necessary.
[44]	Necessity of human intervention.	

Scalability, adaptability, and generalization: A cross-cutting challenge is the scalability of the frameworks and the generalization of the results to real and complex domains [9], [29], [38], [39], [41], [42], [52]. Several works propose extending systems to multi-domain or cross-domain scenarios and adapting to non-static dynamics (e.g., mobile

TABLE 6. (continued) Limitations and proposed solutions.

Study	Limitations	Solutions
[22]	Conceptual paper, limitations are general for Hybrid AI. (i) difficulty in determining who is accountable for the actions of an autonomous system. (ii) risk of unfair outcomes amplified by agent autonomy. (iii) Transparency, users must understand how decisions are made, especially in critical areas. (iv) management of sensitive information with significant risk. (v) data scarcity, computational constraints, and social acceptance.	
[27]	(i) Data scalability: the temporal graph and vector-store grow with interactions, slowing context retrieval. (ii) Challenges of the reflection phase: dependence on the quality of the first retrieval, risk of over-iteration, and increasing computational costs. (iii) Inherent limitations of LLMs: risk of hallucination, bias, and inconsistencies when combining multiple models.	
[19]	Scalability problems due to the limitation of extracting significant patterns from real-time data	Trying to use Quantum AI to address these issues.
[9]	The main limitation is that Hybrid AI strategies have been used in simulated environments, including some simplifications.	Conduct simulations in realistic environments.
[53]	(i) Hallucinations of LLMs, (ii) Limited context window.	No solutions are proposed.
[57]	(i) Conversations that are not very "human" or engaging. (ii) Answers based on limited knowledge. (iii) Technical complexity in implementing distributed collaboration,	It does not specify possible solutions.
[28]	Limitations related to LLMs for construction, updating, and validation.	Collaboration with domain experts.
[38]	(i) Lack of an agent who evaluates and gives feedback. (ii) The graph is designed only for specific cases and is not extensive enough. (iii) The input is only textual. (iv) The system is slow due to numerous API calls.	(i) To develop an evaluation agent. (ii) Extend the database to add more domains. (iii) Multimodal integration.
[18], [21], [23], [24], [39], [45], [46], [48], [50]	No limitations are given; it is purely conceptual work.	
[25], [26], [31], [37], [43], [52], [54]–[56], [59]	NO LIMITATION	

obstacles, patients with evolving profiles). This requires using more heterogeneous and multimodal datasets and introducing adaptive learning strategies that maintain robustness and consistency even in high-variability environments.

Interpretability, explainability, and transparency: Integrating Explainable AI (XAI) forms is an emerging priority [1], [2], [32], [54]. Systems must provide personalized recommendations and motivate the user in a clear, understandable, and verifiable way. Solutions such as expert agents and post-hoc explanations integrated with KGs and symbolic models are proposed. These approaches aim to enhance user trust and human control in decision-making loops.

Validation, benchmarking, and evaluation of impact over time: Validation of hybrid frameworks at full scale and in longitudinal studies is another critical issue [29], [32], [48]. Many systems have been tested only in simulated environments, and there is a lack of standardized benchmarks that can jointly assess performance, accuracy, transparency, and user impact over the long term.

Human-computer interaction and context adaptation: Enhancing human-agent interaction is a fertile area of research [28], [47], [50]. Future frameworks must support more natural, conversational, and personalized interactions by integrating dialog-based agents, multimodal interfaces, and co-adaptation mechanisms (e.g., interactive explanations and collaborative reasoning). Systems capable of learning from user behavior over time and dynamically adapting to new needs or goals are also needed.

Ethical challenges, governance, and security: Some contributions, such as [31] and [36], focus on ethical and governance aspects, including privacy protection, mitigation of bias in models, and tracking the origin of information. The design of human-centered and accountable frameworks is essential to ensure these systems' security, reliability, and

TABLE 7. Future challenges categorized by area.

Area	Studies
Architectural integration and communication between modules	[2], [3], [18]–[21], [23], [24], [27], [35], [37], [40], [44], [46], [49]–[51], [53], [55], [56], [59]
Scalability, adaptability, and generalization	[2], [9], [22], [26]–[30], [33], [38], [39], [41], [42], [45], [49], [51]–[56], [58]
Interpretability, explainability, and transparency	[1], [2], [21], [31]–[33], [53], [54]
Validation, benchmarking, evaluation of impact over time	[1], [3], [29], [30], [32]–[35], [46], [48], [55]
Human-machine interaction and context adaptation	[26], [28], [34], [45], [47], [50]
Ethical challenges, governance and security	[22], [27], [31], [33], [36], [50]
No future challenges	[25], [43]

social acceptability, particularly in sensitive contexts such as health, finance, and education.

The categories presented do not establish isolated challenges; however, the solution to different problems is in the intersection of solutions from other challenges. As can be seen in Table 7, some items belong to more than one category (e.g. [2], [3], [55]). Finally, two jobs do not directly present the future challenges [25], [43].

V. DISCUSSION

In the context of Hybrid AI, integrating heterogeneous components is an increasingly popular strategy to address the complexity of real-world problems. The analysis of the 46 selected studies shows a growing interest in custom agent-based Hybrid AI models, with a significant acceleration of publications in recent years, especially in 2024. This trend suggests that research is rapidly converging on hybrid approaches to address the increasing complexity of human interactions with intelligent systems. Specifically, the momentum visible in 2024 may be related to the simplification in the use of APIs of the most popular LLMs, as well as the decrease in the cost of API access and the availability of new free and open-source LLMs.

Reflection on the results of this systematic review is organized along the four main thematic areas that guided the formulation of the 10 SRQs employed: (i) the evolution of personalized hybrid models, (ii) application contexts, (iii) evaluation modalities, and (iv) open challenges. These areas constitute the analytical framework through which the body of selected studies was interpreted. The aim is to summarize what emerged and offer a critical reading of the main trends, highlighted gaps, and theoretical and practical implications in the rapidly evolving context of Hybrid AI.

One of the main observations is that most studies are still at the conceptual or prototype stage, with only a small percentage of systems thoroughly tested (9%). This confirms that the field is at an intermediate stage of maturity, where hybrid modular architectures are beginning to be implemented, but have not yet reached whole operation or deployment and need further exploration and rigorous validation in real environments or shared benchmarks.

The healthcare sector emerges as the most frequent application domain due to the strong demand for personalized decision support systems, virtual assistants, and coaching based on clinical data. However, much research remains ''domain-agnostic'', as presented in Section IV-C, which indicates how the scientific community focuses on the general definition of technologies and architectures. As confirmed by the study of the proposed objectives of the articles, Section IV-D, where we find the category ''Hybrid Architectures, Frameworks and Theoretical Contribution'' ranked first. Creating specific and customized frameworks is essential to put the theorized guidelines into practice.

Regarding hybridization models, analysis has shown three main integration strategies between rule-based and data-driven models: (i) concatenated architectures, (ii) shared representations, and (iii) agent-level orchestration. However, few studies formalize such integrations at the semantic or logical level. This presents an open challenge, since functional coexistence does not necessarily imply a true cognitive fusion between the paradigms involved.

On the evaluation side, only a few papers present rigorous quantitative evaluation methods. Many studies make use of ad hoc metrics or subjective evaluations [30], [37], [42], [47], [52], [58], and there is a lack of standardization of benchmarks. This limits the comparability and reproducibility of results, hindering solid research advancement.

A. LIMITATIONS AND DISADVANTAGES OF THE ANALYZED HYBRID MODELS

The systematic review highlights how hybrid models offer significant application advantages, as presented in subsection IV-G; however, analysis of the 27 publications that explicitly report critical issues reveals recurring problems that hinder their maturity and transferability to critical contexts. Below, we discuss the main categories of limitations that emerged.

Several papers point out that current architectures have scalability problems. Li et al. [41] records performance decay beyond 80 agents, while Rasal [27] reports progressive slowing of queries as temporal and vector-store graphs grow with user interactions. In general, scalability, along with the ability to generalize to multi-domain and dynamic scenarios, is cited as a challenge across most hybrid frameworks. These technologies cannot be exploited adequately without

load partitioning mechanisms, knowledge compression, and appropriate hardware.

From the point of view of architectural complexity, some authors mention the lack of formal mechanisms of integration between symbolic and sub-symbolic modules, resulting in difficulties of inferential consistency and conflict management [53]. In addition, several projects state limitations related to datasets, which, in many cases, are reduced or unbalanced, and they limit the robustness of the models: multimodal data are missing, and resources are collected in only one clinical or industrial context. This limits generalization and amplifies bias problems. Only a minority of studies evaluate their systems on public datasets; many use simulators or subjective metrics, making comparing approaches and reproducibility of results difficult. The absence of shared suites that measure performance, transparency, and impact longitudinally slows the progress of the entire area.

Accountability, privacy, explainability, and bias mitigation are often mentioned but rarely addressed systematically. Some works point out the absence of ethical impact analysis or decision tracking mechanisms [22], [31], [36], others propose combining LLMs with knowledge & Causal Graphs to offer verifiable explanations and safeguard the confidentiality of sensitive data. However, shared guidelines and regulatory frameworks are needed to ensure operational, and decision making transparency throughout the system life cycle.

Some studies [27], [31] label as ''hybrid'' systems that exclusively combine data-driven modules, e.g., through cooperation between multiple ML models or coordination of LLMs through MAS architectures, i.e., multi-LLM systems. These approaches indeed offer concrete advantages in terms of adaptive capacity, flexibility, and, in some cases, performance, but they also have some limitations that deserve to be critically analyzed.

A first consideration concerns the issue of transparency. However sophisticated, data-driven models, especially LLMs, generally operate as ''black box'' systems, making it difficult to track and understand the reasons for the decisions produced. While this is not necessarily critical in all application contexts, it can be an obstacle in areas where strong traceability of decision-making is required, such as health care, law, or finance. The lack of native mechanisms of explanation also makes it complex to identify systematic errors or biases that are not obvious.

A second disadvantage concerns the handling of structured reasoning. Although modern data-driven models, particularly those based on LLMs, show remarkable competence in linguistic and predictive tasks, they may encounter difficulties in scenarios requiring multi-step planning, conditional rules, or complex logical inferences. In some cases, the ability of these systems to generalize beyond the scope of the data on which they were trained is limited, especially under conditions of ambiguity, incompleteness, or unanticipated variability.

From a computational perspective, a further critical issue is related to scalability. Hybrid architectures based solely on data-driven modules, particularly when they involve multi-model orchestrations, can be resource-intensive. The cost associated with contextual memory management, parallel inference, and dynamic integration of results can become a non-negligible constraint, especially in real-time or distributed applications.

In addition to these considerations, an emerging issue related to dependence on external services should be noted. Many architectures employing LLMs rely on APIs provided by third parties, often through cloud services not directly controlled by the developer or system provider. In contexts where sensitive data, such as health, legal, or financial data, are handled, this dependency can raise significant security, privacy, and regulatory compliance issues. In particular, external models may be impractical or even prohibited in environments where data cannot be transferred outside of proprietary facilities for legal or regulatory reasons. This scenario highlights the need to investigate solutions that ensure local data processing, adopt controllable models, and adhere to stringent directives such as GDPR or industry-specific regulations, such as those imposed by healthcare providers or regulators.

In scenarios where reliability, adaptability, transparency, and local control over processing are needed, exploring truly heterogeneous hybrid models that integrate symbolic elements alongside data-driven components may be advantageous to ensure greater reliability, security, and rigor concerning rules defined.

B. FUTURE CHALLENGES AND RESEARCH ROADMAP

Based on the synthesis of the results discussed across the previous sections, this subsection proposes a research roadmap aimed at addressing the main limitations and open challenges identified. The roadmap highlights a strong interest in solutions combining adaptive flexibility and semantic robustness. However, it also revealed that, in practice, effective integration between data-driven and rule-based modules is still limited or treated superficially. Hybrid architectures described in the literature often focus on functional orchestrations or modular coupling, but rarely address the complex challenge of semantic integration between heterogeneous cognitive paradigms. One of the main thrusts of the future roadmap emerges: structured hybridization approaches that operate at intra-agent and inter-agent levels, combining at least two different paradigms.

[S1] Intra-agent Integration: Towards Deeper Hybrid Architectures. A first line of research concerns internal integration within individual agents. In this context, it aims to overcome the simple coexistence of symbolic and sub-symbolic components within the same agent by developing internal integration mechanisms that allow seamless data exchange and combination.

This perspective also implies the definition of shared semantic interfaces between symbolic components and data-

driven modules. The agent must be able to translate learned results into interpretable conceptual representations and vice versa. In this sense, KGs and flexible logical formalisms can play a key role as intermediate layers.

[S2] Inter-agent Integration: Hybrid MAS. A second focus is agent integration, where each hybrid agent can specialize in a different task to cooperate within a complex hybrid MAS. In this scenario, the challenge is not to integrate logic within a single component, but to design protocols for communication, coordination, and consensus among cognitively heterogeneous agents.

This requires adopting adaptive orchestration mechanisms that can dynamically assign roles, delegations, and decision weights according to the context and expertise of individual agents. In dynamic and distributed environments, ensuring that different agents can contribute to collective deliberation becomes crucial while maintaining semantic consistency and traceability of final decisions.

[S3] Cross-cutting Challenges. These two levels of integration are accompanied by cross-cutting challenges that run through the entire system's life cycle:

- Validation and benchmarking: there is a lack of standard evaluation protocols to measure the effectiveness of hybrid solutions, especially concerning nonperformance metrics such as transparency, explainability, or conceptual robustness.
- Ethical and legal sustainability: LLM or external APIs raise control, accountability, and privacy issues. Careful research is needed on hybrid ''privacy-preserving'' models compatible with regulations such as GDPR.
- Continuity of learning: systems must adapt to changing contexts, maintaining consistency between new knowledge learned and existing rules. Lifelong learning in hybrid contexts is still an open issue.

VI. CONCLUSION

This article presents an SLR on personalized Hybrid AI. Following the Kitchenham-GQM protocol, 46 studies published between 2018 and 2025 were reviewed and mapped to 10 SRQs covering methodological, technological, and application dimensions. The analysis returns (i) a coherent mapping of the hybrid architectures proposed today, (ii) a critical reading of limitations and challenges, and (iii) a roadmap to transform prototypes into operational solutions, offering both research and industrial development a shared scaffolding.

The results highlight an expanding field; over half of the articles have been published in the past two years. The application areas are distributed over heterogeneous domains, while a significant share of the studies remain domain-agnostic. This confirms the need for continued exploration, with studies defining guidelines for designing hybrid systems. Only a few studies employ established quantitative benchmarks, highlighting how many are in the prototype stage. The most recurrent limitations relate to transparency, scalability, computational cost, dependence on cloud services, and the inherent limitations of LLMs. More robust cognitive interoperability models, privacy-preserving architectures for regulated domains, and greater system reliability are needed to bridge the gap between prototype and practice. This presents a significant opportunity for future research and development in personalized Hybrid AI. The insights gained from this study will be valuable for both the theoretical and practical aspects of upcoming research initiatives.

REFERENCES

[1] B. Buzcu, M. Tessa, I. Tchappi, A. Najjar, J. Hulstijn, D. Calvaresi, and R. Aydoğan, ''Towards interactive explanation-based nutrition virtual coaching systems,'' *Auto. Agents Multi-Agent Syst.*, vol. 38, no. 1, p. 5, Jun. 2024.

[2] Y. Kwon and Z. Lee, ''A hybrid decision support system for adaptive trading strategies: Combining a rule-based expert system with a deep reinforcement learning strategy,'' *Decis. Support Syst.*, vol. 177, Feb. 2024, Art. no. 114100.

[3] H. Al Shukairi and R. C. Cardoso, ''ML-MAS: A hybrid AI framework for self-driving vehicles,'' in *Proc. Int. Joint Conf. Auto. Agents Multiagent Syst.*, May 2023, pp. 1191–1199.

[4] G. G. Towell and J. W. Shavlik, ''Knowledge-based artificial neural networks,'' *Artif. Intell.*, vol. 70, nos. 1–2, pp. 119–165, Oct. 1994.

[5] A. S. Garcez, L. C. Lamb, and D. M. Gabbay, *Neural-Symbolic Cognitive Reasoning*. Berlin, Germany: Springer, 2008.

[6] A. Grattafiori et al., ''The llama 3 herd of models,'' 2024, *arXiv:2407.21783*.

[7] A. Liu et al., ''DeepSeek-V3 technical report,'' 2024, *arXiv:2412.19437*.

[8] G. Team et al., ''Gemma: Open models based on Gemini research and technology,'' 2024, *arXiv:2403.08295*.

[9] R. San-Segundo, L. Angulo, M. Gil-Martín, D. Carramiñana, and A. M. Bernardos, ''Hybrid artificial intelligence strategies for drone navigation,'' *AI*, vol. 5, no. 4, pp. 2104–2126, Oct. 2024.

[10] S. Anjomshoae, A. Najjar, D. Calvaresi, and K. Främling, ''Explainable agents and robots: Results from a systematic literature review,'' in *Proc. Int. Joint Conf. Auto. Agents Multiagent Syst.*, Montreal, BC, Canada, May 2019, pp. 1078–1088.

[11] E. Tabassi, ''Artificial intelligence risk management framework (AI RMF 1.0),'' U.S. Dept. Commerce, Nat. Inst. Standards Technol., Gaithersburg, MD, USA, Tech. Rep. NIST AI 100-1, pp. 1–42, 2023.

[12] B. Kitchenham, O. Pearl Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, ''Systematic literature reviews in software engineering—A systematic literature review,'' *Inf. Softw. Technol.*, vol. 51, no. 1, pp. 7–15, Jan. 2009.

[13] M. Galster, D. Weyns, D. Tofan, B. Michalik, and P. Avgeriou, ''Variability in software systems—A systematic literature review,'' *IEEE Trans. Softw. Eng.*, vol. 40, no. 3, pp. 282–306, Mar. 2014.

[14] E. Gagliardi, G. Bernardini, E. Quagliarini, M. Schumacher, and D. Calvaresi, ''Characterization and future perspectives of virtual reality evacuation drills for safe built environments: A systematic literature review,'' *Saf. Sci.*, vol. 163, Jul. 2023, Art. no. 106141.

[15] M. Atalay, U. Murat, B. Oksuz, A. M. Parlaktuna, E. Pisirir, and M. C. Testik, ''Digital twins in manufacturing: Systematic literature review for physical–digital layer categorization and future research directions,'' *Int. J. Comput. Integr. Manuf.*, vol. 35, no. 7, pp. 679–705, Jul. 2022.

[16] E. C. L. Yang, C. Khoo-Lattimore, and C. Arcodia, ''A systematic literature review of risk and gender research in tourism,'' *Tourism Manage.*, vol. 58, pp. 89–100, Feb. 2017.

[17] D. Calvaresi, A. Dubovitskaya, J.-P. Calbimonte, K. Taveter, and M. Schumacher, ''Multi-agent systems and blockchain: Results from a systematic literature review,'' in *Proc. Adv. Practical Appl. Agents*, Toledo, Spain. Cham, Switzerland: Springer, Jun. 2018, pp. 110–126.

[18] M. van Bekkum, M. de Boer, F. van Harmelen, A. Meyer-Vitali, and A. T. Teije, ''Modular design patterns for hybrid learning and reasoning systems: A taxonomy, patterns and use cases,'' *Appl. Intell.*, vol. 51, no. 9, pp. 6528–6546, Sep. 2021.

[19] A. Safari, H. Kharrati, and A. Rahimi, "A hybrid attention-based long short-term memory fast model for thermal regulation of smart residential buildings," *IET Smart Cities*, vol. 6, no. 4, pp. 361–371, Dec. 2024.

[20] H. Fan, J. Huang, J. Xu, Y. Zhou, J. Y. H. Fuh, W. F. Lu, and B. Li, "AutoMEX: Streamlining material extrusion with AI agents powered by large language models and knowledge graphs," *Mater. Design*, vol. 251, Mar. 2025, Art. no. 113644.

[21] A. Meyer, W. Mulder, and M. D. Boer, "Modular design patterns for hybrid actors," in *Proc. 35th Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2021.

[22] S. Pavani and H. Shwetha, "Agentic AI: Redefining autonomy for complex goal-driven systems," 2025, doi: 10.13140/RG.2.2.24115.75047.

[23] T. Schmid, "A systematic and efficient approach to the design of modular hybrid ai systems," in *Proc. AAAI Spring Symp. MAKE*, 2023.

[24] N. Spivack, S. Douglas, M. Crames, and T. Connors, "Cognition is all you need - the next layer of AI above large language models," 2024, *arXiv:2403.02164*.

[25] X. Wang, J. Chen, N. Li, L. Chen, X. Yuan, W. Shi, X. Ge, R. Xu, and Y. Xiao, "SurveyAgent: A conversational system for personalized and efficient research survey," 2024, *arXiv:2404.06364*.

[26] S. Mankari and A. Sanghavi, "Adaptive conversation recommendation systems: Leveraging large language models and knowledge graphs," in *Proc. 2nd DMIHER Int. Conf. Artif. Intell. Healthcare, Educ. Ind. (IDICAIEI)*, Nov. 2024, pp. 1–6.

[27] S. Rasal, "A multi-LLM orchestration engine for personalized, context-rich assistance," 2024, *arXiv:2410.10039*.

[28] Y. Yao, H. González-Vélez, and M. Croitoru, "Explanatory dialogues with active learning for rule-based expertise," in *Proc. 8th Int. Joint Conf. Rules Reasoning Companion*, 2024, pp. 1–15.

[29] S. Amin, M. I. Uddin, A. A. Alarood, W. K. Mashwani, A. O. Alzahrani, and H. A. Alzahrani, "An adaptable and personalized framework for top-N course recommendations in online learning," *Sci. Rep.*, vol. 14, no. 1, p. 10382, May 2024.

[30] O. Bourkoukou and E. El Bachari, "Toward a hybrid recommender system for E-learning personnalization based on data mining techniques," *JOIV Int. J. Informat. Visualizat.*, vol. 2, no. 4, pp. 271–278, Aug. 2018.

[31] A. M. Easin, S. Sourav, and O. Tamás, "An intelligent LLM-powered personalized assistant for digital banking using LangGraph and chain of thoughts," in *Proc. IEEE 22nd Jubilee Int. Symp. Intell. Syst. Informat. (SISY)*, Sep. 2024, pp. 625–630.

[32] B. Buzcu, Y. Pannatier, R. Aydoğan, M. I. Schumacher, J.-P. Calbimonte, and D. Calvaresi, "A framework for explainable multi-purpose virtual assistants: A nutrition-focused case study," in *Proc. Int. Workshop Explainable, Transparent Auto. Agents Multi-Agent Syst.* Cham, Switzerland: Springer, 2024, pp. 58–78.

[33] D. Calvaresi, M. Schumacher, and J.-P. Calbimonte, "Agent-based modeling for ontology-driven analysis of patient trajectories," *J. Med. Syst.*, vol. 44, no. 9, p. 158, Sep. 2020.

[34] R. Das, K. Maheswari, S. Siddiqui, N. Arora, A. Paul, J. Nanshi, V. Udbalkar, A. Sarvade, H. Chaturvedi, and T. Shvartsman, "Improved precision oncology question-answering using agentic LLM," *medRxiv*, pp. 9–2024, 2024.

[35] E. Mitchell, N. Elhadad, and L. Mamykina, "Examining AI methods for micro-coaching dialogs," in *Proc. CHI Conf. Human Factors Comput. Syst.*, Apr. 2022, pp. 1–24.

[36] M. Nzomo, "A hybrid AI framework for sensor-based personal health monitoring towards precision health," in *Proc. AAAI Conf. Artif. Intell.*, Mar. 2024, vol. 38, no. 21, pp. 23405–23406.

[37] A. Subramanian, Z. Yang, I. Azimi, and A. M. Rahmani, "Graph-augmented LLMs for personalized health insights: A case study in sleep analysis," in *Proc. IEEE 20th Int. Conf. Body Sensor Netw. (BSN)*, Oct. 2024, pp. 1–4.

[38] H. Yu et al., "Simulated patient systems powered by large language model-based AI agents offer potential for transforming medical education," 2024, *arXiv:2409.18924*.

[39] Z. Zhang, "RAG for personalized medicine: A framework for integrating patient data and pharmaceutical knowledge for treatment recommendations," *Optimizations Appl. Mach. Learn.*, vol. 4, no. 1, Dec. 2024.

[40] S. Hong, L. Xiao, X. Zhang, and J. Chen, "ArgMed-agents: Explainable clinical decision reasoning with LLM disscusion via argumentation schemes," 2024, *arXiv:2403.06294*.

[41] C. Li, O. Petruchik, E. Grishanina, and S. Kovalchuk, "Multi-agent norm perception and induction in distributed healthcare," 2024, *arXiv:2412.18454*.

[42] J. Liu, W. Wang, Z. Ma, G. Huang, Y. Su, K.-J. Chang, W. Chen, H. Li, L. Shen, and M. Lyu, "Medchain: Bridging the gap between LLM agents and clinical practice with interactive sequence," 2024, *arXiv:2412.01605*.

[43] S. Montagna, S. Mariani, E. Gamberini, A. Ricci, and F. Zambonelli, "Complementing agents with cognitive services: A case study in healthcare," *J. Med. Syst.*, vol. 44, no. 10, p. 188, Oct. 2020.

[44] C. Zhang, D. G. X. Deik, D. Li, H. Zhang, and Y. Liu, "Planning with multi-constraints via collaborative language agents," in *Proc. 31st Int. Conf. Comput. Linguistics*, 2025, pp. 10054–10082.

[45] S. I. Ali, M. B. Amin, S. Kim, and S. Lee, "A hybrid framework for a comprehensive physical activity and diet recommendation system," in *Proc. Smart Homes Health Telematics*. Singapore: Springer, Jul. 2018, pp. 101–109.

[46] I. D. S. Portugal, P. Alencar, and D. Cowan, "An agentic AI-based multi-agent framework for recommender systems," in *Proc. IEEE Int. Conf. Big Data (BigData)*, Portugal, Dec. 2024, pp. 5375–5382.

[47] A. Antunes, J. Campos, M. Guimarães, J. Dias, and P. A. Santos, "Prompting for socially intelligent agents with ChatGPT," in *Proc. 23rd ACM Int. Conf. Intell. Virtual Agents*, Sep. 2023, pp. 1–9.

[48] M. H. de Boer, J. van der Waa, S. van Gent, Q. T. Smit, W. Korteling, R. M. van Stokkum, and M. Neerincx, "A contextual hybrid intelligent system design for diabetes lifestyle management," in *Proc. Int. Workshop Model. Representing Context ECAI*, vol. 23, 2023.

[49] A. Borkowski and A. Ben-Ari, "Muli-agent AI systems in healthcare: Technical and clinical analysis," *Preprint*, 2024.

[50] J. Hu, Y. Dong, S. Ao, Z. Li, B. Wang, L. Singh, G. Cheng, S. D. Ramchurn, and X. Huang, "Position: Towards a responsible LLM-empowered multi-agent systems," 2025, *arXiv:2502.01714*.

[51] E. Pacioni, F. F. De Vega, and D. Calvaresi, "Towards a meaningful communication and model aggregation in federated learning via genetic programming," in *Proc. 17th Int. Conf. Agents Artif. Intell.*, 2025, pp. 1427–1431.

[52] M. Rahbar, M. Mahdavinejad, A. H. D. Markazi, and M. Bemanian, "Architectural layout design through deep learning and agent-based modeling: A hybrid approach," *J. Building Eng.*, vol. 47, Apr. 2022, Art. no. 103822.

[53] A. Sharma, "Merging paradigms: The synergy of symbolic and connectionist AI in LLM-powered autonomous agents," *J. Artif. Intell. Gen. Sci. (JAIGS) ISSN:-*, vol. 6, no. 1, pp. 138–150, Oct. 2024.

[54] Z. Yang, I. Azimi, M. J. Zaki, M. Gaur, O. Seneviratne, D. L. McGuinness, S. M. Rashid, and A. M. Rahmani, "Transforming personal health AI: Integrating knowledge and causal graphs with large language models," in *Proc. ISCAP*, 2024.

[55] V. Aggarwal, "Empowering large language model reasoning: Hybridizing layered retrieval augmented generation and knowledge graph synthesis," *TVET Policies in the Digital Age. Calitatea Vieii*, vol. 36, no. 1, pp. 80–92, Dec. 2024.

[56] E. Y. Chang, "MACI: Multi-agent collaborative intelligence for adaptive reasoning and temporal planning," 2025, *arXiv:2501.16689*.

[57] S. Varitimiadis, K. Kotis, D. Pittou, and G. Konstantakis, "Graph-based conversational AI: Towards a distributed and collaborative multi-chatbot approach for museums," *Appl. Sci.*, vol. 11, no. 19, p. 9160, Oct. 2021.

[58] C. Jeong, "A study on the implementation method of an agent-based advanced RAG system using graph," 2024, *arXiv:2407.19994*.

[59] X. Zhao, M. Blum, R. Yang, B. Yang, L. M. Carpintero, M. Pina-Navarro, T. Wang, X. Li, H. Li, Y. Fu, R. Wang, J. Zhang, and I. Li, "AGENTiGraph: An interactive knowledge graph platform for LLM-based chatbots utilizing private data," 2024, *arXiv:2410.11531*.

[60] R. Socher, D. Chen, C. D. Manning, and A. Y. Ng, "Reasoning with neural tensor networks for knowledge base completion," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst.*, vol. 26. Red Hook, NY, USA: Curran Associates, 2013, pp. 926–934.

[61] R. Barbudo, S. Ventura, and J. R. Romero, "Eight years of AutoML: Categorisation, review and trends," *Knowl. Inf. Syst.*, vol. 65, no. 12, pp. 5097–5149, Dec. 2023.

[62] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-T. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, vol. 33. Red Hook, NY, USA: Curran Associates, 2020, pp. 9459–9474.

[63] H. Chase. (2022). *Langchain*. [Online]. Available: https://github.com/langchain-ai/langchain

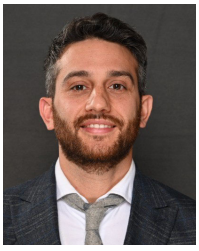
[64] J. Achiam et al., "GPT-4 technical report," 2023, *arXiv:2303.08774*.

[65] Chroma Core Team. (2023). *Chroma: The Open-source Embedding Database*. [Online]. Available: https://github.com/chroma-core/chroma
[66] OpenAI. (2022). *New and Improved Embedding Models*. [Online]. Available: https://openai.com/index/new-and-improved-embedding-model/
[67] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017, *arXiv:1707.06347*.
[68] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*. New York, NY, USA: Association for Computing Machinery, 2016, pp. 1135–1144.
[69] S. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates, 2017, pp. 4768–4777.
[70] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–9.
[71] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Language Technol.*, vol. 1, J. Burstein, C. Doran, and T. Solorio, Eds., Minneapolis, MI, USA: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423/
[72] Neo4j. (2012). *Neo4J–The World's Leading Graph Database*. Accessed: Dec. 26, 2025. [Online]. Available: https://neo4j.com/
[73] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. R. Narasimhan, and Y. Cao, "ReAct: Synergizing reasoning and acting in language models," in *Proc. 11th Int. Conf. Learn. Represent.*, 2022.

ELIA PACIONI (Graduate Student Member, IEEE) received the bachelor's degree in computer and automation engineering from the Università Politecnica delle Marche, Ancona, Italy, and two master's degrees from the University of Extremadura, Mérida, Spain: one in research in engineering and architecture, specializing in information and communications technology, in 2023, and another in technology innovation management, in 2024. He is currently pursuing the Ph.D. degree with the University of Extremadura, Mérida, Spain. He is a Research Assistant with the University of Applied Sciences Western Switzerland (HES-SO) Valais-Wallis. In 2024, he was nominated as an outstanding student at the Evostar conference. His research interests include evolutionary computation, hybrid AI, federated learning, and computer vision. He was awarded the best master's student in the Master in Research Engineering and Architecture.

ANDREI C. COMAN received the bachelor's and master's degrees in computer science from the University of Trento, Italy, and the Ph.D. degree in electrical engineering from the École Polytechnique Fédérale de Lausanne (EPFL), in collaboration with the Idiap Research Institute, where his research focused on deep learning methods for natural language processing, particularly integrating text and graph representations within Transformer-based models. He has also gained industrial research experience as an Applied Scientist Intern at Amazon Science, contributing to retrieval-augmented generation systems and contextual reward modeling for large language models. He is currently a Postdoctoral Researcher with the Applied Intelligent Agents Laboratory (AISLab) with the University of Applied Sciences Western Switzerland (HES-SO) Valais-Wallis, Switzerland. His research interests include natural language processing, representation learning, retrieval-augmented generation, text-graph learning, and large language models.

DAVIDE CALVARESI received the Ph.D. degree in emerging digital technologies-real-time embedded systems from the Sant'Anna School of Advanced Studies, Italy, in 2018. He has been an Associate Professor with the School of Engineering (HEI), University of Applied Sciences Western Switzerland (HES-SO) Valais-Wallis, since mid-2024. Previously, he has been a Senior Researcher with the AISLab Group, HES-SO, Valais-Wallis, from 2018 to 2024. His research interests focused on real-time multi-agent systems, explainable artificial intelligence, agent-based simulations, distributed learning, blockchain, and assistive/rehabilitative technologies. He has been the Chair of several workshops including RTcMAS2018, BCT4MAS 2018-2020), and the EXTRAAMAS 2019-2025). Moreover, he has been the primary investigator of the SEAMLESS project, aiming at enforcing timing compliance in MAS, technical PI of the European project EXPECTATION, aiming at bridging sub-symbolic and symbolic AI to foster interpretability and explainability in Multi Agent Systems, and taken part in several other (inter)national projects.



GAETANO MANZO received the Ph.D. degree in computer science from the University of Bern, Switzerland. He has been a Staff Scientist with the National Institutes of Health (NIH), Bethesda, MD, USA, since February 2022, where he works within the National Center for Biotechnology Information (NCBI) with the National Library of Medicine (NLM). Prior to this, he was a Postdoctoral Researcher with the computational health branch, NIH, and before that, the University of Applied Sciences Western Switzerland (HES-SO). His primary research interests include the development and application of artificial intelligence models within the health and biological sciences. His focuses on leveraging AI to deepen our understanding of complex biological processes, enhance disease prediction and diagnosis, and advance personalized medicine approaches. His work aims to integrate cutting-edge AI technologies with biological research to drive innovation and improve health outcomes at both clinical and research levels. He continues to collaborate with HES-SO on related topics.

MICHAEL IGNAZ SCHUMACHER received the M.Sc. degree in computer science and biology and the Ph.D. degree in computer science from the University of Fribourg, Switzerland. He has been an ordinary Professor with the Institute of Information Systems, HES-SO, Valais-Wallis, since 2007, where he leads the Applied Intelligent Agents Laboratory (AISLab) and the Health Technology Innovation Center (HTIC). Previously, he was a Senior Researcher with the Ecole Polytechnique Fédérale de Lausanne (EPFL). His team has developed several generic software platforms based on multiagent technologies and used in healthcare. His team has been financed by InnoSuisse, FNS, EU H2020, EU CHISERA, and RCSO ISNet. He was involved in several institutional innovations of HES-SO, Valais Wallis, such as the Applied Ethics Service and the Data Acquisition Unit or the Living Laboratory of the Handicap. His research interests include distributed information systems and artificial intelligence applied to healthcare, with a special emphasis on eHealth interoperability, chronic disease management and personalized coaching.